Advances in Network and Communications Engineering

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

2002 - 2003

Edited by

Dr Steven M Furnell
Dr Paul R Filmore

Advances in Network and Communications Engineering

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

2002 - 2003

Editors

Dr Steven M Furnell
Dr Paul R Filmore

School of Computing, Communications & Electronics
University of Plymouth

ISBN 1-84102-118-0

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 2002/03 academic year. A total of 19 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 Filmore

School of Computing, Communications and Electronics
University of Plymouth, January 2004

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 nearly 70 academic staff and has 1300 students enrolled on its portfolio of taught courses. In addition there are 95 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 new 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 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

Communications Research

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

Research interests:

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

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

Network Research Group

Head: Dr S M Furnell, BSc, PhD, CEng, MBCS
Email sfurnell@plymouth.ac.uk

Research interests:

1) Information system 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
Email 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

A MIP-SIP Macro-Mobility Management Scheme for VoIP across Wired and Wireless Domains

G. Gonzalez Lopez[1], Q. Wang[1], M.A. Abu-Rgheff[1] and A. Akram[2]

[1]Mobile Communications Networks Research Group, University of Plymouth, Plymouth, UK.
e-mail: <mosa, qwang>@plymouth.ac.uk
[2]Panasonic Mobile Communications Development of Europe Ltd., Thatcham, UK.
e-mail: Ammad.Akram@panasonicmobile.co.uk

Abstract

With the convergence of the Internet and cellular networks, Voice over IP (VoIP) applications across wired and wireless domains are of increasing importance. Two major mobility management protocols, Mobile-IP (MIP) and Session Initiation Protocol (SIP), have been proposed to handle macro-mobility in the network layer and the application layer respectively, either independently or in a joint way for different scenarios. This paper explores the capabilities of MIPv6 and SIP in supporting VoIP applications from the view of "mobility awareness" or "mobility unawareness" and proposes a MIP-SIP hybrid macro-mobility management scheme for this case. It compares the hybrid scheme against MIPv6 and SIP in terms of handoff disruption time and the user perceived Quality of Service (QoS) through software simulations. The AAA (Authentication, Authorisation and Accounting) functions are also incorporated for inter-domain handoffs.

Keywords

Inter-domain handoff, Mobile-IP, Session Initiation Protocol, Voice over IP, speech quality.

1. Introduction

Terminal mobility refers to a change in the point of attachment of a given host with ongoing Internet connections. Thus, it comprises handoff operations and location management. Terminal mobility can be divided in two main categories: macro-mobility and micro-mobility. While micro-mobility concerns about the host's movement inside a given domain, macro-mobility is related to the movement of a host among different administrative domains. Macro-mobility is enabled as a result of service level and roaming agreements that exist between network operators. Therefore, macro-mobility involves authentication, authorisation and accounting (AAA) mechanisms.

To provide seamless macro-mobility and Voice over IP (VoIP) services in wireless/mobile environments, handoff disruption time must be minimised. Noticeable disruption time will have a direct impact on the perceived speech quality. Even thought MIP is not directly related to VoIP applications, mobility support for VoIP services can be achieved via MIP (Q.Wang and M.A. Abu-ReRgheff, 2003). MIP allows packets to be routed to a mobile host (MH) with out changing the host's network layer address. Thus, MIP is well suited to support applications that are “mobility unaware”, were a change in the network layer address implies connection management issues and does not has to be detected at an application level. On the other hand, SIP seeks to cope with terminal mobility using application layer addresses (M. Handley et al., 2002). However, a SIP handoff implies a change on the MH’s network layer address. Hence, SIP-based VoIP applications are “mobility aware” since the have to detect any change on the network layer address at an application level. However, there may be a need to support terminal mobility for “mobility unaware” VoIP applications, especially where there is a gap in before the full deployment of SIP-based applications. Previous work (N. Nakajima et al., 2003) and (T.T. Kwon et al.,2002) has compared the Mobile-IP and SIP disruption time based on signalling delay during handoff operations. However, (T.T. Kwon et al., 2002) did not consider the network’s jitter. On the other hand, MIPv6’s Return Routability (RR) method was not considered in (N. Nakajima et al., 2003).

This paper proposes and describes a hybrid MIP-SIP mobility management scheme suitable for “mobility unaware” VoIP applications. The hybrid model is contrasted and compared, in terms of speech quality and handoff disruption time against MIPv6 and SIP.. The rest of the paper is organised as follows. Section 2 introduces the Hybrid MIP-SIP Model. The network architecture for the hybrid scheme is described in section 2.1. Section 2.2 describes the necessary extensions to MIPv6 specification. The registration process and mobility scenarios, pre-call and mid-call mobility are described in sections 2.3, 2.4 and 2.5 respectively. The network modelling and simulation are described in section 3. Section 4 gives the simulation results. And finally concluding remarks are offered in section 5.

2. The Hybrid MIP-SIP Model

The main philosophy behind this approach is to combine both, application and network, mobility schemes in support for “mobility unaware” VoIP applications. The Model uses MIP signalling between the foreign network and the MH, and SIP signalling between the foreign network, the MH’s home network and the wire-line correspondent host (CH). The model, however, should also be able to support wireless CHs running “mobility unaware” VoIP applications. In this way, a MH running “mobility unaware” VoIP applications, can roam into a SIP domain and still maintain its home address. Thus, the only concern lays on supporting network layer mobility or not.

2.1 Hybrid MIP-SIP Model architecture

The proposed Hybrid MIP-SIP Model introduces a new network entity called Mobility Manager (MM), which translates MIP signalling to SIP signalling and vice versa. The MM has the following functions.

1) Acts as local MIPv6 HA for the visiting MH.
2) Acts as a SIP User Agent Client (UAC) for the visiting MH.
3) Acts like a Mobility Anchor Point (MAP) as defined in (H. Soliman et al., 2002), limiting the amount of signalling outside the domain when dealing with subnet handoffs (micro-mobility).

Figure 1a shows the architecture of a visited the hybrid MIP-SIP model. Here, the SIP server comprises a SIP registrar, a SIP outgoing and incoming proxy server, and a SIP User Agent Server (UAS). The visitor register (VR) acts like a location server for visiting hosts. The rest of the network elements (i.e. AAA servers) do not suffer any alterations.

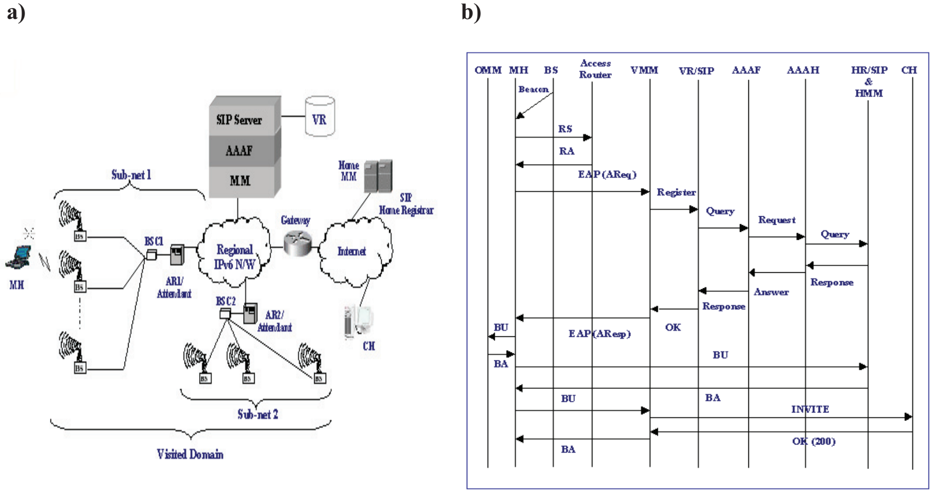


Figure 1 : a) Hybrid MIP-SIP Model visited domain. b) Hybrid MIP-SIP mid-call mobility signaling scheme

2.2 Extensions to MIPv6

The proposed MIP-SIP Signalling Scheme requires minimal extensions to the standard MIPv6 (C. Perkins et al., 2003) and Neighbour Discovery (T, Narten et al., 1998). The model defines two types of care-off addresses (CoA), Regional CoA (RCoA) and On-link CoA (LCoA) as in (H. Soliman et al., 2002). In addition, the MM has to distinguish between three types of binding update (BU) messages, local BU, BU to home network and BU to CH. In order to do so, this model introduces three new flags to the standard format of BU and BA messages. Figure 2 shows the format and extensions proposed to the BU and BA messages. Here, flags A, H, L, and K remain the same as in (C. Perkins et al., 2003), while flags M, C and I are introduced. To follow a description of the aforementioned flags.

- *M*. When set indicates a local BU to the MM.
- *C*. When set indicates a BU to the CH.
- *I*. When set indicates that the BU message was originated from a SIP `INVITE` message sent by a CH that wishes to engage communication. The I flag in the correspondent BA message must be set, so it can be translated to a SIP `OK` message.

2.3 Registration

Upon arrival in a visited network the MH discovers the global address of the visited MM (VMM) via a router advertisement (RA). The MM IPv6 global address is contained in a

Neighbour Discovery option, which this work proposes as an extension to (T, Narten et al., 1998). The MH, then, needs to configure its RCoA and LCoA by stateless means. The RCoA is formed based on the prefix received on the MM Global Address, while the LCoA is formed based on the prefix advertised by the AR. If both addresses prove to be unique (DAD) (T, Narten et al., 1998), the MH must send a local BU to the VMM with M, H and A flags set.

The local BU message must contain the MH NAI option (T, Narten et al., 1998) and all the minimum information to for the VMM to negotiate sessions on behalf of the MH i.e. Codec type. This local BU specifies the MH's RCoA in the home address option defined in (H. Soliman et al., 2002) and the LCoA is used as the source address. The local BU allows the MM to bind the MH's LCoA, RCoA and the MH's NAI.

The SIP UAC located in the VMM, using the information contained in the NAI (user name and realm) (B. Aboba and M. Beadles, 1999) and the MH's RCoA, generates a SIP `REGISTER` request message. This request is sent to the SIP/VR outgoing proxy server on behalf of the MH. The `from` field contained in the SIP `REGISTER` request is the MH's NAI, which will be used as a SIP URI. The `Via` field of the `REGISTER` request is formed by the VMM address. In this way al responses are bound to travel through the MM. The MM will only generates SIP `REGISTER` requests when both flags, M and H, are set. If the registration process succeeds, including all AAA functions, the visited MM will generate a BA message and return it to the MH. However, the MH still has to register its new RCoA with its home MM (HMM) by sending a BU that specifies the binding (RCoA, home address). In this flag H must be set to indicate the HMM to serve as HA to the MH.

2.4 Pre-call mobility

This is when the MH acquires a new address *a-priory* and is usually referred as location tracking (H. Schulzrinne and E. Wedlund, 2000). It is assumed that the MH has already performed the registration and binding processes described above.

The model can distinguish two cases.

1) In this case the CH is running a SIP UAC that generates SIP `INVITE` requests addressed to the MH's home network. The SIP server on the MH's home network returns a redirect `302`-class response giving the address of the VMM which acts as the MH's UAC. The CH will generate a new `INVITE` message and send it to the VMM which is serving the MH to negotiate the session establishment. Once the VMM has negotiated the session establishment, it will generate a BU messages addressed to the MH's LCoA with the I flag set. If the MH is in the claimed location and wishes to accept the session, it will react with a BA message, which then be translated to a `200`-class response by the VMM and sent to the CH.

2) In this case the CH acts as an IPv6 node. In this scenario the Hybrid MIP-SIP Model behaves like HMIPv6 (H. Soliman et al., 2002) since no SIP signalling would be involved. The HMM will act as HA for the MH. Due to the nature of IP, the CH will first send packets to the MH's home address. Packets are intercepted by the home HMM and tunnelled to the MH's RCoA. The HMM will then inform the CH of the MH's RCoA, providing route optimisation.

In both cases, if the MH changes its current address within a VMM domain (LCoA), it only needs to register the new address with the VMM. Hence, only the RCoA needs to be registered with correspondent hosts and the HMM.

2.5 Mid-call mobility

In this scenario the MH is in an ongoing session with a CH, when it changes its RCoA. Once the MH has obtained a new RCoA and performed the registration process, in order to continue communications, the associated signalling has to be handed off to the MH's new RCoA address. Mid-call mobility addresses this problem by having the MH send a BU to the VMM with the C flag set and its RCoA as source address. This BU must also contain the NAI option and the CH address (cached in the MH). The C flag indicates the VMM to use the SIP UAC to generate a SIP `INVITE` request message, addressed to the CH, using the information received in the BU message. The VMM will use the MH's new RCoA address for the `contact` field, the MH's NAI on the `from` field and its own address on the `via` field. If the CH is OK with the handoff, it returns `200`-class response addressed to the MH's RCoA. However, it will be intercepted by the VMM due to the content the `via` field and be translated to a BA sent to the MH. Still, if the MH has not received a BU with the I flag set, it means that no `INVITE` message was issued by the CH in the first place, i.e. the CH is acting as an IPv6 node. Then the MH only issues a BU message directly to the CH with its RCoA as the source and the C flag cleared. Only when the MH receives with the I flag set, it must address any future BU messages, intended to the CH through its VMM. For smooth handoff, the MH can also issue a BU with only the M flag set to is former MM (OMM), so it can forward packets to its new RCoA.

Figure 1b shows the signalling for mid-call mobility in an interdomain handoff. It is assumed that the CH is running a SIP UAC and the MH is running a simple VoIP application that relays exclusively on MIPv6 mobility.

3 Analysis' Methodology

For analysis and comparison, this work assumes a network model which, for simplicity, consists of four access points. To evaluate inter-domain handoff each access point is considered as a different administrative domain. Voice sessions are generated by the CH assuming a G.729a Codec. However, the session establishment is not modelled here, since the focus of this study lays on terminal macro-mobility. Each session has a random duration given by an exponential distribution with a mean of 120sec. Furthermore, sessions are composed by talk spurts followed by a silence periods, both having a random exponentially distributed duration of 352ms and 650 ms, respectively. The CH's and the MH's position as well as the domain acting as the MH's home network are randomised with a uniform distribution at the start of each session. The MH stays in one domain for an exponential distributed time with a mean of 87.63sec. and randomly moves to one of the adjacent domains with a uniform distribution.

The end-to-end packet delay is calculated as the sum of the network delay and the equipment delay. The network delay for each packet is calculated by adding the variable latency in the core network (assuming that the network jitter follows an identical random process with exponential probability density function with a mean of 50ms), the queuing delay at the access router, and the transmission and propagation delay in the wireless link (assuming a free space propagation model). For the case of the hybrid MIP-SIP model, this work models the delay introduced by the MIP-SIP signalling translation as twice the queuing delay at the VMM. The equipment delay refers to the packetisation delay, Codec delay (10ms window rate for G.729a) and jitter compensation delay. To compensate for the network jitter a Jitter Buffer is introduced. Hence, the end-to-end delay for each voice packet is assumed as 193.2 ms to allow for 5% of packet loss (minimum acceptable level (ITU-T Recommendation G.114, 1996) due to network jitter.

This work does not models the delay introduced by processing SIP messages, and it does not considers the delay components added by duplicate address detection (DAD) and delay introduced by switching lower medium to access network Moreover, it also assumes aggressive router selection (N. Nakajima et al., 2003). The parameter to evaluate the speech quality is based on a non-intrusive method called "E-model". This method is able to map the latency and packet losses of a voice session into a 1-5 scale (MOS) which, claims to represent the speech quality as perceived by users. The mean opinion score for speech quality (MOS) is done as shown in equations 1 to 3 (T, Narten et al., 1998)

$$MOS = \begin{cases} 1, R \leq 0 \\ 1 + 0.035 \cdot R + R \cdot (R - 60) \cdot (100 - R) \cdot 7x10^{-6}, 0 < R < 100 \\ 4.5, R \geq 100 \end{cases} \qquad (1)$$

Assuming echo cancellation *R* is given by

$$R = R_O - I_d - I_e + A \qquad (2)$$

Here, I_d is the impact of the overall delay, as the end-to-end delay is a constant value of 193.2 ms, I_d can be reduced a constant value of 2.39. I_e is the impact of packet loss in the Codec performance for a G.729a Codec as shown in equation 3. as default value according to the E-model. Finally R_O=93.2 and *A=10* (T, Narten et al., 1998).

$$I_{e(G.729a)} = 0.0071 \cdot x^3 - 0.2727 \cdot x^2 + 4.9756 \cdot x + 10.0532 \qquad (3)$$

Where *x* represents the packet loss in percentage (%). While (1) and (2) are given by the aforementioned ITU-T recommendation, (3) is derived using a 3rd order polynomial fitting for the typical values for a G.729a.

4 Analysis Results

Figure 4 illustrates the outcome from the analysis described above. The results presented here are based on eight voice sessions (calls) initiated by the CH. Figure 2a. shows the fluctuations of the MOS during the eight sessions using the Hybrid MIP-SIP Model for handoff signalling. Figure 2b graphically represents the impact of packet loss on the G.729a Codec (I_e), using the same signalling scheme. Both MOS and I_e are calculated in intervals of 5 seconds. If a handoff occurs in-between to intervals, both intervals are used in the calculation. It is fair to say that the bigger the intervals, the less is the impact of packet loss in the MOS and I_e.

The same analysis was made using SIP with DHCPv6 stateful address configuration and rapid commit option (R. Droms et al., 2002) and MIPv6 using stateless address configuration (T Narten et al., 1998). Figures 2c and 2d compare the MOS and the disruption time per handoff operation, respectively, between SIP, MIPv6 and the proposed Hybrid MIP-SIP Model signalling schemes. Disruption time refers to the time when packets are routed to the MH's old CoA before the binding messages. Here the MOS is calculated assuming only packet loss due to disruption time.

5 Discussion and Conclusions

Results show a better performance for the three schemes when both MH and CH are located in the same domain. As shown in figure 4a it is undeniable that disruption time has an impact on the speech quality as perceived by users (MOS), at least in short intervals. However, even without packet loss the Codec *per-se* impacts the MOS. Each Codec, due to compression, speech prediction and decompression has default penalty on the MOS that will vary according to the coding algorithm used. For the case of G.729a, I_e = 10.0532 when packet loss tends to 0%.

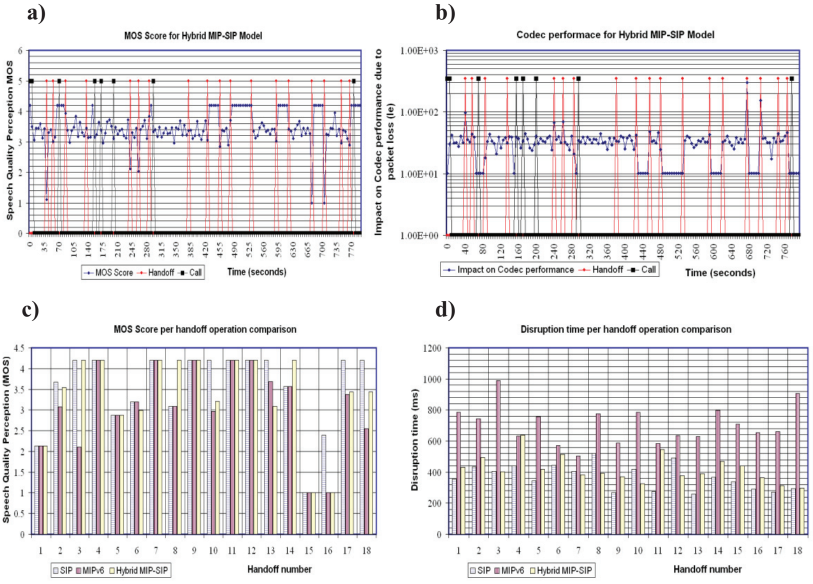


Figure 2 : Analysis results and schemes comparison

Disruption time is a direct function of the latency in round trips for signalling between the MH, the CH and the MH's home network. Due to the Return Rutability (RR) method, MIPv6 signalling adds more round trips for handoff signalling than SIP, enlarging the handoff disruption time. This is mainly the reason why MIPv6's disruption time is longer than SIP's. When actually the length for MIPv6's signalling messages are smaller than for SIP's signalling messages e.g. an `INVITE` message has an average length of 768bytes (including IPv6 and UDP headers) while a BU message to the CH has a length of 178bytes (including IPv6 and routing headers).

Results illustrate a better performance for SIP, with higher MOS and lower disruption times per handoff operation. The Hybrid MIP-SIP Model presents better handoff performance than MIPv6, mainly due to the ability of the OMM to forward packets to the MH's new CoA (smooth handoff). According to (N. Nakajima et al., 2003) the processing time for SIP messages can vary according to particular implementations. Consequently, the processing time for the Hybrid MIP-SIP Model to process SIP messages and translate MIP to SIP signalling and vice versa will also vary according to particular implementations. Despite this, the Hybrid MIP-SIP Model is able to comprise both mobility approaches allowing "mobility unaware" applications and "mobility aware" applications (e.g. SIP) to interact with each other, reducing the disruption time if compared to MIPv6. However, this approach has the following disadvantages. The MM can prove to be a bottleneck in the system. SIP's ability to initiate and release sessions is constrained. The MH has to inform both, its home MM and its home SIP server, about its current location. Finally, translating MIP to SIP signalling and vice versa will also add delay to the signalling enlarging the handoff disruption time, especially if the VMM has to serve a large number of MHs

6 References

B. Aboba and M. Beadles, "The Network Access Identifier," RFC 2486, Jan 1999.

C. Perkins et al., "Mobility Support in IPv6," Internet draft, draft-ietf-mobileip-ipv6-21.txt, work in progress, Feb 2003.

H. Schulzrinne and E. Wedlund "Application-Layer Mobility Using SIP," Mobile Comp. And Commun. Rev., vol 4, no. 3, July 2000.

H. Soliman et al, "Hierarchical Mobile IPv6 mobility management (HMIPv6)", draft-ietf-mobileip-hmipv6-07.txt, IETF Mobile IP Working Group, October 2002, work in progress.

ITU-T Recommendation G.107, "The E-model, a Computational Model for Use in Transmission Planning", 1998.

ITU-T Recommendation G.114, "One-way Transmission Time", 1996.

M. Handley et al., "SIP: Session Initiation Protocol," RFC 3261, June 2002.

N. Nakajima et al., "Handoff Delay Analysis for SIP Mobility in IPv6 Testbed", Accepted for ICC 2003.

Q. Wang and M. A. Abu-Rgheff, "Integrated Mobile IP and SIP approach for advanced location management", Proc. IEE 4th International Conference on 3G Mobile Communication Technologies (3G 2003), London, UK, Jun 2003.

R. Droms et al., "Dynamic Host Configuration Protocol of IPv6 (DHCPv6)" Internet draft, draft-ietf-dhc-dhcpv6-28.txt, work in progress, Nov 2003

T. T. Kwon, M. Gerla, and S. Das, “Mobility management for VoIP service: Mobile IP vs. SIP”, IEEE Wireless Communications, Vol. 9, No. 5, Oct 2002, pp. 66-75, 2002.

T. Narten et al., “Neighbor Discovery for IP Version 6,” RFC 2461, Dec. 1998.

A Cross-Layer Design for Wireless VoIP: Playout Delay Constrained ARQ with ARQ aware Adaptive Playout Buffer

Z. Li, L. Sun, Z. Qiao and E.C.Ifeachor

Centre for Signaling Processing & Multimedia Communications
University of Plymouth, Plymouth, United Kingdom
eifeachor@plymouth.ac.uk

Abstract

Packet error recovery techniques such as Automatic Repeat on reQuest (ARQ) is crucial to speech quality of Wireless VoIP, which is suffered from impairment factors introduced in the wireless channel, such as packet error, delay and jitter. The main aim in this paper is to make use of cross-layer techniques to improve the performance of ARQ. In this paper, we propose a cross-layer design in which 1) retransmission procedures of the link layer ARQ protocol is constrained by the available playout delay 2) If the retransmission procedure is terminated prematurely, received noisy copies of a speech packet are presented to application layer and finally played out. 3) Delivery delay in the wireless channel is constrained to avoid delay accumulations in the transmitting queue. Simulation results show that the perceptual speech quality of a wireless VoIP system can be significantly enhanced since retransmission delays, playout buffer losses, and queuing delays are reduced by this design.

Keywords

ARQ, Perceived Speech Quality, Cross-Layer, Wireless VoIP

1. Introduction

Wireless VoIP, that is, delivery of Voice-over-IP (VoIP) service over wireless/mobile Internet seems to be one of the typical pictures of ‘beyond 3G’. However, due to the unreliable and error-prone features of wireless channels, assuring acceptable perceived speech quality has been a challenging task for Wireless VoIP. Automatic Repeat on reQuest (ARQ) is one of the packet error recovery techniques for Wireless VoIP. ARQ has been widely used because of its efficiency and simplicity.

In ARQ, the sender sends packets or Protocol Data Units (PDUs) consisting of payload and checksums. According to the result of checksum validation, the receiver sends back acknowledgment messages (e.g. ACK or NACK) to the sender. In this paper we considered the Stop&Wait (SW) ARQ in IEEE 802.11 Media Access Control (MAC) Layer [1]. In the 802.11 SW-ARQ, the transmitted packet must be acknowledged before the next packet can be sent. If in a certain timeout period an acknowledgement for a packet is not received by the sender, the sender will retransmit this packet until a maximal retry limit is reached. With this procedure, corrupted packets may be recovered by the retransmitted copies.

However, ARQ schemes also bring a series of problems that impacting the perceived speech quality. The retransmission procedure may introduce excessive delays, especially when the packets have to traverse a high delay wireline network before it reach the wireless part. Further, the layered protocol architecture, which puts ARQ and the playout buffer in different layer, makes things go from bad to worse. Firstly, if the retransmission procedure is only constrained by a fixed maximum retry limit, a retransmission procedure with high retry limit may exceed the playout delay, leading to unnecessary retransmissions and subsequent packets postponed. With low retry limit, the retransmission procedure may be terminated prematurely before it exhausts the playout delay. Secondly, considering a transmitting queue exists in the sender, too much of retransmission delay can make queuing delay quickly climb up. Thirdly, in current protocol stack, packets that failed in transport or MAC layer checksum validations are discarded, despite noisy voice packets may be considered useful at the upper layer [2].

These problems have been addressed in previous works. In [3][4], the retransmission attempts for a packet can be aborted when it reaches its deadline (e.g. presentation time). Nevertheless, these works are not flexible enough and cannot avoid the prematurely terminating of a retransmission procedure when there is still some delay budget left for more retry attempts. In [5] UDP-Lite, a modified UDP protocol with partial checksum, has been developed to allow corrupted UDP packet to be reused at application level. This idea can be also implemented at link layer.

We extended these ideas in a cross-layer design for Wireless VoIP. In our design, 1) retransmission procedure of a packet is constrained in the available playout delay 2) Delivery delay in the wireless channel is constrained within the mean inter-arrival delay of the transmitting queue. 3) Speech data is not covered by the link layer or transport layer checksums and noisy packets are still played out.

The rest of this paper is organized as follows. Section 2 describes components of the proposed cross-layer design. The simulation model and analyze the experimental results are presented in Section 3. Section 4 concludes this paper.

2. The Cross-Layer Design

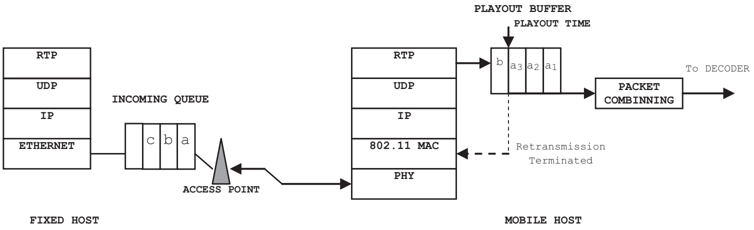


Figure 1 : The System Model

2.1 The System Model

The system model of the proposed cross-layer design is described in Figure 1. We considered the last-hop scenario in an IEEE 802.11 wireless network. Our design is composed of two correlated components: playout delay constraint ARQ and ARQ aware playout buffer. As speech data is not covered by the link layer and transport layer checksums, the playout buffer may receive several noisy versions of a packet. We employed the Majority-Logic packet combining to further reduce the damaged part before sent a packet to the decoder. We avoid a detailed description and refer the reader to [6] for details of this technique. The two key components of the cross-layer design are described hereafter:

2.2. Playout Delay Constrained ARQ

The playout delay constrained ARQ is a specific optimization of current protocol stack for Wireless VoIP. In the receiver, the 802.11 MAC layer presents every received packet to the upper layer, whether it's corrupted or not. In the application layer, the playout buffer can terminate a packet's retransmission procedure at its playout time hence to avoid unnecessary retransmissions. If a corrupted packet hasn't been recovered by the retransmission procedure, the received noisy copies are processed by the packet combining module to get a more reliable version, which is then decoded and played out.

2.3 ARQ Aware Playout Buffer

2.3.1 Queue Model

Assume there is a per flow transmission queue at the sender with a large enough queue length, so we can focus on the queuing delay. With the IEEE 802.11 SW-ARQ, the transmission queue can be seen as an M/M/1 queuing system with Poisson distribution of packets arrivals and exponential distribution of packets departures [7]. Let α be the average inter-arrival delay and s the average packets departure delay. We have $\lambda = \frac{1}{a}$, $\mu = \frac{1}{s}$ where λ and μ are the mean arrival rate and mean service rate. The mean waiting delay in the queue TQ can be computed as $TQ = \frac{1}{\mu - \lambda} = \frac{a \cdot s}{a - s}$

We can deduce that when $s \to a, TQ \to \infty$ which means if the mean delivery delay in the wireless channel is not constrained within the mean inter-arrival delay of incoming packets, TQ will quickly climb up.

2.3.2 Adaptive Playout Buffer

The size of a playout buffer can be fixed or adjustable. Fixed playout buffers cannot adapt readily to changes in network delays as a result are not practical in real VoIP systems. In [8], Ramjee et. al. had proposed several algorithms (e.g. *'fast-exp'*, *'min-delay'* and *'spike'*) to adjust playout delay according to predicted network delay performance. These algorithms estimate mean and variation of network delay $\hat{d}_i$ and $\hat{v}_i$ on the arrival of the i^th packet. In these algorithm $\hat{v}_i$ is given by $\hat{v}_i = \alpha \cdot \hat{v}_{i-1} + (1-\alpha)\cdot abs(\hat{d}_i - n)_i$. But they differ in the computation of $\hat{d}_i$, and may be suitable for different transmission environment. In [9], Sun et al had proposed an *'adaptive'* algorithm that can adapt to preferred algorithm (e.g. *'min-delay'* or *'fast-exp'*) when transmission environment changes. The playout delay is adjusted at the beginning of each talkspurt. Let t_i be the timestamp of packet *i* which is the first packet in a talkspurt, the playout time p_i is computed as $p_i = t_i + \hat{d}_i + \mu \cdot \hat{v}_i$, where μ is a constant. The playout time for the subsequent packets j in the same talkspurt p_j is computed as $p_j = p_i + t_j - t_i$.

2.3.3 ARQ Aware Playout Buffer

The main responsibility of the ARQ aware Playout Buffer is to constrain delivery delays in the wireless channel within the mean inter-arrival delay of the transmitting queue and to predict playout delay for each talkspurt.

Since any transmitted copy was not discarded at link layer, there may be several copies of a packet existing in the playout buffer. Figure 4 gives some timing notations associated packet *i*. Let a_i be the receiver timestamp of the first arrived copy of ith packet, and t_i be the sender timestamp. We can estimate delivery delay in wireline network for packet *i* (denoted by nl_i) by $nl_i=a_i-t_i$. Let r_i be the receiver timestamp of the last arrived copy. The total delivery delay in wireless channel of packet *i* (denoted by nw_i) can be estimated as $nw_i=r_i-a_i$. If no retransmission required $r_i=a_i$, $nw_i=0$. It should be noted that the propagation delay of the first copy has been included in nl_i. To constrain nw_i, the playout buffer keeps running estimations of mean inter-arrival delay (denoted by σ_i) as the upper bound of nw_i. The constrained delivery delays in the wireless channel may reduce available time to fully recover a corrupted packet. But the benefit is the great reduction of queuing delay. We simply estimate σ_i by: $\sigma_i = \alpha \cdot \sigma_{i-1} + (1-\alpha)\cdot abs(a_i - a_{i-1})$, where α is the same constant as used in the estimation of $\hat{v}_i$ and it is set to be 0.99802 in the simulation. The algorithm can be summarized as:

$$n_i = nl_i + nw_i = \begin{cases} nl_i + r_i - a_i & r_i - a_i < \sigma_i \\ nl_i + \sigma_i & r_i - a_i \geq \sigma_i \\ nl_i & r_i = a_i \end{cases}$$

As the ARQ aware playout buffer is only differed in the way of computing network delay n_i. We can estimate mean network delay $\hat{d}_i$ by present algorithms, i.e. the '*adaptive'* algorithm proposed in [9].

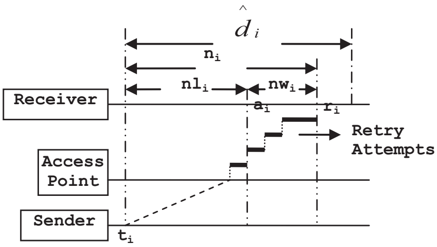


Figure 2 : Timing associated with Packet *i*

3. The Simulation Model and Experimental Results

As presented in Figure 3, our simulation model is comprised of the following components: a voice traffic model, the Adaptive Multi-Rate (AMR) encoder and decoder, a playout buffer, and a wireless network simulator that integrated the 802.11 SW-ARQ and a simple Bernoulli bit error model.

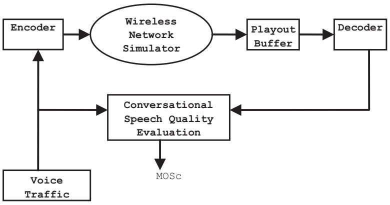

Encoder
Wireless Network Simulator
Playout Buffer
Decoder
Conversational Speech Quality Evaluation
MOSc
Voice Traffic

Figure 3 : The Simulation Model

3.1 Wireless Channel Model

We employed a simple Bernoulli model to simulate bit errors introduced in the wireless channel. The probability of PHY layer packets corrupted by bit errors (denoted by *PER)* can be computed as:

$$PER = 1-(1-BER)^{ph+pl}$$

where *BER* is the Bit Error Rate and *ph* is the packet overhead size from physical level. For our simulations we have used a value of 784 bits for *ph*: 24, 34, 20, 8, 12 bytes at the PHY, MAC, IP, UDP and RTP layer respectively (no header compression is used). *pl* is the payload size, which is set to be 256 bits (32 bytes) corresponding to an AMR 12.2K voice frame.

Let ω denote the estimated playout delay, and R_{ϖ} be the corresponding maximum retry limit constrained by ω. We can also compute the probability of a packet being recovered after R_{ϖ} times of retransmissions *PKR* as:

$$PKR = PER^{R_{\varpi}-1} \cdot (1-PER)$$

And the probability of the bit errors happen in the packet header *PHE* can be given by:

$$PHE = 1-(1-BER)^{ph} \cdot \frac{1}{pl+1}$$

3.2 Voice Traffic Model

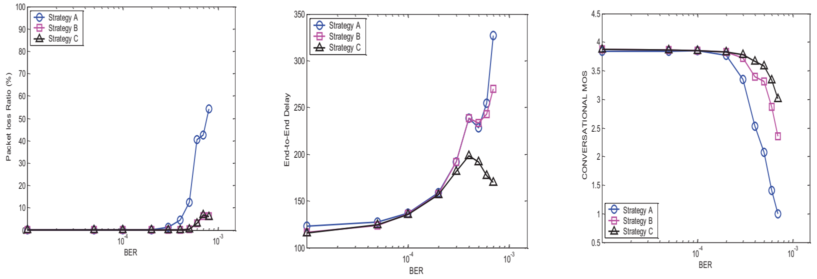
The voice traffic model can be simply represented by the on-off model [10]. In the on-off model, a two-state chain is assumed, one corresponds to the talkspurt and one for the silence periods. The holding time in the two states is assumed to follow an exponential distribution.

In our simulation we selected a mean of 1.0 sec and 1.5 sec for talkspurt state and silence state respectively as suggested in [11]

3.3 Speech Quality Evaluation

In our simulation model, we employed the conversational speech quality evaluation method [9] to qualify the performance of simulation strategies. This method combined E-Model [12] and PESQ [13] to measure the perceived speech quality, the results is represented by MOSc (Conversational Mean Opinion Score). In this method, the impact of bit errors in the payload, packet losses and delay all contribute to the degradation of final evaluated speech quality.

3.4 Experimental Results and Analysis



***a*. Overall packet losses** ***b*. End-to-end delay** ***c*. Conversational MOS**

Figure 4 : Experimental results

We considered three strategies in the simulation study: *Strategy A*. 802.11 SW-ARQ [1] and ‘adaptive’ playout buffer [9]; *Strategy B*. playout delay constrained ARQ with ‘adaptive’ playout buffer; *Strategy C* playout delay constrained ARQ with ARQ aware playout buffer. The simulation results were obtained by averaging results of 30 trials with different random seeds to avoid the impact of packet loss or bit error locations. Each trial continued for 200 seconds corresponding to 10,000 PDUs (one PDU encapsulated one RTP packet).

Figure 4a shows the overall packet loss ratio comparison for these strategies. When BER increases, *Strategy A* discards many corrupted packets that can not be fully recovered before their playout time. *Strategy B* and *C* are the same policy regarding packet losses. Both of them reuse noisy packets and only discard those packets that cannot reach the receiver before their playout time. The result is that *Strategy B* and *C* only discard a small percentage of packets compared to *Strategy A*, even when the wireless channel is very noisy.

We also plotted end-to-end delays as a function of BER in Figure 4b with a fixed 100ms delay in the wireline network. In Figure 3b, we can see that the end-to-end delays of these strategies are climbing with the increasing of BER. *Strategy B* performs slightly better than *Strategy A* when BER become worse, as *Strategy B* has the capacity to terminate unnecessary retransmission. *Strategy C* outperforms *Strategy A and B* with a more stable curve, as it managed to avoid queuing delay accumulations. It should be noted that the delay curves descended at some points where the ‘adaptive’ playout buffer switch to the ‘min-delay’ algorithm more frequently.

The performance enhancement achieved by the cross-layer design in terms of conversational Mean Opinion Score (MOSc) is presented in Figure 4c. From Figure 4c, we can see that the curve of *Strategy A* and *B* deceases significantly after BER 10^{-4}. At a BER of around $10^{-3,}$ *Strategy A* already reaches 1.0, the worst MOSc. On the contrary, *Strategy C*, or the cross-layer design, still achieves MOSc 3.0 at the same BER.

4. Conclusions

We investigated problems introduced by the IEEE 802.11 SW-ARQ protocol in a Wireless VoIP system upon the layered protocol architecture. We also proposed a cross-layer design as a solution for the presented problems. The proposed cross-layer design is composed of two correlated components: 1) playout delay constrained ARQ, in which a packet’s retransmission procedure is terminated at its playout time, and noisy copies of a packet can be combined and then played out. 2) ARQ aware playout buffer, in which requirements for the delivery delay in wireless channel (e.g. not to advocate queuing delay) is considered in playout delay estimation. Through simulations, we show that the perceived speech quality of a Wireless VoIP system is improved by the proposed cross-layer design at the expense of breaking the layered protocol architecture.

5. References

[1] IEEE Standards Department, 1999, *IEEE 802.11 Standard for Wireless LAN, Medium Access Control (MAC) and Physical Layer (PHY) Specification*

[2] Florian Hammer, Peter Reichl, Tomas Nordstrom, 2003, Gernot Kubin, Corrupted Speech Data Considered Useful, Proc. *First ISCA Tutorial and Research Workshop on Auditory Quality of Systems*, Mont Cenis, Germany

[3] Christos Papadopoulos, Gurudatta M.Parulkar, 1996, Retransmission-Based Error Control for Continuous Media Applications, *Proc. of NOSSDAV*

[4] Guijin Wang, Qian Zhang, Wenwu Zhu and Ya-Qin Zhang, 2000, Channel-Adaptive Error Control for Scalable Video over Wireless Channel, *Proc of the 7th Momuc*, *Japan*

[5] Larzon, L. et al, 2002, The UDP-Lite Protocol, *Internet Dratft, Internet Engineering Task Force*, Work in Progress.

[6] Stephen B.Wicker, 1991, Adaptive Rate Error Control Through the Use of Diversity Combining and Majority-Logic Decoding in a Hybrid-ARQ Protocol, *IEEE Transactions on communications*, VOL.39, NO.3

[7] E. PAGE, 1972, *Queuing system in OR*, the Butterworths Group, ISBN 0408702370

[8] R.Ramachandran, J.Kurose, D.Towsley and H.Schulzrinne, 1994, Adaptive playout mechanisms for packetized audio applications in wide-area networks, *Proc. IEEE INFOCOM*

[9] L Sun, E.C.Ifeachor, 2003, Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms, *Proc. of IEEE ICC*

[10] P. Brady, 1965, A Technique for Investigating On-Off Patterns of Speech, *Bell System Technical Journal*, 44(1):1-22,.

[11] *ITU-T Recommendation P.59*, Telephone transmission quality objective measuring apparatus: Artificial conversational speech.

[12] *ITU-T Recommendation G.107 (05/2000)*, The E-model, a computational model for use in transmission planning.

[13] *ITU-T Recommendation P.862*, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs

The Impact of Adaptive Playout Buffer Algorithm on Perceived Speech Quality Transported over IP-based Networks

P. Hu, L.F. Sun, Z. Qiao & E.C. Ifeachor

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

Abstract

Perceived conversational speech quality is an important metric in Voice over IP (VoIP) applications. As IP networks are not designed for real-time applications, the network impairments such as packet loss, jitter and delay have a severe impact on speech quality. Playout buffer at the receiver side is used to compensate jitter at a trade-off of delay and loss. Different playout buffer algorithms can have a different impact on the achieved end-to-end perceived speech quality. It is important to design a playout buffer algorithm which can achieve an optimum perceived speech quality.

The objective in this study is to understand how network impairments and existing adaptive buffer algorithms affect perceived speech quality and further to design a modified buffer algorithm to obtain an optimised perceived voice quality.

In this paper, we firstly studied different objective speech quality evaluation methods and chose a simplified and efficient method for jitter buffer algorithm optimisation in the study. Then we investigated seven representative buffer algorithms and propose a new adaptive buffer algorithm, which can directly maximise the MOS index for given network parameters and can also efficiently adjust playout delay during diverse spikes. We implemented these algorithms and compared their performance under different network conditions with high or low network delay variations. Preliminary results show that the new algorithm can enhance the perceived speech quality in most network conditions and it is more efficient and suitable for real buffer mechanism.

Keywords

VoIP, Perceived speech quality, MOS, Buffer algorithm, delay, loss, delay variation (jitter)

1. Introduction

Perceived conversational speech quality is an important metric in Voice over IP (VoIP) applications. As IP networks are not designed for real-time applications, the network impairments such as packet loss, jitter and delay have a severe impact on speech quality.

From the user's point of view, delay and packet loss are the two important characteristics of VoIP technique. Previous research shows packet lossratio exceeds 5% (Javant, 1980) and one-way end-to-end delay exceeds 400ms (ITU, 1993), holding an IP-based voice communication becomes difficult. However, recently, with numerous researches by (Ramjee, Kurose, et al, 1994)(Moon, Kurose et al, 1998), loss rates of up to 10% can be tolerated when good loss concealment techniques are employed. This improvement has extended the leeway of modifying buffer algorithms to greatly advance the "trade-off" between delay and packet loss.

The aim of producing a mechanism is to be satisfied by human, and is not just to obtain such perfect data. In early researches on buffer algorithm, the aims of designs are simply based on the ratio of delay and packet lossratio. However, several research groups (Fujimoto and Murata, 2002)(Sun and Ifeachor, 2002)(Markopoulou and Tobagi, 2002) has recently proposed that the choice of the best buffer algorithm for a given situation should be determined by the likely perceived quality. The work so far has touched on certain points of optimising the "trade-off" from the aspect of maximising perceived quality, but is not comprehensive or relevant to VoIP applications.

In VoIP communication, as delay variation is an important aspect in real networks, most existing buffer algorithms have researched on it. Delay spike is a major part of delay variation, and can hugely influence the perceived speech quality. Previous studies (Ramjee, Kurose, et al, 1994)(Moon, Kurose et al, 1998) have indicated the presence of "spikes" in end-to-end Internet delays. A spike constitutes a sudden, large increase in the end-to-end network delay, followed by a series of packets arriving almost simultaneously, leading to the completion of the spike. If ignoring the happening of spikes, it might well cause the loss of a huge number of packets, and then results in the low perceived speech quality.

The analysis and experiment results show each existing playout buffer algorithm has its own limitation, so cannot be well utilised in diverse network circumstances. Even in such high variable networks, delay adjustments of those algorithms are not able to properly follow tendencies of delay variations. There thus needs to further work on it.

In this paper, we investigate seven representative buffer algorithms, in the face of both steady and varying network delays. As one important issue is to map network parameters straight to the perceived speech quality, we propose a new adaptive buffer algorithm, which can directly maximise the MOS index for given network parameters and can also efficiently adjust playout delay during diverse spikes.

The remainder of the paper is structured as follows. Section 2 presents a methodology for the objective conversational speech quality measurement. In section 3, we present and analyse different buffer algorithms. In section 4, we propose a new adaptive buffer algorithm to achieve the realisably best perceived speech quality, and compare the performance among the proposed and existing buffer algorithms. Finally, section 5 concludes the paper.

2. Introduction to An Objective Speech Quality Measurement Method

A methodology, proposed in (Sun and Ifeachor, 2003)(Cole and Rosenbluth, 2001), is to be possibly simplified and greatly accurate in evaluating the perceived speech quality. Both concepts of PESQ algorithm (ITU-T, 2000) and ITU-T E-model (ITU-T P.862) are utilised in this methodology.

As the impairment rating scale for the E-model, *R*, is expressed as (ITU-T, 2000)

$$R = Ro - Is - Id - Ie,eff + A \tag{1}$$

If ignoring the effects of other impairments, the factor *R* is simplified as

$$R = Ro - Id - Ie, \tag{2}$$

where *Ro* is the optimum quality value (the default value for *Ro* can be set to 93.2 (ITU-T, 2000)), *Ie* is known as the equipment impairment factor and accounts for impairments due to non-linear codec and packet loss. *Id* accounts for echo and delay. If the cancellation condition is perfect, *Id* can be calculated as (Sun and Ifeachor, 2003)(Cole and Rosenbluth, 2001)

$$Id = 0.024Ta + 0.11(Ta - 177.3)H(Ta - 177.3), \tag{3}$$

where $\begin{cases} H(X) = 0 & \text{if } X < 0 \\ H(X) = 1 & \text{if } X \geq 0 \end{cases}$

Ta represents absolute delay (playout delay).

As mentioned in ITU-T Recommendation (ITU-T, 2000), if given the *R* value, the corresponding MOS can be obtained as

$$MOS = 1, \quad \text{for } R \leq 0 \tag{4a}$$

$$MOS = 1 + 0.035R + R(R - 60)(100 - R) \times 7 \times 10^{-6}, \quad \text{for } 0 < R < 100 \tag{4b}$$

$$MOS = 4.5. \quad \text{for } R \geq 100 \tag{4c}$$

For the *Ie* value, we choose to use PESQ to derive this value. For example, as codec G.723.1, the equation of *Ie* can be obtained from the curve of MOS (PESQ) vs. random packet lossratio as

$$Ie = 20.06 \times \ln(1 + 0.1024 \times lossratio) + 25.63 \ . \tag{5}$$

This method can be extended to other speech codes. It can be easily used to monitor and predict conversational speech quality from the network impairments (Sun and Ifeachor, 2003)(Cole and Rosenbluth, 2001).

3. Existing Adaptive Buffer Algorithms

Playout buffer can be fixed or adaptive. However, fixed buffer cannot adapt to changing network delay, thus it usually results in weak speech quality. For our purpose, we thus just focus on adaptive playout buffer algorithms. The adjustment of buffer is set to the beginning of each talkspurt, as discussed in (Sun and Ifeachor, 2002)(Ramjee, Kurose, et al, 1994)(Moon, Kurose et al, 1998).

In this section, we review and analyse four adaptive buffer algorithms proposed in (Ramjee, Kurose, et al, 1994) as from Algorithm 1-4, and three in (Moon, Kurose et al, 1998), (Fujimoto and Murata, 2002) and (Markopoulou, 2003), sequentially. In order to clearly describe buffer algorithms, we define certain parameters as presented in Figure 1. We refer to t_i as the time of sending *i*th packet. a_i and p_i represent the packet arriving time and playout time, respectively. *ni* is the packet's network delay and *di* is the playout delay (end-to-end delay). The buffer delay is referred to as *bi*.

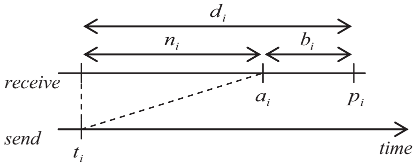


Figure 1 : Timing associated with *i*th packet

<u>Algorithm 1</u> **Exponential-Average (Exp-Avg)** (Ramjee, Kurose, et al, 1994)**:**
This algorithm uses the mean $\hat{d}_i$ and variance $\hat{v}_i$ to estimate the playout delay $\hat{p}_i$ of *i*th arriving packet.

<u>Algorithm 2</u> **Fast-Exponential-Average (F-E-Avg):**
This algorithm is similar as the previous one, but is more sensitive to variable delay.

<u>Algorithm 3</u> **Min-Delay (Min-D):**
To minimise the delays is the purpose of this algorithm.

<u>Algorithm 4</u> **Spike Detection (Spike-Det):**
This algorithm contains a spike detection algorithm. When a spike is detected, the algorithm changes to SPIKE mode, the delay estimate tracks the delays closely. If it is not the SPIKE mode, then the concept of this algorithm can consult the first algorithm.

<u>Algorithm 5</u> **Window** (Moon, Kurose et al, 1998)**:**
As the algorithm 5, this algorithm also includes the spike detection part. During a spike, it simply uses the first packet of current talkspurt in the spike as the playout delay for this talkspurt. If it is not the SPIKE mode, the algorithm calculates the distribution of the *q*th quantile of the last *w* received packets to estimate the playout delay.

<u>Algorithm 6</u> **Enhanced-MOS-based (E-MOS)** (Fujimoto and Murata, 2002)**:**
The main aim of this algorithm is to maximise the perceived quality. It uses equation as

$$MOS = 4.10 - 0.195l + 2.64\times10^{-3} d - 1.86\times10^{-5} d^2 + 1.22\times10^{-8} d^3 \tag{6}$$

to obtain the optimal value of playout delay *d* to maximise the *MOS* value. The loss-ratio *l* is derived as

$$l = l_n + l_d, \tag{7}$$

where l_n is the loss-ratio caused by packet drops within the network, and l_d is the loss-ratio caused by late-arriving packets discarded by jitter buffer as the discard-loss. Using the Pareto distribution, the factor l_d can be expressed as

$$l_d = 100\left(\frac{\hat{k}}{d}\right)^{\hat{\alpha}}, \tag{8}$$

where the value $\hat{k}$ is the minimum packet network delay of the latest w received packets (the maximum w is set to 10000 according to (Fujimoto and Murata, 2002)). Then $\hat{k}$ can be expressed as

$$\hat{k} = \min_{j \in w}\{n_i\}, \tag{9}$$

$\hat{\alpha}$ is derived from the equation as

$$\hat{\alpha} = w\left[\sum_{i=1}^{w} \log\left(\frac{n_i}{\hat{k}}\right)\right]^{-1}. \tag{10}$$

Algorithm 7 **Maximise-MOS (M-MOS)** (Markopoulou, 2003)**:**
This algorithm utilizes the *MOS* function to achieve the best trade-off based on the delay history. The method of playout delay adjustment is similar as Algorithm 5.

4. Proposed Buffer Algorithm to Optimise Perceived Speech Quality

For previous sections, we have reviewed and discussed the previous work on playout buffer algorithm. To develop the function of buffer should be the main purpose for all relative researchers. In this section, we propose our design on adaptive buffer algorithm related with perceived speech quality.

4.1 Modelling Methods for MOS Functions

As described in section 2, the factor *R* can be simply expressed as equation (2), which contains two parameters: *Ie* and *Id.* If we map the factor *R* into equation (4b), then we can obtain a new equation as

$$MOS = 4.409 - 0.0194 \times (Id + Ie) - 0.837 \times 10^{-3} \times (Id + Ie)^2 + 7 \times 10^{-6} \times (Id + Ie)^3, \tag{11}$$

where the function of *Id* + *Ie* works well while $6.8 < Id + Ie < 86.7$. The parameter *Id* can be described as equation (3), and *Ie* as equation (5) when the codec is G.723.1.

The packet loss consists of two parameters: one is the loss caused by packet drops within and another results from packets discarded by jitter buffer due to that the arrival time exceeds the playout time. Thus we can present the loss-ratio factor *l* as

$$l = l_d + l_n, \tag{12}$$

where l_d is the loss-ratio of packets discarded by jitter buffer, and l_n is the loss-ratio of packet drops within the network.

In real network traffic, there is a relationship between the packet playout delay d and the packet lossratio l_d. This is referred to as Pareto Distribution. The relation between these two factors has been presented in equation (8), also the calculation of $\hat{k}$ and $\hat{\alpha}$ has been illustrated in equation (9) and (10) as in section 3.

As only containing an unknown parameter as playout delay d (other parameters as described above can be derived or calculated by packet network delays, packet sequence number, etc. known data), we can find the value d when MOS is the maximum. Then only when a talkspurt starts, the value d is set to as the corresponding talkspurt's playout delay.

For real networks, the playout delay d can be set up in a specified but adaptive range. The limitation of the value d is to ensure the high calculation speed and to avoid the overestimated playout delay. For the possible collision between two successive talkspurts, we choose to increase the playout delay to avoid the loss of packets of the latter talkspurt as (Moon, Kurose et al, 1998).

4.2 New Method on Adjusting Playout Delay During Spikes

For the previous designs as Spike-Det, Window and M-MOS algorithm, the spike detection mechanisms may not be able to follow the change of network delays. To overcome this problem, we propose a new algorithm on adjusting spikes. We refer to the whole concept of our proposal as the Spike-delay Adjustment and MOS-based playout buffer Algorithm (**SAMOSA**).

For ease of understanding, we use the C-language-like pseudo code in Figure 2 to present the algorithm. For the new algorithm, the threshold as the ***ENTER*** is set to as

$$ENTER = \hat{k} - 0.006\alpha^2 + 118 + T(\hat{k}) \text{ ms,} \qquad (13)$$

where
$$\begin{cases} T(\hat{k}) = 0 & \text{if k} \le 150 \text{ ms} \\ T(\hat{k}) = 150 \times \ln(k/150) & \text{if k} > 150 \text{ ms} \\ \hat{\alpha} = 100 & \text{if } \hat{\alpha} > 100. \end{cases}$$

The value ***Y*** is referred to as the current packet network delay.

```
$d_i$: ith packet playout delay
$n_i$: ith packet network delay
$E_i$: estimate playout delay of ith packet -- $d$

IF (mode == NORMAL)
{   IF (**Y** > **ENTER**  && it is the first packet of a talkspurt) *//detected/the beginning of a spike*
    {   var = 0;
        mode = SPIKE;
        $d_i$ = para* $n_i$;
        Recollect data except $\hat{\lambda}$, but save the previous w packets' data  for reference; }}
ELSE
{   var = var/2 + $fabs(2n_i - n_{i-1} - n_{i-2})/8$;
    IF (**X** < **EXIT**)      *// detected the end of a spike*
{   mode = NORMAL;
    IF ( $n_i$ ≤ door × previous NORMAL mode talkspurt playout delay)
      Recollect data except $\hat{\lambda}$, saved reference  data is used for distribution; }}
Measure network delay, $\hat{\lambda}$ and $\hat{\alpha}$ ;
Deduce $E_i$ using *MOS* function;
*IF ( $d_i \neq$ para* $n_i$ && it is the first packet of a talkspurt)*
$d_i$ = $E_i$;
```

Figure 2 : SAMOSA algorithm

When a spike has been detected, the algorithm restarts collecting network delays, and then recalculating the parameter $\hat{\alpha}$, but remains the value of the parameter $\hat{k}$ and saves the collected data (network delay and sequence number) of the latest *w* packets for reference. The playout delay is set to as *para × current packet network delay*, where the parameter *para* is a variable.

The value of parameter *para*, we prefer to use a Linear Regression Trendline (LRT) method to predict the network delay. If the first packet of a talkspurt meets the prediction, then the value of *para* is calculated according to the time distance between the current packet and the last packet of the former talkspurt. We set up this relationship as

$$\begin{cases} para = 1.7 - 0.4\times10^{-3}T & \text{if} \quad T \le 1500ms \qquad (14a) \\ para = 1.1 & \text{if} \quad T > 1500ms \qquad (14b) \end{cases}$$

where *T* is the time distance. While either of these conditions is not satisfied, the *para* is set to 1.1.

The method of detection of the end a spike is similar as the Spike-Det algorithm. We also use the factor *var* as the ***X***. The calculation method is the same as (Ramjee, Kurose, et al, 1994). The value is set to 20.

4.3 Implementation and Performance Comparison of Buffer Algorithms

For our experiments, we directly use trace data proposed by (Sun and Ifeachor, 2002). All trace data was obtained from the Internet connections, as from Plymouth University to Beijing University (PU2BU), Beijing University to PU (BU2PU), PU to Columbia University

(PU2CU) and PU to Darmstadt Univeristy (PU2DU). These traces have different characteristics, such as small end-to-end delay (jitter) in the traces of PU2CU and PU2DU and large/medium delay between BU and PU. Thus we can utilise these traces with diverse network characteristics in our experiments.

The size of the probe packets is set to 32 bytes, and the interval between successive packets is set to 30 milliseconds (ms), similar to the codec of G.723.1. A mean of 1.5 seconds for both talkspurts and silences is chosen as in (Jiang and Schulzrinne, 2000). The playout delay (buffer size) is adjusted only at the beginning of each talkspurt. We also use 30 minutes as the reference tracedata from each trace as above.

Table 1 compares the Average Playout Delay (Avg-P-D), the Amount of packets Discarded by late-arrival (A-Discard), Packet Lossratio (P-Lossratio) and Mean Opinion Score (MOS) of each algorithm in four traces. The first two traces as BU2PU and PU2BU have more variable network circumstances, and the last two traces are more steady networks.

Trace	Algorithm	Avg-P-D (ms)	A-Discard	P-Lossratio (%)	MOS
BU2PU (total voice packets: 29695)	Exp-Avg	312.14	1091	4.60	1.94
	F-Exp-Avg	737.65	74	1.17	1.00
	Min-D	265.58	1890	7.29	2.08
	Spike-Det	247.74	2715	10.07	2.05
	Window	368.08	761	3.48	1.68
	E-MOS	294.49	1296	5.29	2.01
	M-MOS	261.12	1396	5.62	2.21
	SAMOSA	220.28	2137	8.12	2.34
PU2BU (total voice packets: 29683)	Exp-Avg	252.21	1029	6.37	2.22
	F-Exp-Avg	673.96	8	2.93	1.00
	Min-D	184.87	2418	11.05	2.42
	Spike-Det	195.42	2935	12.80	2.27
	Window	277.19	402	4.26	2.20
	E-MOS	212.35	1750	8.80	2.35
	M-MOS	212.90	1491	7.93	2.40
	SAMOSA	179.58	2535	11.45	2.44
PU2DU (total voice packets: 29916)	Exp-Avg	21.84	464	5.36	3.01
	F-Exp-Avg	53.32	91	4.11	3.06
	Min-D	20.61	413	5.18	3.02
	Spike-Det	21.37	949	6.98	2.90
	Window	30.84	335	4.92	3.03
	E-MOS	31.88	164	4.35	3.07
	M-MOS	35.93	131	4.24	3.07
	SAMOSA	24.25	189	4.44	3.07
PU2CU (total voice packets: 29933)	Exp-Avg	52.55	489	1.64	3.26
	F-Exp-Avg	60.10	0	0.01	3.41
	Min-D	47.86	1260	4.22	3.06
	Spike-Det	48.11	1204	4.03	3.07
	Window	48.88	292	0.99	3.33
	E-MOS	76.80	0	0.01	3.39
	M-MOS	50.47	107	0.37	3.39
	SAMOSA	51.01	31	0.11	3.41

Table 1 : Comparison of simulation results of each algorithm

From the results in Table 1, the F-E-Avg algorithm can perform well in the latter two traces, but cause huge playout delays in the former two traces. Thus it is not recommended to use in such variable networks. For other algorithms, some work better in stable networks and some in uneven networks. However, the SAMOSA algorithm performs well in both cases. It can automatically adapt different network circumstances. As for the variable network, it allows slightly increase of packet loss to obtain the considerable decrease of playout delay. In stable networks, it ensures the smaller packet lossratio, simultaneously, achieves the lower playout delay. In other words, it can accomplish a good trade-off as between playout delay and packet lossratio.

5. Conclusion and Future Work

In this paper, we analysed some characteristics of such variable delays. Then we investigated some existing buffer algorithms. As we linked the delay characteristics and algorithm analysis, we proposed a new buffer algorithm based on diverse network features and the view of perceived speech quality. The comparison of simulation results shows that our new algorithm can enhance the perceived speech quality in most network circumstances, and is more efficient than others, so is suitable for any real buffer mechanism.

Due to the fact that the implementation of these algorithms is only based on codec G.723.1, in the future, we will further research on other codecs to prove and enhance our work. Moreover, the work on buffer algorithm will be further deep, other researchers' work may well need to be consulted and studied.

6. References

Cole, R. G. & Rosenbluth, J. H. (2001) *Voice over IP Performance Monitoring.* Journal on Computer Communications Review, vol. 31, no.2.

Fujimoto, K., Ata, S. & Murata, M. (2002) Adaptive Playout Buffer Algorithm for Enhancing Perceived Quality of Streaming Application. IEEE Globecom, 2002

ITU-T Recommendation G.107 (2000). *The E-model, A Computational Model for Use in Transmission Planning.*

ITU-T Recommendation P.862. *Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs.*

Jayant, N. S. (1980) *Effects of packet loss on waveform coded speech*. In: Proc. Fifth Int. Conference on Computer Communications, Altanta, Ga., pp. 275-280

Jiang, W. & Schulzrinne, H. (2000) *Analysis of on-off Patterns in VoIP and their Effect on Voice Traffic Aggregation*. Proc, of ICCCN 2000.

Markopoulou, A. P., Tobagi, F. A. & Karam, M. J. (2002) *Assessment of VoIP Quality over Internet Backbones.* Proc. Of IEEE Infocom

Markopoulou, A. P. (2003) *PhD Dissertation: Assessing the Quality of Multimedia Communications over Internet Backbone Networks.* Stanford University, US.

Moon, S. B., Kurose, J. & Towsley, D. (1998) *Packet Audio Platout Delay Adjustment: Performance Bounds and Algorithms.* Acm/Spring Multimedia Systems, vol. 5, pp. 17-28

Ramjee, R., Kurose, J., Towsley, D. & Schulzrinne, H. (1994) *Adaptive Playout Mechanism for Packetised Audio Application in Wide-area Networks.* IEEE, Toronto, Canada

Sun, L. F. & Ifeachor E. C. (2002) *Prediction of perceived Convestional Speech Quality and Effects of Playout Buffer Algorithms*. IEEE, ICC 03, Anchorage, USA, May 2003, pp. 1-6.

Sun, L. F. & Ifeachor E. C. (2003) *New Methods for Voice Quality Evaluation for IP Networks.* ITC 18. Berlin, Germany, 2003, pp. 1201 - 1210

Telecommunication Standardisation Sector of ITU (1993) ITU ITU-T Recommendation G.114. *Technical report*, International Telecommunication Union.

Security analysis tools – Do they make any difference?

S. Bolakis, S.M. Furnell and P.S. Dowland

Network Research Group, University of Plymouth, Plymouth, United Kingdom.
e-mail: nrg@plymouth.ac.uk

Abstract

Network security is becoming more and more important, as today's large-scale, highly distributed networked systems improve the efficiency and effectiveness of every modern organization. As this integration is accompanied by elevated risks of intrusion and compromise, system administrator's day to day duties are critical. They can keep organisation's networked systems protected against threats, mitigate threats, and improve the overall security level, by adopting various security practices. The security analysis tools, that are available and practically used for implementing these security strategies, is the main field that this paper is dealing with, by surveying views from administrators for the security analysis tactics they currently follow. The results from this survey confirm that information security is the first organizations' priority, and a tactically usage of security analysis tools in daily basis is already adopted in the majority of them.

Keywords

Security Analysis Tools, System Administrator Survey, Security approaches.

1. Introduction

The explosive growth and worldwide reach of the Internet, and the fact that the number of network and computer attack incidents, as well with the exploitable vulnerabilities, are highly increased (Allen, 2001), expanded effort is required for keeping today's information systems secure and protected. Moreover, design errors of software, as well with common software implementation flaws, including lack of input validation or buffer overflows, are discovered and published daily. As a result, system administrators' needs for more efficient ways for keeping secure their networks, have never been in biggest demand, trying to introduce further strategies, enforcing their systems.

A variety of available security analysis tools can give strong preventive steps to protect computer systems and place a high degree of confidence in their security level. At the same time, however, the open availability of such tools provides an opportunity for misuse by unauthorized users, enabling an intruder to get access to a network and control as though he was the administrator (Chiliarchaki, 2000).

The intent of this paper is to evaluate the extent to which administrators are currently aware of security analysis tools, and indeed the extent to which they use them in practice. The investigation involved a survey research, where questionnaires distributed to administration staff within companies and organizations worldwide, exploring their views and securing approaches. The remainder of the paper describes the survey method that was employed, followed by a discussion of the results that observed.

2. Survey Methodology

The survey gathered data from 160 system/security administrators, during June to August 2003, from organizations of different sizes through all over the world. In order to enable collection of administrators' views, a flexible questionnaire had been designed and distributed to the appropriate persons in organizations. This questionnaire was simple to answer, requiring only some pull-down menu choice responses, and able to protect the anonymity of the respondent with optional questions. Specifically, three sections were included:

- **Section A (*System & Personal information*)** concerned with the IT infrastructure of the company being surveyed, and some more profile focused questions. Specifically, questions about the organization, the Network Operating Systems that are used, as well with the number of PCs, users, and servers that the Information System supports, were contained. Moreover, questions about the frequency that administrator's responsibilities are taking place and if system's documentation is available to them, were parts of this section.

- **Section B (*Security analysis tools*)** contained a series of questions that attempted to assess the system administrator's awareness of the security analysis tools. Potential reasons for the existence of exploitable vulnerabilities in information systems and possible gains that an administrator can obtain by using these tools were pointed out, too. Respondents were additionally requested to indicate the frequencies that use and up-date each tool.

  As security tools can be classified in different ways (Macmillan, 2002), for the purposes of this research, they were categorized into seven main categories, which probe for system vulnerability scanning, and enforce the risk assessment of a system's security. These categories are summarized in Table 1.

Network Monitoring Tools	Create maps of the networks and System, and monitor critical parts (*servers, routers, network equipment, workstations and other devices*) for fault, performance, and inspection purposes.
Port/Network Scanners	Automated auditing of a large number of networked hosts, by detecting active hosts on a network, involving mapping and discovery of host based services, for identifying potential security or mis-configuration problems.
Vulnerability Scanners	Identifies any associated vulnerability automatically, and attempt to provide information mitigating and guidance to the user, before an adversary.
Password Crackers	Evaluate weak password selections, verify that users are employing sufficiently strong passwords, and used to attempt privilege escalation.
Packet/Password Sniffers	Employed to capture specific network content. Data collected can be used to evaluate the risks associated with various network protocols, applications and user activities.
Sniffer Detectors	Software that monitors networks for Sniffers. It can monitor networks by scanning for patterns that a cracker might leave while discreetly monitoring data stored on or passing through a network
Intruder detection Tools	Alerts user to attacks on the network in real time, by inspecting the traffic on the wire, and generating alerts if suspicious activities are identified from the audit data generated by the operating systems.

Table 1 : Security analysis tool categories

Respondents' opinions were also gathered concerning the benefits that Attackers may get from the freely obtainable security tools.

- **Section C (*Security policy & other Issues*)** identifies the security measures the organization has employed (*Security policies, contingency/disaster recovery plan, etc.*), as well with the respondent's opinion about the administrator's main priorities, and reasons that may prevent them from keeping today's systems secure enough.

In terms of questionnaire requesting, the strategy that was followed for conducting the survey was through e-mails, requesting for participation. Administrators in universities, organizations and companies around UK, Greece, USA, Spain, China, as well with regional businesses in Plymouth. Posts, requesting participation, were also sent to mailing lists (*Hellug, Full-Disclosure, and SecurityFocus*), newsgroups, web-sites and relevant security forums. As with all web-site polls, the results are not necessarily indicative or representative, and so should be treated with some caution.

3. Survey Results

In each of the organizations that were surveyed, the intention was to obtain the views from one ore more administrators. Accordingly, the achieved answers are covering only the topology of the sub-network that the respondent individual is responsible administrator for.

3.1 Profile of participants

The profile of the companies that participated in this survey is showing in the Figure 1, illustrating the representative distribution, indicated by the number of PCs connected to their organization's network, and the number of users that are currently supported. It is significant that the majority of the respondents (*38,8%*) are using over 20 servers in their systems, indicating the complexity of their administrative systems.

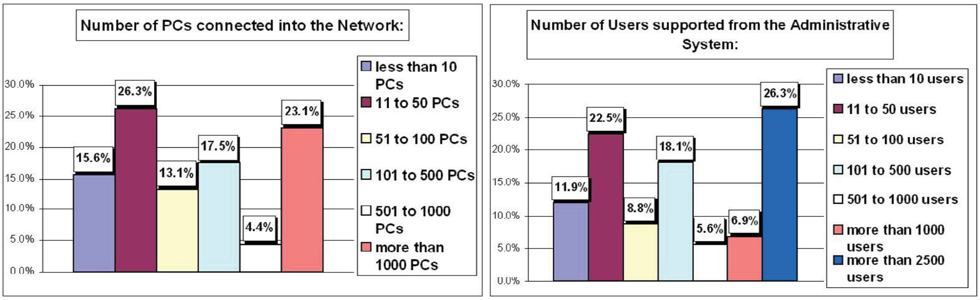


Figure 1 : Distribution of respondent organizations sizes

As system administration is a widely varied task, these individuals have a lot of responsibilities for systems under their control (Allen, 2001). Some of the most common responsibilities are listed in Table 2 followed by the distribution of frequencies that each of them is taking place.

System Administrator's Responsibilities	Daily	Weekly	2 Weeks	Monthly	2-3 Months	Quarterly or Biannually	Annually or More rarely	Never	Not my duty
Maintain file servers and workstations	55.0%	21.3%	5.6%	5.6%	1.3%	1.9%	0.0%	0.6%	8.8%
Install network and application software	17.0%	31.4%	11.9%	18.9%	6.9%	3.1%	3.8%	0.6%	6.3%
Maintain security patches current on servers*	44.4%	31.3%	6.3%	10.0%	1.3%	3.1%	1.3%	0.6%	1.9%
Update system for security holes found	46.9%	25.6%	5.0%	15.0%	1.3%	1.3%	2.5%	0.6%	1.9%
Review activity logs for suspicious activities	58.8%	20.6%	5.0%	7.5%	1.3%	1.3%	0.6%	3.1%	1.9%

* Approximately the period of time between the patch release date and the installation date.

Table 2 : Frequency of administrator responsibilities

Almost 3 of 4 of the respondents are maintaining file servers and workstations, or maintain security patches current on servers, or update system for security holes found, or review activity logs for suspicious activities, in daily and weekly basis. When a patch is published, 44.4% of the administrators make it a priority to apply it in the same day. Quite high the percentage of the respondents who are applying immediately the updated applications for security holes found in the same day, too. Firewall reports or host activity logs, are also reviewed by the majority of administrators frequently (*58.5% in daily basis*). In large-scale organizations with more complicated installations though, these responsibilities can be shared between several people.

From a security perspective, positive common findings were that, 52.3% of the respondents are using both firewall and content filter in their networks, and 33.5% only firewalls. The majority of the organizations (*56.3%*) are providing also their administrators with documented information, helping enlighten their duties. When there are several administrators in a system, it is more important to have everything documented, for preventing technical confusions and achieving effective security policies, as was mentioned by respondents' given commends.

3.2. Security analysis tool findings

Taking into account the plethora of security analyzer tools available to help administrators do their job more quickly and effectively, almost the total of them (*95%*) are aware of their existence and use them in practice (Figure 2). Organizations comprehend the value that security analysis tools can give to their information technology protection.

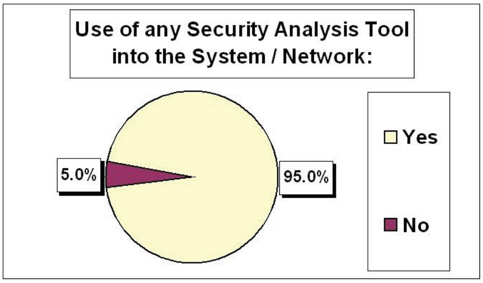


Figure 2 : Use of security analysis tools

Most administrators have developed their knowledge from experience and word of mouth, not from their organization's training procedures. In this survey, only 38.1% of the respondent organizations provided any training for how to use security tools, indicating the lack of security awareness programmes.

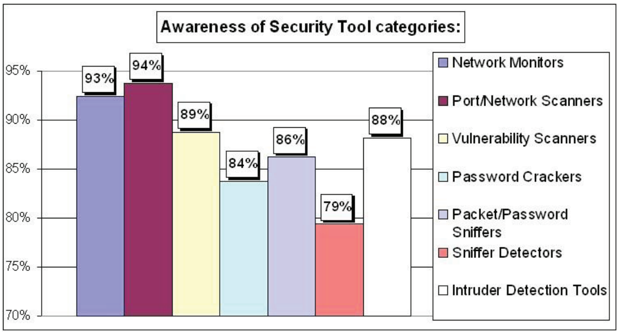


Figure 3 : Awareness of security tools

Although, almost every administrator is aware at least of one of the seven major tool categories. The percentage distributions of administrators' awareness are illustrated in the above figure (Figure 3). These percentages indicate that at least 79% are aware in the total of security analysis tools that are available today. More analytical results about the frequencies at which the respondents are using and updating each category of tool is provided in Table 3. Respondents were restricted to indicate their answers, by select only one of the suggested frequency's choices, for each of the tool categories.

Security Analysis Tool Category	Frequency of	Daily	Weekly	2 Weeks	Monthly	2-3 Months	Quarterly or Biannually	Annually or More rarely	Never	Not my duty
Network Monitoring Tools	Use	59.7%	6.5%	5.2%	5.8%	0.6%	2.6%	3.2%	10.4%	5.8%
	Update	15.2%	13.9%	5.3%	17.2%	6.6%	13.2%	10.6%	11.3%	6.6%
Port/Network Scanners	Use	23.2%	28.4%	6.5%	17.4%	4.5%	5.8%	6.5%	5.2%	2.6%
	Update	10.0%	15.3%	5.3%	21.3%	8.0%	13.3%	14.7%	9.3%	2.7%
Vulnerability Scanners	Use	16.3%	16.3%	5.2%	23.5%	3.9%	11.8%	5.9%	12.4%	4.6%
	Update	10.0%	17.3%	4.7%	22.7%	6.7%	10.0%	8.7%	17.3%	2.7%
Password Crackers	Use	6.8%	6.1%	2.7%	14.2%	7.4%	14.2%	18.9%	25.7%	4.1%
	Update	4.2%	5.6%	1.4%	12.0%	6.3%	16.9%	21.8%	28.2%	3.5%
Packet/Password Sniffers	Use	20.8%	10.7%	3.4%	8.7%	4.7%	8.7%	14.1%	24.2%	4.7%
	Update	9.0%	8.3%	4.2%	11.1%	4.9%	15.3%	18.8%	25.0%	3.5%
Sniffer Detectors	Use	24.5%	3.5%	0.7%	4.9%	3.5%	6.3%	15.4%	37.1%	4.2%
	Update	8.6%	5.0%	1.4%	11.4%	1.4%	7.9%	20.0%	40.0%	4.3%
Intruder Detection Systems	Use	53.0%	8.6%	0.7%	4.0%	2.0%	1.3%	6.0%	21.9%	2.6%
	Update	18.4%	17.7%	3.4%	15.0%	2.0%	6.8%	8.8%	24.5%	3.4%

Table 3 : Frequencies that administrators use and update each tool category

The two most used tools on a daily basis, are Network Monitoring Tools (*59.7%*) and Intruder Detection Systems (*53.0%*), indicating that the first priority is attack detection. Port/Network Scanners is the next most used category, as more than a half (*51.6%*) are using them daily or weekly, when vulnerability scanners are used more rarely by administrators. Password crackers seemed to be the less preferred security analysis method. Sniffer Detectors is the category of tools that surprisingly 37.1% of the respondents are never using in their systems, assuming that the high knowledge of networking that is required to get the maximum value out, is the preventing reason. It is also remarkable that their users never update five of seven security tools categories. Whereas, some tools do not need to be updated frequently, as the technique they use will not alter (e.g. password crackers follows the same methodology, with no need for keep them updated), the success of other tools (such as vulnerability scanners) depends upon up-to-date information.

The responses regarding the technical problems that administrators are often facing by using security analysis tools also have significant findings. More than a half of the respondents have faced problems during the installation of a tool, whereas 71.3% have faced problems during their attempts to configure one. Along with this, 3.1 of 5 declared that security tools can sometimes be oversensitive, and 3.6 of 5 felt that they can detect problems that not exist. In more technical subject-matter, 57.3% of the respondents agree that the main reason for existence of vulnerabilities in today's systems, and the weakest point of technology at the same time, is the software/protocols implementation (Figure 4), while other responses proclaim system or network configuration weaknesses, as potential reason.

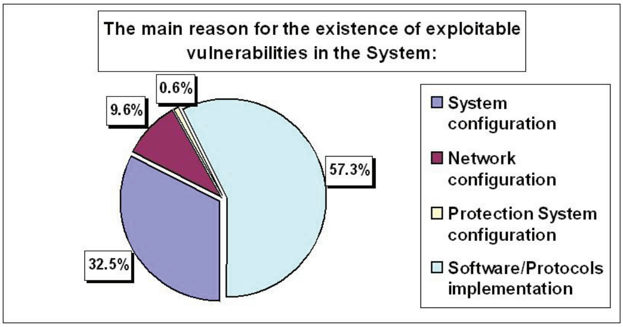


Figure 4 : Practical challenges of keeping a system secure

In the meantime, by using security analysis tools, administrators may have a lot of gains, including technical skills, better performance, and handle security risks and threats more efficiently, as their answers represented in the next figure (Figure 5).

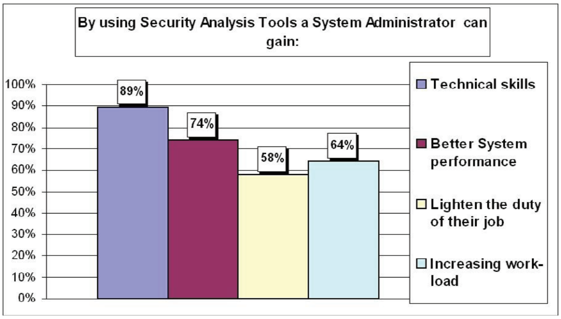


Figure 5 : The possible gains by using security analysis tools

The survey found that the administrators' main priority is to keep abreast of new vulnerabilities and attack methods, according to 54.1% of the respondents, while 45.9% maintain a basic level of security on their systems/networks, regarding to their security policies. Furthermore, administrators seemed to strongly agree that insufficient time (3.9 of 5) and knowledge (3.6 of 5) are the main barriers for sensibly keeping a system secure, when half of the respondents agree that awareness is a problem. Most, however, seem to have no strong views on these subjects.

3.3. Related issues

For information security to be effective, policies can establish guidelines for how the company handles and protects its data, from who makes sure software patches are installed to how often passwords should be changed (Allen, 2001). This survey confirms that information security is very high in profile at board level, as 81.1% of the respondent's businesses are following and often updating their security policies. Moreover, the majority of organizations (*79.9%*) are going one step further, by introducing disaster recovery plan in their security strategy.

Finally, 6 of 10 respondents believe that security analysis can help more in assessing and managing the risks of their organizations, following proactive approaches of security. As concern the views of the respondents in the question '*Who is getting more benefits from the freely obtainable security analysis tools, Administrators or Attackers?*' answers were definitely divided. Benefits can be equal for both of them. As it is pointing from respondents, a tool is just that a tool, and is up to the user to make effective (and appropriate) use of it.

4. Conclusions

This paper has considered the use of security analysis tools by system administrators in practice, in relation to relevant security issues. Although in the survey presented here, the proportion of responses from security-aware administrators is likely to have had significant impact, the results provide some interesting indicators, which are worthy of further consideration and investigation.

The most significant point is that the majority of the administrators are frequently using plethora of tools in their daily duties, indicating that the awareness and the effective countermeasures is the solution for keeping an information system secure. Hopefully, the answer to the question of whether such tools make any difference can now clearly been given, through the survey results. Today these tools are not used only as a single solution in the organization's needs, but the idea to their usage have been integrated to further utilization of security tools compilation, for meeting better its needs.

Information security threats and vulnerabilities, as well as their countermeasures will continue to evolve at any security plan implementation, as security is a process. Most of the 160 respondents appear to be highly aware of the weaknesses of their systems, and the challenges they have to face. System administrators can clearly identify that the lack of knowledge, insufficient awareness, and poor security strategies can be the main reasons that lead to undesirable impacts for a company, even more to serious security incidents. Concluding, as information security is more about striking a balance between people, processes and technology than it is about maintaining a relevant network security toolkit, attention has already be given in this direction.

5. References

Julia H. Allen (2001). 'The CERT Guide to System and Network Security Practices'. CERT Coordination Center. ISBN: 0-201-73723-X. Addison Wesley Professional.

Macmillan (2002). Macmillan Computer Publishing. 'Maximum Security: A Hacker's Guide to Protecting Your Internet Site and Network'. *Network Security Library*. URL: http://www.secinf.net/.

Pelagia Chiliarchaki (2000). ‘Security Analysers-Admin Assistants or Hacker Helpers?’. *MSc Project’s Thesis*. University of Plymouth. Plymouth. United Kingdom. September 2000.

A Generic Framework for the Prevention and Detection of Insider Misuse

M.J. Coussa, A.H. Phyo, and S.M. Furnell

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: nrg@plymouth.ac.uk

Abstract

Security professionals and government survey results have addressed the external threat to information systems on numerous occasions. However, relatively little consideration has been given to the more perilous threat instigated by insiders. The main characteristic that distinguishes insiders from any other category is their possession of legitimate access, and they are capable of misusing their privileges in a way that is not permitted by the organisational policy. This paper discusses the techniques employed by operating systems and intrusion detection systems and confers about their applicability to the problem of detecting insider misuse. It was realised that the majority of misuse cases were due to the inherent organisational structure and insufficient internal controls. As a result, a generic security framework was presented, building upon new methods of preventing and detecting misfeasance based on the granular approach to granting permissions, using role-based access controls. The suggested framework also realised the significant need for application-level audit trails, which can be used in conjunction with operating system audit trails to assist in providing an enhanced role-profiling technique.

Keywords

Insider misuse, misfeasor detection, intrusion detection, role-based access control, auditing, audit trails, profiling.

1. Introduction

Threats to information systems have existed for many years, arguably since the invention of the electrical digital computer in 1939. While the media has been concentrating its attention on the threat brought about by external intruders, relatively little consideration has been given to what can be alleged as a greater threat; insider misuse. In an attempt to differentiate between the two threats, Anderson's seminal paper categorised internal users into three distinct classes, namely the masquerader, the misfeasor, and the clandestine user (Anderson, 1980). Prior to the publication of Anderson's paper, system security procedures focused on denying access to sensitive data from an unauthorised source. Suffice to say, such critical information as bank records and payroll databases need not only be protected from external threats, but even from the more perilous misfeasor. A misfeasor can be defined as an individual or member of an organisation who is granted legitimate access to the information and/or information system by the organisation that he/she belongs to, yet misuses these privileges in a way that is not permitted by the organisational policy. The key trait that differentiates an external intruder from a misfeasor is the misfeasor's possession of *legitimate access* to the information system. The different technologies and mechanisms that have been introduced to the field of intrusion detection, and available to current operating systems, have primarily targeted their mechanisms towards preventing, monitoring, and/or detecting external intruders. This paper examines these techniques from the perspective of insider misuse, and puts forward a generic framework to assist in the prevention and detection of misuse.

2. Insider Misuse – A Perspective

The history of insider misuse possibly dates back to the 19th century when employees in France secretly opposed to the new technology of an automatic loom using punched cards, and as a result changed pattern specifications in the Jacquard loom, thus causing failures and economic loss (United Nations, 1990). Ever since this incident occurred, insider misuse crimes have been committed in nearly every environment, and this extent of insider threats has been addressed on numerous occasions in corporate and government survey results. The CSI/FBI security survey is just one report that portrays a picture of the damage caused by insiders. In 1998, the CSI/FBI Computer Crime and Security Survey reported the average cost of an outsider (hacker) penetration at $86,000, while the average insider attack cost a company $2.8 million (Richardson, 2003). This was followed by the 1999 survey which stated that 55% of the reported attacks were caused by insiders. These figures have changed since then, only to show that the cost of insider abuse surpasses the cost of system penetration by outsiders (Table 1).

	System penetration by outsider	**Insider abuse of net access**	**Unauthorised insider access**
2003	2,754,400	11,767,200	406,300
2002	13,055,000	50,099,000	4,503,000
2001	19,066,600	35,001,650	6,064,000
2000	7,104,000	27,984,740	22,554,500
1999	2,885,000	7,576,000	3,567,000
1998	1,637,000	3,720,000	50,565,000
1997	2,911,700	1,006,750	3,991,605

Table 1: The cost of computer crime (figures in US$).

While statistics prove that insider misuse activities remain the most common source of computer crime, it is the special characteristics the insiders exhibit that allow them to perform such intentional, and sometimes unintentional, acts of misuse. Being the member of an organisation, the foremost such characteristic inherited by an insider is *legitimate access*. This type of authoritative access enables an insider to build up a combination of skills, knowledge, resources, authority, and motive, which can potentially be the climax of threats to an organisation, especially if this insider possesses more access than is required to fulfill his/her duty. Regardless of the type of authority a legitimate insider is granted, this user would remain capable of misusing the assigned privileges. Misuse is further made possible due to the inherent structure of an organisation and the lack of sufficient internal controls. In this case, a classic example of insider misuse would be fraud, which can be easily committed if internal controls are weakly designed and inadequately placed within an organisation. Paradigms of fraudulent acts of misuse include:

- Ghost employees – A payroll clerk can create a ghost employee in the payroll database. This ghost employee would then be paid a normal salary.
- Falsified payments – Initiation and authorisation of transactions, such as monetary payments in a bank, should be carried out by two separate individuals so that unauthorised modifications can be detected.

To mitigate general security threats to their information systems, organisations pursue a series of steps that constitute a security management framework: prevention, detection, reporting, and response.

3. Security Management

The four independent functions that constitute security management in an organization are illustrated in Figure 1.

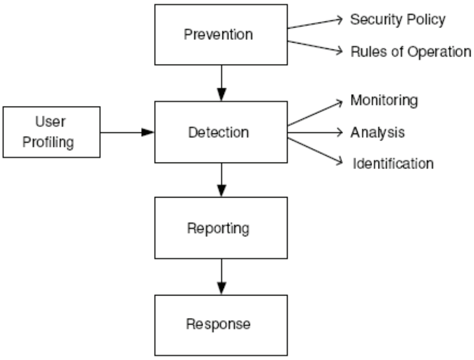


Figure 1: A security management framework.

The two most important functions are *prevention* and *detection*, without which latter processes cannot perform accurately. Prevention mechanisms such as access controls have been, by large, implemented in Operating Systems (OS), although they have also been addressed along with detection mechanisms in current Intrusion Detection Systems (IDS). While the security framework has been primarily used to counter outsider threats, designing and implementing the first two functions, for any form of intrusion or misuse, is a task that requires insight and careful deliberation. The following sections detail the applicability of access controls and IDS mechanisms to the prevention and detection of insider misuse.

3.1 Access Control

Access controls are a means by which a user's actions, operations, and access to objects are controlled. Over the years, various models for access control have been recognised and implemented in operating systems. The three foremost methods of access control are discretionary, mandatory, and role-based.

3.1.1 Discretionary Access Control

In a system based on Discretionary Access Control (DAC), access permissions are governed by the owner of the object. This grants the owner the right to assign permissions to other users at will. The owner of the object is responsible for specifying the subjects that are permitted to access the object, and hence is responsible for maintaining the Access Matrix (AM) that is associated with each object. The DAC models implemented in many operating systems have very limited capabilities when it amounts to preventing insider misuse. Using the AM, the owner can grant permissions to other subjects at his/her discretion. This presents a threat because the granted users might exploit their privileges to misuse the object. However, objects are usually owned by the organisation itself and not the actual object creator. In DAC, since the creator of the object becomes the owner, if any privileges have been accidentally granted to other users, these users can then misuse their privileges. Thus saying, DAC allows for the most basic form of security and usually most misfeasors can bypass this stage to be in a position to misuse their privileges.

3.1.2 Mandatory Access Control

While DAC is identity-based and owner-controlled, the Mandatory Access Control (MAC) model is lattice-based and policy controlled. With MAC, access rights are based on the classification of subjects and objects into different security levels. Every subject or object is marked with a sensitivity label consisting of an access class label and a list of categories label. Combining these two labels allows for a secure access control implementation. The access class labels can be used to restrict an insider's access to information based on the 'need-to-know'. For instance, a user cleared at level C (classified) would not be allowed to read anything tagged with a higher access class label, or even write to anything tagged with a lower access class label. While this might be useful for controlling the flow of information for secrecy, the fact remains that a misfeasor cleared at level C might misuse their granted privileges, given that this user has legitimate access to the data. Furthermore, MAC allows for the integrity of information to be ensured by controlling information flow. Yet again, a legitimate insider might still misuse any granted privileges on the same level of integrity. The MAC model has been developed for multilevel security systems, which has also caused it to be a nonflexible approach especially in commercial environments.

3.1.3 Role-Based Access Control

RBAC was developed by Ferraiolo and Kuhn (1992) as a viable alternative to the traditional access control mechanisms of DAC and MAC. The underlying principle of this model is that the decision to grant access permissions to objects is based on the job function of the individual, otherwise termed as the 'role', and is dictated by the presence of a security policy (defined by the organisation). In this way, assigning access to objects and applications within the organisation could be based on the role of the user. The RBAC model can accommodate both military and non-military organisations because it takes on the structure of the organisation and the policy governed.

Although the access control methods discussed herein are widely utilised for the prevention of intrusions, they cannot be effectively associated with misfeasor prevention. The reason for this lies behind the fact that insiders possess legitimate access to the object(s) they are trying to access, thereby bypassing the protection mechanisms. Any misuse activity beyond this prevention stage can compromise security and would have to be detected.

3.2. Intrusion Detection Technologies

The notion of intruder monitoring instigated by Anderson (1980) led to a dramatic rise in the popularity of surveillance and security systems, generally dubbed intrusion detection systems. Since then, research institutes, universities, and commercial companies have put forward many IDS prototypes which are aimed primarily at detecting vulnerabilities and access violations. The field of intrusion detection has followed two complementary methods to detect intruders:

- **Anomaly detection** – Otherwise known as behavior-based detection, anomaly detection is a technique based on the observation of deviations from a 'normal' usage pattern. Initially, a profile is built for the system or user that is being monitored. A profile could define the behavior of subjects and/or objects, such as users, groups, files, and other resources. Detection is then performed by continuously monitoring the activity of the subject/object, attempting to detect significant deviations from the original profile.

- **Misuse detection** – This form of detection, also termed knowledge-based, differs from anomaly detection in the sense that intrusions are detected by comparing activities against a collection of known attacks. However, this requires the system to have a knowledge base of well-defined attacks on vulnerabilities. The concept behind misuse detection is that there are many ways of representing an attack in the form of a pattern or signature, so that variations of the same attack can be detected.

Anomaly detection is usually pursued by means of statistical profiling. This method deals with generating behavior profiles for users (through the collection of audit trails) over a period of time (can range from short to long). The analysis engine then calculates the variance of the current profile and compares it to the original profile. Statistical analysis is aimed at detecting any deviations from the 'normal' behavior of a user, and can additionally detect known and unknown forms of misuse, thus deeming it a reasonable solution for insider misuse detection. Nonetheless, building a 'normal' profile for an insider is not an easy task because the insider can misuse the system without exhibiting any abnormal deviation. Contrary to detecting misuse through statistical profiling, misuse-based techniques depend on the presence of a knowledge base containing rules/signatures of specific attacks and system vulnerabilities. However, misfeasors are considered legitimate users who "misuse" their privileges instead of exploiting vulnerabilities. This technique can be applied to misfeasor detection; yet, the chances of detecting a 'misuse' pattern in an audit trail are relatively modest.

Although anomaly detection can prove to be a good approach to detecting insider misuse, its weaknesses can be strengthened by improving user-profiling and improving the accuracy of the system so that false-alarms could be minimised. Current detection engines rely mainly on the presence of OS audit trails; however, insiders tend to misuse their privileges in the context of the application they are using, thus justifying the need for collecting audit trails at the application level (the area of application-level auditing is new and has been the subject of little research (Almgren and Lindqvist, 2001)). The anomaly detection engine can use application-level audit trails to monitor the behavior of the insider in the context of an application. Rule-based techniques on the other hand (another form of misuse detection), could be used to detect fraud linked with insider misuse from the application perspective. However, this again requires that the knowledge base be frequently updated with misuse patterns that can be discovered in applications. These drawbacks, in addition to previously discussed weaknesses that are inherent in an organisation, i.e. weak internal controls and the lack of granular permissions, attribute to the growing scale of undetected misuse. This requires a generic security framework that can be matched to the structure of an organisation, and modeled in such a way as to minimise misfeasance.

4. Framework for Preventing and Detecting Insider Misuse

The proposed framework illustrated in Figure 2 comprises of elements that are arranged to match the flow of a typical computer session. The framework assumes that all users that gain access to the system are legitimate users. Following the login process is an OS with an RBAC implementation, which controls the access permissions to objects and applications. Throughout the session, an auditing subsystem audits the user's actions at two levels: OS and application level. Considering the fact that insiders tend to misuse their privileges in the context of applications, there is a mandatory need for high-level application auditing. The anomaly/misuse detection engines make use of the generated audit trails in an attempt to detect misuse in applications. Finally, to support the detection of misuse, a file integrity checker ensures the overall integrity of the system by monitoring for specific file/directory modifications. The following sections detail the main elements of the framework.

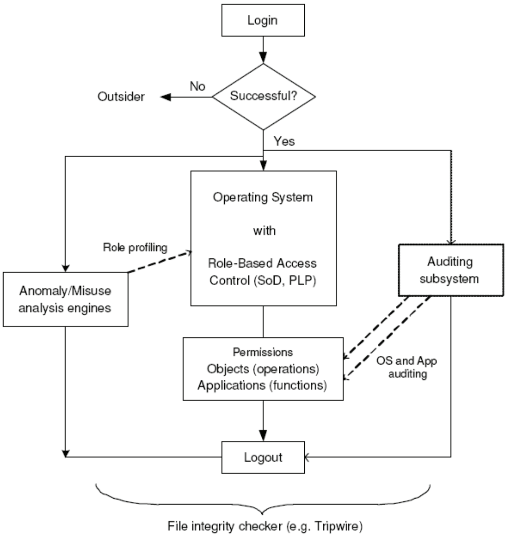


Figure 2: The generic framework for insider misuse prevention and detection

4.1 Role-Based Access Control

At the heart of framework is the RBAC model. The RBAC architecture provides many improvements over DAC and MAC, with the principal benefit being granularity of access control. From the perspective of insider misuse, the role assignment step can limit the insider to performing functions which are solely defined in the security policy as necessary to perform the required job. Furthermore, by deploying RBAC on a different level of abstraction, in an application for instance, a granularity of access permissions can be achieved. For example, in the OS, operations would include read, write, execute, while in an application, operations would include add, delete, modify.

Other advantages of RBAC include flexibility and low management overhead. Flexibility deals with RBAC's ability to enforce the Principle of Least Privilege (PLP), and static or dynamic Separation of Duties (SoD). The PLP states that users should be granted the least amount of privilege required to accomplish their tasks, which is very similar to operating on a 'need-to-know' basis. One might argue that an insider already has the privileges to perform operations. This, however, is not true, since an insider is usually granted all or a subset of privileges which are liable for misuse. RBAC ensures that the least number of privileges are provided (this does not rule out the possibility of an insider misusing the least of granted privileges). The SoD feature ensures that no single individual is permitted to execute all the transactions within a particular set of transactions. To some extent, this can act as a deterrent to fraud because collusion with another individual is required to complete the fraudulent act.

4.2 Enhanced Profiling Techniques

An anomaly engine can detect future events based on what is happening or what has already happened, thus deeming it a proactive solution to insider misuse. However, as mentioned earlier, building a 'normal' profile of an insider is not an easy task. One solution to this problem is to increase the threshold, which will result in increased false alarms. Alternatively, an anomaly engine can be used in this security framework to build a profile of each role. It can be argued that all users within the same role can exhibit a similar activity and behavior. This leads to the suggestion of profiling the user in the context of the role. A role profile could include objects/directories accessed, object operations, application usage, time of day, session duration, and CPU, I/O, and memory usage. Again, deviations can be expected, hence it has to be assumed that the role profile might change. This can be resolved by allowing the statistical engine to calculate an aggregate value for all the user profiles for that specific role. This would then assist in detecting whether all the users in that role exhibit a similar deviation. In the latter case, the deviation threshold could be lowered so that fewer false alarms may be generated.

4.3 Application-Level Auditing

The presence of OS and enhanced application level audit trails can ease the detection of insider misuse. OS audit trails are reasonably good for normal command-line monitoring, although insiders have a tendency to misuse their privileges within an application (e.g. fraud), therefore it is important to have specialised, high-level audit trails. The suggested application audit trail could contain:

- User ID (real and effective) and role ID
- Fields and functions accessed (when this log was created) and field ID
- Level of severity
- Application details (e.g. name, version, etc.)
- Type of error and error ID generated by application (e.g. invalid data type or input range)
- Name of network host on which this application was invoked

Such an audit trail could be used in conjunction with the OS audit trail to assist the SSO in holding a person accountable for any acts of misuse. Application-level audit trails could also be utilised to profile applications used in each role (users in one role are likely to access similar functions and fields).

5. Conclusion

A major security requirement in organisations is the presence of a framework that can detect both external and internal threats. External threats have been addressed on several occasions by intrusion detection systems, yet, these systems cannot detect insider misuse proficiently. The liability falls partially on the detection systems because no serious effort has been presented to detect this class of internal users. The framework presented in this paper identifies a model for preventing and detecting insiders based on the granular approach to granting permissions, using role-based access controls. The suggested framework also realises the significant need for application-level audit trails, which can be used in conjunction with OS audit trails to assist in providing an enhanced role-profiling technique. Misuse patterns, however, are countless, and the absolute prevention of insider misuse can be considered unachievable. The fact that some insiders can still pass through undetected leads to additional tasks that have to be carried out to mitigate the threat. These include enforcement of strong physical security parameters, efficient management, conducting employee checks, and the formulation of strong policies.

6. References

Almgren, M. and Lindqvist, U. (2001), Application-Integrated Data Collection for Security Monitoring, *in* 'Recent Advances in Intrusion Detection (RAID)', Davis, California, pp. 22-36.

Anderson, J. P. (1980), Computer Security Threat Monitoring and Surveillance, Technical report, James P. Anderson Co., Fort Washington, PA.

Ferraiolo, D. and Kuhn, R. (1992), Role-Based Access Control, 'Proceedings of 15th National Computer Security Conference'.

Richardson, R. (2003), 'Eighth Annual CSI/FBI Computer Crime and Security Survey'

United Nations (1990), 'International review of criminal policy – United Nations Manual on the prevention and control of computer-related crime'. Date visited: 23/06/03.
http://www.uncjin.org/Documents/EighthCongress.html

Analysis of insider misuse in commercial applications

F.J. Portilla, S.M. Furnell and A.H. Phyo

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: nrg@plymouth.ac.uk

Abstract

Insider misuse is one of the main concerns of modern organisations, as insider abuse can cause great damage to an organisation. Its detection is a complex task that requires an adequate characterisation of the elements involved in it, including users and applications. A proper characterisation can simplify to the identification of abnormal behaviour that denotes a misuse, therefore increasing the effectiveness of detection tools. This paper focuses in the analysis of the capabilities provided by commercial applications and how insiders can possibly use them in an inappropriate way, exploiting flaws in these applications or abusing of their legitimate rights. In the same line, a special attention is given to the misuse of data access facilities, regarding to different kinds of data and the way it is accessed. Besides, general guidelines for prevention and detection are given, whereas traditional detection mechanisms are assessed as methods to satisfy the requirements of insider detection.

Keywords

Insider misuse, intrusion detection.

1. Introduction

In the last years, security in computer systems has become a main concern of organisations due to the increasing number of incidents. Although the number of external attacks is larger than the insider misuse, insiders cause more impact to systems and organisations worldwide. Unauthorised insider access and net abuse caused around $12 million combined losses in the U.S. in 2003, whereas system penetration by outsiders caused $2.7 million (Power 2003).

Insider security is mainly defined by access control, enforced by administrators, to restrict user privileges regarding to the organisation's policy. However, configuration errors in these controls and flaws in operating systems and commercial applications make the use of additional security mechanisms necessary, in the form of intrusion detection tools.

Insider misuse is more difficult to detect than external penetration, because the differences between legitimate and illegitimate operations are not always clearly identifiable. Intrusion detection systems (IDS) analyse monitored data in order to determine the existence of a misuse (Phyo and Furnell 2003). Some misuses involve design, implementation or configuration errors in applications. IDS can apply misuse detection techniques to identify these misuses, considering known patterns of abuse (attack signatures). However, this approach lacks the adequate level of abstraction that is required to recognize any possible misuse, as new vulnerabilities appear. Anomaly detection, in contrast, identifies abnormal patterns in user interaction. Although it is more difficult to implement (it requires training the system, for example), it represents a more general approach, and it could possibly detect more kinds of misuse, involving vulnerabilities or not. These two approaches can be implemented using several techniques, including artificial intelligence and data analysis mechanisms (Noel 2002).

The abuse of legitimate rights represents a challenge to traditional detection systems. This kind of misusers can often be effectively identified through anomaly detection if the patterns of user behaviour differ in some extent. However, in some cases there is no essential difference between normal and abnormal uses of applications. This is particularly observable in applications involving data access such as databases. The accessed data is meaningful for the user, but not to the underlying system, where detection is based. Misuses that do not involve a direct violation of access controls or cause identifiable effects on the system can hardly be detected. Surveillance seems to be necessary, but inefficient and technically difficult to achieve in most of systems.

It is essential to know in which inappropriate ways insiders can use the capabilities of commercial applications, to elaborate mechanisms that could successfully detect these misuses. This paper describes how features in common applications can be misused, and suggests a functional classification. Recommendations for prevention and detection derived from the analysis of these misuses are given. The study pays special attention to misuse of meaningful data access, which is considered as a feature of databases.

2. Methodology

The analysis is exclusively focused in actions that insiders can perform using legitimate privileges in commercial applications. Hence, operating system and internal penetration effects, like bypassing authentication and possibly taking administrative control of systems and applications, are not considered. A user in this situation is a ‘masquerader’, from Anderson’s classification (Anderson 1980), and differs from the profile of ‘misfeasor’ this study regards to. Misfeasors abuse of their privileges within the system, using available tools and applications in a way they were not designed for. However, misuses previous to penetration, performed by legitimate users who intend to gain additional privileges, are actually considered.

A part of this study of insider misuse is based upon the identification of vulnerabilities that can exist within the application. The analysis of information about known vulnerabilities is freely available from many Internet sources. Besides, details about the features that common applications provide are necessary to evaluate the abuse of privileges in them.

An effective control of insider activities in an organisation requires an adequate definition of a security policy and its implementation in the system. Policies are an expression of the organisation’s security requirements, so they can vary from one to another environment. A misuse is any operation that is not allowed by policy, even if a user has the rights in the system (Chung 1997). Policies impose constraints to the users, but its translation into the context of the system may not be totally successful, and the threats to a system may change regarding to the established policy (Neumann 1999). There are several reasons that explain why access control can fail in providing the security specified by policy:

- Access control cannot be enforced selectively to each system object to guarantee the exact level of security for each user in any possible situation or scenario, and therefore there is a trade-off between security and flexibility imposed by traditional approaches (mandatory and discretionary access control).
- Applications inevitably have flaws that can be used to evade the controls imposed by the system, even if their implementation was perfect. In addition, applications cannot always be configured to provide control over the features that can be accessed by different users.

In response to the necessity of adequate access control mechanisms, Role-Based Access Control (RBAC) is a potential solution, since it defines more meaningful mechanisms of access control, with many advantages over the traditional approaches. However, it requires the effective identification of roles within a system, and a proper assignation of rights (Ferraiolo et al. 1995). Besides, detection mechanisms are a requirement to solve the gap existent between the definition and the enforcement of the organisation’s policy within the system. Anomaly detection techniques can incorporate the concept of roles to be more precise in their profile specifications.

There are three main aspects that determine the scope of a misuse: actions that are taken by the user, the data accessed, and the manner it is accessed. This study refers to the manner common applications can be used and the way data is accessed to perform illegitimate actions.

3. Misuse of features in common applications

The following classification establishes a functional division of applications, regarding to the typical features that can be object of misuse by insiders. Commercial applications nowadays, however, normally implement many of the features that fall into several of these categories.

Table 1 presents this classification, regarding to facilities provided to conventional users by the application. Although some of the operations that involve vulnerabilities are common to insider and outsiders, the features that can be exploited only by outsiders, or the ones that are provided by the operating system, are not considered.

LEGITIMATE ACTION	MISUSE	
Client/Server applications		
Message exchange	Unusual exchange of messages that degrades performance	V
Connectivity to server	Exceeding possible number of connections to cause a denial of service	L
Execution of tasks	Executing privileged procedures	V
Word processors		
Writing a document	Insertion of illegal content	L
	Insertion of malicious code	L
	Link to restricted information in other document	V
Reading a document	Bypassing permissions to obtain privileges in the document	V
Mail clients		
Sending and receiving emails	Inserting illegal content	L
	Setting up a remote attack	L
	Private use/gain	L
	Overload of emails to degrade network performance	V
Browsers		
Browsing the Internet	Access to illegal content	L
Access to cached files and history	Displaying other user's viewed files and previous accesses	V
Multimedia players		
Playing video/audio file	Viewing illegal content	L
Programming tools		
Developing programs	Creation of malware	L
Displaying memory segments	Access to memory segments with sensitive info	V
General purpose applications		
Reading a file	Opening a temporary files with sensitive information	V
Writing a file	Modifying temporary files to change program flow	A
Input strings	Buffer overflow for elevation of privileges	A
	Buffer overflow for code execution	A
	Buffer overflow for denial of service	A
Database Applications		
Data access	Use of legitimate access rights to access data improperly	L

L: Use of legitimate rights V: Exploitation of vulnerability through common interaction
A: Exploitation of vulnerability through advanced mechanisms

Table 1 : Misuse of typical application features

The analysis of these illegitimate operations allows identifying similar practises among different kind of applications, involving:

- *Illegitimate content*: users could insert or visualise inappropriate content, from pornographic material, deception, and defence of criminal actions to jokes and spread of malicious code.
- *Support of private activities*: users can utilise provided resources to support personal operations, such as covert businesses or simply for entertainment.

- *New attacks*: users could deliberately weaken system defences or provide information that makes posterior attacks (internal or external) easier.
- *Weak permissions*: users could take advantage of vulnerabilities in access control in applications that establish trusted paths to restricted objects within the application or in the system.
- *Saturation of service*: users can collapse an application (normally server) or the system by deliberately consuming their resources through an excessive demand or complex requests.
- *Input strings with embedded code*: users can insert malicious code to cause a buffer overflow and execute it in a privilege environment. This is a very common vulnerability.
- *Unforeseeable interaction*: users can exploit particular bugs, whose nature is very diverse. Unhandled exceptions can be provoked through an unusual interaction (fast scrolling, display of modal windows, repetition of events).

Some of these misuses can be detected in different ways, depending on their characteristics:

- Logging unusual events, such as application crashes, access to particular files (e.g. password files), excessive use of processing cycles and network resources, rapid spawning of child processes, etc.
- Misuses related with unusual operations, can be detected by comparing the normal interaction patterns for a user in a particular application. For example, an employee using the mail facilities for private activities can produce an excessive amount of mail.
- Manipulation of illegal content can be detected when the possibility of filtering it exists.

The previous features consider the misuse of features of common applications. One of them is data access, which is considered as a feature typical of a database system, since, in contrast with other applications, it provides a meaningful access. Threats to these systems are normally a primary consideration of organisations because its misuse is associated especially with fraud and data theft.

4. Improper data access

A single database can hold many different types of data interrelated among them. Besides, the nature of data is directly related with motivations of improper access and its impact.

Motivations. Depending on his motivations, a user accesses to the different kinds of data in a particular manner or another. Motivations of insiders are normally different and more specific than in the case of outsiders, being fraud and data theft the most typical, since they involve an unearned gain. Other misuse may involve: private activities, sabotage, vandalism, voyeurism and the elaboration of other misuses/attacks.

Impact. The impact of improper access can have different grades of severity, and it can affects to individual privacy, confidentiality and integrity of activities (services and transactions), and, in general, integrity of the content of the database.

Some data is self-contained, representing an asset for the organisation and a consequence of their activities. In this case, data theft is the main threat. Other data may be involved in operations carried out by the organisation, giving support to services and transactions. Outcomes from an improper access to this information are a direct or indirect damage to their confidentiality (in the case of disclosure) and integrity. A user could possibly perform improper data access in this way to commit fraudulent actions. As computer fraud is a major concern, intrusion detection systems start to be applied to fraud detection, and therefore these two areas are meant to converge (Noel 2002).

4.1 Misuse associated with viewing rights

Despite controls established in databases, authorised users may utilise their legitimate access rights in several ways to bypass them. Possible misuses associated with legitimate visualisation rights are:

- Data aggregation: users could try to collect information about one or more individuals, transactions or products for different purposes. Privacy and confidentiality of data can be compromised.
- Displaying data in an improper way (conditioned or sorted): when information is not displayed in a manner that exclusively serves the purposes of the database system, it can provide additional information and capabilities. For example, displaying a telephone directory sorted by number.
- Retrieval of a large amount of data: users could attempt a partial reconstruction of the database by retrieving a large amount of information. This reconstruction could possibly provide more operations over the data that were initially restricted.
- Discovering the existence of restricted information: unsuccessful attempts to display restricted fields could allow users to identify records with sensitive information or to guess part of them.

Besides, restricted information can be obtained by means of inference mechanisms. In some situations, particular actions can be used to unveil part of the structure of the database that was hidden for security reasons.

4.2 Misuse associated with modification rights

Misuses associated with legitimate creation and modification rights are:

- *Deliberate insertion of false data*: users can insert erroneous content in the database in order to damage its integrity or to corrupt the supported procedures.
- *Misuse of coherence mechanisms*: users can exploit mechanisms that check for coherence and compatibility of related values in the database. They may be able to discover the structure of the database, by displaying error messages when attempting to perform a writing operation. Besides, inserting false information into particular fields might be used to change the values of initially restricted fields.
- *Scheduling modification/destruction of information*: in many cases, users are not able to modify or erase a particular record. However, modifying related information could be used to deceive a user or a procedure that actually owns those rights.

4.3 Prevention and detection mechanisms

Prevention of improper data access requires proper access controls in order to restrict the operations that a particular user does not require for its functions, including reducing its knowledge of the structure of the database and its mechanisms. From the types of users that can access a database, casual, naive (or parametric), stand-alone and sophisticated (Castano et al.1995), the last of them represents the higher risk. Users that are able to develop and customise their own applications can potentially access the database in many ways, and it is more difficult to control. Access control mechanisms like Role-Based Access Control need to be properly defined and adapted to the requirements of each user's interaction.

Several detection approaches can be combined to effectively identify improper access to the database: analysis of user access (query and update filtering), monitoring user's access patterns and data analysis of the content of database (to detect anomalous situations). Detection can be oriented to the single evaluation of an access, or its combination with several past accesses.

In general, abuse of legitimate access can be detected by:

- Filtering individual queries/updates for abnormal content and incompatibilities among values.
- Analysing access patterns in a period of time, such as the amount of data accessed and signs of automation, to evaluate if they fit into the user profile.
- Checking the coherence of multiple accesses. It includes correlation between current and past transactions, and current and past data (for example, a radical change in a value).
- Evaluating the context of the access, trying to discover existent links between the user and an individual or with parameters of a transaction.
- Evaluation of a set of accesses in order to determine if they were justified and if the access sequence denotes common interaction.

As it can be observed, detection of an abuse of data access rights often requires some knowledge about the relations between different sets of data. This may require the analysis of large segments of data, which is technically unviable. New techniques, such as data mining, could increase the efficiency of data analysis in the future.

5. Current state of intrusion detection systems

An effective detection of the abuse of privileges requires the collection of information at multiple levels, including monitoring a large amount of data at the level of the application. This way, more accurate information about user interaction can be given (Phyo and Furnell 2003). However, this requires the capability of analysing large amounts of data. The current state of intrusion detection systems based in anomaly detection is limited by the development of data analysis and artificial intelligence techniques that can solve the high false alarm rate of current statistical methods (Noel 2002). The most viable approach in the medium term is the combination of misuse and anomaly detection systems order to provide better detection on known and unknown misuse scenarios.

6. Conclusions

This paper has evaluated the problem of insider misuse of applications, presenting a relation of features that commercial applications provide to common users. Misusers are very versatile when carrying out their activities in the system. Therefore, detection mechanisms must be very specific when defining different access patterns of a user, considering different features of the application, normal interaction patterns and the assigned functions of the user to assess the existence of misuse. It is important to identify these characteristics in order to limit the range of possible values that these patterns could take and to decrease the risk of undetected misuses. However, in many cases a misuse cannot be identified by abnormal patterns, because there is no intrinsic difference between some legitimate and illegitimate actions in the system, though it represents a clear misuse at the organisation level. The absence of identifiable patterns makes anomaly detection very difficult, especially in first stages of the profiling process.

Special attention has been paid to data access, as a feature typical of database systems, but that can be found in commercial applications under other forms. Abuse of legitimate access rights in a database requires a great effort in detection, and it often requires the analysis and knowledge of the content and the structure of the database, which can result very resource consuming and technically unaffordable.

Traditional security approaches do not provide a satisfactory level of security and efficiency at the same time. Besides, the existing gap between the organisation policies and its enforcement in the system requires methods to reduce this distance. Inevitably, this leads to the definition of mechanisms that translate efficiently the knowledge of the organisation into the systems. Role-based access control, an improvement of profiling mechanisms in detection systems, and standard definition of metadata in terms of database security could possibly decrease the insider threat in the future.

7. References

Anderson, J.P. (1980), “Computer Security Threat Monitoring and Surveillance”, Technical Report, James P Anderson Co. , Fort Washington, April 1980

Castano, S., Fugini, M., Martella, G. and Samarati, P. (1995), “Database security”, *Addison Wesley*, 1995

Chung, C.Y., (1997), "A survey of Misuse Detection Systems", http://seclab.cs.ucdavis.edu/~chungy/doc/MDS.htm , December 1997

Ferraiolo, D.F., Cugini, J.A. and Kuhn, D.R. (1995), “Role-Based Access Control (RBAC): Features and motivations”, In the Proceedings of Computer Security Applications Conference, December 1995

Neumann, P.G. (1999), “The challenges of Insider Misuse”, *SRI Computer Science Laboratory, Paper prepared for the Workshop on Preventing, Detecting, and Responding to Malicious Insider Misuse*, 16-18 August 1999, at RAND, Santa Monica, CA.

Noel, S., Wijesekera, D. and Youman, C. (2002), “Modern Intrusion Detection, Data Mining, and Degrees of Attack Guilt”, Applications of Data Mining in Computer Security (Advances in Information Security), Kluwer Academic Publishers, ISBN:1402070543, June 2002

Phyo, A.H. and Furnell, S. (2003), “Data Gathering for Insider Misuse Monitoring”, In the Proceedings of the 2nd European Conference on Information Warfare and security, pp. 247-254, University of Reading, UK, 30th June-1st July, 2003

Richardson, R. (2003), “2003 CSI/FBI computer crime and Security Survey”, Computer Security Institute, http://www.gocsi.com/forms/fbi/pdf.jhtml , Spring 2003

Artificial Impostor Profiling for Keystroke Analysis on a Mobile Handset

J. Lecomte, N.L. Clarke and S.M. Furnell

Network Research Group, University of Plymouth, Plymouth, United Kingdom.
e-mail: nrg@plymouth.ac.uk

Abstract

Keystroke Analysis is a biometric approach that utilises the typing characteristics of a user to perform identity authentication, and has two key advantages in a mobile context – the necessary authentication hardware (i.e. the keypad is already present) and the technique can operate transparently. Although studies have proved the feasibility of such an approach on a mobile handset, a failing exists in the practical deployment of the system. Classification is performed by neural networks that are trained using both the authorized users samples and impostors as a means of comparison. However, in the real world, the availability and suitability of impostor samples will be limited. This paper proposes a means of artificially creating impostor data directly based upon samples from the authorized user in order to provide optimally configured classification engines. These artificial impostor approaches have not only solved the availability issue but have improved the system performance (in comparison to the traditional approach) by up to 25%.

Keywords

Biometrics, Keystroke Analysis, Keystroke Dynamics, Impostor Profiling, User Authentication

1. Introduction

The mobile telecommunications industry has experienced formidable growth in recent years with in excess of 1.3 billion subscribers worldwide (Cellular Online, 2003). In order to capitalise on possible revenue, network operators have moved from a voice centric telephony device to a multimedia communications device capable of providing a wide variety of data based services. These services will permit the subscriber to access a number of potentially sensitive locations, including, corporate networks, personal bank accounts and share dealing services (Giussani, 2001). In parallel with this increase in ownership there has been a corresponding increase in mobile handset theft, with over 700,000 stolen in the UK during 2001 (BBC News, 2002). With the increase in data sensitivity and handset misuse, the need to ensure subscriber identity becomes paramount.

Subscriber authentication is currently provided by the PIN (Personal Identification Number), which has been shown to be an inconvenient and frequently unused technique (Clarke et al., 2002) with over 45% of respondents not using any security for their handset. Conversely however, 81% of respondents thought it a good or very good idea to have additional security, even though they do not use what is currently available. With this apparent contradiction in security requirements and the need for more secure approaches, research has recently focussed on the use of biometrics.

One biometric of particular interest, due to its non-intrusiveness is keystroke analysis. This technique utilises the typing characteristics of subscribers to differentiate between them and can therefore (in principle) authenticate users during their normal handset interactions, such as when they are dialling telephone numbers and entering PINs. Although feasibility studies have demonstrated promise in utilising such a technique (Clarke et al, 2003; Clarke et al, 2004) an issue arises in the practical implementation and evaluation of the method. The current classification process utilises neural networks, where data from the authorised subscriber is used alongside impostor data to teach the network the difference in input characteristics. So the network is taught what input data belongs to the authorised user and what belongs to impostors. In practicality however, the suitability and availability of the impostor data limits the performance and implementation of the technique, for the following reasons:

- Availability – a bank of impostor data will always be required for each user to teach the neural network
- Suitability – the bank of impostor data may or may not be similar to the authorised users' dataset. Impostor data that is able to surround an authorised users dataset would be the ideal.
- Performance – the network is evaluated using the same impostor users with which it was explicitly trained to reject (although the particular samples have not been used in the training), giving rise to possibly skewed performance rates.

This paper presents a number of algorithms design to artificially create impostor data, based specifically on the authorised user. Creating impostors that closely imitate (but not duplicate) the authorised user's input distribution, should result in removing the availability issue and improve the suitability of impostor data and increase the performance of the overall classifier.

2. Keystroke Analysis Investigation

The experimental procedure to evaluate the impostor algorithms sought to duplicate the investigations described in Clarke et al (2002). This permits a comparison between the original results and those generated using the impostor algorithms. To this end, the impostor algorithms were tested against three types of input data:

1. Entry of a fixed four-digit number, analogous to the PINs used on many current systems.
2. Entry of a series of telephone numbers. The classification of dynamic inputs is expected to increase intra-user variance, and thereby make it harder for the network to classify.
3. Entry of a fixed telephone number in order to facilitate a comparison against the results from the second experiment.

A total of thirty two test subjects provided the input data required for all three investigations. Table 1 illustrates the dataset sample sizes after outliers have been removed. For the traditional tests, each user are taken in turn as the authorised user with all the remaining users acting as impostors and trained using the training dataset. The evaluation of the neural networks is then performed by a validation dataset – containing data not used in the training procedure.

	Total # of Samples	# of Training Samples	# of Validation Samples
4-Digit PIN	25	16	9
Varying Telephone	38	26	12
Fixed Telephone	21	14	7

Table 1 : Number of Samples in Investigations

The artificial impostor tests will only utilise the authorised users training samples during the training stage and to create the impostor data. The networks will again be validated using the identical validation dataset as before in order for a fair comparison of results to be made.

A specially written application was used to collect the sample data. However, it was considered that the standard numerical keypad on a PC keyboard would not be an appropriate means of data entry, as it differs from a mobile handset in terms of both feel and layout, and users would be likely to exhibit a markedly different style when entering the data. As such, the data capture was performed using a modified mobile phone handset, interfaced to a PC through the keyboard connection.

Due to the limitations of data collection, the input data required for training and testing of the authentication system had to be collected in a single session. Ideally, the data would be collected over a period of time, in order to capture a truer representation of the users typing pattern. For example, by asking the user to type in 50 telephone numbers all at once, could result in an exaggerated learning curve.

3. Artificial Impostor Algorithms

Impostor algorithms were created using traditional statistical tools used normally to study natural phenomena and pattern recognition in general (Jain et al., 1999). Each approach attempts replicate the authorised users profile, but by adding and subtracting a noise content in order to bound all authorised samples with unauthorised samples. Three approaches are described and evaluated in this paper:

- bootstrap sampling of vector components with noise;
- weighted intervals with noise;
- manipulation of normal distribution parameters.

3.1 Bootstrap Sampling of Vector Components

The concept behind bootstrapping involves choosing at random samples with replacement from a dataset and analysing each sample in an identical manner. In this particular approach, rather than taking complete samples each component of the sample or vector is taken independently. Given the 4-digit PIN input where there are 16 samples, each of the 4 latencies will be taken in turn, bootstrapping from the 16 available samples from that placement, creating a new four latency sample using the authorised user's data. By subsequently adding noise to each of the four latencies this will shift the sample away from the authorised users' distribution. The reason sample components are taken from their same respective positions rather than from the complete dataset, are in order to conserve inherent typing characteristics within the sample, as illustrated in Figure 1. For instance, the fifth sample component in the telephone input investigations tends to be typically larger than the others as it represents the point between the end of the area code and start of the individual number (in a UK format telephone number) – a normal stage to pause.

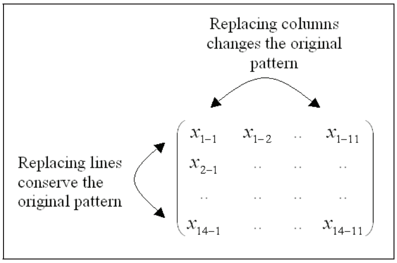

Replacing columns changes the original pattern
Replacing lines conserve the original pattern
x_{1-1}
x_{1-2}
x_{1-11}
x_{2-1}
x_{14-1}
x_{14-11}

Figure 1 : Typing Pattern Conservation

An important consideration in utilising this approach is the size of the noise added to the vector components. Care must be taken not to generate the same distribution or even a very close distribution to the authorised user, so that the intra-user space (space between an authorised users own input samples) is not affected. The noise is used to create some sufficient distance from the authorised user. The noise is randomly chosen over a bounded interval, for each sample component. The algorithm can be optimised by monitoring the evaluation stage and increasing or decreasing the noise, thereby moving the distance of the impostor samples from the authorised user's distribution. The equation for the algorithm is illustrated in Equation 1.

$$\text{Impostor vector} = (\text{Random}(x_{1-1} \;\; \ldots \;\; x_{14-1})+\text{n}_1, \text{Random}(x_{1-2} \;\; \ldots \;\; x_{14-2})+\text{n}_2, \ldots, \text{Random}(x_{1-11} \;\; \ldots \;\; x_{14-11})+\text{n}_3)$$

Equation 1 : Bootstrapping Sampling of Vector Components

3.2 Weighted Intervals with Noise

This approach takes a more pragmatic approach than the first by assigning probabilities to vector components in defined intervals. As before, this technique splits the sample into its constituent latencies in order to conserve any typing characteristic, as illustrated in figure 2, however, instead of subsequently performing a bootstrapping method, this approach takes all samples of that vector component and sorts them into ascending order.

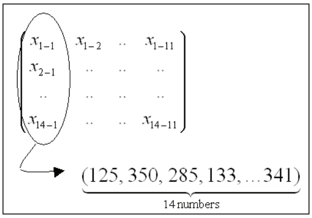


Figure 2 : Extracting a samples' constituent parts

A noise is then added to each latency to surround the value, thereby ensuring values inside these bounds are authorised and outside are classified as impostors. Probabilities are then calculated based upon what latencies are observed with predefined boundaries. These probabilities are defined as the weight of that particular interval. Samples are then generated by picking at random, ensuring however the weights are maintained thereby ensuring the impostor samples closely mimic the authorised users data overall, but with the addition of noise. Equation 2 illustrates this approach.

$$\text{Impostor vector} = (\text{Random}(\text{limit\_min}; x_{1-1} - n), \text{Random}(x_{1-1} + n \;;\; x_{2-1} - n), \ldots, \text{Random}(x_{14-1} + n; \text{limit\_max}))$$

Equation 2 : Weighted Intervals

This concept is thus very similar to the probability density function, where each interval is assigned a probability of having one of its value represented in the final vector.

3.3 Manipulation of Normal Distribution Parameters

The third algorithm moves away from the raw data to utilising the parameters that describe the authorised users input data. The input data generated from users can be approximated to a normal distribution where the mean and standard deviation parameters can be used to describe the distribution. A noise is again added to the vector components to ensure a desirable distance away from the user distribution. The equation of this algorithm is illustrated below, where the noise is a random coefficient x (0.1<x>2) to preserve a user space, but stay close enough to the distribution.

$$\textbf{Impostor\_matrix} = \left(\mathit{impostor\_vector}1(\bar{x}, \sigma, 11, 1) \times \mathit{noise}\,1 \;\; \ldots \;\; \mathit{impostor\_vector}14(\bar{x}, \sigma, 11, 1) \times \mathit{noise}\,14 \right)$$

Equation 3 : Normal Distribution Parameter Manipulation

Where $\mathit{impostor\_vector}1(\bar{x}, \sigma, 11, 1)$ is a pseudo random vector of dimension (11×1) based on the standard deviation σ and mean $\bar{x}$ of the authorised user.

In all three approaches an important consideration concerning the amount of impostor data generated must be taken into account. With too much impostor data and the network will respond by rejecting all input samples, but with too little, too many impostors will be able to gain access. With the noise, this gives rise to a second variable that can be altered in order to optimise the performance of the algorithms.

4. Results and Discussion

A comparison of the results achieved by all three approaches, as illustrated in table 2, indicates that in general the *manipulation of normal distribution parameters* technique proved most successful achieving the lowest Equal Error Rate (EER) for both telephone input scenarios, and only 1% off the lowest EER for the 4-digit PIN. A reason for this can be conjectured to be due to the more general classification boundaries that are produced using just two measures of the authorised user's distribution, instead of the large manipulation of actual raw data which the remaining techniques utilise. The manipulation of normal distribution parameters also represents the simpler approach, in terms of both time and computation.

Additionally, a descriptive statistical analysis of the input data unsurprisingly reveals that both telephone input scenarios have a larger intra-user variance (i.e. the spread of input samples within a user's collection of data) than the 4-digit PIN, indicating the classification boundaries created with the telephone inputs scenarios will be more general as the samples vary so much. This also helps to argue the reason as to why the more general normal distribution parameter technique proved more useful – as the other techniques followed user's input data too closely not allowing for the more general pattern.

	Bootstrap Sampling of Vector Components		Weighted Intervals		Normal Distribution	
	EER (%)	Parameters (Noise/# Impostors)	EER (%)	Parameters (Noise/# Impostors)	EER (%)	Parameters (Noise/# Impostors)
4-Digit PIN	18	150-250/ 50	21	0.5-0.6/ 30	19	0.5-0.6/ 80
Varying Telephone #	41	200-250/ 50	44	0.5-0.6/ 30	35	0.4-0.5/ 80
Fixed Telephone #	25	200-300/ 80	24	0.2-0.3/ 30	21	0.5-0.6/ 80

Table 2 : Impostor Algorithm Results

The results in table 2 illustrate the best achievable results after both noise and amount of impostor data parameters had been varied. For the bootstrap sampling of vector components, the noise parameters are in milliseconds, however the other two techniques measure noise in terms of standard deviation about the authorised user's mean. The number of impostors is a measure of the amount of artificial data that was utilised, with 1 impostor equating to the number of samples provided by the authorised user (e.g. for the 4-digit PIN, 1 impostor = 16 samples).

A problem with the traditional approach to keystroke analysis is that impostor's used in training the classification engine are also the users that are subsequently used to evaluate the performance. Although the data has never been used by the engine before, the neural network has been specifically trained to reject that particular user's input data. As these artificial impostor algorithms do not use real impostor data during the training procedure, the results given here permit a more accurate representation of the achievable classification that could be expected.

However, the principle objective of this research was to artificially created impostor data that performed as well as, if not better than, utilising real impostor input samples. Table 3 illustrates a comparison of the best artificial impostor results against the traditional technique of using actual impostor data.

	Traditional Approach			Artificial Impostor Algorithms			
	FAR	*FRR*	*EER*	FAR	FRR	EER	Technique
4-Digit PIN	*9*	*39*	*24*	27	9	18	Bootstrap sampling of vector components[*2]
Varying Telephone #	*9*	*71*	*40*	28	41	35	Manipulation of Normal Distribution Parameters[*1]
Fixed Telephone #	*10*	*38*	*24*	23	18	21	Manipulation of Normal Distribution Parameters[*1]

Table 3 : Artificial Impostor Algorithm Results (with a comparison versus the traditional approach)

[*1] Noise parameters set at +/- 0.4-0.5 of standard deviation about mean with the equivalent of 80 impostors worth of input data.

[*2] Noise parameters set at +/- 150-250 mS with the equivalent of 50 impostors worth of input data.

As the results show, the artificially created impostor data has outperformed the traditional approach in all three input scenarios, with a 25% improvement in the 4-digit PIN and 12.5% improvement in both telephone input scenarios.

5. Conclusions & Future Work

The use of artificially created impostor data over real impostor data has a number of advantages; a true representation of the performance, optimised neural networks for all compatible authorised users, no requirement for a database of users to be used as impostor data and a self-contained authentication technique with small storage footprint – as only the authorised user's data need be kept. Moreover the use of artificially created impostor data has improved the performance of the technique over the traditional approach, indicating stronger classification boundaries have been created using impostor data generated directly based on the authorised user's input samples.

However, with a view of improving the performance still further, a number of areas have been identified for further research. The first minor improvement would be to dynamically adapt the noise level on a individual user basis thereby optimising the performance, as the current approach sets the noise level of all users to the same level, which might on average be the best level but might not be the case for individual users.

The basic assumption throughout this paper, and used explicitly in the manipulation of normal distribution parameters algorithm, has been the approximation of user's input samples to a normal distribution. Although this stands true, the approximation can be quite general in a number of users, so an argument exists for implementing another more complex distribution to model the input data. Figure 3 below illustrates theoretically a user's distribution with the dotted line and the attempt to model the distribution more accurately through the use of multiple normal distributions.

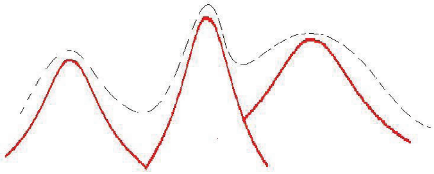


Figure 3 : Model of a Complex Distribution using a mixture of Normal Distributions

Finally, it would be interesting to study more advanced algorithms to create impostor data. For instance, some biometrics utilise Hidden Markov Models to create impostor files (Rabinier, 1989). They are used to model events that have a probability to occur depending on a previous event. This technique could be used to more intelligently create impostor samples.

Given previous research projects have identified the usefulness and promise of keystroke analysis, this research has successfully identified a means of solving the issue of creating a classification engine using only data supplied by the authorised user.

6. References

BBC. 2002. "Huge surge in mobile phone thefts", BBC News Report, 8th January 2002.

Cellular Online, Stats Snapshot 8/2003, http://www.cellular.co.za, Sep 2003. http://www.cellular.co.za/stats/stats-main.htm.

Clarke, N.L., Furnell, S.M., Lines, B., and Reynolds, P.L., 2003. "Using Keystroke Analysis as a mechanism for Subscriber Authentication on Mobile Handsets", Proceedings of the IFIP SEC 2003 Conference, Athens, Greece, May, pp97-108.

Clarke, N.L., Furnell, S.M., Rodwell, P.M., and Reynolds, P.L., 2002. "Acceptance of subscriber authentication methods for mobile telephony devices", Computer and Security, vol.22, n.3, pp.220-228.

Clarke, N.L., Furnell, S.M., Lines, B., and Reynolds, P.L., 2004. "Application of Keystroke Analysis to Mobile Text Messaging", to be presented at IsOneWorld conference, April 2004.

Giussani, B., 2001. Roam; Making Sense of the Wireless Internet. Random House Business Books.

Jain, A. K., Robert P.W. Duin, Jianchang Mao, 1999. "Statistical Pattern Recognition: A Review". IEEE Transactions on Pattern Analysis and Machine Intelligence.

Rabinier, L., 1989. "A Tutorial on Hidden Markov Models & Selected Applications in Speech Recognition". Proceedings of the IEEE, 77(2): 257-285.

Privacy Implications of Network Monitoring

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

Network Research Group, School of Computing, Communications and Electronics,
University of Plymouth, Plymouth, UK.
e-mail: nrg@plymouth.ac.uk

Abstract

This paper investigates the security concerns that relate to network monitoring and proposes a set of guidelines to improve the amount of information that can be extracted from the network traces while respecting the current legislation. Investigation was approached by understanding and analysing the current existence legislation such as the Directive 95/46/EC and Directive 2002/58/EC. Furthermore, investigation was extended towards analysing TCP/IP stack, where the captured packet's headers are analysed and studied. Knowledge gained from both investigations was implemented in developing a simple packet capturing tool. The uniqueness of this tool is; it focuses on the privacy issues that other tools like TCPdpriv and TCPurify fail to address. The advantage of using this tool is it retains as much information in the headers of the captured packets, especially in HTTP header. A set of guidelines was developed which defines the procedure appropriate for Offline Network Data Analysis.

Keywords

Online privacy, Data Protection, Network Monitoring, Sniffing.

1. Introduction

The relentless growth in Internet usage and online businesses have boost most of the organization with rapid development in computer and data networking technology to supports the needs. Due to the popularity of the Web, it is crucial to understand how usage relates to the performance of the network, the server and the clients.

In order to support the Internet growth, the network infrastructure needs to be maintained. Usually this task is carried by monitoring the network performances. Most of the organisation would analyse their network by capturing packets that travels in the network. These packets have the capability to reveal personal information of the user. According to the current legislation, collecting user's personal information without the knowledge of the user is against the law (European Commission 1995).

This research focuses on the security issues raised by network trace analysis and studying the vulnerabilities of the user data at all levels. This includes from the application of data exchanged to the identity of the user, based on the Internet Protocol (IP) details. In addition, legislation related to the network trace analysis are studied and applied in developing a simple packet-capturing tool, which focuses on the privacy of the user data. This paper is organised as follows: Section two discusses about the existing tools for packet capturing. Section three briefly presents the EU Directives whereas section four discuss about TCP/IP layer. The developed tool is presented in section five and section six concludes the paper.

2. Existing Tools

There are two different software available on the Internet which is primarily focuses on the privacy of the user, which are TCPdpriv and TCPurify. TCPdpriv is a program for removing sensitive information from tcpdump files. In other words it is a program for eliminating confidential information from packets collected on a network interface. TCPdpriv has the capability to capture data as well as the analysis of existing files. The captured packets are written as a tcpdump file, which can be subsequently processed. Besides that, TCPdpriv allows some degree of control over how much of the original data is removed or scrambled. For example, it is possible to have an IP address scrambled but it retains its class designation. The number ninety-nine (99) generally means *"release the information as is"*. This level is available for IP unicast and multicast addresses, classless, TCP and UDP port numbers. TCPdpriv provides a mechanism for outputting *randomised* IP addresses (using the *-A50* option). By this method, the amount of information encoded in the outputted IP addresses is larger than the amount of information encoded in the options that output IP addresses as sequential numbers (but, less than the amount of information encoded in the *-A99* option that causes the IP addresses on the output side to be the same as those on the input side).

TCPurify is a packet sniffer or capture program similar to tcpdump. This packet sniffer program has much reduced functionality in comparison with tcpdump but it focuses more to user's privacy. TCPurify erase most of the packet after the last recognised header like Ethernet, IP or TCP. In other words it removes all data payload and HTTP header before storing the packet. There are some of the packets kept such as ICMP packets and daytime. One of the interesting features is; it has the capability of randomising some or all IP address. This randomisation is reversible with the help of a one-shot generated file which is created at capture time. In a nutshell, TCPurify is intended to be a security auditing and intrusion detection tool, which in turn made a poor general purpose packet sniffer compared to tcpdump.

Both tools, TCPdpriv and TCPurify, are more interested in randomising the IP address. These tools do not process any of the HTTP headers, where most of the user's privacy related information is stored. Basically, the HTTP content is removed. Analysing the end result from these tools would not aid much on the network monitoring purposes.

3. Data Protection

The European Union (EU) passed Data Protection Directive Protecting personal information and harmonising privacy laws among its member states in 1995. The directive has been effectively applied across EU which has lead to enactment of legislation among all EU member states. The Directive establishes basic requirements and does not prevent national laws from imposing higher requirements. The Directive seeks to establish an equivalent level of protection for personal data in all Member States, as to facilitate the transfer of personal data across national boundaries within the European Union. The Directive applies to personal data processed wholly or partly by automatic means and to manual data held in filling systems structured by reference to individuals, but it does not apply to activities which falls out the scope of EU law. There are number of other Community regulations that deal with some aspects related to the internet. There are two Directives plays a major role in data protection issues; which are Directive 95/46/EC (provide protection to individuals with regard to the processing of personal data) and Directive 2002/58/EC (processing of personal data and protection of privacy in the electronic communication sector).

Privacy laws and regulations in the United States are a mixed with constitution protections, laws governing information protection in different industries (such as health care, banking and banking) and socially oriented laws that effect both the disclosure of information and the infringement of personal freedoms. Safe Harbour Privacy Principle was issued by U.S Department of Commerce. This was the end result of the discussion held between United States and European Union on transatlantic data flow issues. Overall privacy laws in the United States are a fragmented mess that no single government entity has attempted to make sense of or is empowered to control or guide.

4. TCP/IP Layer Analysis

As the name suggest TCP/IP is a protocol stack that provides two kinds of services, packaging data and routing the packaged data. The TCP/IP protocol is similar to the packaging and moving service except for the data is not only packaged and moved to the correct destination but it is also unpacked and delivered to the correct application. TCP provides packaging, reassembling, flow control and error detection service. IP manages the determination of the shortest possible path to the destination computer.

4.1 Ethernet Header

The term Ethernet is usually referred to a standard published in 1982 by Digital Equipment Corporation, Intel Corporation and Xerox Corporation (Stevens 1994). It is the predominant form of local area network technology used with TCP/IP. This research is more concentrated towards Ethernet header which contains IP datagram for the IP module. Each Ethernet has its own Media Access Control (MAC) address which is assigned to the machine when it was manufactured. Every Ethernet packet has a 14-octet header that allows the receiver to identify the sender. In other words, every each Ethernet has a unique identity. According to Article 6 of Directive 2002/58/EC, the *MAC destination address* and the *MAC source address* should be anonymised. These information can be used to reveal the originator of the packet with the help of other information contained on the IP headers like the IP address.

4.2 Internet Protocol (IP) Header

IP works as the glue that holds the whole Internet together. IP takes care of delivering a data packet irrespectively of the location of the destination. Location in this context refers to the network on which the computer is located. Every host and router on the internet has an IP address, which encodes its network number and host number. The combination is unique, where in principle there would not be two machines on the internet have the same IP address.

All the IP address are 32 bits long and are used in the Source address and Destination address field of IP packets. Every IP address comprises of 4 bytes. The way in which the IP address is represented is called dotted-quad.

Referring to Article 6 of Directive 2002/58/EC, Article 6 and Article 8 of Directive 95/46/EC, IP addresses of a captured packet should not be kept for network performance monitoring. Specified articles are related to processing traffic data and processing personal information (European Commission 2002) and (European Commission 1995). The IP addresses are a unique characteristic of a packet and it can lead to the user's identity. The source address and the destination address should be anonymised in order to respect the current legislation. IP address reveals the location, where the packets were created. With this information, the network administrator could identify busy or congested networks which might cause problems towards overall infrastructure. These addresses of the packets do provide valuable information towards analysing network performance; especially monitoring traffic and congestion of the network. If the entire IP addresses are anonymised, it would make monitoring network performance a harder task. If the IP address is anonymised partially, it would respect the legislation and still could be used for network performance analysis. Partial IP address anonymisation means the last 2 bytes of IP address are anonymised. The remaining IP address would be the Network IP address of the particular packet. For instances, (refer Figure 1) when user y from Organisation "AA" requesting a page content from an HTTP server, the server would have 141.16.24.3 as an IP source address. When the partial anonymisation process takes place, the IP source address in IP header would be 141.16.0.0.

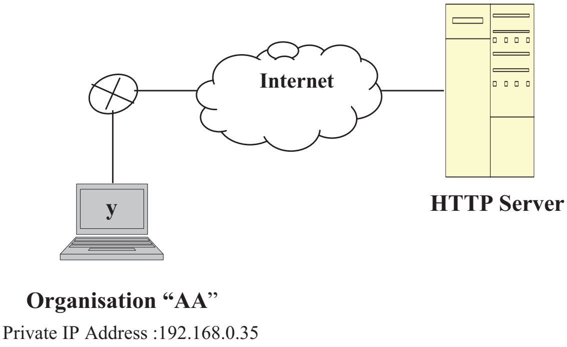


Figure 1: Communication between Computer and an HTTP Server

4.3 Transmission Control Protocol (TCP) Header

Exchange of data can be described as one of the primary function of internetworks. Transmission Control Protocol (TCP) was specifically designed to provide a reliable end-to-end byte stream over an unreliable internetwork. The size of the TCP header is usually 20 bytes; if Option field is not present.

The TCP headers do provide useful information for network monitoring. The Sequence and Acknowledge number from the TCP header, can be used together to identify any flaws on the network. Acknowledgement number specifies the sequence of a successfully received segment incremented by 1. The number is sent with an acknowledgement segment as a conformation of the receipt of another segment. The incremented sequence number indicates that the segments with sequence number less than the acknowledgement numbers have been successfully received. The number also indicates the sequence number of the next expected segment in the sequence of segment being transmitted from the sender. Besides that, the source and the destination port reveal the application of the packets. Certain ports are assigned to certain protocols such as port 80 is assigned for HTTP protocol and port 21 for FTP protocol.

Option segment header field plays an important role in the efficient transmission of data. There are four kinds of option in this field; there are End of Field, No Operation, Maximum Segment Size and Timestamp. Timestamp option does provide information time when the segment was first sent to the receiver machine. The option can also be used to record the time of sending and acknowledgement for the successful receipt of a segment. Since the timestamp contain information of the packet was created, this particular information should be erased from the TCP header once this data are no longer needed to maintain the communication.

4.4 The Hypertext Transfer Protocol

There are headers that are specified for each type of message. Headers fall into five main classes; they are *General, Request, Respond, Entity* and *Extended* headers. General headers are used by both client and the server which contain useful information for both parties. Request headers are used for only request messages. This header provides extra information to servers, whereas Respond headers provide information to the client. Entity header refers to headers refer to headers that deal with the entity body. For instances, this header will tell the type of the data in the entity body. Extension headers are non-standard headers that have been created by application developers. These five classes of headers contain parameters that provide specific information to the user or client. Some of these attributes contain information that will violate the privacy of the user.

Attributes like Client-*IP* and *From* reveals the users identity. These attributes need to be erased when it no longer needed for establishing communication between the client and the server. According to Article 6 of the Directive 97/66/EC, any data such as the session login data, time stamp of the packet should be erased as soon as website is not accessed by the internet user. HTTP attributes such as *Date, Age, Location, ETag, Expires and Last Modified* falls under this specification. As a result, these attributes should be made anonymous. Furthermore, attributes such as *Host, Location, Referer* and *Title* contains information related to the activity of the user at the internet. According to Article 8 of Directive 95/46/EC, Article 6 of Directive 2002/58/EC and Directive 95/46/EC, personal information and sensitive issues First principle of Data protection Act 1998, (Sensitive Data) should not be kept without the user's consent and should be made anonymous upon termination of the Internet session.

5. Developed Tool

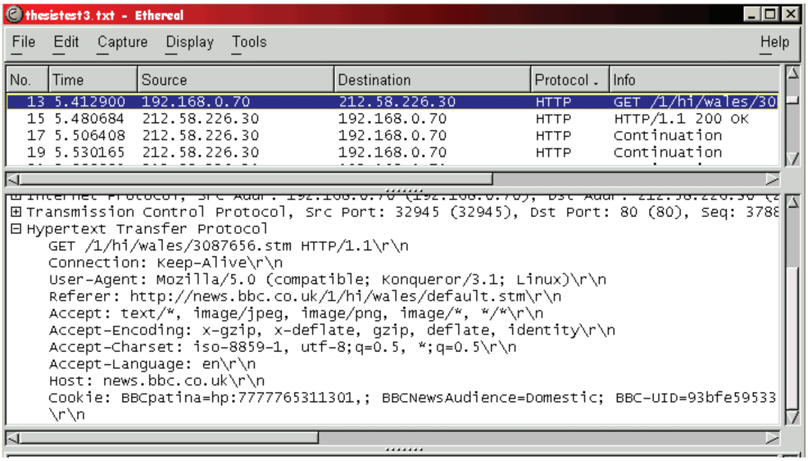
The developed tool can be categorised under packet capturing family, which is also known as sniffer. This tool has reduces features compared with other packet capturing software such as TCPDump. The uniqueness of this tool is, it focus and emphasis towards privacy in network monitoring. This particular program is specialist in processing HTTP application packets. It retains as much information as possible in the captured packet for network monitoring purposes, while respecting the current legislation. In HTTP header, there are lot different types of attributes exist depending on type of the messages; Respond massage or Request Message. The only common characteristics that exist in HTTP header are every each attributes ends with the character "\r\n", whereas character "\r\n\r\n" indicates that it is the end of the HTTP header without including the payload. Base on these two characteristic, attributes are erased. Those attribute that need to be erased are declared at the beginning of the program. All these attributes are searched in the HTTP header content. Once it has been identified, character 'x' is replaced with the actual content. Figure 2 and Figure 3 illustrate the results of the process. This program was developed based on the rules and regulations that been discussed earlier in previous sections.

6. Guidelines for Network Monitoring

Online Network Monitoring means, analysing the traffic data while maintaining the originator of the captured packet are kept anonymous. These captured packets should not be stored. These packets should be erased immediately after the traffic analysis process is completed. Typically, a monitor operates on network in 'promiscuous' mode. It means viewing every packet on the network. The monitor will produce summary information, including error statistic, performance statistic between server and the client, number of collisions, packet size, bandwidth consumption, traffic congestion, application usage, number of server that been used and more. This information would give a good view to the network administrator to understand the chain of, performance problems that current Web users face.

Offline Network Monitoring means, analysing the traffic data while maintaining the originator of the captured packet anonymous. The captured packets can be stored for later analysis. In order to store the captured packet certain information should be erased. Table 7.1 illustrates the information should be erased or made partially anonymous from the packet header. The tool developed in this research do erase (or some attributes are made partially anonymous) all the attributes that been specified in Table 1.

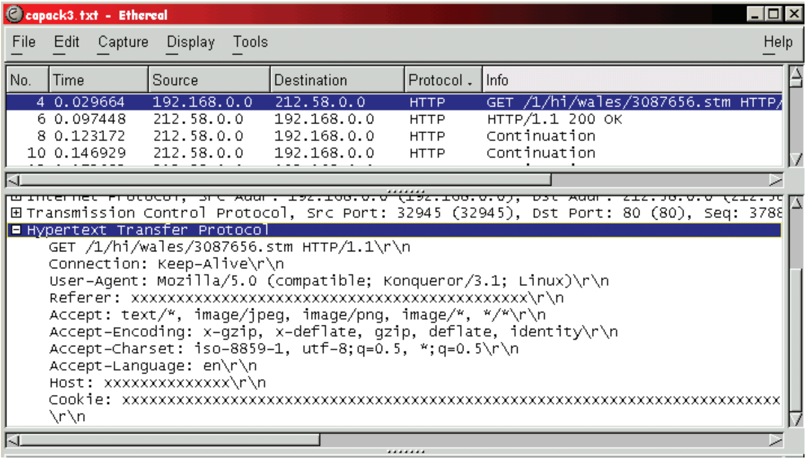
Even though, the timestamp has been erased, the relative time between packets are still maintained, which still can be used to analyse the delay between the packets. Besides that, in HTTP header, attributes like Accept, Accept-Encoding, Accept-Language provides information or characteristic of the web page that been analysed. For instance, those attributes that been specified on the previous paragraph, provides information about language been used, the content of web page either text or picture and coding method been used. Moreover, User-Agent attribute reveals type of operating system and the browser, the client been used. HTTP Status code is one of the characteristic used to reveal the condition of the requested web page, whether the request was successful, redirected, error on web page or error on the server. These attributes can be used to analyse the traffic congestion or about the condition or performance of the particular server. There are more attributes in HTTP header where it can be used to gain more information about the network.



```
 13 5.412900  192.168.0.70     212.58.226.30    HTTP      GET /1/hi/wales/30
 15 5.480684  212.58.226.30    192.168.0.70     HTTP      HTTP/1.1 200 OK
 17 5.506408  212.58.226.30    192.168.0.70     HTTP      Continuation
 19 5.530165  212.58.226.30    192.168.0.70     HTTP      Continuation
```

```
⊞ Internet Protocol, Src Addr: 192.168.0.70 (192.168.0.70), Dst Addr: 212.58.226.30 (2
⊞ Transmission Control Protocol, Src Port: 32945 (32945), Dst Port: 80 (80), Seq: 3788
⊟ Hypertext Transfer Protocol
    GET /1/hi/wales/3087656.stm HTTP/1.1\r\n
    Connection: Keep-Alive\r\n
    User-Agent: Mozilla/5.0 (compatible; Konqueror/3.1; Linux)\r\n
    Referer: http://news.bbc.co.uk/1/hi/wales/default.stm\r\n
    Accept: text/*, image/jpeg, image/png, image/*, */*\r\n
    Accept-Encoding: x-gzip, x-deflate, gzip, deflate, identity\r\n
    Accept-Charset: iso-8859-1, utf-8;q=0.5, *;q=0.5\r\n
    Accept-Language: en\r\n
    Host: news.bbc.co.uk\r\n
    Cookie: BBCpatina=hp:7777765311301,; BBCNewsAudience=Domestic; BBC-UID=93bfe59533
    \r\n
```

Figure 2 : Before HTTP attributes are made anonymous



```
  4 0.029664  192.168.0.0   212.58.0.0     HTTP     GET /1/hi/wales/3087656.stm HTTP/
  6 0.097448  212.58.0.0    192.168.0.0    HTTP     HTTP/1.1 200 OK
  8 0.123172  212.58.0.0    192.168.0.0    HTTP     Continuation
 10 0.146929  212.58.0.0    192.168.0.0    HTTP     Continuation
```

```
⊞ Internet Protocol, Src Addr: 192.168.0.0 (192.168.0.0), Dst Addr: 212.58.0.0 (212.58
⊞ Transmission Control Protocol, Src Port: 32945 (32945), Dst Port: 80 (80), Seq: 3788
⊟ Hypertext Transfer Protocol
    GET /1/hi/wales/3087656.stm HTTP/1.1\r\n
    Connection: Keep-Alive\r\n
    User-Agent: Mozilla/5.0 (compatible; Konqueror/3.1; Linux)\r\n
    Referer: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx\r\n
    Accept: text/*, image/jpeg, image/png, image/*, */*\r\n
    Accept-Encoding: x-gzip, x-deflate, gzip, deflate, identity\r\n
    Accept-Charset: iso-8859-1, utf-8;q=0.5, *;q=0.5\r\n
    Accept-Language: en\r\n
    Host: xxxxxxxxxxxxxx\r\n
    Cookie: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
    \r\n
```

Figure 3 : After HTTP attributes are made anonymous.

Header	Attributes
Ethernet Header	MAC Source Address
	MAC Destination Address
IP Header	IP Source Address
	IP Destination Address
TCP Header	Timestamp
HTTP Header	Host , Client-IP, From, Location, E-Tag, Referer, Title, Cookie, Expires, Date, Age, and Last-Modified

Table 1 : Attributes that been erased for Offline Network Monitoring

Packet capturing tool like TCPurify, do not provide any information from the HTTP header since the HTTP header is truncated in the anonymisation process. Even though it has the capability of randomising the IP addresses but the major disadvantages of TCPurify is no information can be extracted from the HTTP header. In a nutshell, the end result from the developed tool can provide more useful information in comparison with the TCPurify. However, online network monitoring does provide far greater information of the network, but those captured packet are not allowed to be stored.

7. Conclusion

The aim of the research is to implement privacy in network monitoring. The European Union legislation which is related to data protection laws has been studied. With the knowledge of TCP/IP protocol suite, a simple packet capturing tool has been developed. This tool complies with the European Legislation, which are Directive 95/46/EC and Directive 2002/58/EC. The tool is been develop in C code under Linux environment. This program was developed Offline Network Monitoring. The captured packet's The IP address are made partially anonymous and certain attributes Ethernet, IP, TCP and HTTP headers are erased. Any attributes related to time, location and the personal details of the user are made anonymous. The future development of this research lies in enhancement of more extra features on the developed tool. Further investigation need to be carried out on international law such as in Australia, China and South East Asia countries.

8. References

European Commission (2002), *'Directive 2002/58/EC, concerning the processing of personal data and the protection of privacy in the electronic communications sector'*, Commission of the European Communities.

URL: *http://europa.eu.int/eur- ex/pri/en/oj/dat/2002 /l_201/l_20120020731en 00370047.pdf [December 15th, 2002]*

European Commission (1995), *'Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data',* Commission of the European Communities. URL:*http://europa.eu.int/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=EN&numdoc=31995L0046&model=guichett [December 10th, 2002]*

Information Commissioner (1998), *'Data Protection Act 1998',* United Kingdom, Her Majesty's Stationery Office. URL:*http://www.hmso.gov.uk/acts/acts1998/19980029.htm [December 2nd, 2002]*

Pfitzmann, A. and Waidner, M., (1985), ' *Network without user Observability-Design Options,'* Eurocrypt 85, LNCS 219, Springer-Verlag, pp 245-253.

Shanmugam, R., Padmini, R. and Nivedita, S., (2002), *"Special Edition Using TCP/IP"*, QUE Publishing, Second Edition.

Stevens, W.R., (1994)"*TCP/IP Illustrated Volume 1 The Protocols*" (1995). Addison-Wesley Publishing Company.

EduCQ : An Educational Tool for Information Security

B. Larchevêque and S.M. Furnell

Network Research Group, University of Plymouth, Plymouth, United Kingdom.
e-mail: nrg@plymouth.ac.uk

Abstract

During the last decade, the dependency of most kinds of companies upon information systems has increased as never before. In most business domains, small and medium sized structures generally suffer from a lack of skills in information security topics, because of their smaller budget that often leads them to have some more pragmatic investments. It is indeed expensive for a company to have their own specialized staff, and expensive to educate their existing employees in charge of the information system. This paper proposes a suitable solution to enable such companies to increase their information security: a computer based-tool that could help their staff to be aware of certain dangers they could face and to improve their knowledge in this critical subject. Some research on the best way to assess the employees has been carried out with the concern that this way would have to be implemented in a computer-based tool. The outcome of those investigations is the development of a Computer Based-Tool: the *EduCQ Project*. This tool allows the user to assess his level in information security topics through three series of questions from different types (questions asked about dialogs, multiple-choice questions and image based assessment).

Keywords

EduCQ, Education, ISO 17799, Security, Awareness, Computer Based-Tool.

1. Introduction

The information security need is a major concern in companies that possess critical data -such as hospitals or banks-, but also in others services companies whose wealth rests in a numerical format. These concerns bring with them plenty of questions that few people can appreciate: one of those could be *"will the problem inherent to a type of network be transmissible to the rest of the infrastructure?"* Computer networks, and more generally information systems have their inherent problems, and as the trend seems to designate their network as a potential medium for the future of network some good security habits from its users could be a good starting point to enter this era. In the scope of companies, which usually have some requirements more related to their business, this matter of security of information systems is also a crucial point they should bear in mind. Actually, as we will explain in the following part, networks still presents some weaknesses, which can be exploited by a group of skilled people but they also suffer from some misuses caused by trusted users –e.g. human errors or thefts-. More over, some other kinds of macroscopic acts arise more often than we could imagine.

SMEs which do not necessarily have somebody aware on this deep issue in their staff, and which would maybe not have the time to implement a security policy, are particularly at risk from information security threats. In the next section we will see the actual situation of data protection in IT companies, and why the project has an industrial interest. The research will be lead in order to provide an accurate answer to the lack of security in IT companies. The state of the art provided in this section will help us to identify those needs, while a quick survey will demonstrated how much an improvement of all categories of security issue is necessary for companies that expect to stop paying for misuser and/or cyber-criminals. This paper identifies the need and importance of information security training, presents the existing tools and usable scenarios and implements a solution that could be a base for such a tool.

2. The need for security training

There is actually a lack of security in most companies: even if the percentage of companies that report some unauthorized acts on their system seems to stagnate since one or two years, the situation is worrying. Moreover the real size of the issue is hard to know since all the companies that suffer from attacks do not necessary report those acts, most of the time because of the bad publicity that could result from this confession. Figure 1 shows the main reasons that explain the lack of information security.

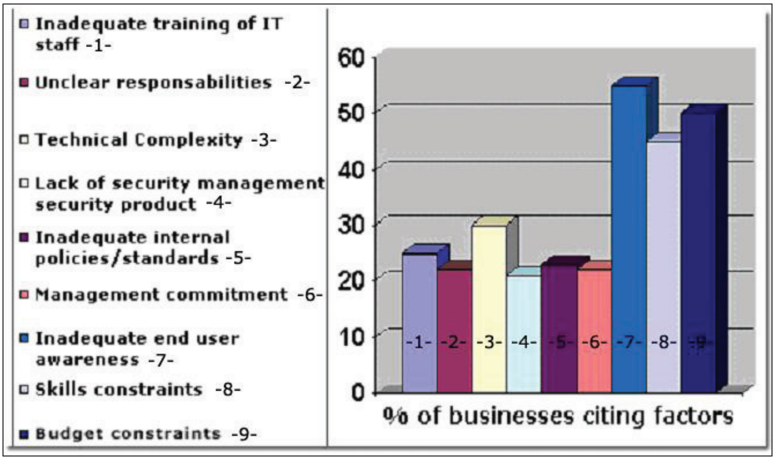


Figure 1 : Obstacles to Information Security (National Computing Center, 2000)

It appears that some parts that can be unified, to identify the responsibilities, are due to the own staff of a company. For example we find unauthorized access by insiders, laptop theft, insider abuse of net access or others malwares installation. This issue shows that there is an obvious need for security training.

The idea to use a CBT (Computer Based Training) as assessment tool has been identified through several researches. The cost for security staff teaching is one of the most expensive in comparison with others staff formations. The budget a company should provide for example just for a day seminar about introduction to information security management and *BS 7799/ISO 17799* can be up to £270 + VAT[2], per employee! A market report even shows that the cost for security breach prevention is slightly more expensive than the security breach cost… A software can be much cheaper for a company than a formation, especially because the update are generally quite affordable (in our case it would even be free for use). Moreover, when you know that this software will promote awareness it can represent a relevant alternative to inform your staff about IT security.

3. The EduCQ system principle and implementation.

The aim of the system has always been to possess the widest scope of functionalities as possible, while keeping a reasonable range of complexity and abstraction required by the modest size of the project. The objective would be to allow:

- The user to run the application under all the main type of systems.
- The user to share the application with others users on the same machine.
- The user to set the software environment to its preferences (color, look and feel,...)
- The administrator or any granted user to manage the database -so the software content- remotely, from any system.
- The administrator to manage the user database, enclosed in the forum tables.
- The users to write their own classes to improve or set the software more precisely to their need.
- The user and administrator to use everything, from the technology to the software functionalities, freely.

As a result, the *EduCQ Project* owns a common database that allows the user to dispose of constantly up to date assessments set. As the technology should stay away from the system administration, any administrator can manage the database which gives more chance to *EduCQ* to survive to the end of this project. Figure 2 summarizes the overall system architecture.

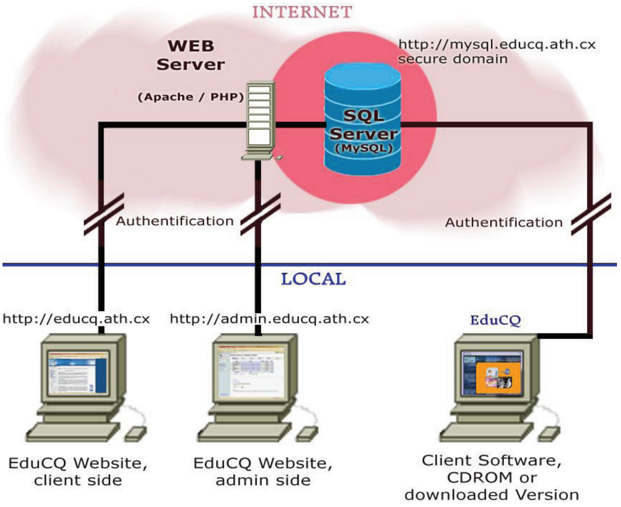

INTERNET
WEB Server
(Apache / PHP)
http://mysql.educq.ath.cx
secure domain
SQL Server (MySQL)
Authentification
Authentification
LOCAL
http://educq.ath.cx
http://admin.educq.ath.cx
EduCQ
EduCQ Website, client side
EduCQ Website, admin side
Client Software, CDROM or downloaded Version

Figure 2 : The EduCQ overall system architecture

During a tool development, the developer usually has to take care of the technology he will use, to see if this choice is relevant for the tool's functionalities, but also for the environment in which he wants to run the tool. The *EduCQ* software, in its actual version 0.22, allows the user to add some other users, to save their profile, to setup the directory they want to use for the reporting, to export their score in several format –in text files and web pages so far-. The assessment type allows also this user to read some mp3 files, or to make him discover the *ISO 17799* guidelines.

To warrant the spreading of the project, a website has been create, with both user and administrator sides. The user side website proposes the classes and sources of the software, a support (both technical and functional, respectively through the Javadoc and the User Guide) but also a forum which is a great added value for this kind of tool. The administrator website proposes an intuitive interface that allows a granted user or an administrator to keep the database up to date without any SQL knowledge.

4. Potential assessment methods

There are only a few educational or assessment tools that exists in the information security area. To build the *EduCQ* Software, some examples of those tools were studied that help not to reproduce certain mistakes, or to improve the ergonomics of the software. Some former students' guidelines and researches results were also really useful to progress. The software was finally implemented using a web based scenario database, enclosing some Case study Scenarios and some Multiple-Choices Questions. A preference for multimedia has been chose for the first assessment mode (audio) and for both assessment the *ISO 17799* standard (ISO committee, 2000) has been implemented as question type. This standard, after an overview of the different security standard, appears to be the most relevant for my project: indeed the similar tools were all using this standard, also known as BS7799, provides best practice recommendations for information security management. It helps identify, manage and minimise the range of threats to which information is regularly subjected.

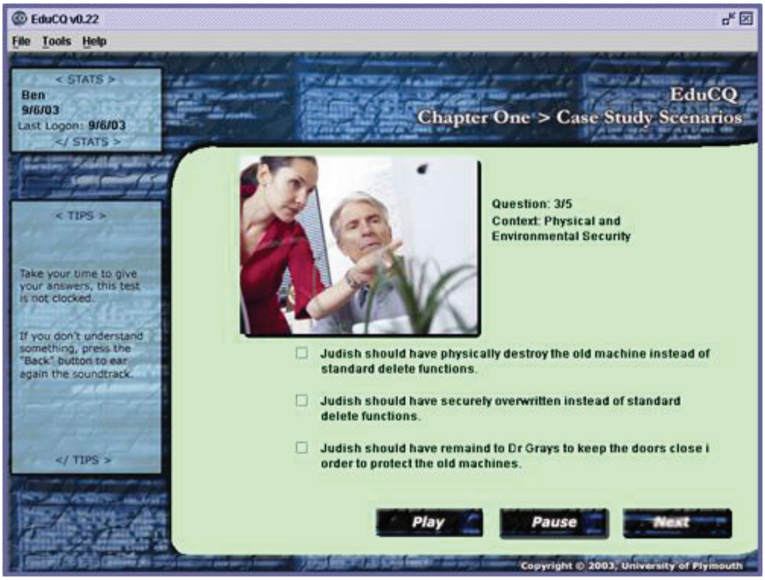
4.1 Case Study Scenarios

The case study scenarios are a complete set of situations that imply fictive employees from most background and profile, working in different kind of institutions that can be compared very easily to the reality. The format chosen for those scenarios (Warren, 2002), is comparable to some theater dialogs. The advantage of such a format is that it is particularly designed for a multimedia implementation. We can notice here that the guideline of Warren has been reached.

The simplest scenarios have been used, since it is much easier to keep the user attention with some dialog involving 2 or 3 peoples rather than 6. Those scenarios have been chosen for their relevance to the industry needs, identified in the first part. The principle here is a sound track to listen with three related propositions potentially true. The user has to choose the right ones. Five scenarios were implemented, which is not enough to propose a complete tool, but which is appropriate to start a product as *EduCQ* is.

4.2 Multiple-Choice Questions

The multiple-choice questions present in the system were completely retrieved from the *ISO 17799* standard. According to the leadership of this standard in information security promotion, it was capital to dedicate it a part of EduCQ. This part of the test is a question based on one of the *ISO 17799* assets, followed by 4 propositions. One of those propositions is wrong; the other one has also been extracted from the standard. 84 questions were implemented, covering about 80% of the standard scope: this standard is indeed cut in 10 parts, and 8 were implemented. The two last parts could be implemented to enhance the scope of the software.



< STATS >
Ben
9/6/03
Last Logon: 9/6/03
</ STATS >

EduCQ
Chapter One > Case Study Scenarios

< TIPS >

Take your time to give your answers, this test is not clocked.

If you don't understand something, press the "Back" button to ear again the soundtrack.

</ TIPS >

Question: 3/5
Context: Physical and Environmental Security

- □ Judish should have physically destroy the old machine instead of standard delete functions.
- □ Judish should have securely overwritten instead of standard delete functions.
- □ Judish should have remaind to Dr Grays to keep the doors close i order to protect the old machines.

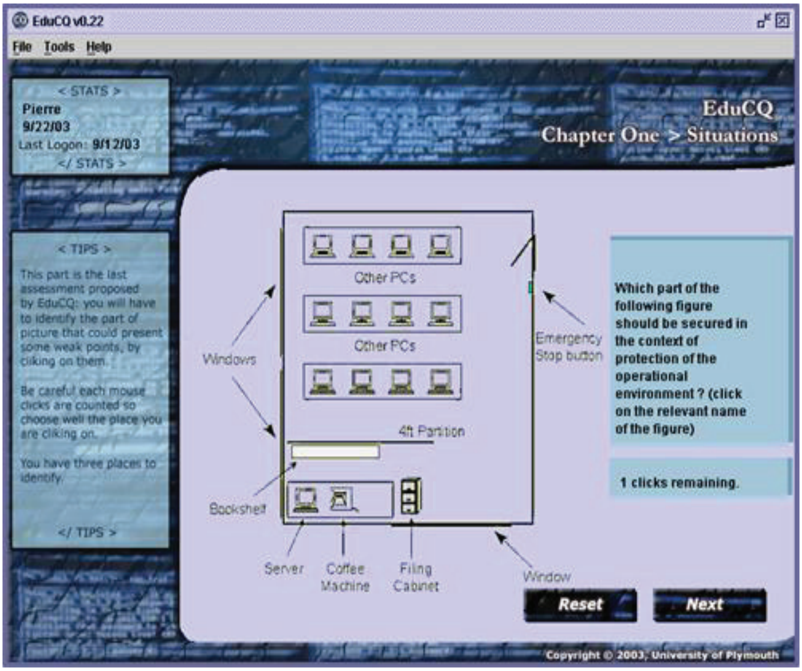
Copyright © 2003, University of Plymouth

Figure 3 : The EduCQ dialogs interface

To select and create the questions the election mode has seek to target the short questions, still with the idea to provide a pleasant assessment to the user. A second important point was the topic of the subject. As we already know the most important point of interest for companies involve with the "Communications and operations management" section of the *ISO 17799* standard. It is not surprising then to see that the Multiple Choices Questions enclose 21 questions of this type, which represents exactly 25% of all the Multiple Choices Questions.

4.3 Situations

The idea of the situation part is to provide another multimedia assessment type to the user. One aim was also to try to introduce more technical aspect of security in information system. Thus this part proposes an assessment which involve evaluation of potential weak points in a network. There is a clickable picture where the user has to identify the potential critical point of a topology. He has the right to click 3 times on the picture. If the places clicked are wrong the point are not given. This part of the software was implemented from the *"E-Tool Project"* (Gennatou, 2002). Figure 4 shows this interface



< STATS >
Pierre
9/22/03
Last Logon: 9/12/03
</ STATS >

< TIPS >

This part is the last assessment proposed by EduCQ: you will have to identify the part of picture that could present some weak points, by cliking on them.

Be careful each mouse clicks are counted so choose well the place you are cliking on.

You have three places to identify.

</ TIPS >

Which part of the following figure should be secured in the context of protection of the operational environment ? (click on the relevant name of the figure)

1 clicks remaining.

Figure 4 : The EduCQ situations interface

A reporting procedure is charged to collect the assessment information, so that it can give at the end the user a score. The 5 case study scenarios are marked over 30%, the multiple choices questions over 45% and the topology over 25%.

5. Further areas of research

During this project the adjustment between the diversity of ideas or discoveries and the technology restrictions has been a constant target to bear in mind. In term of pure research, the implementation of the scenarios has been performed at a quite good level for multimedia presentation, with a fair compromise with the educational requirements. We could think that soundtracks, MCQ or clicking areas on pictures are a bit limited in term of multimedia implementation, but as we have tried to be compliant with most systems and configurations, those tools were more or less the more advanced we could implement.

In term of quantity, the database has already been reasonably filled, as there are 84 multiple choices questions that were added to the previous student's researches. The level of quality of those multiple choices questions appears substantial even if they have not been properly tested with the regular protocol -benchmark, report... -. We can argue that the *ISO 17799* standard has been implemented in the software, according to the fact that 80% of the paper as been converted -when it was relevant for an educational tool- into multiple choices questions. Finally the software limitation in this case is more related to the *CSS* and obviously to the *Situations* which as only been introduced at a basic level.

The framework provided for this project is really a proper one to build an evolutionary project: indeed the technologies used are free and emergent, the system implementation has been traced wielding the best engineering practices, and a consequent technical documentation has been written.

There are several particular points that could be good to develop in the future in order to improve the *EduCQ* potential.

- **Create some web based analyzing tools to plot the reporting:** The *reporting* class developed in *EduCQ* allows the user to save his results under different file formats. This is quite useful but it could be much better to use some web tools to plot those results, such as *PHP* or *XSL*. One aspect could be to do some transparent reporting on our base in order to trace some user profile, to know more about the users' weaknesses. Another good improvement could be the possibility for the system to propose some more advanced views of their results -in term of layout-.

- **Increasing the Database:** This guideline could appear as obvious, but it has to be mentioned. The database encloses a short range of *scenarios* and *Situations* that needs to be populated. As the framework is working properly now, it could be quite an easy task -in term of technical complexity at least- to fill the database, especially with the administration website. No multimedia files are stored into the database, only the paths to those files, so that it even makes things easier. The 20% remaining multiple choices questions has to be extracted from the *ISO 17799*

- **Enhancing the third part multimedia functionalities:** The spirit of the software was to create a dynamic tool independent of the content, to implement an up to date content as each use, and also to avoid to distribute some heavy and precarious sounds files to the user. This goal has not been reached because as I had been unable to setup a streaming server on my personal system: the first reason is the narrow bandwidth I get, and the second one is the complexity to implement it in *Java*. As the *EduCQ* system should be installed on the University network, this problem could be fixed if somebody manages to implement it.

6. Conclusion

The dependency of most companies from any size upon information systems is not anymore to demonstrate. According to the most recent and reliable studies carried out on the information security issues, the situation seems to get worst, especially in some areas such as insider abuse of internet use or laptop theft. Those domains of misuse are highly related to miseducation of the company staff. As for any security matter you can have different policies to try to fix it. The *EduCQ Project* proposes an element of response to this issue by providing an educational tool to promote awareness and understanding of information security. Usually companies' staff suffers from a lack of time and even from a lack of resources to get an expensive formation. *EduCQ* solve the latter problem, since it is a free tool. In an area almost emptied of direct concurrence, the credibility of the tool has been consequently equipped by using an international standard as a basis for the software content implementation. The *ISO 17799*, coupled to an evolutionary CBT, allow the software to meet in many points the purpose that was previously set for *EduCQ*. Thanks to its predecessor -A. Warren and M. Gennatou project-, but also thanks to my personal researches, the database enclose about one hundred potential questions from different range of complexity and different domains of security concern that should by asked to the user. The multimedia implementation will give to the project a real added value: in combination with the serious fashion provided by the more regular MCQ completely based on the *ISO 17799* standard, the general aspect of the software remains professional. The framework provided during the software implementation, unifying the best engineering practice, and a wide scope of complementary technology, ensured that this new project start a solid basis, oriented towards the future. This tool, which is completely relevant to the actual situation, can be enhanced to be properly called a professional free solution. Despite this unfinished aspect, this tool is relevant to the company needs. There is actually a need on one hand, a lack on the other hand, *EduCQ* acts at this node. Hopefully this solution will know the brilliant posterity that the actual conditions and trend seems to offer.

7. References

S.M.Furnell, A.G.Warren, P.S.Dowland. 2003. "Improving Security Awareness through Computer Based Training", to appear in Proceedings of WISE 3 – 3rd World Conference on Information Security Education, Monterey, USA, June 26-28 2003.

CSI / FBI. *"Computer Crime and Security Survey"*. page 10, 2003.

ISO committee. *"ISO/IEC 17799:2000: Information technology; Code of practice for information security management"*. International Organization for Standardization, 2000.

S.M.Furnell, M.Gennatou and P.S.Dowland. 2002. "A prototype tool for information security awareness and training", Logistics Information Management, vol. 15, no. 5/6: 352-357.

SGS Group, *"Training Directory; 2001/02 course calendar",* 2001,
http://www.sgs.co.uk/training/publications/SGS2002broch.pdf, Accesed: 21th August 2003

Assessing Global Internet Accessibility

M. A. Alharbi, S.M. Furnell and B.V. Ghita

Network Research Group, University of Plymouth, Plymouth, United Kingdom.
e-mail: nrg@plymouth.ac.uk

Abstract

The Internet provides benefits to individuals and organizations as it serves all sectors of society. It is a powerful, fast, and inexpensive way of communication and exchange of information. But for many it is still just a luxury. This paper attempts to identify the limitation behind low connectivity by assessing the level of accessibility and its cost.

A web-based questionnaire was used in order to gather information from individuals and organizations about their means of connecting to the Internet and how much it costs. It also sought their opinions about their connections. The total number of responses was 412 from 41 different countries. However, the number of responses per country was not balanced. Thus profile were built to countries with meaningful set of responses, those countries are as listed by the highest number of responses and in descending order ; Saudi Arabia, United Kingdom, China, India, France, Greece, United States, Egypt and Mexico. The profile includes the means of connection and the cost of access.

The research showed that the telecom policies are very important to the development of the telecommunication infrastructure as well as the adoption of the new technologies.

Keywords

Internet Access, Digital Divide.

1. Introduction

The Internet has grown to an extent were it become essential to the lives of many people and in many cases influenced their life-style. On the other hand, there are people who do not even have a telephone line, let alone a PC. The gap between those who has access to the internet and those who do not is widening causing what so called Digital Divide (Norris, 2000).This gap can be measured by different models. The model adopted by this research was the measures of the Internet limitation by the assessing the level of accessibility and its cost.

The model used helps in drawing a general picture of the Internet situation in terms of availability of the means of connections the typical used methods and the charges associated with the connection. The information is gathered from the end-user.

The same approach was adopted by study carried out by Ngini (2001), The study concluded; that the levels of Internet access in the developed economies are generally higher, there is consequently a deepening uneven distribution of quality access infrastructures between developed countries and developing economies of the word, notable, the Africans and middle Eastern , the Asian countries appear to be winning the war on infrastructure development and low levels of Internet access, and the telecommunication industry in most of the developed nations is fully deregulated, and the governments provide enabling environments for private sector participation in capacity building (Ngini, 2001)

2. Questionnaire and Methodology

In order to achieve the objective of the research, the method of open web-based questionnaire was utilized to help gathering information from individuals and organizations about their means of accessing the Internet, how much it costs to be connected d their opinions on the connection. The questionnaire was available online between 16 July 2003 and 29 August 2003.

The address was emailed to personal contacts of the author requesting them to take part in the survey and direct others to it. The URL of the survey was sent to 128 Internet users in 22 countries, only 116 emails were delivered.

The questionnaire consisted of 21 questions divided into three groups;

- Personal (to help establishing the user's profile)
  - Country and place of residence.
  - Gender, age and highest educational qualification.
  - Employment status and whether the user is private, corporate or both.
  - Monthly income.
- Internet connectivity and level of access
  - Where does the user access the internet from (i.e. home, work, public library, café, or university/collage)?
  - Access Technology (i.e. Dial-up, ADSL, Cable, ISDN, Mobile, Satellite, or Wireless).
  - Cost of Access.
  - Weekly hours spent online.
  - What does the participant use the Internet for?
  - Awareness of faster access technologies availability and what are the obstacles towards obtaining them.
- Opinions
  - Internet speed adequacy to own requirements.
  - What is considered as limitations in the users region.

The survey was available in six different languages; Arabic, Chinese, English, French, Greek and Spanish. The variety of languages was aimed to obtain a larger and more balanced response from over the world.

The total number of response was 412 from 41 countries across the world. Unfortunately, no responses were received from Australia/Oceania and only one response for South America. Therefore, Brazil was included with North America in the analysis and both represented as the Americas. The responses received are detailed in table 1.

Africa	**17**
Country	*Responses*
Egypt	11
Libya	2
Morocco	1
Nigeria	2
Reunion	1
Asia	**217**
Country	*Responses*
Bahrain	3
China	34
India	34
Indonesia	1
Iran	1
Jordan	5
Kazakhstan	1
Kuwait	14
Lebanon	1
Malaysia	2
Oman	3
Pakistan	8
Russia	1
Saudi Arabia	93
Syria	2
Taiwan	1
Thailand	1
Turkey	1
United Arab Emirates	8
Yemen	3

Europe	**146**
Country	*Responses*
Albania	1
Czech Republic	1
Denmark	1
France	28
Germany	7
Greece	27
Italy	5
Netherlands	2
Poland	4
Slovakia	1
Spain	7
United Kingdom	62
North America	**31**
Country	Responses
Canada	3
Mexico	9
United States	19
Australia/Oceania	**0**
Country	*Responses*
South America	**1**
Country	*Responses*
Brazil	1

Table 1 : Total responses

3. Summary of infrastructure and cost

Representative countries were chosen from each continent included in study. Those countries were chosen due to their high number of responses. Those countries are; Saudi Arabia, United Kingdom, China, India, France, Greece, United States, Egypt and Mexico.

3.1 Saudi Arabia

The Telecommunication infrastructure of Saudi Arabia is modern. Yet, most users complain about the speed and the cost, which could be blamed on the telecom policies. However, with the establishment of the new regulatory body called the Saudi Communication Commission, the situation could improve. Since the government has realized the importance of telecommunication factor in the growth of the economy.

The cost of using dial-up in Saudi Arabia varies between £15-25 for the unlimited monthly access; there is an extra £0.50 per hour charged by the Saudi Telecom company (STC). While the ADSL charges were set by STC to £50 (SAR 300) installation fee and £37 (SAR 220) monthly charges.

3.2 United Kingdom

The UK has one of the most advanced telecommunication infrastructures in the world, yet the cost of the connection is regarded as high. This is due to the high taxation by the government (Carmel, 2000). On the other hand, their lead in privatization the telecom sector should be accredited. It allowed competition which is a key factor in providing good telecom services.

UK's typical home means of connection are dial-up, ADSL, and cable. Unlimited dial-up connection would cost between £ 6.49 to £18 per month and the ADSL between £15-35 per month.

3.3 China

China's domestic network and services are unevenly distributed (CIA- The world fackbook, 2003). Low teledensity, which is the number of phone lines per 100 inhabitants, and GDP both characterize the Chinese telecommunication infrastructure. Nonetheless, methods of access like satellite, cable and wireless are available.

The typical home connection would be either dial-up or ADSL. It would cost about £8.64 to have a monthly unlimited dial-up access; there is £ 0.10 per minute charged by the telecom provider. ADSL connection in China would cost between £ 34.24 – 36.64 per month.

3.4 India

The policies introduced by the Indian government have helped accelerate the development in the Information Technology sector. The users can gain access using satellite, cable, mobile, and wireless. However, dial-up, ADSL, and cable are the typical home means of connection.

The cost of unlimited dial-up access is between £ 12.5 – 14 monthly excluding the telecom provider charge which would be between £ 0.01 -0.06 per min. While, the ADSL connection would cost between £13.80 and £27.80.

3.5 France

France has one of the advanced telecommunication systems (CIA- The world fackbook, 2003). However, it fell behind in the area of Internet and Internet connectivity as perceived by the public as not useful or unnecessary ref. Nonetheless the cost plays an important role and also the government policies.

Cost of dial-up unlimited range from £ 10 to 17 a month and the ADSL would go up to £ 40 from £20.83 per month but on average it would be £26.72. Users from France were the only among the users from the countries listed in this paper to use wireless to access the Internet from home.

3.6 Greece

The Greek government has adopted new policies towards telecommunication in general which is expected to boost the telecommunication sector. These policies have led to roll out ADSL, after a long wait.

Greece was the only country with ISDN home connection among the list countries and one of three from whole list in the survey. The cost of unlimited dial-up access is about £10 to 30.50 per month excluding the £ 0.24 -0.69 charged by the telecom provider. The telecom provider charges goes down if the monthly subscription is high. ADSL monthly subscription is about £40.

3.7 United States

United States has very large, technologically advanced communication systems (CIA- The world fackbook, 2003) The estimated number of Internet users by April 2002 is 165.75 million (Nau, 2003). Despite the large number of users, there are still 40.9 % with no Internet. America is like any other country suffers the accessibility diffusion.

The typical home methods were dial-up, ADSL, and cable. The cost of unlimited dial-up access was between £ 6.34-19 per month. While ADSL connection would cost between £18.99-37.41 per month.

3.8 Egypt

Egypt had opened its doors to foreigner investors, started privatization, and made the Internet free for everyone, the only charges subscribers have to pay is the phone call. The model was meant to encourage the population to use the Internet; however, this would work perfectly to those who do not access the Internet frequently and do not use the premium numbers.

Although, ADSL is available to users in Egypt only dial-up was used by home users to connect to the Internet. Cable was not available at all. There are two different rates for the dial-up; the premium rate cost £ 0.60 per hour and the normal rate would cost about £ 0.12 per hour.

3.9 Mexico

Mexico witnessed a tremendous growth in its telecommunication sector, since it has entered the competition phase and ended the era of the duopoly. Cost of access has gone done while the quality of service has increased (Thomasson 2003)

ADSL was not on the home users menu in Mexico and the dial-up was the only method used home users in Mexico; users would pay between £ 10.33-28 to get unlimited dial-up access.

4. Findings and Discussion

The survey showed that 66.50% of the respondents were males, 79.78% of the total respondents aged between 21-30 years old, 94.13% lived in urban areas, 71.15 % held a degree or postgraduate qualification, and the average monthly income was £1,082 . These characteristic comply with ITU Internet users' profile, which states that the Internet user is male, young, educated, urban, and wealthy.

The survey also reveled that users from Africa tend to access the Internet from Internet cafés more than any other users. Whereas, Internet users from Asia were more of a home-based users. Europe was the leading in providing Internet access to staff and students at universities as shown in figure 1.

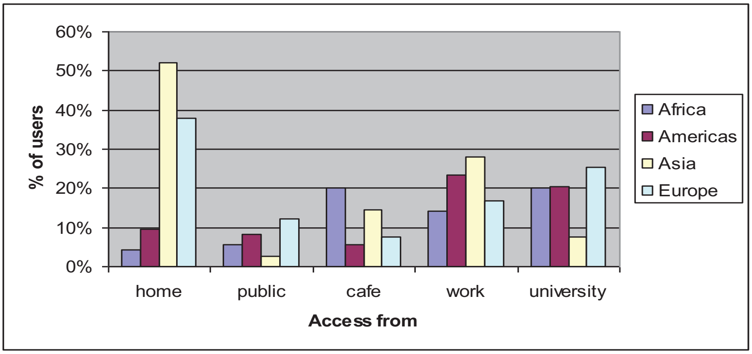


Figure 1 : Access locations

Africa was the least in access the Internet from home, which can be related to the low teledensity, the cost of personal computers and the cost internet subscription. Africa has teledensity of 2.70 (ITU, 2002) and 1.23 personal computer per 100 inhabitants (ITU, 2002).

It can be noted that none of the respondents from Africa lived in a village and they were mainly living in Capital cities. As Internet access in Africa seem to be privilege to the elite, those who accessed the Internet from other cities or towns were holding degrees or postgraduate qualification. In Asia only 1% of users lived in villages.

In general, dial-up is the dominant home connection, as it is the most affordable among the other means, as shown in figure 2. However, 43% of dial-up users were unhappy and complained about the speed. ADSL came in second place despite it high cost as regarded by many users as the main reason behind stopping them to use the faster method. ADSL was more popular in the Americas and Europe than Asia. In contrast, there was no single home user with ADSL in Africa. Participants from the Americas and Europe were satisfied with the speed of their connection with 75% and 70% respectively, though that their connection speed was adequate to their requirements.

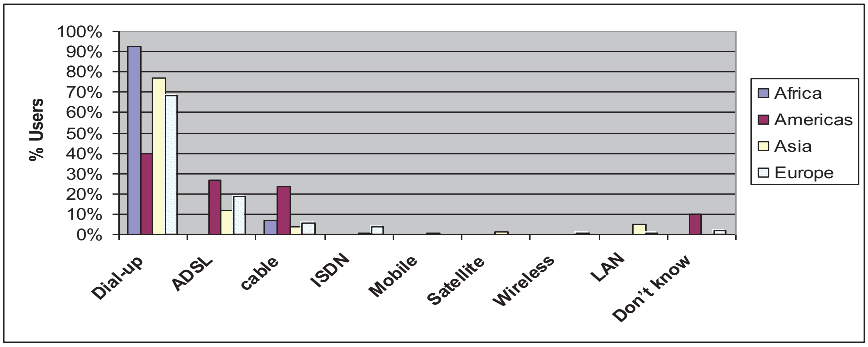


Figure 2 : Typical home connection

The results showed that 40% of the participants would spend between 2-10 hours a week surfing the Internet. Users from the Americas were more accustomed to surf the Internet for 3 -5 hours a day more than the other users as shown in Figure 3. Interestingly, 13 % of respondents from Africa tended spend more than 7 hours a day online, and this percentage was the highest among other users.

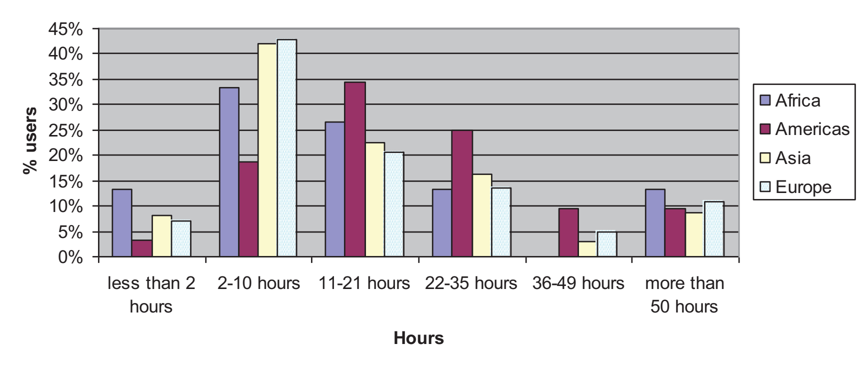


Figure 3 : Hours spent on-line

Participants were mostly keen on email, then downloading and news. Shopping has not yet gained a wide acceptance, as just 29.10% were online shoppers. Surprisingly, gambling was the last on list and adult materials came third from the bottom.

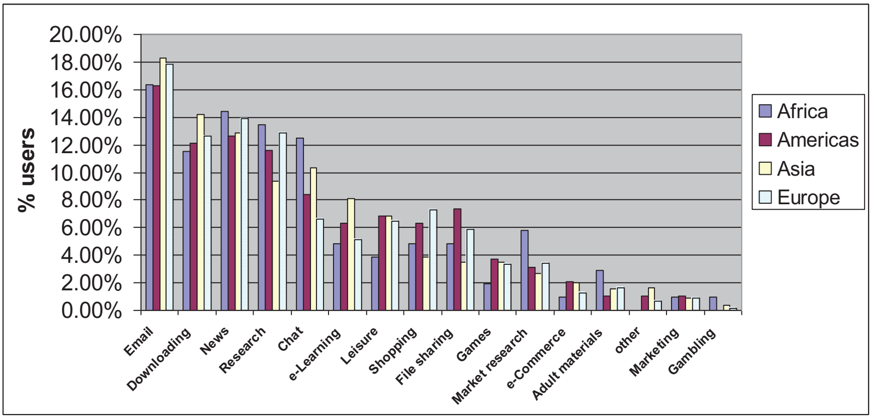


Figure 4 : Use of Internet access

5. Conclusion

This paper has presented profiles of 9 countries from four continents and highlighted the survey's results. The results from the survey showed that Internet users from Europe and America enjoy the variety of choice from methods of connection, place of connection and speed of connection. On the contrary, Internet users from Africa would have to live in urban areas in order to gain access.

Although varieties of connection means were available in the counties presented, their availability everywhere and the cost were the main issues. Users were reluctant to subscribe to ADSL, for instance, because of the cost associated. Others were welling to subscribe but the service was not provided in their area.

Developing countries seem to get a grasp of what has to be done to accelerate the development of their telecommunication infrastructure and the spread of using Internet. Thus, Regulatory bodies have been established, telecom sector has been privatized, and projects to provide citizens with internet access or facilitate the use of Internet have been launched. However, this is just the beginning of the journey and unless their execution is monitored, no real improvements would be achieved.

6. References

Carmel E. (2000). *UK: Internet Diffusion*[online]. American University in Washington D.C. Available from: http://www.american.edu/carmel/dh2982a/uk4.html [Accessed 5 august 2003].

CIA.(2003).*The world fact book* [online].America. Available from: http://www.cia.gov/cia/publications/factbook/ [Accessed 5 august 2003].

NUA. (2003). *How Many Online?* [online]. Available from: http://www.nua.ie/surveys/how_many_online/ [Accessed 6 July 2003].

Ngini, C., (2001).*The Internet-assessing the Global Accessibility*. Thesis [MSc.]. University of Plymouth, UK..

Pippa Norris. (2000). The Worldwide Digital Divide:Information Poverty, the Internet and Development [online]. Harvard University. Available from: *http://ksghome.harvard.edu/~.pnorris.shorenstein.ksg/acrobat/psa2000dig.pdf* [Accessed 4 august 2003].

Service Enforced Handover Algorithm

B. Murzeau[1], B. Lines[1], and P. Reynolds[2]

[1]Network Research Group, School of Computing, Communications and Electronics,
University of Plymouth, Plymouth, United Kingdom
[2]Orange Personal Communications Limited, St. James Court, Almondsbury Park, Bristol,
United Kingdom
e-mail: nrg@plymouth.ac.uk; paul.reynolds@orange.co.uk

Abstract

In a near future, standalone handover managers will become major actors within a mobile telecommunications network. Indeed, with the development of new network generations, users will need to switch from one generation to another transparently according to certain policies. Moreover, the emergence of new multimedia services will require new handover facilities. One of the most important features will be to disperse the traffic streams between a number of mobile networks (GSM, UMTS…) so that an optimum arrangement of traffics per network is achieved. This optimisation will be based upon the services QoS requirements and will be managed by a new handover algorithm, described in this paper. Hence, an architecture has been defined to implement the concept of this new handover algorithm, and to define the communication between the different entities of it represents. The results of this research are the creation of two new description languages; the *Mobility Manager Markup Language*, and the *eXtended Algorithm Markup Language*. This last language has been used to write the new handover algorithm, based on policy rules checked by a policy server.

Keywords

Handover, service, QoS class, policy, GSM, 3G, COPS, synchronous, asynchronous, XML, XML SCHEMA

1. Introduction

Anyone who has ever tried to make a telephone call from a mobile handset may have encountered a denied access because the network was too busy. Over the past few years, mobile phone use has expanded, and the existing networks have needed significant enhancements to meet the demand. ‘3G’ will introduce new IP-based services for mobile users: some of the new services (e.g. video streaming) will need a guaranteed level of QoS from the network in order to work properly and satisfy the users. In the current IP world QoS is handled by simply increasing the bandwidth, this approach does not fit into mobile communications, where resources are limited. Nevertheless, an IP approach to end-to-end QoS (between any two points in the world) is required to be able to build an ‘all-IP’ world of networking. With a QoS solution based on different QoS classes, also called traffic classes, linked with a quality manager (policy server), the use of the network resources can be optimised. This represents the aim of this project.

The first part of this research paper is composed of a brief review about the state of the art in handover. The second part resumes my researches on the development of the handover algorithm. It includes a section on *requirements*, where the parameters of the handover algorithm are determined. Next, the primordial notion of *architecture* is broached. Finally, in the *algorithm* section, a handover algorithm solution is proposed.

2. Overview of Handover

2.1 Basis

In mobile communications, *handover* is the mechanism that transfers an ongoing call from one cell to another as the mobile handset moves through the coverage area of a cellular system without any disconnection and to improve the general quality of service during the call (Hoi Leung Kung, 2002). Two kinds of handover are distinguished (see Figure 1): Intracell Handover & Intercell Handover, (within these two sets of handover are the two subsets called Hard and Soft handover).

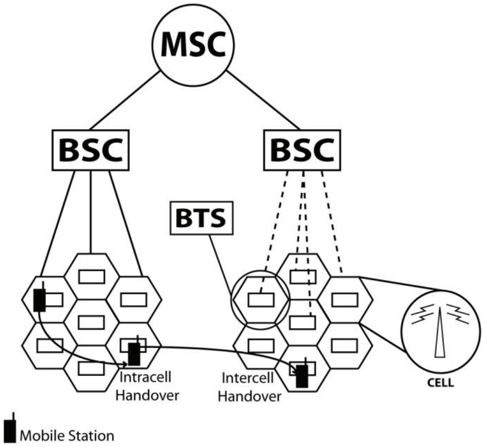


Figure 1: Intracell / Intercell Handover

2.2 State of Art in Handover Solutions

Handover is an essential element of mobile network engineered solution. It is composed of two elementary parts:

- Radio network controlled handover: This kind of handover is based upon the evaluation of criteria, such as *carrier to noise ratio*, *interference*, *and power control*. There are two basic algorithms used, both closely tied in with power control: the Minimum acceptable performance algorithm, and the Power budget algorithm (Schwarz, 2001).
- Network managed handover: Network managed handover relies on such criteria as ongoing network maintenance operation, or more particularly on the current *traffic load* of the cells. Traffic reason handover is a condition that has been added to the existing handover algorithms.

Recently, a new form of handover arose.

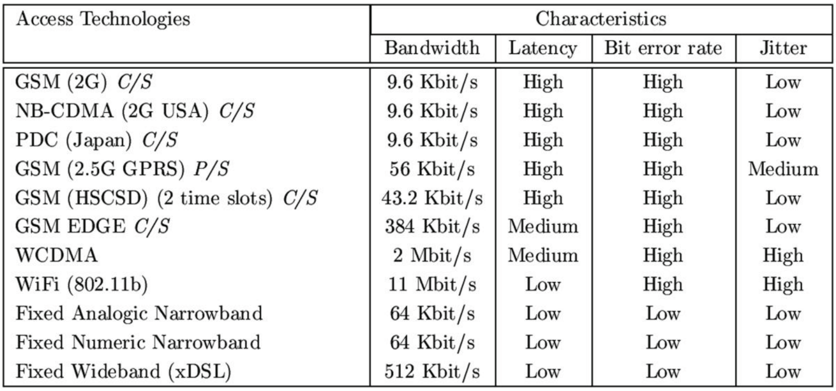
2.2.1 Handover Manager & Network Policy

Professor Paul Reynolds, from Orange PCS Limited, has invented and patented a new concept of handover: *Handover in accordance with a network policy* (Reynolds, 2003) and an associated handover mobility manager architecture (Reynolds, 1997). His invention provides a method of controlling handover of a mobile station conducting a communications session in a mobile communications network; the network including a plurality of radio access domains. The method is based on a handover trigger received by the handover manager indicating a requirement of a handover decision.

3. Requirements

3.1 Access Technologies

Access technologies represent all the technologies of mobile or fixed telecommunications networks through which services might run. Each access technology has its own characteristics; therefore common features must have been chosen in order to compare them. Bandwidth, Latency, Bit Error Rate, Jitter are the most relevant characteristics for access technologies. Table 1 lists the relevant access technologies and defines the values of their characteristics.



Access Technologies	Characteristics			
	Bandwidth	Latency	Bit error rate	Jitter
GSM (2G) *C/S*	9.6 Kbit/s	High	High	Low
NB-CDMA (2G USA) *C/S*	9.6 Kbit/s	High	High	Low
PDC (Japan) *C/S*	9.6 Kbit/s	High	High	Low
GSM (2.5G GPRS) *P/S*	56 Kbit/s	High	High	Medium
GSM (HSCSD) (2 time slots) *C/S*	43.2 Kbit/s	High	High	Low
GSM EDGE *C/S*	384 Kbit/s	Medium	High	Low
WCDMA	2 Mbit/s	Medium	High	High
WiFi (802.11b)	11 Mbit/s	Low	High	High
Fixed Analogic Narrowband	64 Kbit/s	Low	Low	Low
Fixed Numeric Narrowband	64 Kbit/s	Low	Low	Low
Fixed Wideband (xDSL)	512 Kbit/s	Low	Low	Low

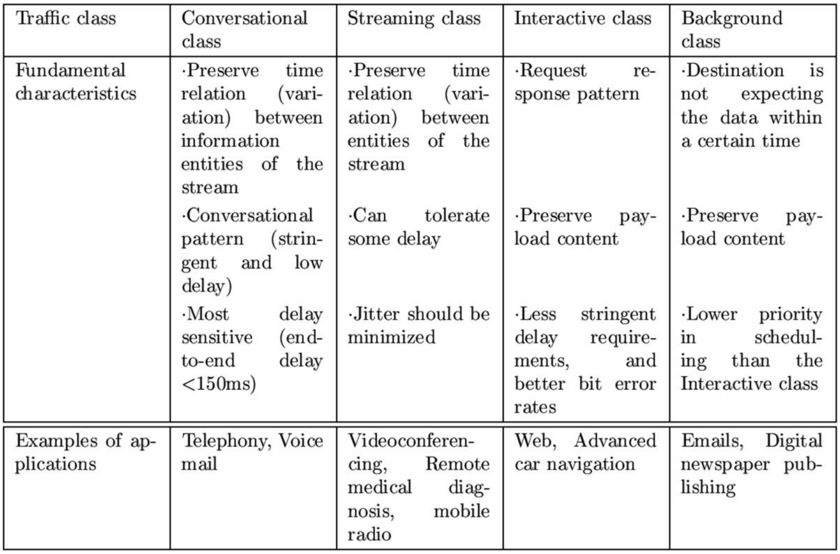
Table 1: Access Technologies Characteristics

3.2 Services

3GPP has defined four UMTS QoS classes (3GPP, 2003), also referred to as traffic classes. In fixed networks complex mechanisms have been defined, but for UMTS restrictions and limitations of the air interface have to be taken into account. The QoS mechanisms provided in cellular networks have to be robust and capable of providing reasonable QoS resolution (3GPP, 2003). The four different QoS classes are summarised in Tables 2 and 3.

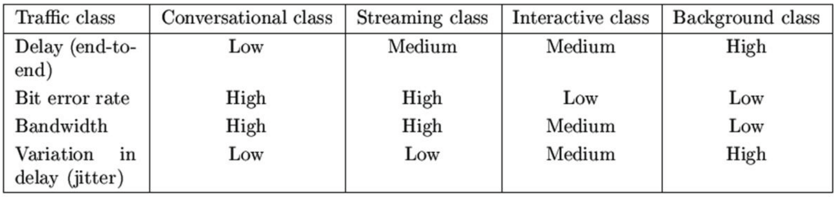
3.3 Protocols

In the future sophisticated applications, such as multimedia applications, will use several flows, i.e. protocols involved in the communication. These protocols are needed in different phases of multimedia sessions, and they may have different traffic characteristics. Therefore, a careful consideration is needed for each of them to identify their properties, resumed in Table 4.



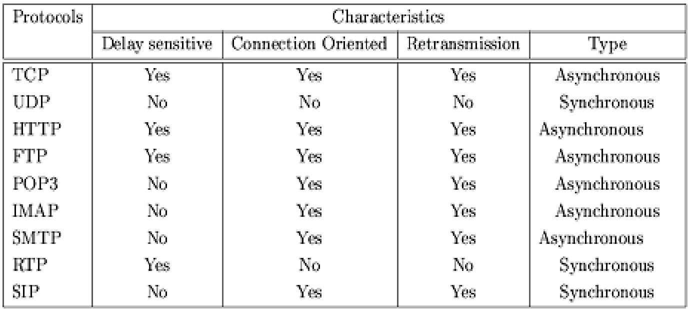
Traffic class	Conversational class	Streaming class	Interactive class	Background class
Fundamental characteristics	·Preserve time relation (variation) between information entities of the stream ·Conversational pattern (stringent and low delay) ·Most delay sensitive (end-to-end delay <150ms)	·Preserve time relation (variation) between entities of the stream ·Can tolerate some delay ·Jitter should be minimized	·Request response pattern ·Preserve payload content ·Less stringent delay requirements, and better bit error rates	·Destination is not expecting the data within a certain time ·Preserve payload content ·Lower priority in scheduling than the Interactive class
Examples of applications	Telephony, Voice mail	Videoconferencing, Remote medical diagnosis, mobile radio	Web, Advanced car navigation	Emails, Digital newspaper publishing

Table 2: QoS Classes



Traffic class	Conversational class	Streaming class	Interactive class	Background class
Delay (end-to-end)	Low	Medium	Medium	High
Bit error rate	High	High	Low	Low
Bandwidth	High	High	Medium	Low
Variation in delay (jitter)	Low	Low	Medium	High

Table 3: QoS Classes Requirements



Protocols	Characteristics			
	Delay sensitive	Connection Oriented	Retransmission	Type
TCP	Yes	Yes	Yes	Asynchronous
UDP	No	No	No	Synchronous
HTTP	Yes	Yes	Yes	Asynchronous
FTP	Yes	Yes	Yes	Asynchronous
POP3	No	Yes	Yes	Asynchronous
IMAP	No	Yes	Yes	Asynchronous
SMTP	No	Yes	Yes	Asynchronous
RTP	Yes	No	No	Synchronous
SIP	No	Yes	Yes	Synchronous

Table 4: Protocols Characteristics

4. Architecture

4.1 Handover Model

A simple handover model comprises three phases, the *information gathering* or measurement phase, the *decision* phase and the *execution* phase. The phases' have been defined by Orange PCS Limited (Reynolds, 1998).

4.2 Physical Diagram

The architecture is a fundamental element of this project. Everything is organised around it, and implementation choices are made thanks to it. It is composed of a set of building blocks, and does not provide any functionality. The following architecture, represented by Figure 2 has been proposed by Orange PCS Limited (Reynolds, 2000).

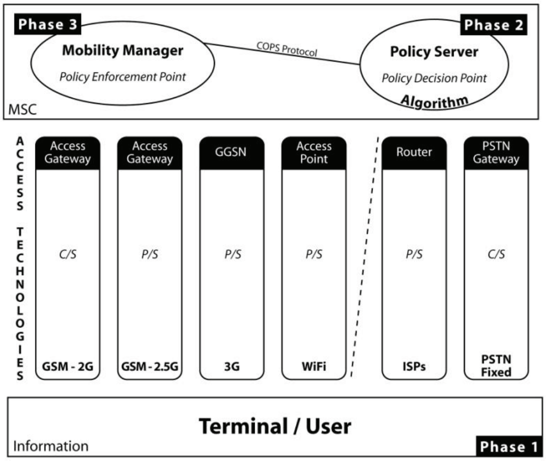


Figure 2: Mobility Management with a Standalone Policy Server

4.3 Syntax & Semantic

The mobility manager communicates with the policy server by way of the COPS protocol. Two questions arise: what kind of information they will exchange, and how they will understand each other. The mobility manager sends “local” information to the policy server after a handover request from the mobile station, or a request from the policy decision point. Hence, this flow of data information should be formatted to enable the dialogue. Therefore, the creation of a specific language appears to solve out this issue. We have named this language the *Mobility Manager Markup Language*, or *3ML*. The use of the XML technology with a XML Schema is designed to be self-descriptive and thus represents the easiest solution to create a proprietary descriptive language.

4.3.1 Description in XML

The semantic is described here. A typical characteristic definition entry is (in XML):

```
<bandwidth>
   <low> &lt; 30 kbit/s </low>
   <medium> &gt; 30 kbit/s &amp;&lt; 150 kbit/s </medium>
   <high> &gt; 150 kbit/s </high>
</bandwidth>
```

4.3.2 Language: MMML

The 3ML language is composed of a XML schema, which describes how the syntax of the XML files based on this schema is. It represents the format of the messages transferred from the mobility manager to the policy server. COPS will transport transparently 3ML messages (XML files).

5. Algorithm

An algorithm will be used for *inter* handover, which means handover between access technologies in our case. Intra handover will be managed by the access technologies themselves. The original objective of the handover algorithm is to spread the traffic streams between access technologies in order to minimise the QoS deltas. To do so, it has to resolve conflicts beget for example by the network policy, or the business criteria.

5.1 Interactions & Conflicts

Several conflicts will have to be dealt by the algorithm. For example, different network operators have different strategies when operating the systems. One may prioritise coverage and voice, and another real-time applications with guaranteed quality, but limited coverage, etc.

5.2 Algorithm Description Language

The algorithm is implemented in the PDP. The PDP is primarily composed of a policy server, and a policy repository where the policy rules are recorded. The algorithm is deployed by creating policy rules. Handover policy rules will be written in a second new language we have developed called eXtensible Algorithm Markup Language (XAML) XAML.

5.2.1 eXtensible Algorithm Markup Language

XAML is a pure invention as the previous Mobility Manager Markup Language. It defines the common information model for handover policy rules. This new language is created thanks to the XML Schema technology. The policy rules and data will be encapsulated in a XML file, and then the logic will be checked by the valid XML Schema to certify that the policy rules are XAML compliant.

5.2.2 XAML Rule Example

```
<!-- Outclass Policy -->
<!-- Last Modified 8 September 2003 -->
<!-- To satisfy the user, the network will handover
a session, if a better QoS is available -->
<policy_rule label="outclass">
  <head>
    <action>handover</action>
    <parameter>BaseStationID</parameter>
    <parameter>ServiceQoSClass</parameter>
  </head>
  <body>
    <event>GetNetworkParameter</event>
      <condition>
        <and test="ServiceAvailability">
          <parameter>BaseStationID</parameter>
          <parameter>ServiceQoSClass</parameter>
          <not test="LowerQoS">
            <parameter>BaseStationID</parameter>
            <parameter>ServiceQoSClass</parameter>
          </not>
        </and>
      </condition>
  </body>
</policy_rule>
```

6. Conclusion

The development of the handover algorithm has not given so relevant results directly, because all the rules are not written, and hence I could not test it. In my approach, I think that I did not enough use the Tables of chapter *requirements* for the definition of the policy rules. However, the statements of the main facts of my findings are:

- A solution for implementing a handover algorithm in a policy-based network constituted of a plurality of different access technologies.

- A proposition about how communicate between a mobility manager and a policy server (COPS + MMML).

- Definition of a language (XAML) to describe policy rules of a handover algorithm.

As a further work it could be a particular point to improve, by realising XAML updates. Indeed, there is no hierarchical notion in the current XAML. It could be useful to have some root rules that rely on lower rules. This can be achieved by using "nesting" in XML, afterwards new policy rules should be added to the policy repository. Another feature of the algorithm has also been missed: *degradation* of service QoS requirements. In case of no network available to meet the service QoS requirements, a degradation of service could be processed, to improve the probability of handover the communications session successfully. Finally, it should have been good to represent the handover algorithm in SDL/GR, to compare those two representations and see which one is the best (clarity, time of development, etc.).

7. References

Hoi Leung Kung, L. (2002), "Handover Mechanisms in 2G and 3G Networks", *University of Technology, Sidney*. Available: http://members.fortunecity.com/shit130878/capstone/ [Accessed: 10th August 2003]

Reynolds, P (1997) Patent - Remote Management of Handover'

Reynolds, P (1998) 'Mobility in long-term Service Architectures and Distributed Platforms'

Reynolds, P. (2000) 'Mobility Management for the Support of Handover on Heterogeneous Mobile Environment'

Reynolds, P. (2003) Patent - Network Handover Policy

Reynolds, P. (2003) IETF Draft - Motivation for Network Controlled Handoff using IP mobility between heterogeneous Wireless Access Networks

Schwarz (2001) "GSM: Network Aspects", University of Duisburg Essen. Available: http://www.fb9dv.uni-duisburg.de/education/comnet4/ [Accessed: 18th August 2003]

3GPP (2003) *Quality of Service (QoS) Concept and Architecture*, Technical Specification TS 23.107 – v5.9.0.

Section 2

Communications Engineering
& Signal Processing

Modelling Human Perception for Objective Prediction and Assessment of Video Quality in Packet Networks

E. O. Madu and E. C. Ifeachor

School of Computing, Communications and Electronics, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, United Kingdom.
e-mail : Edwin.Madu@plymouth.ac.uk and E.Ifeachor@plymouth.ac.uk

Abstract

Complex and non-linear processes in modern multimedia transmission, such as low bit rate codecs for data reduction have been recently popular. It therefore limits the use of conventional engineering performance metrics in predicting perceived performance because of its non-linearity. There is therefore a need to have a reliable objective assessment method based on perceived performance for optimal design, commission and monitoring of quality. This paper looks at how quality can be measured, optimised and propose a method of measuring quality. It also suggests a method for transmitting video such that the quality will be improved.

Keywords

Quality, Human perception, Video packet, MOS, MPEG-2.

1. Introduction

The success in streaming media business depends on vital factors like quality and cost. Signal compression is widely used to fit media clips into available bandwidth. Yet compression can damage quality and cause jerky video, blurred images, and choppy audio. Media service providers then find it hard to offer on-line media subscriptions and promote billable video-on-demand services. It is therefore important to:

- To understand behaviour of key quality parameters in IP video communication.
- To select and improve on a novel and robust method for predicting quality which takes into account user perceived quality.
- To investigate the relationship between system parameters and subjective quality.
- Understand the compound problem of transmitter video through packet network.

The use of Human Perceptual models have long have been used to study the quality of video transmissions, starting from analogue television. With the move to packet transmission and wireless networks, transmission quality has been impaired by various means including compression and decompression algorithms, delay, and loss in transmitted packets, noise, and other distortions. Its results are bit error rate, fading, and diminishing bandwidth. Network providers and equipment manufacturers have been in search for a way to quantify these impairments in a consolidated, unified, and "end-to-end" manner. This paper looks at the different procedure and recommendation on these methods; it is with the aim to ascertain their effectiveness and which concepts are relevant in test measurements.

1.1 Transmission of Video over IP network.

Video transmission involves sending data, which are usually too large for raw transmission or storage, so the video streams are mostly compressed. All the popular compression techniques like H.261, H.263, MPEG-1 and MPEG-2 are lossy. MPEG (Motion Picture Expert Group) is a popular standard used today, which means to achieve a higher compression rate some information in the original image may be lost during the compression and cannot be recovered when decoded. Thus, the compressed video streams may have lower quality than the original ones. The higher the compression rate, the lower the size of the frame, and vice versa. To achieve a high compression rate, temporal redundancies of subsequent pictures must be exploited. MPEG distinguishes 3 mainframe types of image coding: I-frame, P-frame, and B-frame. To support fast random access, intra-frame coding is required. I-frame stands for Intra-coded frame. They are self-contained. The compression rate of the I-frames is the lowest. P-frame stands for Predictive-coded frame. The encoding and decoding of P-frames requires the information of previous I-frames and/or all previous P-frames. Compression rates of P-frames are higher than that of I-frames. B-frame stands for Bi-directionally predictive-coded frame. The encoding and decoding of B-frames requires the information of the previous and following I- and/or P-frame, but achieves the highest compression rate. The encoding pattern of this stream is IBBPBBPBB, where the last two B-frames depend on both the second P-frame and the next I-frame, the loss of one P-frame can make some other P- and B-frames useless, while the loss of one I-frame can result in the loss of a sequence of frames. In MPEG encoded video streams, I-frames and P-frames are more important than B-frames.

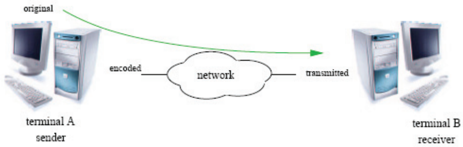


Figure 1 : Video transmission over Packet Network

And has led to the use of high compression techniques, and transmission over packet network has very small control on the arrival time for transmitted packet. For this reason, real time transmission, delayed packets could be included or discarded on arrival, and in each case will cause errors at the receiver. Although standards with give priority to certain packets are currently under development and it will help to reduce delays.

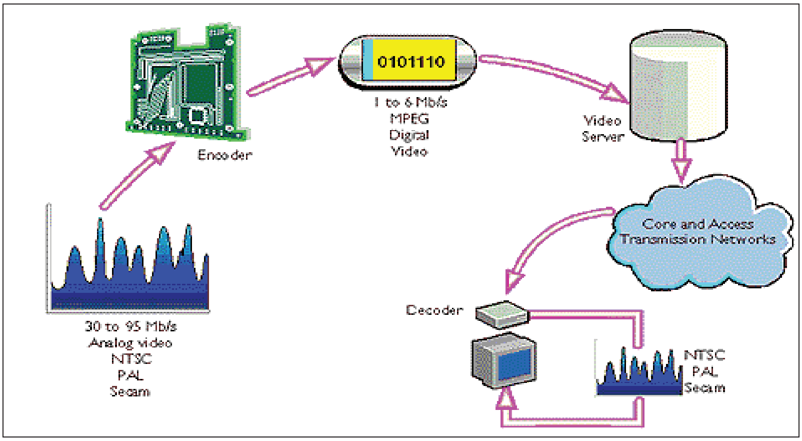


Figure 2 : Basic video architecture [1]

The basic procedure of video communication over packet network involves the analysis of the analogue video signal. This will include operations like filtering, analogue to digital conversion, computation of transform coefficients, or correlation of the pixels with a particular vector quantization pattern. The accuracy of the output depends on the number of bits used to represent it, and typically 8 to 12 bits. Usually no compression is done with the analysis. Data is only transformed to a format that is more compressible than the original signal format.

The quality of transmission of a video communication depends essentially on the bit rate, travel delay, jitter and loss [3]. The project is focused more on the bit rate and packet loss and this will be explained:

- **Bit Rate**
  This relates to how the frames are refreshed per time. A screen refreshed at 30 times per second produces good quality. The image size depends on the particular produce and the selected user option. Considering a case of MPEG-2 with image of 720x 480 and having 256 colours. The ideal bit rate will be calculated as follows:

  = 30 updates per second x 720 pixels x 480 pixels = 2654.208Mb/s

  Clearly, it is high for practical transmission across networks. Therefore, compression and encoding techniques will have to be implemented. This is the where codecs play vital roles. Codecs based on MPEG-2 and MPEG-4 are sometimes used, particularly for packet transmission. The choice of compression algorithm depends on the available bandwidth or storage capacity and the features required by the application

[4]. A choice of MPEG-2 provides the capability to compress in either NTSC[1] or PAL video with average bit rate of 3 – 6 Mbits/s with good quality.

- **Packet loss**
  Having adopted a particular image size, and nominal bit rate, the next issue to be accounted for is its effect on how the bit stream is transported across the network. A major question is the percentage of packets, which gets lost. And this will include both the packets, which are lost in the transport, plus packets that arrive too late to the buffer before being displayed on the audiences' screen. Packet losses are mainly due to router queues that are temporarily filled due to bursts of congestion. The extent of the impact depends on the size of the burst loss and codec type [5].

2. Quality Evaluation

Recently a large number of real-time multimedia applications over packet networks has rapidly increased, and consequently therefore is the need to assess the quality. This quality has continuously been studied but has been focused on addressing Quality of service issues, which is on a network basis rather than from the end-users point of view. This is because it is the end user that determines how successful a service or an application is. Hence quality is not all about providing large bandwidth reservation for transmission [6], but providing optimal conditions for successful transmission. This is because the difference between requirement to provide minimum quality for a particular task and the maximum point beyond which increased quality has no benefit for the user.

In digital communications, there is common use of statistical approach to deal with image content and also to generate a psycho-visual relationship. Also with the use of compression, resulting artefacts tend to be inconsistent or irregular.

Firstly, we note that quality is difficult to define in a straightforward and generic manner [7]. It depends on the application and is wholly dependant on the human user's degree of perception. Two different kinds of quality metrics have been adopted: online and offline. Off-line quality monitoring is used to gain in-depth understanding of the specific application behaviour and requirements in terms of quality. Here, different methods, techniques, disciples and combinations of these may be used:

- User studies and trails, including subjective quality assessment tests, questionnaires, specialised psychology tests, and interviews with user groups, application designers and service providers.
- Objective measures that use computational models to quantify the quality given a certain

International Telecommunications Union (ITU) provides the most widely used method for measuring the subjective quality of video images. These are known as recommendations and are made to address subjective assessment of multimedia applications. Image quality is

[1] NTSC is a standard created by National Television Standards Committee and scans 525 lines per second.

assessed using the single stimulus method, which can be either a quality scale or impairment scale or the double stimulus impairment scale.

2.1 Video Quality Impairments

This section briefly discusses the various types of distortion of compressed digital video transmission, and then a review of work on objective quality metrics. Transmission of video material is subjected to 2 main types of distortion:

(a) **Distortion due to lossy encoding**: when original video is encoded, either in real-time or off-line in order to reduce its bandwidth requirement, distortion at the first level is introduced.
(b) **Delay variation, and packet loss**: Transmitted packet stream over network faces problems like delay, delay-variation and packet loss, which therefore makes the information unavailable to the decoder.

The common types of distortion introduced during encoding and transmission of digital video are:

- **Encoding Artefacts**
  Large numbers of encoder are based on Motion compensation (MC), Discrete Cosine Transform (DCT) of the blocks of pixels, and quantisation of the resulting transform coefficients. The major source of encoding distortion is quantization of the transform coefficients, and encoding parameters such as frame dropping. The main types of artefacts in a compressed video sequences [7] are:

  (a) **Noise**: Noise is the perceptual measure of high frequency distortions in the form of spurious pixels. It is most noticeable in smooth regions and around edges (edge noise). This can arise from noise recording equipment, or in the compression process.
  (b) **Blocking and tiling**: Blockiness is defined as [8] the distortion of the image characterised by the appearance of an underlying block encoding structure. Blockiness is due to the negligence of the interblock correlation, where an independent quantization of DCT coefficients in neighbouring blocks is used [9]. It is caused by independent quantisation of blocks, resulting in discontinuities at the boundaries of adjacent blocks. Tiling creates false horizontal and vertical edges at the block boundaries and it pattern makes it one of the most apparent visual distortion. Blockiness is common to all DCT-based image compression techniques, and because DCT is performed on 8x8 blocks in the frame, it leads to artificial borders between these blocks. It can also be due to transmission errors, which often affect the entire blocks in the video.
  (c) **Blurriness**: Blurring is a total distortion over the entire image, characterised by reduced sharpness of edges and spatial details. It is caused by the suppression of higher-frequency coefficients by a coarser quantisation. Blurriness is one of the main artefacts of wavelet-based compression technique, but DCT based compression are also affected by this artefact, but to a lesser extent.

- **Transmission Artefacts**

  Another source of impairment occurs in the transmission of compressed video stream over the packet network. The bit stream is fragmented into a series of packets, which are then sent out to the destination. Two different types of impairments are noticeable in this area: **Packet loss** and **End to end delay.**

  When packets are lost they are unavailable to the decoder, also long delays of packets are worthless to the application. Therefore both of them have the same impact: data availability. Its impact is dependent on the nature of the video encoder and the level of redundancy present in the compressed bitstream (for example, intra-coded bitstreams are more resilient to loss). For Motion compensated/ Discrete Cosine Transform codecs, like MPEG, interdependencies of syntax information can cause an undesired effect, which the loss of a macroblock may corrupt subsequent macroblock until the decoder can resynchronise. These results in error blocks within the image: these bear relationship to the rest of the image and usually contrast greatly with adjacent blocks.

3. Test methods

3.1 Subjective Assessment

Subjective quality measurement is done to get a user's perception and understanding of quality. This requires a lot of conditions. It can depend on environmental conditions, such as: viewing distance, screens brightness and contrast, observation angle, chromacity of the background and adequacy of room illumination [10]. It can also depend on if the video material is interesting to the user or not or the level interaction between the viewer. So the effectiveness of assessment depends as much on a careful description of what to assess as it does on the technical qualities of the assessment procedures used. This quality is distinct from image quality measured using mathematical procedures or computational models (i.e. the degree of distortion or difference between the original or reconstructed images) or the observed quality or image fidelity. An example is an image with higher contrast or slightly more color appears to appeal more to human viewers, even though, according to a strict mathematical interpretation of distortion.

Three assessment methods have been specified by the ITU-Recommendation BT. 500-10 [11], which provides standards for measurement quality. It specifies the experimental conditions such as the distance of view, viewing conditions (room lighting, display features, etc), selection of subjects and test material, and assessment and data analysis method. All these are done so that subjects can predict accurately their opinion based on their perception. This is the aim of a Mean Opinion Score (MOS). Mean Opinion Score is the average of grading obtained from human subjects from experiments.

3.1.1 SSCQE: Single stimulus Continuous Quality Evaluation

The single stimulus continuous evaluation is the method in which subject are presented with a series of video sequences only once and then they are made to grade the quality instantaneously in real time using a slider with a continuous scale. In order to capture changes in the quality, viewers are shown longer test vide, which is typically 20 -30 minutes. The reference is not presented, and viewers asses the instantaneously perceived quality by continuously adjusting the slider along the scale (from ‘‘bad’’ to ‘‘excellent’’). The slider can be implemented as hardware device or software. Instantaneous quality can therefore be measured by noting the value of the slider at a particular frequency. This allows analysis of the configuration of the system. The problem with the SSCQE is the load imposed on the viewer, and the comparison between different test sequences, since they all have different contents. Program tend to have significant impact on the SSCQE rating because of the non-linear influence of good or bad parts within the sequence can be expressed in the Minkowski metrics [12]. The viewer tends to have load imposed on him since it is quite difficult to trace the momentary changes in the quality, thus making the stability and reliability of the derived results problematic. Another problem also, is the synchronisation of the time between the video and the slider. This project has developed a user interface to enable proper measurement between the slider and the video test sequence. This problem was experienced in [13] but a solution interface will be explained in the Simulation Section.

3.1.2. Double Stimulus Impairment Scale (DSIS)

In this method subjects are allowed to watch multiple sequences, which are made up of the reference sequence and the test sequence. These sequences are made short and subjects rate the overall amount of impairment on a five level scale rating i.e. from Imperceptible clip to ‘very annoying’. The reference sequence is always presented before the test sequence and the method is very useful in evaluating visible impairments, such as noticeable artefacts caused by encoding transmission.

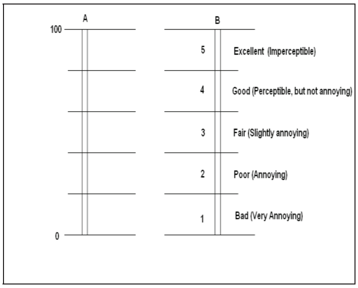


Figure 3 : ITU Quality Scale [14]

The problem with Double Stimulus Impairment Scale (DSIS) is that when using short test sequences, a digital video operating over longer periods of time generate substantial quality variations that may not be uniformly distributed over time.

3.1.3. Double Stimulus Continuous Quality Scale (DSCQS)

In this measurement, the viewers are shown multiple sequence pairs, which will include both the reference sequence and the test sequence. The sequences are shown in alternating manner in a random order. The sequences are usually short (about 8 to 10 seconds). The viewer doesn't know the reference in advance, and subjects will have to rate it in a scale with a range from 'bad' to 'excellent' and it has an equivalent scale from 0 to 100. One of the advantages of the DSCQS is that it provides better result when the quality of the reference and the test sequences are similar, if not it is easier for the subject to spot the difference between the two sequences.

3.2 Objective Assessment

Even though Subjective test is the most reliable means of having an understanding on the performance of digital video transmission systems, it is so expensive to set up and the complexity makes it so unattractive for automating the assessment method. With the use of human subjects in the process, the process is unusable when the quality monitoring systems have to be embedded into practical processing systems. Quality metrics are able to produce objectively obtained ratings present an attractive alternative.

A lot of research has been going on for years in the area of objective quality measurement. The first was in analogue video transmission, but with the development in digital video transmission, a new set of problem was introduced, leading to different impairments. This necessitated the development of quality metrics that consider the impact of encoding and transmission in the digital domain.

One of the common quality measurements is to calculate the distortion at pixel level. The peak signal-to-noise ratio (PSNR) measures the mean squared error (MSE) between the reference and the test sequences. This metric because of its simplicity is still being used, but it cannot describe the complex and multi-dimensional system like human visual system, and does not provide good predictions in many cases.

- **PSNR**: the PSNR (Peak Signal-to-Noise Ratio) metric has been found not to take the visual masking phenomenon into consideration. It means that every single errored pixel contributes to the decrease of the PSNR, even if the error is not perceived. PSNR is useful in some applications, such as image coding, where the encoding process introduces degradation almost everywhere in the image, but the packet lossy channel introduces impairments in some areas of the reconstructed pictures [15]

- **SNR** [16]: Signal-to-noise (SNR) measures are estimates of the quality of a reconstructed image compared with an original image. The basic idea is to compute a single number that reflects the quality of the reconstructed image. Reconstructed images with higher metrics are judged better. In fact, traditional SNR measures do not equate with human subjective perception. Several research groups are working on perceptual measures, but for now we will use the signal-to-noise measures because they are easier to compute. Just remember that higher measures do not always mean better quality.

4. Simulation of Network Loss

The figure shows the simulation test-bed used run to investigate network effect on video, and to suggest an improvement. The improvement will be on a parameter, which does not affect the standard, found in MPEG-2.

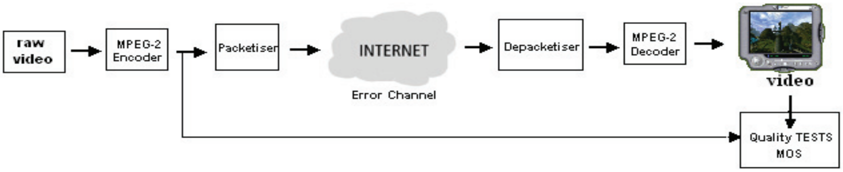


Figure 4 : Block Diagram for the simulation process

Simulation was run on C++ platform using raw video (YUV[2]) sources format. The program provided a tool to simulate packetisation of the video stream, and allow the handling of simulated lost packet loss. In the MPEG parameter-files the bit rate can be controlled, and importantly the size of the Group of Picture. The packetisation process is in conformance with RFC 2250 [17]. The test videos are obtained from the Video Quality Expert Group ftp server [18]. The video is then encoded with MPEG-2 encoder at constant bit rate of 4.0 Mbit/s- the effect of this parameter is given in [19].

The following simplification was made:

- Actual data is replaced by a random byte sequence when a packet loss occurs.
- The header is not affected by the packet loss so that it doesn't experience frame skips.
- Packet size of 188byte was used with 4 bytes reserved for the header.

At the output of the net work the degraded video was compared with the original clip and then a rating of Mean Opinion Score was done using the Genista Media Optimacy (MOS). The output videos were all based on different packet loss rates.

4.1 Simulation Results

The aim of the simulation is to show how a typical network condition can be improved by adjusting the number of Group of Picture (GOP). In this case the PAL system was considered- where initially the number is 12.When the GOP was adjusted from 12 to 4, the mean opinion score was found to gradually to increasing, but at GOP size of 8, the value tends to drop again. Showing that 8 group of pictures is the optimum for an MPEG-2 transmitted at 4.0 Mbit/s.

Here the quality is measure by the Genista Media Optimacy (MOS).Other metrics can be used to study the effect of the GOP [20], like the RMS error and the spearman's correlation. The structure of the GOP has also been studied for H.261 video streams in [21].The MOS presents a user's perception and therefore produces better results compared with the mathematical metrics. The results are shown below:

[2] Y= 0.299Red + 0.587Green + 0.114Blue, U= 0.565(Blue – Yellow), V= 0.713 (Red – Yellow)

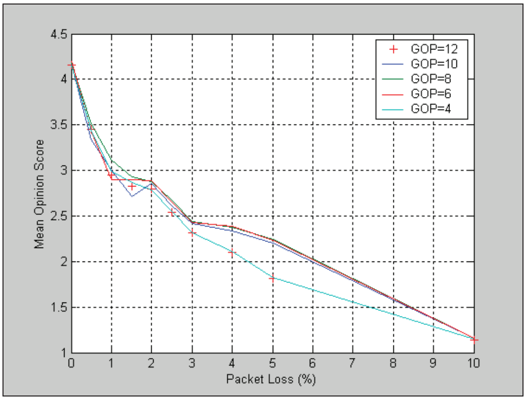


Figure 5 : Variation of quality with GOP

The figure shows the result obtained from the subjective tests for 10 test video sequences converted to MPEG. The video are then distorted on a varying packet loss rate a Mean Opinion Score was obtained for each sequence is then averaged to find the overall Mean Opinion Score. The result is presented in the plot above.

In summary the quality affecting parameters were studied to conform to real-time transmission and can be summarised as:

1. Bit rate: 4 Mbit/s.
2. Number of frame per second (25fps) - this is the regular PAL system used in Europe
3. Packet loss rate (LR): with the C++ program, packets can be dropped randomly and uniformly to satisfy a given percentage loss. This parameter is chosen from a range of 1 – 10%. It has been found that a loss rate higher than 10% will drastically reduce the video quality [22].
4. The burst losses in the packets were not considered.

The result could be used to propose a method in which MPEG-2 video transmission can be improved. In this, there should be a feed back loop to request for adjustment in the number of GOP transmitted. Although this can add some form of delay, but the quality of transmitted video is improved.

5. Developing an Interface for Single Stimulus Continuous Quality Evaluation (SSCQE)

Tests using SSCQE has been done in [23] and [7] and after going through the series of Objective measurements, it became necessary to build a user interface to solve problems due to time synchronisation therefore to enable efficient measurement of user opinion for a video quality test. It is a tool for subjective measure that allows easily loading of video clip and storing of participant data. The .exe file can be obtained on request from (edwin.madu@student.plymouth.ac.uk). Basically it consists of 5 sections: The screen, the slider Name bar And Control buttons

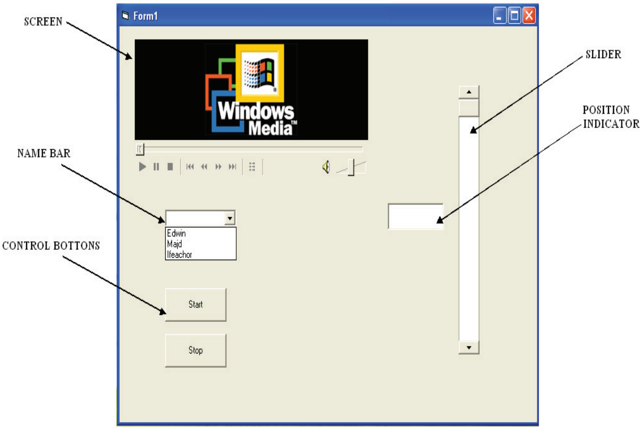


Figure 6 : Interface for measuring objective grading

- **The screen**: this displays the video to be studied and the size of the widow can be changed.
- **The slider**: the slider can be made to move from the lowest position, indicating a low quality video to the highest position, for high quality video. The position is sampled at 10ms and the rate can be adjusted too.
- **Name bar**: this is suitable when the name of each subject is to be noted, and then linked to the quality grading done. Actually a text file is opened to store the data inputted by each participant.
- **Control button**: it has a start and stop button that will enable start and stop of play of the video clips. When the start button is started, it also start recording data into the text file immediately. This is the advantage of the program.

6. Conclusions and Future Work

Video quality evaluation presents a complex analysis, it can be seen that no few parameters can characterise the system. Packet loss as a factor is chosen for it major influence in packet communication. The simulation used is a simple and robust network condition, but provide a reasonable picture of effect of internet video on a network conditions. Although in the simulation, packet loss was assumed to random and random packet losses have been found to have larger macroblock impairment ratio than correlated packet losses do for the same loss probability, without consideration of the use of error concealment techniques. Anyhow, this problem should further be investigated together with error concealment and quality measure of reconstructed pictures.

We have shown that a method based on adjusting the number of Group of Pictures enhances the robustness of video coders to packet losses. The main emphasis of the work was to consider quality improvement in the IP video system, which will be compatible to MPEG-2 and other existing codecs. The investigation shows that better results could therefore be obtained by using feedback loop to the MPEG-2 encoder to adjust the size of the GOP parameter depending on the network condition. In future work we would like to take burst error propagation into account, especially those found in Wireless environment as in [3]. If successful, then it is important for it to be modified to cope, and to continue to providing users with the optimal service provision for the prevailing network conditions.

7. References

[1]. MPEG-2 Testing, Computer Modules Inc. available at http://computermodules.com/broadcast-systems/transport-stream.shtml visited on 5 July 2003.

[2]. Jan Ozer, (1994) Publishing Digital Video, AP Professional available at http://www.fsbassociates.com/books/pubdigvideochp.htm. visited on 2 June 2003.

[3]. NetPredict Inc.: Performance Analysis for Video Streams across Networks available at http://www.netpredict.com/pdfs_all/WhitePaper-Video.pdf. Visited on 19 July 2003.

[4]. Olivier Verscheure, Pascal Frossard and Maher Hamdi: MPEG-2 Video Services over Packet Networks: Joint Effect of Encoding Rate and Data Loss on User-Oriented QoS available at http://ltswww.epfl.ch/~frossard/publications/pdfs/nossdav98.pdf visited on 21 July 2003.

[5]. L. F. Sun, G. Wade, B. M. Lines, E. C. Ifeachor: 'Impact of Packet Loss Location on Perceived Speech Quality' available at http://www.iptel.org/2001/pg/final_program/15.pdf visited on 29th August 2003.

[6]. Anna Watson: Assessing Audio and Video Quality in Multicast Conferencing available at: http://www-mice.cs.ucl.ac.uk/multimedia/projects/pipvic/annas_abstract.html, visited on 27 June 2003.

[7]. Network QoS Needs of Advanced Internet Applications-A Survey, available at http://qos.internet2.edu/wg/apps/fellowship/Docs/Internet2AppsQoSNeeds.pdf visited on 21 July 2003.

[8]. Mylène C.Q. Farias, John M. Foley, and Sanjit K. Mitra: Some Properties of Synthetic Blocky and Blurry available at http://www.engineering.ucsb.edu/~mylene/SPIE_2003.pdf visited on 9 July 2003.

[9]. R. Kutka, A. Kaup and M. Hager: Quality improvement of low data-rate compressed video signals by pre- and post processing, available at http://www.lnt.de/~kaup/paper/digicomp-96.pdf visited on 28 July 2003.

[10]. Johan Berts and Anders Persson: Objective and subjective quality assessment of compressed digital video sequences available at http://www.etek.chalmers.se/~e4joni/xjobb/report.pdf visited on 12 June 2003.

[11]. Recommendation ITU-R BT.500-10: Methodology For The Subjective Assessment Of The Quality Of Television Pictures available at: http://itd.colorado.edu/bors5837/lect3/ITU-BTdot500-10-(03-00)%20.doc visited on 15 July 2003.

[12]. Richard Simik: Improvement of Digital Video Quality Metric available at http://simik.wz.cz/data/radioel2003.pdf visited on 12 August 2003.

[13]. Stefan Winkler and Frederic Dufaux: Video Quality Evaluation for Mobile Applications available at http://ivrgwww.epfl.ch/publications/wd03.pdf visited on 29 July 2003.

[14]. Fletcher John and Prior-Jones Michael: MPEG-2 VIDEO: Quality Versus Quantiser Scale available at http://www.bbc.co.uk/rd/pubs/papers/pdffiles/ibc00jaf.pdf visited on 12th August 2003.

[15]. Tingting Zhang, Ulf Jennehag and Youshi Xu : Numerical Modeling of Transmission Errors and Video Quality of MPEG-2 available at http://www.itm.mh.se/forskning/teleinfo/publications/journalMPEG.pdf visited on 1 August 2003.

[16]. Vanguard Software Solutions, Inc.: PSNR Computation available at http://www.vsofts.com/codec/codec_psnr.html visited on 28 July 2003.

[17]. RFC 2250 (RFC2250), Internet RFC/STD/FYI/BCP Archives available at http://www.faqs.org/rfcs/rfc2250.html visited on 19th August 2003.

[18]. User-Oriented Quality of Service Analysis in MPEG-2 video delivery available at http://ltswww.epfl.ch/~frossard/publications/pdfs/rti99.pdf visited on 29th June 2003.

[19]. V. Vitsas and A. C. Boucouvalas: IrDA IrLAP Protocol Performance and Optimum Link Layer Parameters for Maximum Throughput available at http://dec.bournemouth.ac.uk/staff/tboucouvalas/globe2002.pdf visited on 12th August 2003.

[20]. Antony W. Rix, John G. Beerends, Michael P. Hollier and Andries P. Hekstra: PESQ – the new ITU standard for end-to-end speech quality assessment available at http://www.psytechnics.com/papers/2000-P02.pdf visited on 11th August 2003.

[21]. F. Fitzek P. Seeling M. Reisslein_ M. Rossi M. Zorzi: Investigation of the GoP Structure for H.261 Video Streams available at http://www.eas.asu.edu/~mre/GoP.pdf visited on 27th August 2003.

[22]. Jizhong (Jim) Wu Home Page, School of Computer Science and Engineering, University of New South Wales available at http://www.cse.unsw.edu.au/~jimw/ visited o 17th August 2003.

[23] Anna Watson: Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing available at http://www-mice.cs.ucl.ac.uk/multimedia/projects/etna/watson.pdf visited on 12th August 2003.

A prediction of audio quality for personal audio devices

C. Marzin, B. Hamadicharef and E.C. Ifeachor

Signal Processing and Multimedia Communications Research Group, School of Computing,
Communications and Electronics, University of Plymouth, Plymouth, Devon U.K.
e-mail: e.ifeachor@plymouth.ac.uk

Abstract

Perceived audio quality is a key metric for personal audio devices such as Modern Digital Hearing Aids, MP3 players and mobile phones. The primary aims of the study reported in the paper are to investigate the audio quality assessment for personal audio devices with emphasis on Hearing Aids by developing an audio quality-testing platform and to evaluate objectively the user perceived audio quality. This project presents a novel way to assess the user perceived audio quality using the latest ITU-R BS.1387 the Perceptual Evaluation of Audio Quality (PEAQ) algorithm. A testing quality system has been developed and consists in three main parts: a model of the hearing loss, a simple model of the hearing aids and the audio quality assessment part. The quality assessment part uses PEAQ to measure the audio degradation between the reference sound and the test sound signals going through both the hearing loss (person) and hearing aid (compensator). The whole system has been implemented under the Matlab environment. A series of experiments have been neated in which each individual part of the hearing aid model has been tested in term of audio quality, based on various profiles of hearing loss. The audio materials used for these experiments have been taken from the PEAQ audio CD with tracks of a broad variety of sounds such as saxophone and classical instruments as well as sounds from Hammond organ. Results show that the filter bank and its associated number of bands, related to the original audiogram of the hearing loss of the particular individual, and compression scheme and its particular parameters related to have the most important effect on the perceived audio degradation. Interestingly, it has been found that the type and size of the window used in the framing process has little effect. This work is hoped to contribute to the future design of Hearing Aids based on a user perceived audio quality metric, and furthermore to help developing a new hearing aids quality index used for manufacturers to benchmark their products.

Keywords

Digital Hearing Aids, Hearing Loss, Audio quality, PEAQ algorithm

1. Introduction

In this world of communications, men and women are, everyday speaking to each other, listening to each other and to music such as an opera concert. Music and sounds are part of our life. The speech and ears are important to the well development of humans even more the case in a multi-national society. Sometimes from birth, certainly with age, and with exposure to very loud music in the case of youngsters, the human auditory system can degrade dramatically. Modern hearings aids have been designed to help people to compensate for their own particular hearing loss. Thus, they are at the crossroads of three topics of modern research: Multimedia Communications, Audio Signal Processing and Biomedical.

Today, most of the tests done on audio devices are based on performance measures, which are not directly related to the audio quality as perceived by the user. Most publications and patents on hearing aids are focused on system performance, Digital Signal Processing (DSP) technologies and not based on perceptual audio quality. We should not forget that modern hearing aids are designed to help people to compensate for their own particular hearing loss. Conventional tests that have been performed on hearing aids are principally turned to performance tests (Parsa and Jamieson, 2001). Parsa and Jamieson have based their experiments on hearing aids on the amount of distortion using the Signal-to-Noise Ratio (SNR) measure but as they said: these measures have poor correlation with perceptual judgments of hearing aid sound quality by hearing impaired listeners. Conventional measurements give a global quality, which is to simplistic for perceived audio quality; on the other hand perceptual audio quality of measurements estimates the quality locally. The listener does not perceive the signal processing performance of the system, but the audio quality the user perceived can be estimated and can be used as a key element of the fitting process. Therefore, one of the project challenges is to focus on the true satisfaction of the listener by measuring the perceived audio quality using PEAQ algorithm (Thiede *et al*, 2000; BS.1387, 1998).

The aims of the study reported in this paper are (1) to investigate the latest perceptual audio quality techniques to obtain objective measures of user-perceived quality for hearing aids, (2) to develop and implement a research tool, which should be able to measure and quantify the audio quality of modern digital hearing aids, (3) to evaluate the impact of parameters variation, of each DSP building blocks that constitute the modern digital hearing aid, on the audio quality or audio degradation, and (4) to draw conclusions from the results and suggest new ideas for future work.

The remainder of the paper is organised as follows. In section 2, the assessment of the audio quality as perceived by the users, using the PEAQ algorithm is introduced. In section 3, we move to the heart of the matter of this research with the experimental hearing aids system and with the experiments results setting up to assess the impact of each of the building blocks. In section 4, research avenues for future work are presented. Finally section 5 concludes this research paper.

2. Audio quality assessment with PEAQ

Audio quality can be evaluated in two different ways: subjectively by listening tests, or objectively using PEAQ. To predict the audio quality for Hearing Aids (HAs), we have used EAQUAL and OPERA software (OPTICOM GmbH Germany, 2003). Figure 1 shows a conceptual diagram of the testing system that aims to tackle the problems stated above. This system is divided into three distinct parts: the Hearing Loss Model, the Hearing Aid Model and the audio quality assessment based on the PEAQ algorithm.

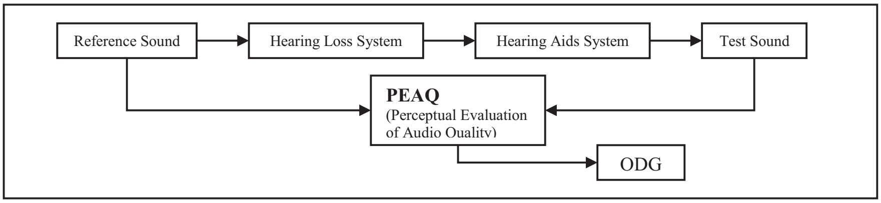


Figure 1 : Perceived audio quality assessment of Hearing Aid

A sound file (standard WAV-file at 48 kHz / 16-Bits PCM), is taken as a reference and fed to the hearing loss model. The hearing loss model degrades the sound in a similar way that corresponds to the user's loss of audition. The result is passed onto the hearing aid model. The hearing aid model is used to compensate the hearing loss model degradation, the results being the degraded sound file used as the test sound. Then the two files, representing the reference sound file and the test sound file are applied to PEAQ returning the Objective Difference Grade (ODG) measurement.

In all this work, the audio quality assessment is using the basic version of PEAQ algorithm. It will indicate the overall performance of the system, i.e. how good the hearing aid model manages to compensate for the hearing loss.

3. Hearing Aids Audio Quality System

The goal of the conceived hearing aid is to amplify non-uniformly a sound over different frequency bands in order to compensate the patient hearing loss.

The first part of this section presents the hearing loss model developed in the project. The second part describes the hearing aid model implemented during the project. Finally, the third part establishes how the audio quality measurements are performed (see Figure 2 which is still the same as Figure 1 but in more details).

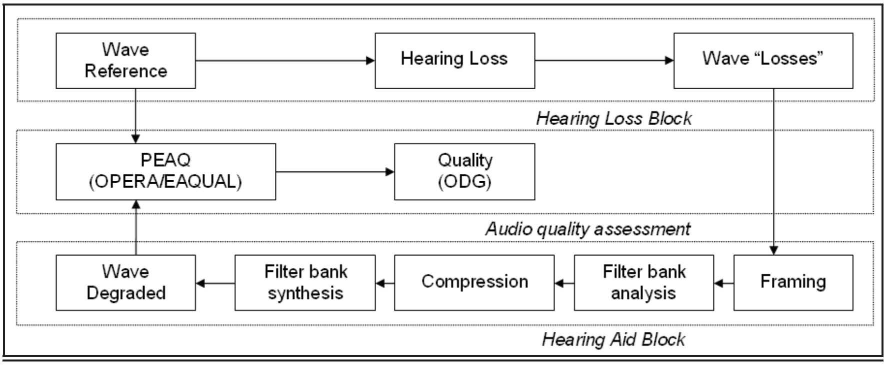


Figure 2 : Block diagram of the system

3.1 Hearing Loss Model

Hearing loss transfer function is based on measurements set up in (Launer, 1995). The frequency spectrum is divided into 14 bands non-linearly spaced from 20 Hz to 12 kHz as shown in Figure 3.

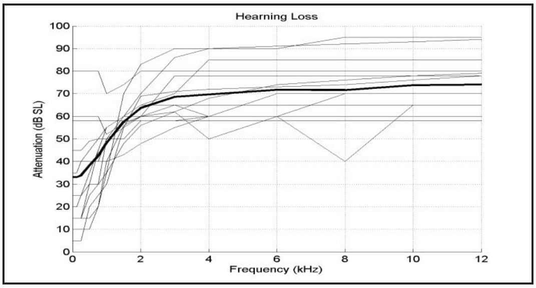


Figure 3 : Hearing loss of 14 hearing-impaired subjects

The values used are based on a study carried out on 14 hearing-impaired subjects (Launer, 1995). The measurements made for the frequency below 250 Hz and above 6 kHz are estimated from the shape of the pure tone audiogram. The results obtained are expressed in Sound Pressure Level. In order to simulate the hearing loss, we have converted hearing losses from Sound Pressure Level to loss factors express in dB from each frequency and each subject. From the 14 spectrum subjects, an average hearing loss transfer function (bold curve in Figure 3) has been calculated in order to have a spectrum that represents most of hearing loss cases. From these values, a filter based on Parks Mc Clellan optimal FIR filter design using *remez* Matlab method has been designed. It gives good approximation looking at the frequency band edges and at the corresponding amplitude. With this hearing loss filter, results are very relevant about the hearing loss: the tested signal (black curve) is attenuated compared to the original one (grey curve) as shown in Figure 4.

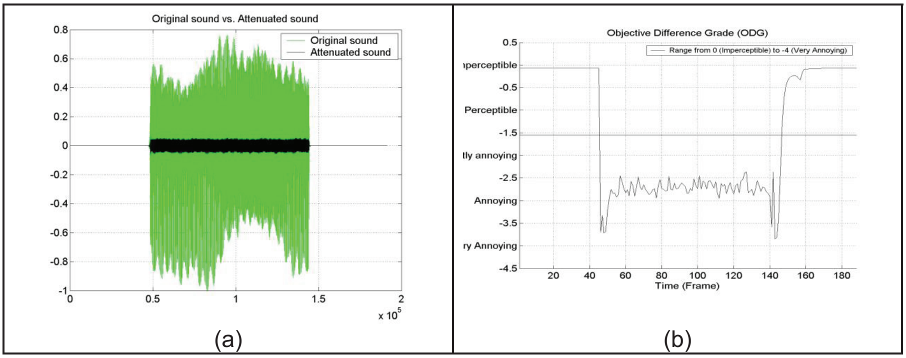


Figure 4 : a. Original sound vs. tested sound & b. Results estimated with PEAQ

Figure 4 shows that the ODG quality is approximately between – 2.5 and – 3, which mean that the sound, perceived by impairment people, is at least “slightly annoying”. We can notice that the average objective quality measurement is estimated at - 1.5. The example used a sound from Hammond organ, very pure sin-wave.

3.2 Hearing Aid Model

The hearing aid model defined in the project consists of framing, filtering, compression and interpolation (Ifeachor and Jervis, 2002). Filters used in hearing aid require a high resolution. The human hearing frequency resolution is complex and depends on frequency, signal bandwidth and signal level. Moreover, human ear is very accurate in frequency analysis. Therefore, sounds going through the hearing aid are analysed using a large resolution by the aid of a filter bank. The filter bank is divided the frequencies into frequency bands that can be independently amplified to match the needs and loss pattern of hearing of the subject. Each set of sub-bands occupies only a small portion of the original signal band.

A filter bank is the association of a set of digital band pass filters. There are two stages in a filter bank (Harteneck et *al*, 1999): on one hand, the analysis filter bank, which is a set of digital band pass filters used to decompose the input wave file into a set of subbands signals, on the other hand, the synthesis filter bank, which is the opposite operation of the analysis. It combines a set of subbands signals into an output wave file that occupies the whole Nyquist range.

Moreover, from the two different approaches (“parallel” and “cascade” implementations) that exist to construct a filter bank, all implementations considered in the study use the parallel topology. Differences between the two methods are that the parallel topology requires fewer filters than the cascade one, where as the cascade implementation requires only two kinds of filters (a low and a high pass filter).

Several different systems have been tested with *M*-channels analysis/synthesis filters: 92 is the optimum number of channels according to the results obtained in Figure 5.

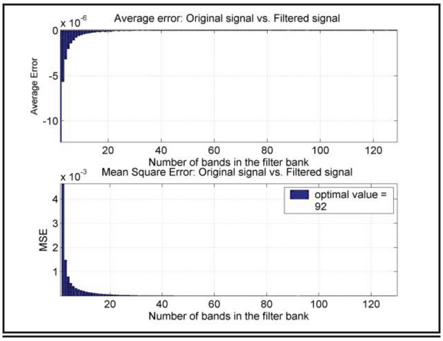


Figure 5 : Impact of number of bands vs. MSE

Actually, we had to face to the problem that the reconstruction of the signal after the filtering was not perfect compare with the signal before the filtering. If the signal is not reconstructed well after the filter bank, the perceptual quality estimated with PEAQ is not very accurate. To solve this problem, the concept of perfect reconstruction has been considered using FIR filters system. FIR perfect reconstruction is based on two important theories: the cosine modulation and the construction of the prototype filter.

The synthesis and analysis filters coefficients are calculated by (Keiler, 2003):

$$H_i(n) = P_0(n) \times 2 \times \cos\left(\theta_i + \frac{\pi}{M}\left(i + \frac{1}{2}\right)\left(n - \frac{N-1}{2}\right)\right)$$

$$, with \qquad \theta_i = (-1)^i \frac{\pi}{4}$$

$$G_i(n) = P_0(n) \times 2 \times \cos\left(\theta_i - \frac{\pi}{M}\left(i + \frac{1}{2}\right)\left(n - \frac{N-1}{2}\right)\right)$$

N is the order of the prototype filter. The prototype filter $P_0(z)$ is obtained using a Kaiser Window approach (windowing depends on two parameters *L,* where L = N+1 and β determined by the attenuation *As*) (Lin et *al*, 1998) and is defined by its stopband attenuation and the number of channels desired for the filter bank. From Figure 6, it can be seen that the stopband attenuation has an impact on the perceived audio quality. When the stopband attenuation increases, the Mean Square Error (MSE) decreases (note: an optimum is reached for 61dB) and the ODG quality measurement improves (from –3 for 30dB up to 0 for 200dB).

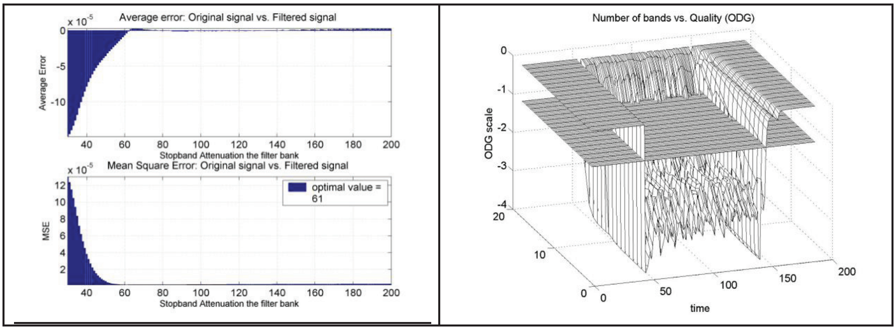


Figure 6 : Stopband attenuation vs. MSE & PEAQ

After the development of the filtering of the hearing aids, it is necessary to take in consideration the compression. The input signal is fed to an analysis filter bank, which divides the signal into N bands from 0 to 24 kHz (Nyquist frequency). In each band, the signal is amplified by a ratio corresponding to the patient hearing loss. This maps the 14 non-linear audiogram values to a more accurate the Hearing Loss model, interpolating up to N bands in the filter bank as shown in Figure 7. Each signal is then fed into the synthesis filter bank and all added together to produce the final output. Results shows that design of filter bank using less than 256 bands, the 14 non-uniform values do not map efficiently to the filter bank (i.e. at least two of non-uniform values end up in one same band of the filter bank).

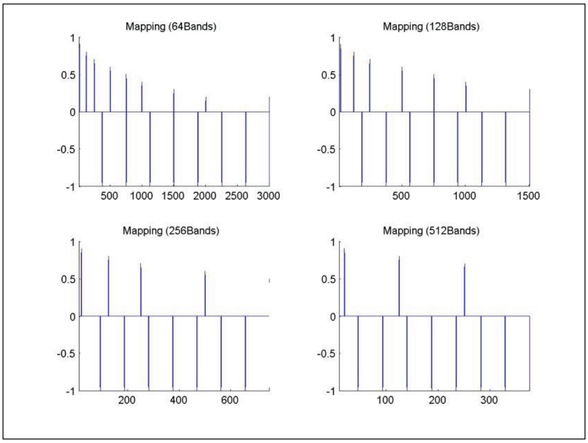


Figure 7 : Hearing loss repartition of non-uniform values vs. filter bank

3.3 Audio quality measurements

The audio quality was assessed with OPERA (OPTICOM GmbH Germany, 2003) and EAQUAL software. To evaluate the quality of the modern digital hearing aid, tests have been carried out on each building block with audio WAV-files that are sampled at 48 kHz on 16-Bits PCM. To estimate the quality of the system, sound files taken from the PEAQ audio CD are analysed with the corresponding tested sound files at each output blocks consider.

Finally, the perceived audio quality of hearing aids is determined on the measures given by Objective Difference Grade at the output of the PEAQ model.

4. Future work

Future work will focus on further analysis of the automation of parameter estimation, the prediction of hearing aids audio quality, the development and audio quality evaluation of an advanced hearing aid model, experiments with real hearing aids devices, looking at the speech side with PESQ (P862, 2000), and finally extending this project with a multi-disciplinary research view.

One could imagine an automated development system in which the user, who could be the engineer of the hearing aid, sets a quality level and the system automatically evolves the hearing aid model until it reaches a configuration for which the audio degradation thought hearing aid is kept under the specified audio quality (see Figure 8). This approach could be called "Audio quality based hearing aids modeling". Techniques like Genetic Algorithms (GAs) have been used in many other engineering fields to search for optimal parameters (Durant, 2002). GAs use a so called "fitness function" that represents the goodness of the solution from the evolved population of solutions. This fitness function could be based on a perceptual audio quality measure, closer to the real user perceived audio quality.

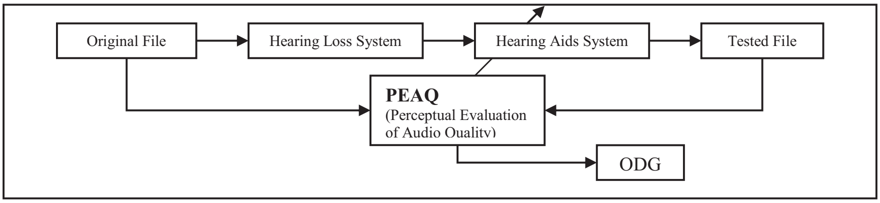


Figure 8 : Optimisation of the perceived audio quality

In another view of the problem, one could imagine that the system would try to predict the hearing audio quality from its parameters. This will be determined with the aid of a Neural Network. The Neural Network will try to predict, from the hearing aid parameters and the signal, a predicted ODG. The performance of the Neural Network will be assessed with a correlation factor, r (between 0 and 1), which correlate the ODG given by PEAQ and the predicted ODG given by the Neural Network.

The MathWorks Matlab environment has many engineering toolboxes that can be used for modelling complex systems. For example, Digital Signal Processing, Time-Series Statistics and Control Systems could be used to create an advanced model of a hearing aid from models described in (DSP Factory url, 2003) and more recently in (Shusina, 2003). This could extend the understanding of the real issues of audio quality related to the perception from the user point of view.

This research, focused on audio quality using the PEAQ algorithm, could be extended to a speech point of view. Similar to the concept of this research project, one could replicate this work with the PESQ algorithm. This could make an interesting link between the audio and speech research. This paper is at the centre of three important research fields: Biomedical, Audio Signal Processing and Multimedia Communications. Biomedical, because the hearing aids are based on studies of the human auditory system, the measurement of hearing loss and thus has a very strong medical aspect to it. Audio Signal Processing because modern hearing aids are "Systems on Chip" (SoC), powerful DSP cores that are trying to compensate for people hearing loss. Multimedia Communications because the audio quality of hearing aids will benefit people in their day-to-day activities in terms of communications and multimedia activities like listening to music, watching television, etc. Hearing aids are real laboratory for psycho-acoustics experiments. Their design is based on modern DSP technologies but also uses human psychology and medicine. With the help of researches from department of psychology and medical doctors, one could hope to get the adequate collaboration for extending the current project and giving it more medical scope.

5. Conclusion

A hearing aid audio quality-testing has been implemented in Matlab environment. Series of experiments have been performed to assess the audio quality degradation of each DSP building block constituting the modern hearing aid.

There is still a lot of work to be done and new research directions have been suggested. One can also question the suitability of the usage of PEAQ as it has been confirmed to be a very sensitive tool when it comes to assess audio quality (Benjamin, 2003). Thus, one could imagine MOS tests of elderly people using hearing aids and evaluate in more detail the improvement of comfort.

It is hoped that this project will help to improve future design of hearing aids, be used as the basis for many of the suggested ideas for investigations and future work. Finally, with maybe a longer term "audio research vision", this work could be the first step towards the development, implementation and evaluation of an audio quality index to benchmark modern hearing aids.

6. References

E. Benjamin, "*Evaluating Digital Audio Artifacts with PEAQ*", 113th Convention of the Audio Engineering Society, Los Angeles, Preprint 5711.

E. A. Durant, "*Hearing Aid Fitting with Genetic Algorithms*", Doctoral Ph.D. Thesis, University of Michigan–Ann Arbor, 2002

E. C. Ifeachor and B. W. Jervis, "*Digital Signal Processing A Pratical Approach*",2nd Ed., Prentice Hall, 2002

M. Harteneck S. Weiss and R. W. Stewart, “*Design of Near Perfect Reconstruction Oversampled Filter Banks for Subband Adaptive Filters*”, IEEE Transactions on Circuits and Systems Part II: Analogue and Digital Signal Processing, Vol.46(No.8): pp.1081-1086, August 1999.

ITU-R Recommendation BS.1387, “*Method for Objective Measurements of Perceived Audio Quality*”, International Telecommunications Union, Geneva, Switzerland (Dec. 1998).

ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs”, 2000 May.

F. Keiler and U. Zolzer, “*DSP Implementation of a Low-Delay Filter Bank for Audio Applications*” Proceedings of International Signal Processing Conference, Dallas, Texas, April 2003.

Stefan Launer, “Loudness Perception in Listeners with Sensorineural Hearing Impairment”, March 1995

Y.-P. Lin and P. P. Vaidyanathan, “A Kaiser Window Approach for the Design of Prototype Filters of Cosine Modulated Filterbanks.”, IEEE Sig. Proc. Letters, Vol. 5, No. 6, June 1998.

V. Parsa and D. G. Jamieson, “*Hearing Aid Distortion Measurement Using the Auditory Distance Parameter*”, Audio Eng. Soc.,111th Convention, September 2001

N. A. Shusina, “Unbiased Adaptive Feedback Cancellation in Hearing Aids”, May 2003

T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M, Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, “*PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality*”, J. Audio Eng. Soc., vol. 48, pp. 3–29, Jan.–Feb. 2000.

http://www.dspfactory.com, visited 18/05/2003

http://www.opticom.de, visited 22/03/2003

Sound Synthesis using Independent Component Analysis

J. Grizard, B. Hamadicharef and E.C. Ifeachor

Department of Communication and Electronic Engineering, University of Plymouth,
Plymouth, United Kingdom.
e-mail: eifeachor@plymouth.ac.uk

Abstract

This paper concerns the area of sound analysis and more specifically try to provide a new method to resolve the sound synthesis problem. Find new methods of synthesising sound is one of the most exciting challenge for musicians today. We are going to consider the sound synthesis problem as a source separation problem. Indeed, a sound is composed of several components called harmonics and our aim will be to recover the harmonics of a sound by applying Independent Component Analysis (ICA). Basically, ICA is a statistical technique for decomposing a complex set of data into independent sub-parts and is very well suited to achieve Blind Source Separation problem. Thus, the idea is to provide a new insight to solve sound synthesis problem by using ICA method. Over the last decades, ICA has received many interests in the Artificial Intelligent and Signal Modelling research communities. ICA has been used to extract independent signals from mixture signals. In Our case, the independent signals are harmonics, strong signal features of an audio signal that are required to be extracted to be analysed. The overall aim of my research is to determine if ICA can be suitable to resolve the sound synthesis problem and so to recover the different harmonics of a sound. Therefore, in this work we present a new audio application of Independent Component Analysis to tackle the task of analysis of sounds recordings of Hammond organ.

Keywords

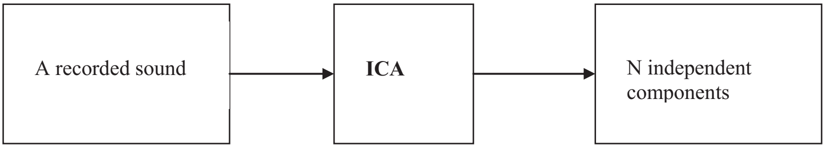
Sound Synthesis, Independent Component Analysis (ICA), extraction, harmonics, new method.

1. Introduction

We can define the sound synthesis problem as a problem of Blind Source Separation where the aim will be to find an efficient and suitable method to extract the harmonics from a sound. Indeed, by definition, a sound is a complex wave form generated by the combination of different harmonics. The extraction of harmonics from sounds will permit to combine some of them to create new sounds, to find similarities between sounds composed of certain same harmonics.

The Independent Component Analysis (ICA) method is well suited to this type of problem and could be used to recover the independent generators from a recorded sound. Indeed ICA is well known to resolve the cocktail party problem where the aim is to recover the different voice of the different speakers from a cacophony speech. Thus, in the cocktail party problem, ICA is able to recover each different voice from different speakers. In my project, this is slightly different. I must recover the different harmonics from a sound (for example one note of the Hammond Organ) by applying ICA to this particular sound. We know that ICA is efficient to resolve the cocktail party problem but we do not know yet if it can be used to extract the harmonics of a sound.

In summary, we have to prove the suitability of Independent Component Analysis to extract the harmonics of a sound and thus to solve the problem of sound synthesis. To achieve this, in a first hand, we will exploit a recorded sound (in my research it corresponds to a sound of Hammond organ) composed of nine harmonics. By using ICA, we will extract nine independent components from this sound.



M different sensors of this recorded sound

Constraint: $M \geq N$

Figure 1 : Independent components extracted by ICA

In another hand, the independent components extracted by ICA will be analysed and compared with the real harmonics of the sound produced by the sound software WinPVOC. The comparison between the real harmonics and the relevant independent components will be realised in terms of wave shape and spectro-temporal distribution to determine how closely they match.

2. Independent Component Analysis

Basically, Independent Component Analysis (ICA) is a statistical and computational technique that permits to reveal some hidden factors which underlie a set of signals, measurements or random variables (A. Hyvarinen et al., 1999).

ICA is one method, perhaps the most widely used, for performing blind source separation. For the moment to simplify ICA and also by analogy with my project, I have omitted noise in this general model.
To explain clearly how ICA works, I am going to analyse a general example. We assume that we observe n linear mixtures $x_1,...,x_n$ of n independent components:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n \qquad (1)$$

Where: s_k is an independent component or source

a_{ij} the i^{th} raw and j^{th} column element of a matrix A.

In my research, $x_j(t)$ would represent one recording of a specific sound of the Hammond organ, s_k would represent one of the harmonics of the recorded sound and a_{ij} would represent some environmental parameters that could depend on the distances of the microphones to record the sound, the recording conditions…
To simplify the problem, we can rewrite equation (1) as follows:

$$x = \sum_{i=1}^{n} a_i s_i \qquad (2)$$

Thus, we obtain:

$$x = As \qquad (3)$$

The only vector we know is x the signal that we observe, and we must estimate A and s (that represents the independent components of the observed signal). Thus, the aim of ICA is to recover first A and then s. To resolve equation (3), ICA assumes that the components s_i are statistically independent and that the unknown mixing matrix A is square. Then, after estimating the so called mixing matrix A, it can compute its inverse W called demixing matrix and also obtain the independent components s_i :

$$s = Wx$$

Nevertheless, the major problem of estimating data model in ICA is how to estimate the mixing matrix A using only the information contained in the mixture signal x. The good results of Independent Component Analysis method depend especially on the mixing and demixing matrix A and W. We call G the global matrix equals to:

$$G=WA$$

In theory, if the global matrix G equals to a generalized permutation matrix, we will obtain a perfect signal separation.

3. Sound Analysis

3.1 SoundICA Matlab Toolbox and Fast ICA algorithm

To make it more suitable for my research, we have adapted the ICALAB toolbox for signal processing that has been designed and developed under matlab the 14 of April 2003 for the last version by Andrzej Cichocki, Shun-ichi Amari, Krzysztof Siwek, Sergio Cruces in cooperation with others members of the laboratory for Advanced Brain Signal Processing. (Cichocki et al., 2003).

Thus, to complete this research successfully, we have used the ICA matlab toolbox where the fast ICA algorithm was implemented in. The fast ICA algorithm implemented by A. Hyvarinen (A. Hyvarinen et al., 2002) exploits the statistic function kurtosis and is based on fixed point iteration in order to avoid slow convergence that depends generally on the choice of the learning rate sequence. Furthermore, with this algorithm, the computations are parallel, distributed, simple and required a small amount of memory. That permits the fast ICA algorithm to extract the independent components not only one by one by using hierarchical decorrelation (A. Hyvarinen et al., 2002) but also it can estimate all the independent components in parallel with symmetric decorrelation. However, the fast ICA algorithm is not adaptive. But globally, this algorithm has a convergence speed superior to the others ICA algorithm.

In summary, the matlab toolbox is able to decompose multi-variable signals into independent components

3.2 Sound C2 and C3 of the Hammond organ

We have chosen to exploit a sound C2 of the Hammond organ in order to determine if Independent Component Analysis is suitable to extract the harmonics of a sound.

By design, the Hammond Organ uses a Tone Wheel generator to generate its characteristic sound. Each sound of the Hammond organ is composed of nine harmonics associated to a specific frequency. Thus, each note of the keyboard triggers a set of 9 frequencies that should be easy to detect using ICA.

Each note (sound) of the Hammond organ is assigned nine harmonics derived from the tone wheel. The proportions with which these harmonics are mixed are controlled by a set of nine harmonic drawbars. The drawbars permit to control the strength of each of the nine harmonics, which compose a sound of the Hammond organ. Basically, the variation of the drawbars does not change the note of the Hammond organ but changes the tone of the note (M. Vail, 2002).

The sound C2 is the note number13 on the 61 notes of the Hammond organ. The sound C2 is also composed of nine main harmonics represented by nine generators. Each harmonic has an own frequency and an own amplitude.

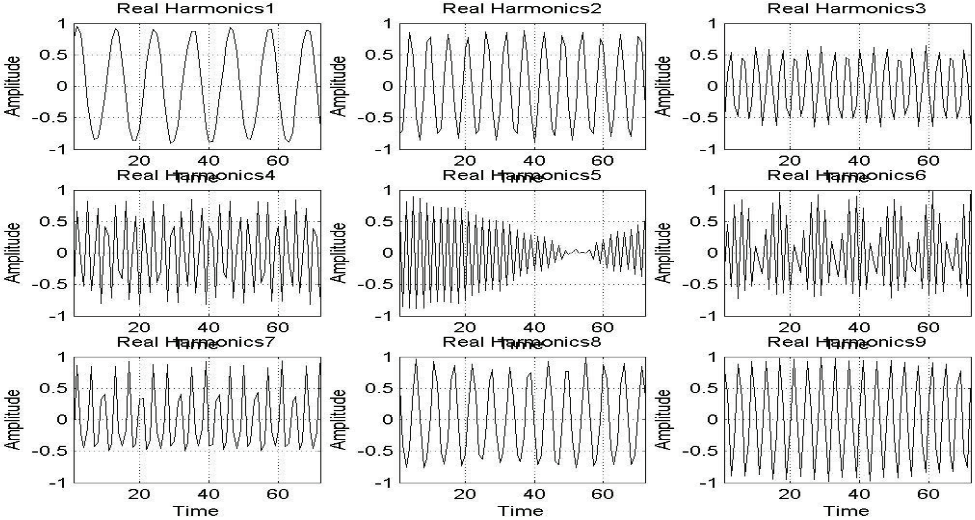


Figure 2 : The 9 harmonics of the sound C2 used as a comparison tool

In order to see if the independent components obtained by ICA are closed to the real harmonics of the sound C2, we will have to compare the relevant independent components with the real harmonics. Thus, the figure above depicts the nine harmonics of the sound C2 plotted by using the same scale, the same length of sound, the same size of matrix as those used to plot the independent components extracted by ICA.

4. Experiments and Results

To lead our research, we have realised several experiments by using fast ICA algorithm to extract independent components from a sound of the Hammond organ with some specific drawbars preset in order to see if ICA is suitable to extract the harmonics from a sound.

We will look at a precise example where the fast ICA algorithm has been applied to a sound C2 with the following drawbars preset: 006876540. To complete our experiment, we have used nine different sensors (recording) of the sound C2 with the drawbars preset 006876540. After transforming each of these recordings into matrix, we have obtained a nine rows matrix where each row corresponded to a specific recording of the specific sound C2. Before applying the fast ICA algorithm, we have preprocessed the data by using two simple lowpass filters with impulse response [1, 1]. The use of two lowpass filters was designed to smooth the independent components that I will obtain by applying ICA to the "sensors matrix". Indeed, the negative effect of ICA on the data is to increase the amplitude of the data. Hence, two simple low pass filters permit to resolve this basic problem. Then, we have applied the fast ICA algorithm to extract nine independent components from the sound C2 with the drawbars preset 006876540. By contrasting the relevant independent components to the real harmonics of the sound C2, we have noticed certain similarities between some of them. In order to see clearly the resemblances between some independent components extracted by ICA and some real harmonics, I have plotted those that presented the greater similitude:

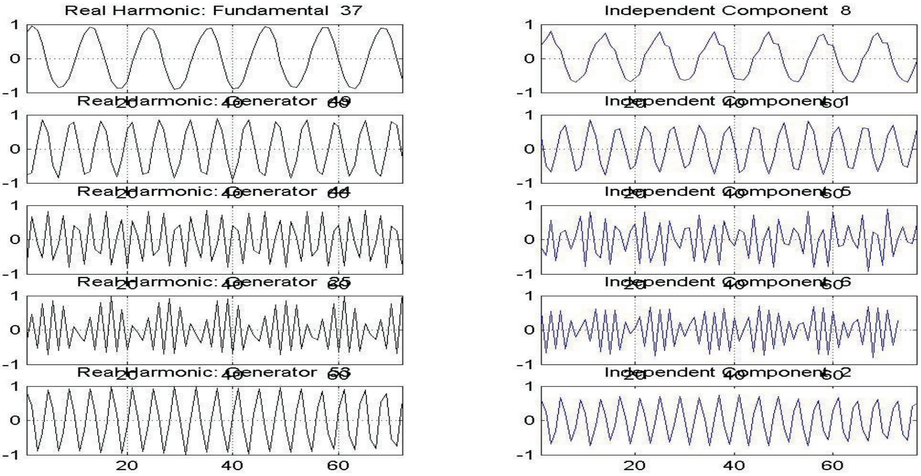


Figure 3 : Comparison between some real harmonics and some independent components

In order to confirm the similarities between the harmonics and the independent components extracted by ICA, I have applied the Fast Fourier Transform (E. Ifeachor et al., 2002) to both of them to inspect the magnitude versus the frequency of these signals. Indeed, each harmonic that composes a sound has a unique frequency (see the arrays in part 3.3.2). Thus, we will look at the frequencies similarities between some independent components and some real harmonics to see how closed they are from each other. The figure below presents one example where we have plotted one independent component extracted by ICA with its corresponding real harmonic in the frequency domain by applying the Fast Fourier Transform. On the figure below, you can see a comparison between the independent component 5 (in grey) and the real harmonic 4 (Generator 44 plotted in black) in the frequency domain:

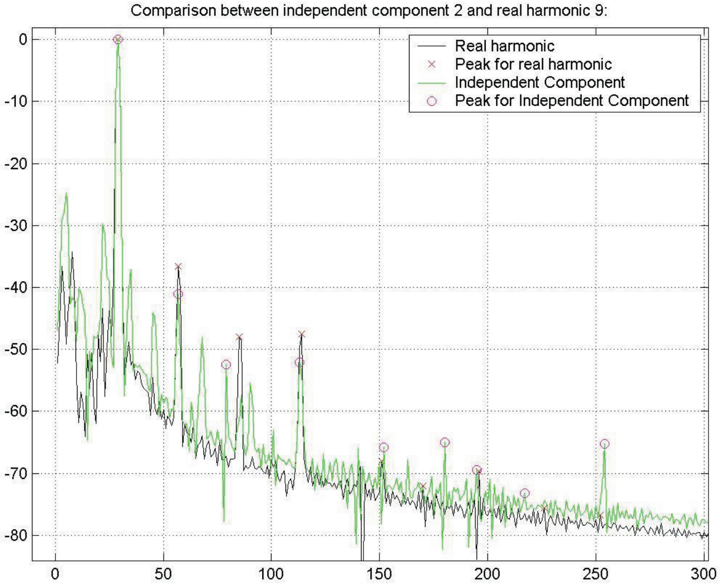


Figure 4 : Comparison between real harmonic 9 and independent component 2 in the frequency domain

As you can notice on the figure 3, certain independent components seems to be identical to certain independent components. Furthermore, they have nearly the same peaks of frequency (figure 4) and have an identical global shape. The fact that the Fast Fourier Transform of certain independent components extracted by ICA and their corresponding real harmonics are nearly identical is a real proof of the important similarities between some independent components of the sound C2 (with the following drawbars preset: 006876540) extracted by using the fast ICA algorithm and the real harmonics of this sound. Finally, the hearing comparison between those independent components and their corresponding real harmonics has confirmed their mutual resemblance. Thus, in this case, we can say that ICA has achieved to extract 5 harmonics upon the nine that compose the sound C3.

5. Further Areas of Research

The example exposed in the part 3 has permitted to demonstrate the suitability of ICA to extract certain harmonics of a sound of the Hammond organ. However, many progresses could be done to improve the quality of the independent components extracted by ICA and to increase the number of harmonics that ICA can extract.

As we have seen in the example (part 3), we have used two simple lowpass filters with impulse response [1, 1] in order to resolve the problem of difference of amplitude created by

ICA. Indeed, the negative effect of ICA on the data is to increase the amplitude of the data. Indeed, the independent components extracted by ICA always have their amplitude higher than those of the real sources even when ICA is suitable and efficient (cocktail problem party). The use of two simple lowpass filters permit to attenuate this problem. However, these simple lowpass filters do not correspond to the most adaptive filters for this specific problem. Thus, the design of a real adaptive filter created specially to resolve the "gap" of amplitude generated by Independent Component Analysis could improve appreciably the results obtained by ICA and make better the extraction of the harmonics from a sound.

Furthermore, the major problem in Independent Component analysis is to estimate the mixing matrix A and so the demixing matrix W that is basically the inverse of the mixing matrix. The difficulty of ICA lies in the evaluation of the environmental parameters which constitute the mixing matrix. If the estimation of these parameters is good, the separation of the independent components will give consistent results. Thus, it seems to be primordial to be able to evaluate the mixing and the demixing matrices. Hence, the creation of a reliable tool that could estimate the quality of the mixing matrix will permit considerably to improve the results obtained by ICA. This tool could permit ICA researchers to use it as a measurement tool to evaluate the quality of the independent components obtained by ICA.

Finally, we have shown that ICA was able to extract some harmonics of a sound of the Hammond organ. These harmonics or generators are created by the Tone Wheel Generator of the Hammond organ. Therefore, we can wonder if ICA would be suitable to extract the components from the Tone Wheel generator. In other words, it could be interesting and useful to resolve sound synthesis problem to decompose harmonics and to see if ICA is suitable to extract the primary components of sound.

6. Conclusion

We claim in this paper that Independent Component Analysis can be used to extract some harmonics of a sound of the Hammond organ and thus to be able to resolve the sound synthesis problem.

As we have clearly seen in the example presented in part 4, Independent Component Analysis is able to extract some harmonics from a sound but presents some limits. Indeed, ICA fails to extract certain harmonics from a specific sound.

The harmonics that we wanted to extract from a sound of the Hammond organ are not sine waves. They are complex signals and some of them are more complex than the others in function of their shape, their amplitude… The difference of complexity between the harmonics can partially explain why ICA has more difficulties to extract certain harmonics than the others.

Furthermore, the different preset of drawbars can also be considered as a rational explanation to the non extraction by ICA of certain harmonics. We have noticed during our experiment that the harmonics extracted by ICA correspond to the drawbars, which have a degree of volume intermediate (not 0 and not 8). Indeed, when one drawbar is set at 0, ICA has never succeeded to extract the harmonic corresponding to this drawbar. In the case of a drawbar set to 8, the strength of the matching harmonic is maximum that can involve a certain saturation of the sound and can make the extraction of the harmonics by ICA very difficult.

By reading the above paragraph, we could draw some conclusions that would not be suitable. Indeed, we could think that a sound C2 with the following drawbars preset 327645222 could be easily extracted by ICA. Nevertheless, this is the combination of nine harmonics that creates a sound of the Hammond organ. In fact, a sound of the Hammond organ is a complex wave form generated by the combination of the nine harmonics. Thus, the closed links between the harmonics make harder their extraction from a sound by ICA.

7. References

E. Ifeachor, B. Jervis, "*Digital Signal Processing: a Practical Approach*", Second Edition (2002), pp.105-171

Andrezj Cichocki, Schun-ichi Amari (April 2003), "ICALAB package". http://www.bsp.brain.riken.go.jp/ICALAB/ICALABSignalProc/

Aapo Hyvarinen and Erkki Ojafrom (2002), "ICA algorithm Matlab package". http://www.cis.hut.fi/projects/ica/

Aapo Hyvarinen and Erkki Ojafrom (April 1999), *"Independent Component Analysis: a Tutorial"*. http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/

Aapo Hyvarinen and Erkki Ojafrom (April 1999), *"Fast and Robust Fixed-point Algorithms for Independent Component Analysis*". http://www.cis.hut.fi/~aapo/pub.html#FastICA

E. Miranda, *"Computer Sound Design: Synthesis Techniques and Programming"*, Second edition (2002), pp.1-18

Mark Vail, "*The Hammond Organ: Beauty in the B"*, Second edition (2002), pp.5-19, 42-43

Perceptual Modelling of Piano Tones

Y. Zhang, B. Hamadicharef and E.C. Ifeachor

Signal Processing and Multimedia Communication, University of Plymouth, Plymouth,
United Kingdom
e-mail: eifeachor@plymouth.ac.uk

Abstract

An automated system of modeling piano sound using excitation/filter sound synthesis technique is presented. High quality piano sounds have been analysed in time and frequency domain. Filter models are designed to match the string resonance and excitation signals have been created using inverse filtering technique. The audio quality of the synthetic sounds is assessed perceptually using the basic version of PEAQ (Perceptual Evaluation of Audio Quality) algorithm. Modifications have been made in synthesis parameters to observe the impact on result sound quality. Issues towards the DSP implementation are also discussed and results of the simulation are presented.

Keywords

Modeling piano, Piano tones, Sound quality, PEAQ.

1. Introduction

Piano is one instrument that has been the centre of attention over many years, for its difficulty in synthesis. As a percussive instrument, Piano sound can be decomposed into two successive parts: the attack part, which is of short duration, and the resonance. The resonance can be shown to a sound composed of exponentially damped sinusoids. However, the attack part of the signal has a much more complex structure and is usually quite difficult to analyse, model, or synthesize. Here, we will try to revive some properties of the attack part in piano sound by changing the synthesis parameter, in detail: excitation signal, in the system, and evaluate the impact on result quality.

The aim of this work is to create a simulation sound modeling system to investigate the potential of perceptual audio quality based modeling of musical instruments. We aim to automate the modeling of Piano using known sound samples by the synthesis technique described by Laroche and Mellier (Laroche and Meillier, 1994). Perceptual Evaluation of Audio Quality (PEAQ) will be introduced to assess the perceived quality of the synthesis sound (Thiede *et al*, 2002). Experiments of changing synthesis parameters will then be carried out in the system, to evaluate the impact of synthesis parameters on result sound quality, which will do help for a better understanding of the synthesis system.

In the following, we detail the modeling approach first in section 2. Then we introduce a new manner to assess the perceived sound quality in section 3. Section 4 presents the results of experiments conducted. Section 5 is the conclusions and the suggestion for further work.

2. Modeling of Piano Tones

Multi-channel excitation/filter synthesis has been used to model piano sounds (Laroche and Meillier, 1994). In simple terms, excitation signal aims to model the impact of the hammer on string, and the piano string is modeled as a filter which corresponds to the properties of string resonating, the string filter. Figure 1 shows the analogy between the physical piano and our synthesis system. In our work, we ignore the soundboard and consider them to be included into the filter resonance.

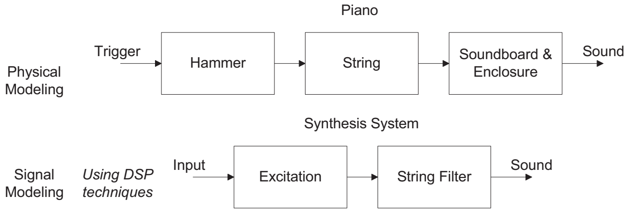


Figure 1 : Analogy between physical piano and synthesis system

Figure 2 illustrates the complete modeling system. The sound is fed into a bank of bandpass filters all centred to the frequency of each harmonic we obtained from analysis using FFT. Energy separation using Teager Energy Operator (TEO) is then applied to each harmonic to obtain their frequency track and amplitude envelope. Then the filter parameters are calculated from these results, and the synthesis could be carried on by inverse filtering the original signal and filtering the excitation. After we get the result, the perceived quality of the result is calculated by PEAQ.

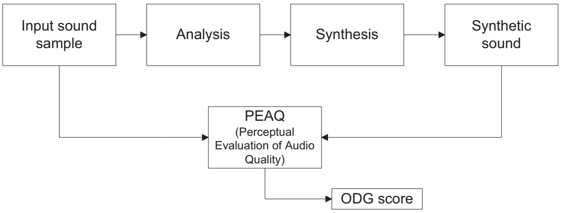


Figure 2 : Synthesis system

2.1 Sound Analysis

The sound analysis is based on a technique combining bandpass filters to isolate each harmonic and applying Teager Energy Operator (TEO) to extract both amplitude and frequency behaviour of each of the sound components (Kahrs, 2001). As shown in Figure 3a, we can extract easily most of the main harmonics of the piano sound as they exhibit large peaks in spectrum. The accuracy is important as it might affect the amplitude and frequency of harmonics in the final sound, thus the frequency result of FFT analysis is used, for there are sometimes slight differences between the theoretical values and practical ones.

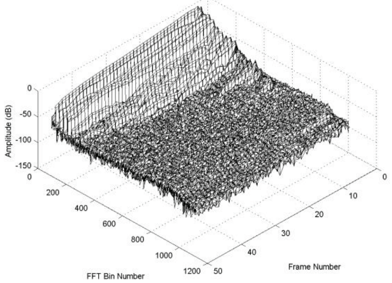


Figure 3a : Power Spectrum of note E5

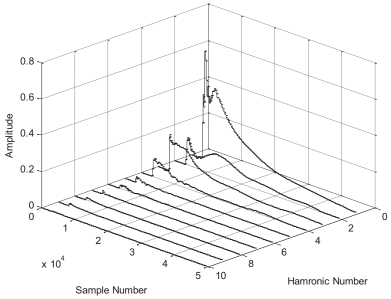


Figure 3b : Amplitude results of TEO on E5

Figure 3b shows the amplitude envelope obtained by energy separation using TEO, and these results will be used in filter modeling later.

2.2 Excitation Signal

The excitation signal is a kind of burst of noise. It characteristic can be observed from Figure 4. Excitation signal is obtained by inverse filtering the sound sample as H in Figure 2. Excitation signals must be stored in memory, when the synthesis system is implemented in DSP hardware. This is the main usage of memory in DSP implementation.

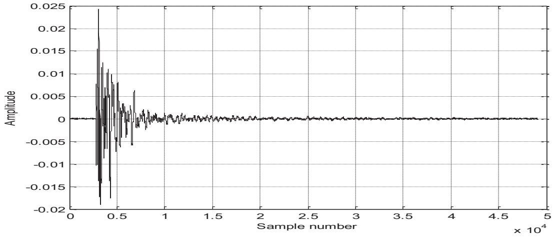


Figure 4 : Sample excitation in time-domain

One important issue is to try to reduce the excitation signal, as to save the memory usage. The reduction can be carried out individually or globally. Individual reduction is to reduce the length of the excitation by applying windowing function. Different windowing functions have been tested and results are presented in section 4.

Global reduction of excitation is to reduce the number of excitations, this is archived by using common excitations for a group of sounds.

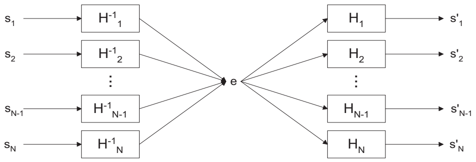


Figure 5 : Synthesis system of common excitation

Figure 5 shows how to generate the common excitation using known sound samples. Technique of least-squares deconvolution is used here, as the common excitation is the IFFT of:

$$E(f) = \frac{\sum_{i=1}^{N} H_i^*(f) S_i(f)}{\sum_{i=1}^{N} H_i^*(f) H_i(f)} \tag{1}$$

where $H_i(f)$ is the complex transfer function of filter H_i, and $S_i(f)$ is the Fourier transform of input signal s_n.

Yet another possible solution exist of reducing the number of excitations, that is using interpolation to generate the excitation of unknown sound samples via combining the known excitations. Of course we need the string filter of the unknown sound, which could be obtained by analysis, and stored before synthesis. The memory space needed for string filters is much less than that of excitations. Different combinations of known sounds are tested in a modified synthesis system.

2.3 String Filter Design

The strings are modeled using filters. For detail of filter design the reader should read (Ifeachor and Jervis, 2002). The string filter is based on second-order cosine sections and its magnitude transfer function is shown in Figure 6. The filter models the string resonance and aims to match the peaks in the frequency responses as in Figure 3a. These peaks are actually the harmonics, and the poles of the filter are obtained by calculating from a section of frequency tracks and amplitude envelope as the result of sound analysis in 2.1 (Sussman and Kahrs, 1996). The stability of the filter must be ensured, that is to say all the poles of the filter is within the unit circle. We need to manually relocate the poles if there are poles outside the unit circle. The magnitude of poles, which are outside the unit circle, have been limited to 0.999. This helps to ensure the stability of the filter.

Quantisation effect on filter coefficients is an important issue related to the string filter, in case of an implementation using a fixed-point DSP hardware like Motorola DSP 56300. The quality of result sounds with different quantisation bits has been compared.

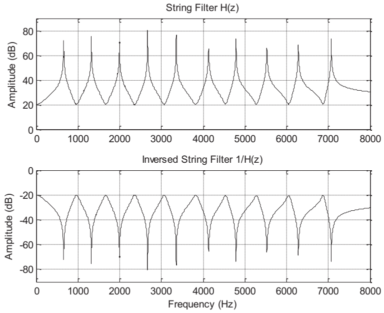


Figure 6 : Magnitude transfer function of string filter, note E5

3. Perceived sound quality

As the main aim of this research work is to create a simulation sound modeling system to investigate the potential of perceptual audio quality based modeling of musical instruments. The perceived sound quality assessment is done in two manners. First is the done with the listening test, carried out by audio experts. The second method is to use the Perceptual Evaluation of Audio Quality (PEAQ) based on the ITU-R BS.1387 (Thiede *et al*, 2002). Last block in Figure 1 shows the test setup for perceived sound quality in the system. PEAQ uses the original sounds as the reference signal and the synthetic sound as the test signal. PEAQ generates 11 Model Output Variables (MOVs) from the perceptual analysis. The final output called Objective Difference Grade (ODG) is used as an overall sound quality index. PEAQ consist of two main parts, the perceptual model, model of the human hear and the cognitive part, model of the judgement behaviour of the test subjects (ITU-R BS.1387).

4. Results

Experiments have been carried out and results are presented in the following sub sections. First, the excitation and filter are kept and the sound quality is assessed.

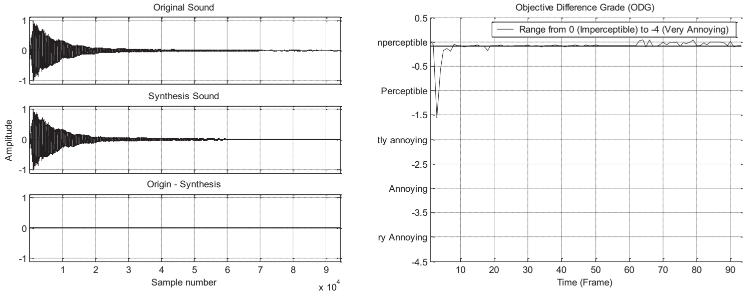


Figure 7 : Result of sound without modifying synthesis parameters with ODG

The mean ODG score is -0.09936 in this case, indicating very good result if no parameters are changed.

4.1 Reduction of the excitation signal

4.1.1 Windowing the excitation

The window functions experiments here are listed in Table 1, with the related ODG scores.

Window Length	4K	8K	16K	32K	64K
Triangle	-3.557	-3.590	-3.495	-3.237	-2.770
Rectangular	-3.653	-3.374	-3.007	-2.120	-0.682
Mixed	-3.559	-3.422	-3.149	-2.594	-1.446
Logarithm	-3.568	-3.505	-3.284	-2.906	-2.041
Exponential	-3.532	-.3.569	-3.589	-3.654	-3.621

Table 1 : ODG results of windowing the excitation

The results of shorten the excitation by applying windowing are not good in view of perceptually assessments, ODG scores between -3 and -4 stand for very annoying, listening tests also confirms this. By shortening the excitation, the components controlling beating in piano sound are eliminated, for these components exist all over the excitation signal. The quality drops as the resonance presents a very unnatural steadiness.

4.1.2 Common excitation

Common excitations generated from 3 notes and 5 notes are presented below with the PEAQ ODG scores.

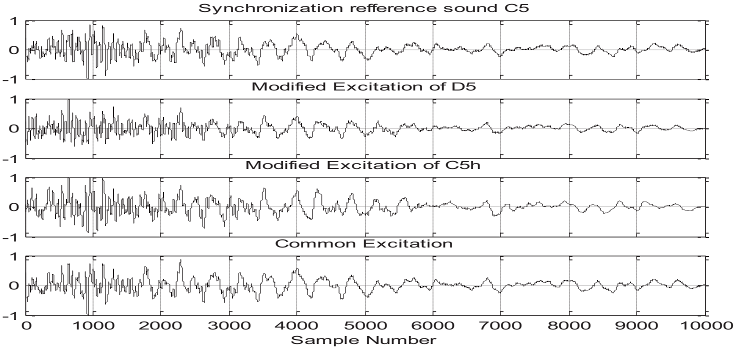


Figure 8 : Excitations of C5, C5#, D5 and the common excitation of them

The ODG scores are listed in table 2 and 3:

Piano note	Mean ODG	MSE of Excitation
C5	-1.7155	0.0012421
C5#	-1.4428	0.0015984
D5	-1.6626	0.0014435

Table 2 : ODG results and excitation MSE of excitation vs. common excitation, 3 notes

Piano note	Mean ODG	MSE of Excitation
C5	-2.3423	0.0023425
C5$^{\#}$	-1.8432	0.0017373
D5	-1.9461	0.0026022
D5$^{\#}$	-2.2632	0.0017150
E5	-2.1640	0.0023920

Table 3 : ODG results and excitation MSE of excitation vs. common excitation, 5 notes

As the number of sounds using in the calculation increases, the quality of result sounds drops. The common excitation is a combination of the known excitations. More sounds are used in the calculation of the excitation; more components are included in the estimated excitation, and more interference when applying to a sound. This leads to the drop in result quality.

4.1.3 Interpolation of excitation

The combinations tested are: comb1) white keys to black keys (e.g. use the excitations of notes F5 and G5 to generate the excitation for F5$^{\#}$), comb2) sound in a chord (e.g. D5+F5$^{\#}$ =>A5), comb3) same tone of different octaves (e.g. C4+C6 =>C5), comb4) sounds closed together (e.g. C5+D5 =>E5). Also the possibility of using common excitation to the sounds close to the group of known sounds is tested in this modified system.

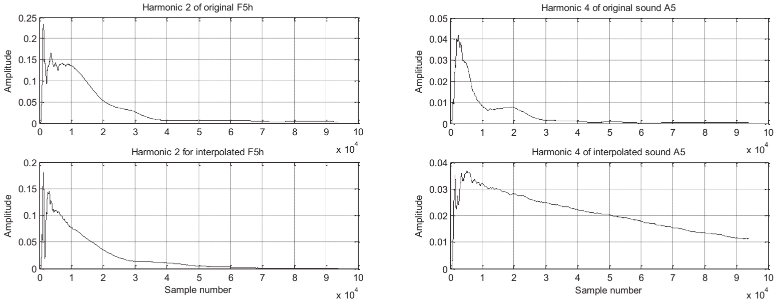


Figure 9 : Comparing harmonic, 2nd in comb1 and 4th of comb2

Result quality ODG scores of comb1): -3.671; comb2): -3.7362; comb3): -3.7138; comb4): -3.8637; using common excitation: -3.7501. All of results show the quality of very annoying.

Observing from Figure 9, the interpolation results in changing decay rate in result sound harmonics. Further study by comparing the spectrum of estimated and unknown excitations indicates that the estimated excitation contains the frequency components from all the known excitations, but lack of the components in the unknown excitation.

4.2 Filter coefficients quantisation

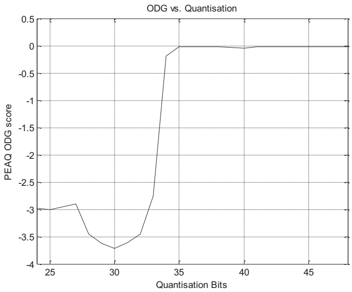


Figure 10a : ODG vs. Quantisation, E5

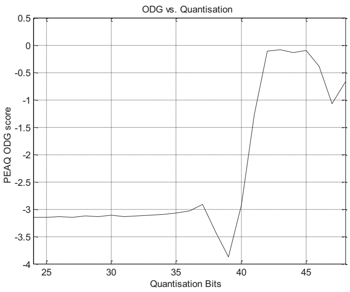


Figure 10b : ODG vs. Quantisation, C4

Figure 10a and 10b show the ODG versus the quantisation defined as word length in bits. It is clear that the quantisation of the filter leads to distortion in resulting sound, and the quality drops dramatically at certain point for which the string filter start to ring. For note E5, at least 34bits are required and 42bits for note C4. Combining 2 words in Motorola DSP56300 results in a 48bits representation could be enough for filter coefficients. Quantisation is also an important issue with the possible future hardware implementation in FPGA and VLSI chips.

5. Conclusions

In this paper, a fully automated piano sound modeling system has been presented. From the recorded piano tones, the analysis extract both filter model parameters and excitation signals, the main components of excitation/filter model. The assessment of the sound quality is carried out using the PEAQ algorithm basic version.

The perceptual effect on sound quality has been investigated by modifying the model parameters: filter coefficients and excitation signals. Quantisation effect on filter coefficients is a major concern in DSP implementation. This has been investigated to obtain the minimum word length required for filter coefficients.

Processing on excitation signal include: shortening the excitation, grouping into a common excitation from notes within one octave, interpolating the excitation for sounds using input samples of different combinations.

The system can model piano sounds with little distortion. Listening tests and PEAQ results shows the quality of sounds satisfy the requirements as for mean ODG score bigger than -1. When there is a change in excitation or filter parameters, the result sound quality drops on all these ways of modification which shows the influence of both excitation and filter parameters.

Future work could add a feedback path to the system to change the analysis/synthesis part in system from the ODG scores of the synthesis sound. A model of the excitation signal could be developed as well as a soundboard Digital Waveguide model to allow modeling of low pitch notes. However, it has been found that the excitation signal has great influence on the result sound quality. There is more work to be done in that direction.

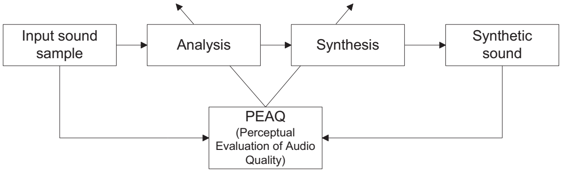


Figure 11 : Synthesis system with feedback path

6. References

Ifeachor, E.C. and Jervis B.W. (2002), *Digital Signal Processing: A Piratical Approach*, Second Edition, Prentice Hall, London.

ITU-R BS.1387 (2001), *Method for objective measurements of perceived audio quality*, ITU Recommendation BS.1387-1, http://www.itu.int/rec/recommendation.asp?type=items&lang=e&parent=R-REC-BS.1387-1-200111-I, Sep, 2003.

Kahrs, M. (2001), "Audio Applications of the Teager Energy Operator", *111th Audio Engineering Society Convention*, New York, USA, Sep, 2001.

Laroche, J. and Meillier, J-L. (1994), "Multichannel Excitation/Filter Modeling of Percussive Sounds with Application to the Piano", *IEEE Transactions on Speech and Audio Processing*, Vol.2, No.2, April 1994, pp.329-344.

Sussman, R. B. and Kahrs, M. (1996), "Analyse and Resynthesis of Musical Instrument Sounds Using Energy Separation", *Proceeding of ICASSP-1996,* Atlanta, GA, May, 1996.

Thiede, T. Treurniet, W.C. Bitto, R. Schmidmer, C. Sporer, T. Beerends, J.G. Colomes, C. Keyhl, M. Stoll, G. Brandenburg, K. and Feiten, B. (2002) "PEAQ-The ITU Standard for Objective Measurement of Perceived Audio Quality", *Journal of Audio Engineering Society*, Vol. 48, pp 3-29, Jan-Feb, 2002.

Low-Density Parity Check Codes and Projective Geometry

P. Mulotte and M.Z. Ahmed

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

Abstract

The LDPC (Low-Density Parity Check) codes were invented by Gallager in 1961 and had been forgotten until recently. Today LDPC and Turbo Codes are the most popular codes used. In this paper the performances of two LDPC codes is assessed. One is the systematic form of a regular Gallager matrix and the other is the appending of a regular matrix to an identity matrix resulting in an interesting systematic form. The LDPC codes are decoded with the Sum-Product Algorithm (SPA) proposed by Gallager (Gallager, 1963) and the Maximum A Posteriori (MAP) decoder (Bahl, Cocke, Jelenik and Raviv, 1974). The complexity of these decoders is also investigated and it was found that the SPA has lower complexity than the MAP decoder; resulting in quicker decoding. A new area in codes design, Projective geometry is also presented. This paper provides an approach to the design of Projective geometric sets. The performances of codes produced by Projective Geometry are tested with both SPA and MAP decoders on a BSC (Binary Symmetric Channel). These have been found to outperform the regular LDPC codes.

Keywords

Error Correcting Codes, LDPC, Projective Geometry, Sum-Product Algorithm, Maximum A Posteriori decoder

1. Introduction

Low density parity check (LDPC) codes were invented by Gallager in 1963(Gallager, 1963). The concept idea is to build matrices containing mostly 0's and a few 1's. Gallager's work was quickly abandoned because of the important need of computation that computers cannot perform at that time. His work was rediscovered and generalised 30 years later by MacKay and Neal (MacKay and Neal, 1997, MacKay, 1999, 2000). Then lots of other people started studying the LDPC codes because of their interesting properties. In only a few years, lots of improvements have been brought to that field. Let us take, for example, the use of finite geometric designs achieving outstanding results.

Because of their near Shannon limit performances, LDPC are considered to be serious competitors to Turbo Codes. More recently, Richardson and Urbanke have developed irregular LDPC codes that perform even better than turbo codes for very large block lengths (n>105) and can come within 0.1 dB of the Shannon capacity (Richardson T., Shokrollahi A., and Urbanke R., 2001). LDPC codes are one of hottest subjects in the Error Correcting Codes field.

This paper investigates the implementation of an LDPC code and to assess its performances with a MAP and an SPA decoder. Projective Geometry is then presented and codes based on LDPC codes based on PG are designed.

Gallager's original design method is used to generate a sparse LDPC decoding matrix. This is then transformed into a systematic form using

1. an algorithm based on a Gaussian elimination and
2. an easy way, concatenating an Identity matrix with an LDPC matrix.

By looking at the characteristics of the code we will see that the second method has much lower complexity.

The decoding part will be made with a MAP (Maximum A posteriori) decoder (Bahl, Cocke, Jelenik and Raviv, 1974) and an SPA (Sum-Product Algorithm) (Gallager, 1963) decoder. Gallager in his thesis proposed the SPA (Sum-Product Algorithm) algorithm. On the other hand we have the MAP (Maximum A Posteriori) (Bahl, Cocke, Jelenik and Raviv, 1974) which has proved its usefulness for years. Both decoders are tested on the Binary Symmetric Channels with AWGN (Additive White Gaussian Noise). The performances of these two decoders and of the two codes are nearly identical except that the complexity of the SPA is much lower.

The next section will be about the explanation and design of LDPCs based on Projective Geometry (Anderson, 1990). A basic code using Projective Geometry, a Type-I PG-LDPC, is designed and compared with the previously designed codes to find an interesting improvement in the performances of the code. The next step would be to go deeper in some Projective Geometry theorem like Singer’s theorem or the k-flats to design optimal codes.

2. Study and statistics of the codes

Two different codes based on a regular LDPC matrix were designed. To encode data, the matrices need to be in systematic form. A systematic form means that the first part is an Identity matrix. To build codes with this structure, first an algorithm based on a Gaussian elimination is used. The second method used consists of concatenating an Identity matrix with a regular LDPC matrix, this is referred to as the *easy way*. The properties of these two is then investigated.

2.1 Density

The concept idea of LDPC matrices is to be composed of lots of 0’s and only a few 1’s. The result of a Gaussian elimination is far from being that sparse. Typical density obtain was around 0.23 for any block length. Whereas the density of the concatenation method is inferior to the density of a regular LDPC matrix, decreasing even below 10^{-2} when the block length increases.

2.2 Minimum row distance

The minimum row distance is the minimum number of differences between any two lines in the matrix. We now have a look at how the minimum row distance is affected by the construction processes:

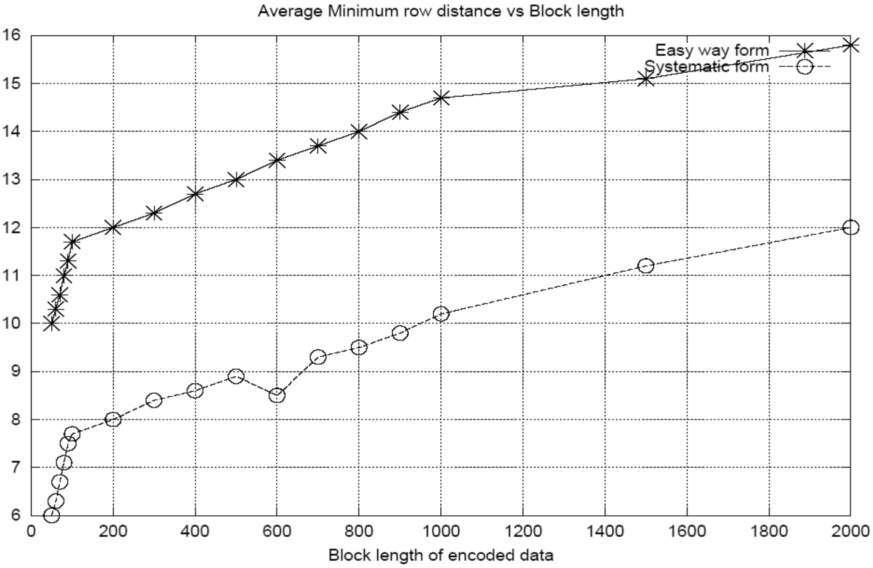


Figure 1 : Comparison of minimum distance

The minimum row distance of the *easy way* will automatically be equal to the minimum row distance of a regular LDPC matrix plus 2. The Gaussian elimination based process results in a very poor minimum row distance whereas the *easy way* gives a result that is nearly linear.

With a lower density and a much better minimum row distance, the concatenation of an LDPC matrix with an Identity matrix is much more efficient. I will use this code to compare the Projective Geometry code.

3. MAP and the SPA decoders

The two decoders used, the MAP (Maximum A posteriori) decoder [5] and the SPA (Sum-Product Algorithm) (Gallager, 1963) decoder are going to be used to evaluate the code performances. These decoders are probabilistic decoders. They evaluate the probability to have received one value conditionally to the transmitted data. The key point of this process is the computation of the Extrinsic Information, which is for each bit the ratio of the input probability (AP) on the probability of each bit after the decoder (APP). This APP is computed from the parity-check equations. Both decoders have their own way of calculating the extrinsic information but their results are numerically nearly identical. These decoders are iterative decoders. They take in input the probability of each bit and the extrinsic information from the previous decoder for each bit. With each iteration more bits are corrected allowing more correction at the next iteration. This concept is the key of the latest iterative probabilistic decoders.

3.1 MAP algorithm

The MAP algorithm is based on the Trellis representation of a code. The basic concept of the trellis is to represent all the sequences of a system in an edge-labeled graph. This representation describes a code with all its constraints. Here is an example of Trellis representation for an LDPC code:

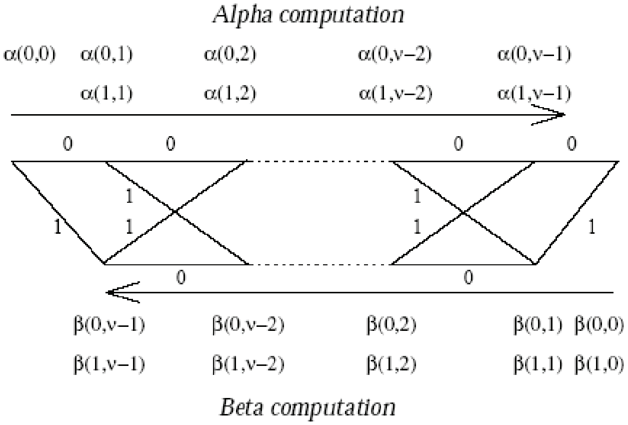


Figure 2 : MAP decoder computations

From the alpha and beta computation, we determine the extrinsic information. This operation is made for each parity-check equation (line of the matrix). Then it results in the extrinsic information of each bit in each equation. To combine these extrinsic informations the product of all the extrinsic information from each parity check equation for each bit is obtained and passed to the next decoder. At the last iteration a threshold detector of the APP is used.

3.2 SPA Decoder

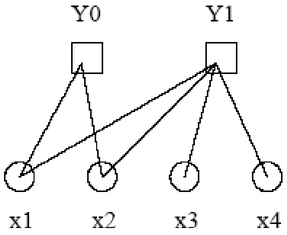


Figure 3 : Tanner Graph

The Sum-Product Algorithm (also called Belief Propagation) is a graphical method proposed by Gallager. A code can be represented by a bipartite graph. The nodes, Yi, represent the parity-check equations and the linked xj are the bits involved in each of these parity-check equations. The decoding uses this graph as a network for information. This decoder is based on equations derived by Gallager simplifying a lot the process. Like for the MAP decoder, it results in the extrinsic information that is passed to the next decoder.

3.3 BER results

The BER (Bit Error Rate) is one of the most important characteristics of codes:

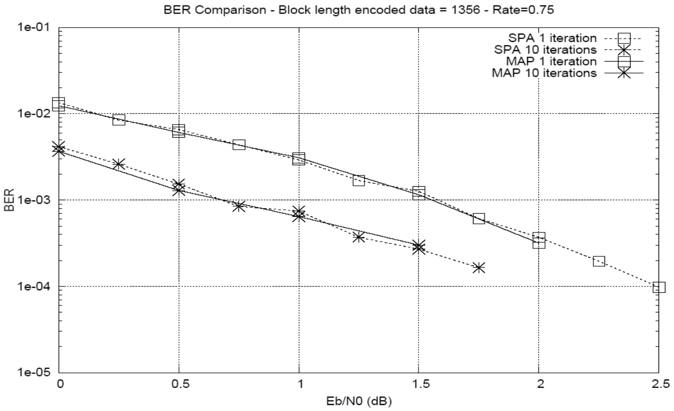


Figure 4 : Comparison of BER results

The results of this simulation are very interesting. The BER achieved by both decoders is nearly identical. Not only for one iteration but also for more like 10. The way the extrinsic information is computed is different but gives the same result. That must mean that both algorithm are different points of view of the same concept.

3.4 Complexity

The complexity is the number of operations required by a process. In this experiment, the complexity is the number of operations needed by the two decoders. The number of multiplications is taken the measure of complexity as the number of additions is insignificant for the SPA and equal to the number of additions for the MAP. These systems are designed to deal with real-time applications so a low complexity is important.

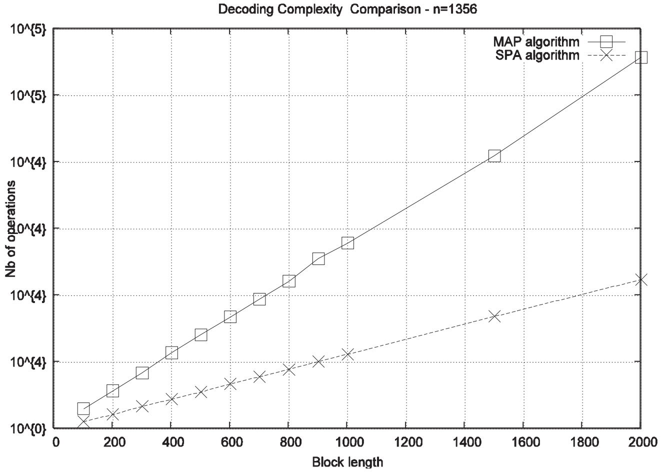


Figure 5 : Complexity comparison

The complexity of the decoder is very important in real-time systems. The MAP decoder complexity is O(n^2) and the complexity of the SPA decoder is O(n). These results are collected relatively to my implementation. It gives the SPA decoder a big advantage.

3.5 Comments about the SPA decoder

The SPA decoder complexity is very interesting but there is a problem with it. The SPA can encounter the presence of cycles resulting in inaccurate results. Cycles are cycles in the Tanner graph. For example on the graph in 3.2, we have a cycle of length 4: Y0-x1-Y1-x2. Most harmful cycles are the cycles of length of length 4 These kind of cycles prevent the involved bits from drawing all the information they could from the other bits of the parity-check equations. There are some ways to avoid these cycles, the splitting of rows and columns, the use of some specific designs like the Kirkman Triple systems or the Projective Geometry (H. Song, J. Liu, and B. V. K. Vijaya Kumar, 2002).

4. Projective Geometry

The goal of the Projective Geometry is to build finite spaces that when translated into a matrix allow us to find the optimal characteristics we need for this matrix.

4.1 Basics

Usually Projective Geometric sets are noted PG(m,q). Where m is the number of dimensions and q is a parameter affecting the number of points, of lines and the number of intersections.

In this geometry each line is a parity-check equation and the points are all the bits.

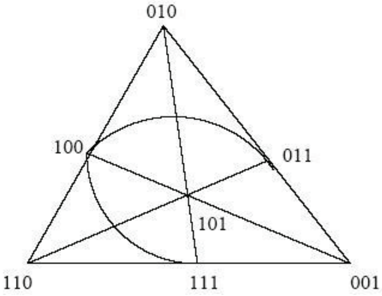


Figure 6 : Projective Geometry Field example PG(2,2)

Here is a quick example of a finite space produced from Projective Geometry. This set is PG(2,2). We have designed it so that for each point three lines intersect, each line hold 3 points and so that 2 given points are contained by one unique line. This will give us a matrix with three 1's per line, three 1's in each column and an optimal minimum distance of 4. This kind of space is a 2-dimensional Projective Geometric code, we use it to generate 2-dimensional PG-codes.

One of the interesting point is that because two points are belonged by only one line, we can't have any cycle of length 4 in this design.

4.2 Performances of Projective Geometry based code

The code implemented is a PG(2,16). It is a matrix (273,273) with 5 points by line and 5 points by column. It is concatenated to an Identity matrix to create my encoding matrix, obtaining a minimum row distance of 10. Another code using the concatenation of an Identity matrix with a regular LDPC matrix using the same number of points by line and by column. As both decoders obtain the same results I used the SPA because of its lower complexity.

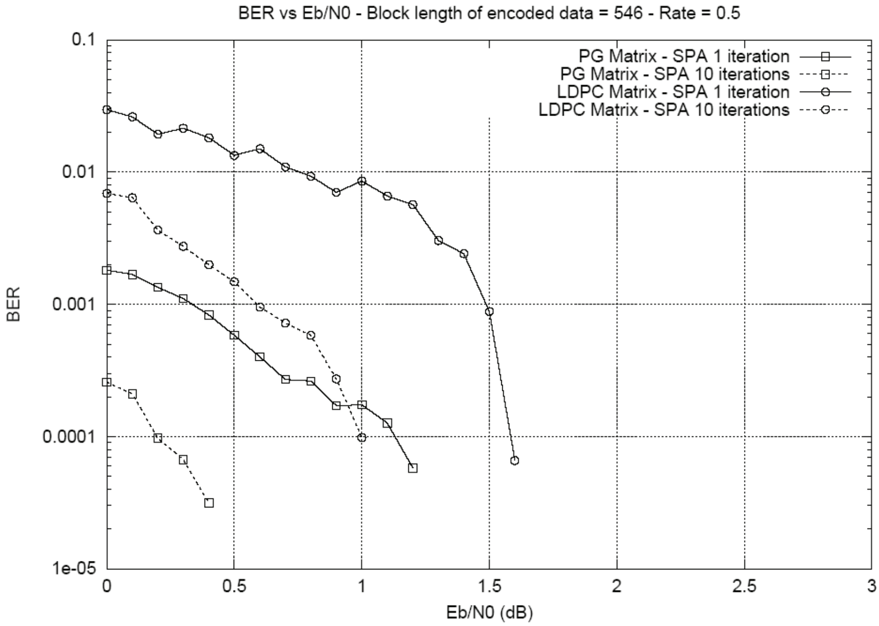


Figure 7 : Comparison of Projective Geometry BER results

The results are very impressive. The results of the PG (Projective Geometric) code are much better than the results of the regular LDPC code. For one iteration, the BER falls straight down after 1dB. And for 10 iterations of the decoder the BER falls after 0.4dB. This is a very interesting proving the efficiency of the codes developed with the Projective Geometry.

In this case, we can see the improvement brought by the 10 iterations of the decoder. These 10 iterations make the Projective Geometric code gain 0.6dB for the PG-code and 0.4 dB at BER of 10^{-4}.

5. Discussion

The two decoders give nearly exactly the same results. Their only difference, for these codes, is the complexity of the MAP which is higher than the complexity of the SPA algorithm. It makes the SPA decoder much more suited to the decoding of such codes.

The use of Projective Geometry is a very new method to design codes. The Results show a huge improvement by the use of this method. But this method uses only the very basic theorems and I think some theorems of Projective Geometry might be used to enhance the designs, like Singer’s theorem or the use of the k-flats (Anderson, 1990) which might help to design more complex pattern giving optimal configurations for more sets and codes. This part of design is still mostly unknown and can offer thousand of optimal configurations.

6. References

Anderson I. (1990), “Combinatorial designs: Construction Methods”, *Ellis Horwood*, pp. 98-112.

Bahl L.R., Cocke J., Jelenik F., Raviv J. (1974), “Optimal decoding of linear Codes for minimising sybmol error rate", *IEEE Transactions in Information Theory*, vol. 20, pp. 284-287.

Gallager R.G. (1963), ”Low-Density Parity-Check Codes”, Phd, MIT.

H. Song, J. Liu, and B. V. K. Vijaya Kumar (2002), “Low complexity LDPC codes for partial response channels,” *Globecom’2002*, Taipei, Taiwan, November 2002.

MacKay D.J.C., Neal R.M. (1997), “Near shannon limit performance of low density parity check codes", *Electronic Letters*, vol. 33, no. 6, pp. 457-458.

MacKay D.J.C. (1999), “Good error-correcting codes based on very sparse matrices", *IEEE Transactions in Information Theory*, vol. 45, no. 2, pp. 399-431.

MacKay D.J.C. (2000), “Turbo codes are low density parity check codes"; *personal communication*, vol. MacKay Draft 0.2, 2000.

Richardson T., Shokrollahi A., and Urbanke R.(2001), “Design of capacity-approaching irregular low-density prity-check codes”, *IEEE Trans. Inform. Theory*, vol. 47, pp. 619-637.

An Investigation into Iterative Turbo Decoding Schemes using Density Evolution

R.H. Laskar and M. Ambroze

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

Abstract

In this paper we analyse the density of the extrinsic information of the iterative turbo decoder. This is done by actual density evolution and also based on the observation that the extrinsic information from component maximum *a posteriori* decoders is well approximated by Gaussian random variables when the inputs to the decoders are Gaussian, this is then approximated by Gaussian density functions. The iterative turbo decoder is analysed by investigating whether a signal-to-noise ratio (SNR) grows with iterations. We define a noise figure for the iterative decoder, such that convergence to the correct codeword is achieved if the noise figure is bounded by a number below 0 dB. Mutual information transfer characteristics of the extrinsic information is also analysed and both the SNR and mutual information can be viewed in the form of an extrinsic information transfer chart (EXIT chart). This allows the exchange of extrinsic information to be visualised as a decoding trajectory and the convergence threshold of a code to be determined, which predicts the turbo cliff position of the bit error rate (BER) chart. Examples are given to show the influence of different code complexities on the convergence behaviour for parallel concatenated convolutional codes.

Keywords

Convergence, iterative decoding, mutual information, turbo codes, density evolution

1. Introduction

Turbo codes were first introduced by Berrou (1993), and are claimed to be one of the most recent significant advances in coding. They offer near-optimal performance close to Shannon's limits and require no more than moderate complexity. So it is no wonder that there have been a great deal of interest since the launch of turbo codes. There have been many studies into the convergence behaviour of iterative turbo decoding schemes and this paper gains an insight into some of the approaches that have been introduced so far.

To analyse the effect of iterative decoding, the method of density evolution, (Richardson and Urbanke, 2001), is an effective approach with the assumption that the block size tends to infinity. This technique is used to trace the probability density function, with successive iterations, of the log-likelihood ratios of extrinsic information messages. Whereas Richardson and Urbanke applied this method to low-density parity check codes (LDPCs), in this paper, we will apply density evolution to compute iterative decoding thresholds for turbo codes over a binary input additive white Gaussian noise (AWGN) channel, as has done by Divalsar *et al* (2001) in their paper, who also analysed turbo-like serially concatenated codes.

Hesham El Gamal (1999), observed that when the input to the constituent maximum *a posteriori* MAP decoders are Gaussian, the log-likelihood of extrinsic information messages from each decoder is well approximated by Gaussian random variables. They considered the MAP decoder to be a SNR transformer and in addition to this, they proposed an approach to analyse the overall turbo decoder. This approach was that the iterative decoder convergences to zero probability of error as the number of iterations increase if and only if the channel Eb/No exceeds the threshold which is implied by the independent Gaussian model. Divalsar *et al* also defined a noise figure for the overall decoder and argues that if the noise figure is bounded by a number strictly lower than 0 dB, then the iterative decoder converges to the correct code sequence. In this study, these analyses will be applied to fully understand the implications that are suggested.

In this study we will first of all simulate the turbo code over a binary (AWGN) channel to assess its BER performance. We will then apply the actual density evolution technique, and in addition, the Gaussian approximation will be implemented to approximate the density function of the log-likelihood ratio of the extrinsic information messages. This will then be plotted on an EXIT chart (Brink, 2001) to analyse the transfer characteristics and decoding trajectories based on each of the SNR and mutual information measures.

Then the convergence threshold will be investigated by studying the theory that if the two curves for each decoder do not intersect, the iterative decoder converges will be analysed to determine the threshold for convergence. This will then be plotted onto to the graph of the BER performance to see whether it does in fact give an accurate prediction of the turbo cliff position

2. Iterative turbo decoder

If we consider the turbo decoder shown in Fig.1, we can see that after being permutated by a pseudo-random interleaver, the sequence of extrinsic information is being passed from component decoder to the next. Here we can define the interleaver by a function *t(m)*, *m* = 1,2,…N, which points to the position in the original sequence to be permuted to position m in the interleaved sequence. Here N is the length of input sequence and hence the length of the interleaver. So we can deduce that the de-interleaver can be defined by $t^{-1}(m)$.

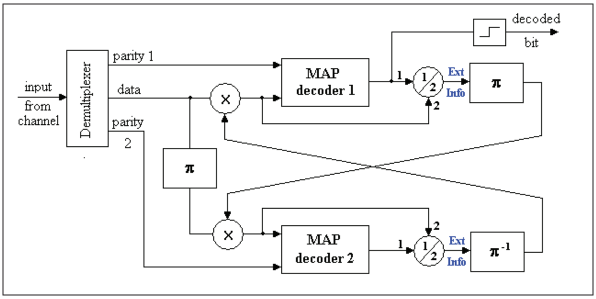


Figure 1 : Turbo Decoder

If we assume that the all-zero codeword is transmitted, then the extrinsic information that each bit, *i*, was a 0, as can be seen in the diagram, is equated as follows

$$Extrinsic\ Information_i = \frac{Output\ of\ MAP\ decoder}{Input\ of\ MAP\ decoder} = \frac{a\ posteriori\ probability_i}{a\ priori\ probability_i}$$

for each bit. Note that the input of the map decoder refers to the probabilities of each bit of the data to be a zero and not the parity. Also at this stage, we define λ_i to be the log-likelihood ratio, which we define as

$$\lambda_i = \log \frac{Extrinsic\ Information\ that\ bit_i = 0}{Extrinsic\ Information\ that\ bit_i = 1}$$

As the interleaver that was used was random, the λ_i are independent and identically distributed. The probability distribution $f(\lambda)$ of these log-likelihoods, where seen to evolve with each iteration. With the RSC(17/13) 4 – state code, the observations shown in Fig.2 were taken.

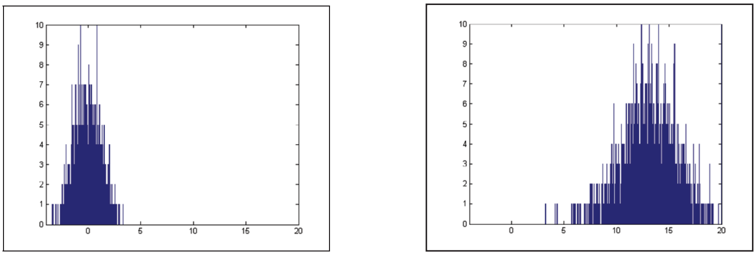


Figure 2 : Distributions of the log-likelihood of the extrinsic information of the 1st and 4th iteration

The length of the blocks were taken to be 2000 bits long and the λ's computed within 200 bits of the edges were discarded. As we can see, with consecutive iterations, the probability densities evolve from narrow densities near zero, with a small mean to wider Gaussian-shaped densities with higher means. Note that the probabilities for the extrinsic information that each bit is a 0 or 1 were both limited to be within $[10^{-20}, 1-10^{-20}]$, so no division by 0 occurred when calculating λ_i. As this was the case, $\lambda_i \in [-20,20]$, which is why as the iterations increase, the densities move towards λ_i =20 and eventually become a narrow strip at λ_i =20.

The distribution $f(\lambda)$ becomes more Gaussian-shaped with iterations as λ_i are i.i.d., so $f(\lambda)$ can therefore be approximated by a Gaussian density function. The Gaussian density function will therefore be dependent on two parameters, its mean, $\mu = E(\lambda)$ and its variance, σ^2 = Var(λ) and for this random variable, a signal-to-noise ratio is defined as $SNR = \frac{\mu^2}{\sigma^2}$. In addition, if f(λ) is both Gaussian and consistent, then the following properties hold, $\sigma^2 = 2\mu$ and SNR = $\mu/2$.

3. Actual Density Evolution

The RSC(17/13) code was used here with a code rate of 1/3 and the channel Eb/No was of 0.5 dB. The SNRs at the input and output of each component decoder for the actual density evolution was evaluated as SNR = $\frac{\mu^2}{\sigma^2}$. Denoting the SNR input and output of the MAP decoder 1 and 2 as SNR1in, SNR1out, SNR2in and SNR2out, despite the first MAP decoder starting with SNR1in = 0, a non-zero Eb/No of the channel enabled MAP decoder 1 to produce a non-zero SNR1out.

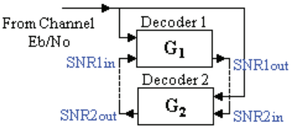


Figure 3 : Analysis of turbo coding as a non-linear dynamical system with feedback using density evolution (Divalsar *et al*, 2001)

So this means that the SNR output for a decoder is a non-linear function of the input SNR for a certain Eb/No. So if we denote the functions for decoder 1 and decoder2 as G_1 and G_2 and noting that SNR2in = SNR1out, we have SNR1out = G_1(SNR1in, Eb/No), SNR2out = G_2(SNR2in, Eb/No) ⇒ SNR2out = $G_2(G_1$(SNR1in,Eb/No), Eb/No). Therefore, by working out the SNR at each stage of the iterations, the SNR1in = SNR2out can be plotted against SNR1out = SNR2in and the change in SNR can be analysed. The actual density evolution model generates input λs directly from the histogram from the previous decoder and the Gaussian approximation generates input λs from a consistent Gaussian density. With the Gaussian approximation, the consistency condition gave a better approximation to the actual density evolution and so the mean and variance were found by the following equations, $SNR = \frac{\mu}{2}$, $\sigma^2 = 2\mu$.

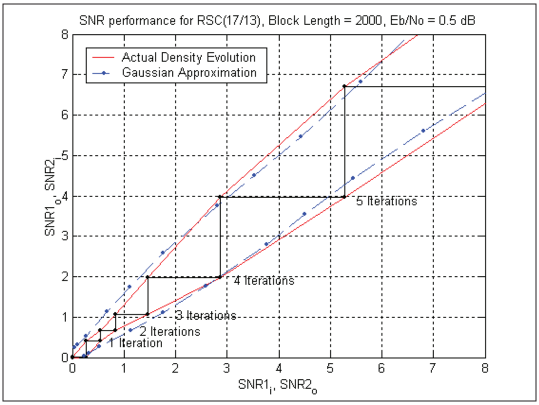


Figure 4 : EXIT chart for SNR of a rate 1/3, RSC(17/13) code

The results in Fig 4.show that the SNR increases with every successive iteration, which means that there is an overall increase in the values for λ and therefore convergence towards the true data sequence for each block is likely to be taking place.

For each iteration, a noise figure was defined as F = SNR1in / SNR2out and clearly if the figure is less than 1, there is an improvement in SNR from the beginning of the extrinsic information to the end. Table 1 shows the noise figures for the first 5 iterations.

	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration	5th Iteration
Noise Figure (4 s.f)	1.309×10^{-6}	0.5992	0.6269	0.5394	0.4982

Table 1 : Noise Figures of 1st 5 iterations of a RSC(17/13) code

As all of these noise figures, are below 1, i.e. above 0 dB, there is a constant increase in SNR for each consecutive iteration and so convergence to the correct codeword is taking place.

4. Mutual Information

If we consider the *a priori* probability*,* the probability that a 0 or 1 was sent and the *a posteriori* probability, the probability that a 0 or 1 was sent given that a 0 or 1 was received for each bit received from the channel, we can view the change in these probabilities as the amount of which the decoder has learned from the acceptance of the received bit from the channel. So as stated by Hamming (1986), the difference between the information uncertainty before (*a priori*) and after the reception of a bit from the channel (*a posteriori*), measures the gain in information due to the reception of that bit.

Now if we consider the log-likelihoods λ_{in}'s, for the *a priori* input to the decoder first, we can evaluate the mean by $\mu_{in} = \mathrm{E}[\lambda]$ and either the empirical variance, σ_{in}^2 for the actual density evolution or based on the consistency conditions for the Gaussian Approximation. The conditional probability density function for λ_{in} is then defined as:

$$p_{in}(\xi|X = x) = \frac{\exp(-((\xi - \mu_{in}x)^2))/2\sigma_{in}^2)}{\sqrt{2\pi}\sigma_{in}}$$

where X is the transmitted systematic bits. So now we can deduce that $p_{in}(\xi|X = x) = p_{in}(-\xi|X = x)\cdot e^{x\xi}$ and we can compute the mutual information, as stated by Brink (2001), is by:

$$I_{in} = \frac{1}{2}\cdot\sum_{x=-1,1}\int_{-\infty}^{+\infty} p_{in}(\xi|X = x)\log_2\frac{2\cdot p_{in}(\xi|X = x)}{p_{in}(\xi|X = -1) + p_{in}(\xi|X = 1)}d\xi$$

$$0 \le I_{in} \le 1$$

So we can also apply the same equation to the output values of λ from each decoder. Now by implementing the actual density evolution and the Gaussian approximation as with the SNR measures, we obtain the graph shown in Fig. 5 for the RSC(17/13) code.

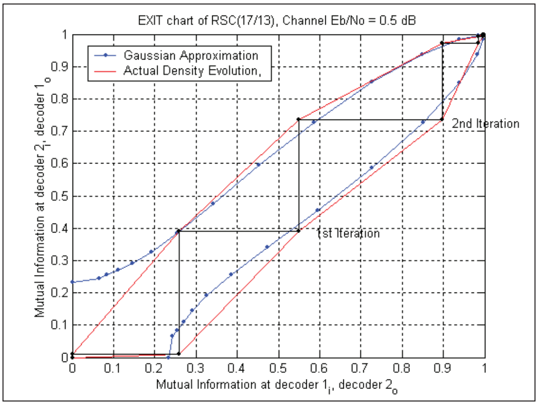


Figure 5 : EXIT chart of the Mutual Information of a rate 1/3, RSC(17/13) code

5. Finding the point of convergence

Divalsar *et al* stated that for the SNR measures and Brink for the mutual information measures, that the point at which the curves touch together, is the point of convergence. In other words when the curves touch, there is no more increase in SNR or mutual information with consecutive iterations and the correct convergence of the code sequence is not obtained. This point was found to be at a channel Eb/No of –0.21 dB for the SNR measures and at –0.18dB for the mutual information measures.

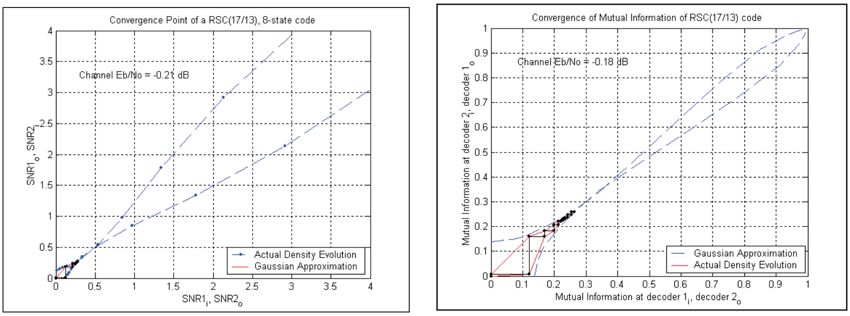


Figure 6 : a) SNR and b) Mutual Information at the convergence point

The actual density evolution shown in both Figs, 6a and 6b, illustrates how the SNR and mutual information cannot improve any further after the point at which the curves touch , with subsequent iterations. It is only when a high SNR is obtained and the mutual information reaches 1, that we are sure that the correct code word has been decoded. The trajectory for the actual density evolution shown in Fig. 6a, for the SNR increases up until the 4th iteration and we can see this fact more clearly by examining the noise figures as shown in Table 2.

	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration	5th Iteration
Noise Figure (4 s.f.)	1.692×10^{-4}	0.7817	0.9317	0.9612	1.03

Table 2 : Noise Figures of the 1st 5 iterations at the Convergence Point

By the 5th iteration, the noise figure is greater than 1 and so the SNR was actually decreasing a little. So to see whether the convergence points found by both methods, give a good prediction of the turbo cliff position, they are plotted on the BER graph for the rate 1/3, RSC(17/13) code in Fig.7. From this we can deduce that both methods are fairly accurate.

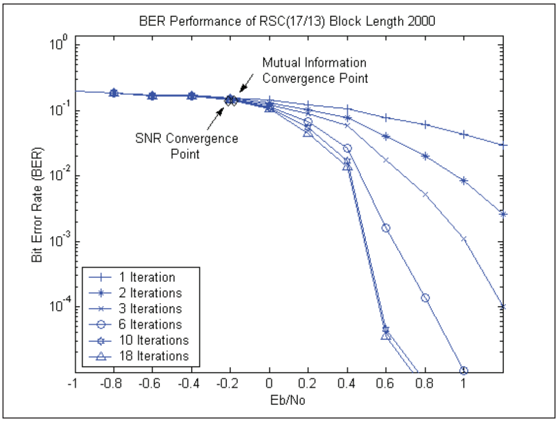


Figure 7 : BER performance of rate 1/3, RSC(17/13)

6. Different Code Complexities

To see the different effects of code complexities, a RSC(5/7) 4 state code, a RSC(17/13) 8 state code and a RSC(33/31) code is compared.

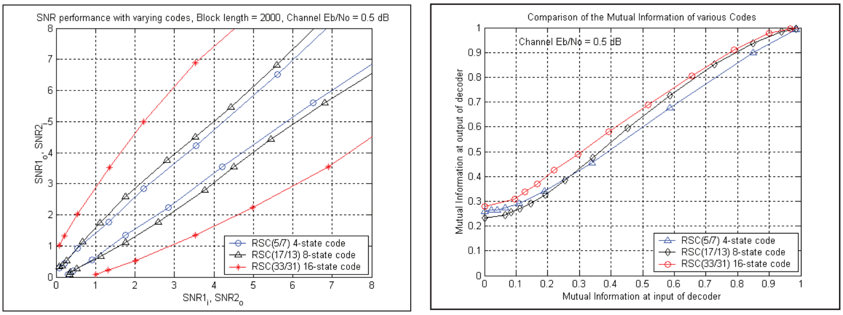


Figure 8 : EXIT charts showing the difference between various codes with different state complexities with a) SNR computations and b) Mutual Information values

For both methods, only the Gaussian approximation is plotted so as not to overload the graphs in Fig.8. For the SNR, the results show that the rate of increase of the SNR with iterations increases with the number of states. For the mutual information, there doesn't seem to be any significant change, but however for both graphs, we can see the 4 state and the 8 state codes seem to behave very similar and the 16 state codes differs from the other two more. So we would expect that the convergence point to be lower for higher state complexity codes. The convergence point for these codes were found and then checked to see if the predictions were correct on the BER graphs and are given inTable 3.

Convergence Points found by:	RSC(5/7) 4 state code	RSC(17/13) 8 state code	RSC(33/31) 16 state code
SNR measures	-0.12	-0.21	-0.33
Mutual Information measures	-0.1	-0.18	-0.36

Table 3 : Convergence Points given by SNR and Mutual Information EXIT charts

The results were verified on the BER graphs and for these particular codes that have been checked, we can see that the higher the state complexities the better the convergence properties.

7. Conclusion

The EXIT charts used in this paper proved to be a very effective approach in evaluating the convergence behaviour of iterative turbo decoding. The analysis of the extrinsic information using the SNR and mutual information was very useful for gaining an insight into this behaviour and the convergence thresholds found, although slightly varied between the SNR measures and mutual information measures, all gave a fairly accurate prediction of the turbo cliff position. For the SNR measures, the noise figure defined for the iterative decoder proved to be a very useful device to detect whether correct convergence occurs and to analyse the rate of convergence through successive iterations.

8. References

Berrou C, Glavieux A, (1996) "Near Optimum Error Correcting Coding and Decoding," *IEEE Trans. Commun*, vol. 44, pp. 1261-1271.

Brink, S. T., (2001), "Convergence Behavior of Iteratively Decoded Parallel Concatenated Codes," *IEEE Trans. Commun*, vol.49, no.10.

Divsalar D., Dolinar S. and Pollara F. (2001), "Iterative Turbo Decoder Analysis Based on Density Evolution," *TMO Progress Report* 42-144.

Gamal H. E.,(1999) *On the Theory and Application of Space-Time and Graph Based Codes*, Ph.D. Dissertation, University of Maryland at College Park.

Hamming. R.W., .(1986), *Coding and Information Theory*. Englewood Cliffs, NJ: Prentice_Hall.

Richardson T, R. Urbanke, (2001) "The Capacity of Low Density Parity Check Codes under Message Passing Decoding," *IEEE Trans. Inform. Theory*, vol. 45, pp .599-681.

Performance of Serial Concatenated Turbo Codes in High Density Multi-Track Magnetic Recording

C.J. Tjhai and M.Z. Ahmed

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

Abstract

Magnetic recording has been driven towards increasing bit and track densities. This paper presents a high density multi-track magnetic recording system disturbed by intertrack-interference (ITI) and additive white Gaussian noise (AWGN), concatenated in serial with high rate convolutional codes. An external Least-Mean-Squared (LMS) based adaptive ITI equaliser is implemented to mitigate the effect of ITI. Results show that the iterative decoding system with adaptive ITI equaliser can reduce SNR loss to approximately 0.1dB and 0.5dB for ITI of 10% and 20% respectively. Unlike Maximum-Likelihood decoding which has different burst error pattern depending on the precoder, iterative decoding results in low burst error regardless of the precoding employed. For future high density multi-track magnetic recording, iterative decoders have to be accompanied by ITI removal in order to achieve an acceptable performance.

Keywords

Equaliser, intertrack-interference, iterative decoding, multi-track recording, partial response, turbo codes

1. Introduction

Magnetic recording at the capacity of Terabit per square inch requires increased in both bit and track densities. While an increase in bit density will result in closer bit transitions, increasing track density will lead to greater interactions between adjacent tracks causing side reading of adjacent information (Wood 2000). Both will cause significant detection problems.

Since the introduction of turbo codes by Berrou, Glavieux and Thitimajshima (1993), iterative decoding technique has attracted interests in many areas including magnetic recording. The research of iterative decoding for magnetic recording was started in 1998 (Ryan, McPheters and McLaughlin 1998, McPheters, McLaughlin and Hirsch 1998) and it has been very active since then. So far, various concatenated systems have been proposed to exploit the benefits of iterative decoding. These systems consist of an interleaved precoded partial-response (PR) channel as the inner code and either parallel concatenated convolutional code (turbo code), convolutional code, product code or low density parity check code (LDPC) as the outer code.

In this paper, we consider the serial concatenation scheme proposed by Souvignier, Friedmann, Oberg, Siegel, Swanson and Wolf (1999) where the outer code is a punctured convolutional code. Instead of having PR channel disturbed by AWGN only, interference from adjacent tracks or intertrack-interference (ITI) is also considered. Only high-order PR channels are of interest in high-density recording and as such the investigations are focused on high-order PR channels.

The paper is organised as follows. Section 2 describes the recording and detection system simulation model. Section 3 describes the adaptive signal processing technique to mitigate ITI. Simulation results are presented in section 4 and section 5 concludes the paper.

2. Magnetic Recording System with ITI

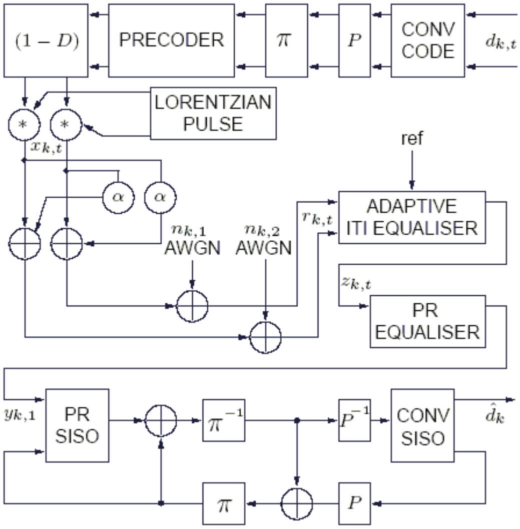


Figure 1 : Magnetic recording model with ITI and iterative decoder

The proposed recording and detection system model is shown on Fig. 1. Input data $d_{k,t}$ of 4096 bits are convolutionally encoded and punctured to achieve code rate of *k/k+1* before applied to a pseudo-random interleaver. The convolutional code is a simple rate 1/2 recursive systematic code with polynomial of [4 5/7]. The subscript *k* and *t* denotes time and track respectively and the label *P* and P^{-1} denote puncturing and depuncturing respectively. The interleaved codewords are fed to a precoded PR channel. Three modified extended EPR4 (MEEPR4) channels are considered as the high-order PR channels: $(3+2D-2D^2-2D^3-D^4)$ (Ghrayeb and Ryan 2001), $(5+4D-3D^2-4D^3-2D^4)$ (Sawaguchi, Kondou, Kobayashi and Mita 1998) and $(2+2D-D^2-2D^3-D^4)$[3] (Conway 1998). Unless otherwise stated, it is assumed that $1/(1\oplus D^2\oplus D^3\oplus D^4)$ precoder is used. The ideal Lorentzian readback samples

[3] For simplicity, the $(3+2D-2D^2-2D^3-D^4)$, $(5+4D-3D^2-4D^3-2D^4)$ and $(2+2D-D^2-2D^3-D^4)$ channels will be referred to as MEEPR4a, MEEPR4b and MEEPR4c channels respectively.

$x_{k,t}$ is independently disturbed by AWGN and ITI. The noisy readback samples on track 1 $r_{k,1}$ is given by:

$$r_{k,1} = x_{k,1} + \underbrace{\alpha x_{k,2}}_{ITI} + \underbrace{n_k}_{AWGN}$$

The amount of ITI is controlled by α and it is assumed that the ITI is linear and symmetrical. The PR equaliser, a 15-tap linear phase finite-impulse-response (FIR) filter, is implemented as a zero-forcing Lorentzian-to-Sinc equaliser with post-filtering to shape the sinc output to a target PR pulse. Ideal equalisation and perfect timing recovery are assumed and this two-track model is adapted from work of Ahmed, Davey, Donnelly and Clegg (2002). The signal-to-noise ratio (SNR) is defined by equation below where σ^2 is the variance of the AWGN:

$$\text{SNR} = 10 \log\left(\frac{1}{2\sigma^2}\right)$$

The equaliser output on track 1 $y_{k,1}$ is fed to an iterative decoder consisting of PR and convolutional codes soft-input-soft-output (SISO) modules. Each SISO module is a Maximum-a-Posteriori (MAP) decoder with Max-Log approximation (Max-Log-MAP decoder) (Robertson, Hoeher and Villebrun 1997). The two SISO modules iteratively exchange extrinsic information and in the final iteration, a hard-decision output d_k is generated which is then used in conjunction with $d_{k,t}$ for bit-error and sector-error measurements.

3. Adaptive ITI Equaliser

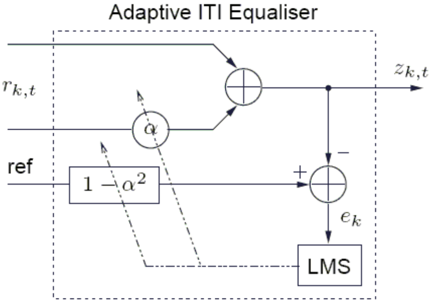


Figure 2 : ITI Equaliser based on LMS algorithm

An adaptive equaliser has been implemented using Least-Mean-Squared (LMS) algorithm to reduce the amount of ITI on the readback samples. It sits at front end of the receiver—right before the PR equaliser. Fig. 2 depicts the block diagram of the implemented adaptive ITI equaliser. The adaptive ITI equaliser predicts the amount of ITI from training sequence and it has been verified that training sequence of 1024 samples is more than enough for equaliser to converge to a considerably closed estimate, see Fig. 3. It is worth noting that ITI estimation is more difficult at high PW$_{50}$/T than at low PW$_{50}$/T. The convergence graph was obtained from simulation on MEEPR4c channel at PW$_{50}$/T of 3.50 for μ of 0.05. The parameter μ is the step size of the LMS algorithm and it is kept at 0.05 in this paper.

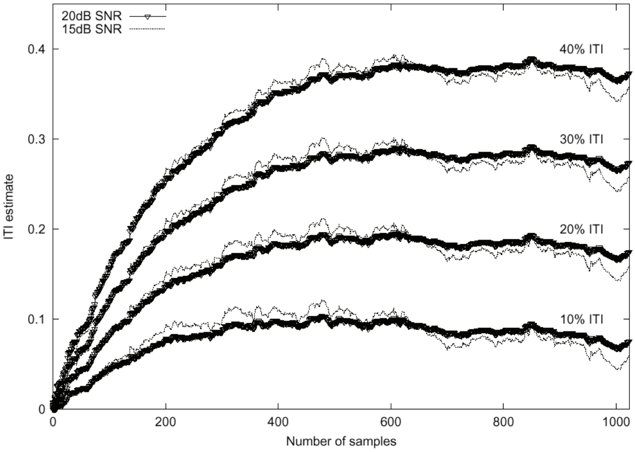


Figure 3 : Estimated ITI at SNR of 15dB and 20dB for various levels of ITI, μ=0.05.

The implemented adaptive ITI equaliser is adapted from the work of Ahmed et al. (2002) where the ITI equaliser is incorporated into trellis decoder resulting in increased decoding complexity. On the other hand, the ITI equaliser in this paper is implemented as an external filter and the decoder complexity is not affected at all. The ITI equaliser in Ahmed et al. (2002) does not require training sequence, instead it trains itself using data samples. This method is ideal at low PW_{50}/T, at high PW_{50}/T however, it is very difficult to distinguish bit transitions as such thresholding the readback samples to generate reference samples will not cause the equaliser to converge and this has been verified by the author.

4. Simulation Results

The performance of the adaptive ITI equaliser under Maximum-Likelihood (ML) decoding is shown on Fig. 4 and this performance is not as good as that of Ahmed et al. (2002). However, this performance is still acceptable considering the fact that it does not introduce further complexity to the decoder.

Fig. 5 shows that ITI can degrade the error performance significantly even when iterative decoder is used. The performance of iterative decoder with 5% of ITI can still be tolerated, but with ITI greater than 20% it is clearly unacceptable. In the presence of adaptive ITI equaliser, the performance difference between no ITI and 5% ITI under iterative decoding is negligible. With 20% ITI, there is improvement with adaptive ITI equaliser but it is not as much as in the former case. These are depicted on Fig. 6. The amount of SNR gain on one of the MEEPR4 channels for a fixed bit-error-rate (BER) can be determined from Fig. 7. At 10% ITI the implemented system can reduce the SNR loss from 0.5dB to 0.1dB with 8/9 outer codes and from 0.6dB to 0.2dB with 16/17 outer codes. At 20% ITI the SNR gain is much higher, rate 8/9 codes can reduce the SNR loss from 2.5dB to 0.5dB and rate 16/17 codes can reduce the SNR loss from 3.0dB to 0.6dB.

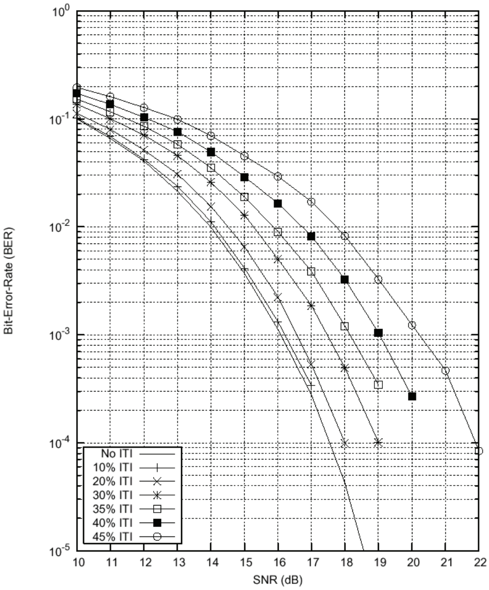


Figure 4 : The performance of ML decoder under the presence of adaptive ITI equaliser at various levels of ITI. (PR4 channel at PW_{50}/T of 2.0)

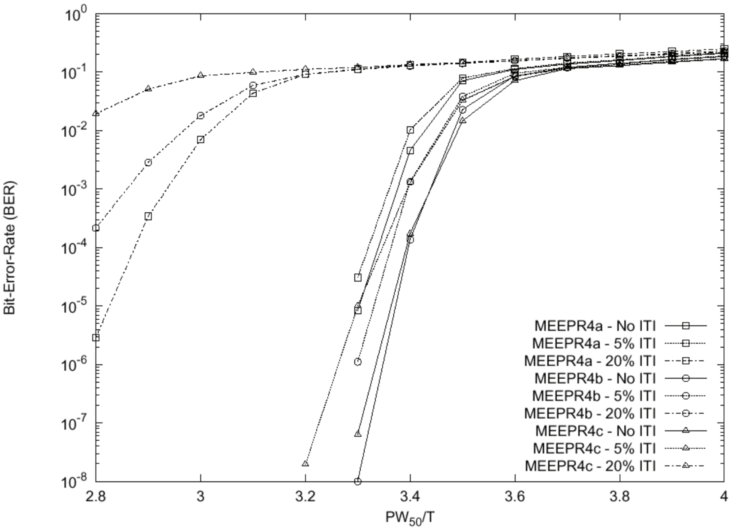


Figure 5 : Iterative decoding of rate 8/9 codes under various levels of ITI without adaptive ITI equaliser

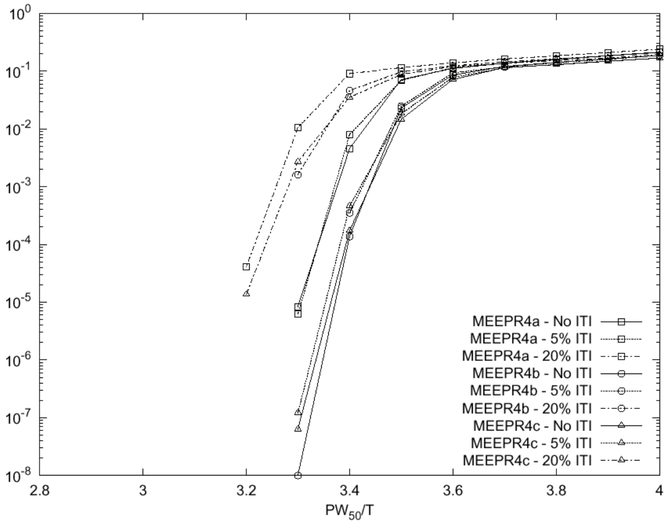
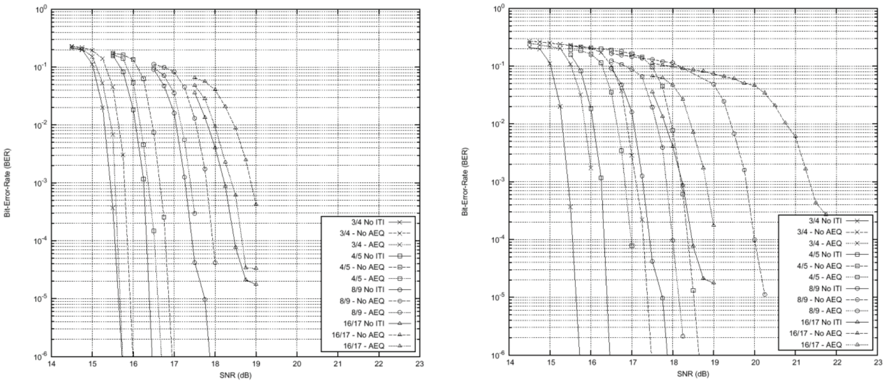


Figure 6 : Iterative decoding of rate 8/9 codes under various levels of ITI with adaptive ITI equaliser



(a) 10% ITI

(b) 20% ITI

Figure 7 : Iterative decoding and adaptive ITI equalisation on MEEPR4c channel with various code rates under 10% and 20% ITI. (PW$_{50}$/T is 3.50 and 10 decoding iterations)

Results from ML decoding on MEEPR4c channel with PW_{50}/T of 3.50, 5% ITI and 17dB SNR have indicated that various precoders can lead to different burst error pattern, see Fig. 8. Precoder of $1/(1 \oplus D \oplus D^3 \oplus D^4)$ for example, has very good burst error pattern. Precoder of $1/(1 \oplus D)$ has a similar burst error pattern, but its burst error of length 6 is quite significant. Low burst errors can be easily corrected with simple Reed-Solomon (RS) codes. Fig. 9 shows the results for the same precoders when iterative decoding technique is employed. With iterative decoding there is not much variation in the burst error profile on all of the precoders, most of the burst error is concentrated on burst of 2 or 3 errors. The normalised burst error histograms below were generated from the results obtained when either 50 sectors in error or 10000 simulation times was met.

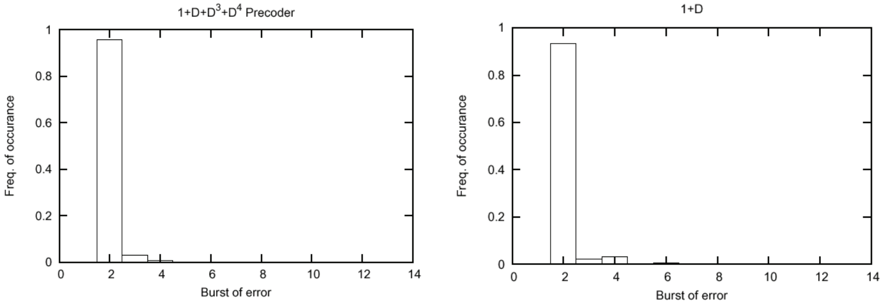


Figure 8 : Normalised burst error on MEEPR4c channel with ML decoding. (5% ITI and no ITI equalisation)

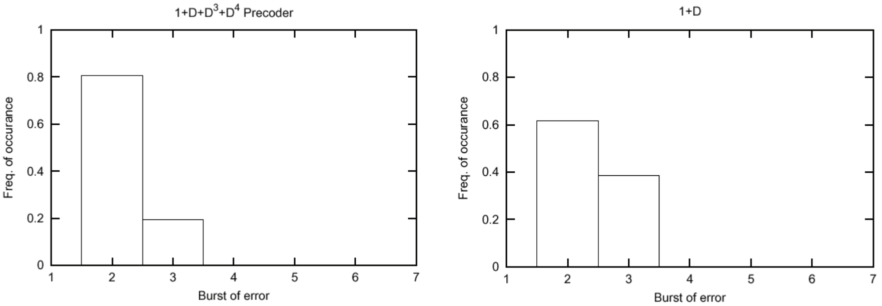


Figure 9 : Normalised burst error on MEEPR4c channel with iterative decoding. (5% ITI and no ITI equalisation)

5. Conclusions

A concatenated system consisting of highly punctured convolutional and precoded MEEPR4 channel with adaptive ITI equaliser is presented to investigate the applicability of turbo codes for future high density magnetic recording. The system considered is a two-track recording model with Lorentzian readback samples with PW_{50}/T greater than 3.00 and disturbed by ITI. Simulation results show that ITI causes significant degradation in BER even if iterative decoding is employed. Introducing an external adaptive ITI equaliser in front of the iterative decoder has improved the system performance. Results show that this system reduces the SNR loss to approximately 0.1dB and 0.5dB for ITI of 10% and 20% respectively. Under ML decoding different precoder produces different burst error pattern, whereas with iterative decoding all proceders seem to produce low burst error pattern. This can be understood as the effect of interleaving.

The performance of the adaptive ITI equaliser is not as good as the one in Ahmed et al. (2002) where it is integrated into trellis decoder. However, an external adaptive ITI equaliser does not increase the trellis, thus decoding complexity. In addition, the performance against the non-ITI-equalised iterative decoding system has been significant.

Turbo codes are clearly applicable for future high density recording. However, a solution with iterative decoding only is not sufficient in the presence of ITI. Advanced signal processing techniques are required to mitigate ITI.

6. References

Ahmed, M. Z., Davey, P. J., Donnelly, T. and Clegg, W. W. (2002), 'Track Squeeze Using Adaptive Interack Interference Equalization', *IEEE Transactions on Magnetics* 38(5), 2331–2333.

Berrou, C., Glavieux, A. and Thitimajshima, P. (1993), 'Near shannon limit error-correcting coding: Turbo codes', *IEEE International Conference on Communication, Geneva, Switzerland,* pp. 1064–1070.

Conway, T. (1998), 'A New Target Response with Parity Coding for High Density Magnetic Recording Channels', *IEEE Transactions on Magnetics* 34(4), 2382–2386.

Ghrayeb, A. and Ryan, W. E. (2001), 'Concatenated Code System Design for Storage Channels', *IEEE Journal on Selected Areas in Communication* 19(4), 709–718.

McPheters, L. L., McLaughlin, S. W. and Hirsch, E. C. (1998), 'Turbo codes for PR4 and EPR4 magnetic recording', *Asil Omar Conference Pacific Grove, CA On Signals Systems and Computers* pp. 1778–1782.

Robertson, P., Hoeher, P. and Villebrun, E. (1997), 'Optimal and Sub-Optimal Maximum A Posteriori Algorithms Suitable for Turbo Decoding', *European Transactions on Telecommun. (ETT)* 8(2), 119–125. Available from http://www-lns.tf.uni-kiel.de/ict/publikationen_hoeher.html, date visited: 10 February 2003.

Ryan, W. E., McPheters, L. L. and McLaughlin, S. W. (1998), 'Combined turbo coding and turbo equalization for PR4-equalized lorentzian channels', *Conference on Information Sciences and Systems* pp. 489–494.

Sawaguchi, H., Kondou, M., Kobayashi, N. and Mita, S. (1998), 'Concatenated Error Correction Coding for High Order PRML channels', *IEEE Globecom, Sydney,* pp. 2694–2699.

Souvignier, T., Friedmann, A., ¨ Oberg, M., Siegel, P. H., Swanson, R. E. and Wolf, J. K. (1999), 'Turbo decoding for PR4: Parallel versus serial concatenation', *IEEE International Conference on Communication* pp. 1638–1642.

Wood, R. (2000), 'The feasibility of magnetic recording at 1 Terabit per square inch', *IEEE Transactions on Magnetics* 36(1), 36–42.

M-Band Wavelet Modulation for Wireless Digital Communication

L. Tang, A. Muayyadi and M. Ali Abu-Rgheff

School of Computing, Communications and Electronics, Faculty of Technology, University
of Plymouth, United Kingdom
m.osa@plymouth.ac.uk

Abstract

In recent years, the construction of wavelet orthonormal bases has been studied extensively. Previous research works have shown that M-channel Perfect Reconstruction (PR) filter banks can be used for computing the discrete wavelet transform by iterating the filter bank on its lowpass output [6]. It has been observed that there is a close relationship between FIR PR filter banks and compactly supported wavelet bases in the 2-band cases (as well as the general M-band case). The PR filter banks can be exploited to construct the orthonormal M-band wavelet bases. In this paper, M-band wavelet-based modulation system simultaneously transmitting M symbols per modulation cycle is proposed. The system consists of a pair of synthesiser/analyser filter banks to generate Orthonormal Cosine-Modulated Wavelets. This novel modulation scheme offers improved message over conventional modulation techniques: For a single user, the serial data can be divided into M parallel streams and then transmitted and received through different orthogonal sub-channel at the same time. The transmission efficiency is found to have dramatically improved. Furthermore, the system performances are guaranteed by the biorthogonal transmitter and receiver. The BER performance of the proposed system is evaluated by simulation and then compared with 16QAM in AWGN and Rayleigh fading channels and the results are presented.

Keywords

CDMA, Wireless, Digital Communication.

1. Introduction

“Wavelets” applications in digital communications DSP are a relatively new but rapidly growing research area. As one potential example, the wavelet-based “fractal modulation” paradigm has been proposed as a novel diversity strategy for communication for a particular class over unreliable channels. Fractal modulation can be viewed as a special form of a class of techniques broadly referred to as multitone modulation. The advantage of this modulation scheme emerges from its multirate diversity strategy, as it allows simultaneous transmission of the data signal at multiple rates. Typical applications of the multitone transmission include the asymmetric digital subscriber line (ADSL), the high-speed digital subscriber lines (HDSL) and the hybrid fiber-coax (HFC) networks [1].

The original Wavelet systems have a scale multiplier, M = 2, to give two-band division at each stage, (the concept of octave), with logarithmic frequency bandwidth. The Mallat fast algorithm could work through the binary tree to decode the data. The main disadvantage of the 2-band wavelet is that it cannot deal with high frequency signals with a relatively narrow bandwidth. Therefore, this 2-band wavelet system was proposed to be extended into a general

M-band system to meet some specific requirements that utilise a uniform bandwidth division rather than a logarithmic bandwidth division.

In this paper, we propose the M-band wavelet modulation *(Figure 2)* that simultaneously sends M streams of data at the same rate through a physical channel. The transmitter and receiver are implemented by a pair of a Wavelet Synthesiser / Analyser filters that are in turn represented by orthonormal M-channel PR filter banks. [2] states that when the number of channels is equal to the downsampling factor M, the wavelet filter has the PR property .

The paper is organised as follows: section 2 introduces the wavelet modulation system and presents the wavelet matched filter detection system. Section 3 presents the mathematical modelling of the wavelet modulation and transmission channels. Analytical and simulation of the BER performance of the proposed wavelet system and an equivalent QAM system, being the significant contributions of the work, are presented in section 4. Finally, the results are discussed in the concluding section 5.

2. Wavelet Modulation System

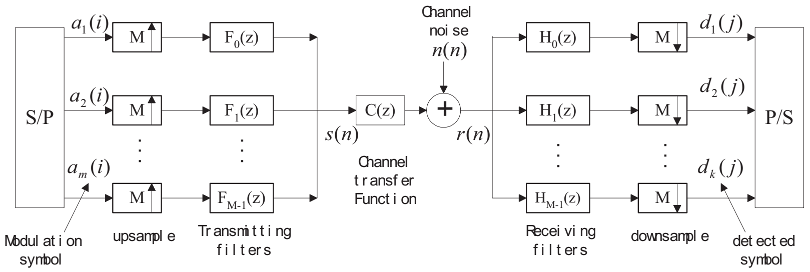


Figure 4 : M-band Wavelet Modulation

The basic M-band wavelet modulation can be described in a transmultiplexer structure as shown in Figure 1. The analyzer filter, $F(z)$, and the synthesizer filter, $H(z)$, are biorthognal filter banks that can be implemented for Perfect-Reconstruction (PR) transmission (no error at the output). Here they act as the transmitter and receiver respectively.

The Wavelet Filters, in other words, the M-channel PR filter banks $H(z)$ and $F(z)$, are also orthogonal filters, which are given by:

$$h_k(n) = 2h(n)\cos\left((2k+1)\frac{\pi}{2M}\left(n-\frac{N-1}{2}\right)+\theta_k \right)$$

$$f_k(n) = 2h(n)\cos\left((2k+1)\frac{\pi}{2M}\left(n-\frac{N-1}{2}\right)-\theta_k \right), 0 \le n \le N-1,\ 0 \le k \le M-1$$

Where *h (n)* is the prototype filter and N is the filter length [4]. Their frequency response can be obtained by shifting the response of an ideal prototype lowpass filter. The frequency shifting is achieved by cosine modulation. The above form $\theta_k = (-1)^k \frac{\pi}{4}$ is used to cancel the first aliasing error from the adjacent subbands. The other aliasing errors are then minimised by the stopband cutoff frequency $\omega_c << \frac{\pi}{M}$ such that there is no overlap between $H_k(z)$ and $H_{k\pm2}(z), F_k(z)$ and $F_{k\pm2}(z)$ [4]. In our simulation system, 8-band wavelet filter banks will be employed and we focus on the performance in every other consecutive channel, that is channel 1,3,5,7 to reduce the inter-symbol interference (ISI) and inter-channel interference (ICI).

The transmitted signal is generated as follows. The data symbol stream is converted from serial to parallel M branches. In the k th branch, the symbol is interpolated (upsampled) by a factor M and then filtered using the synthesis filter $F_k(z)$.

After being transmitted through the channel, the received signal is then filtered using the analysis filters $H_k(z)$, decimated (downsampled) by M factor, and converted from parallel to serial to obtain the original data signal. The filter $H_k(z)$ is exactly matched with the filter $F_k(z)$ so that this process constitutes matched filter detection in an AWGN channel.

3. System Modelling

3.1 Wavelet System Analysis

The serial parallel data symbol conversion gives the following equation:

$$a_m(i) = a(iM + m) \tag{1}$$

where $a_m(i)$ is the symbol at mth branch and $a(i)$ is the incoming data symbols.

The transmitted signal can be written as:

$$s(n) = \sum_{m=0}^{M-1} \sum_{i=-\infty}^{\infty} a_m(i) \, f_m(n - iM) \tag{2}$$

where $f_m(n)$ represents the transmitter synthesis filter of the mth branch.

Substituting equation (1) into (2)

$$s(n) = \sum_{m=0}^{M-1} \sum_{i=-\infty}^{\infty} a(iM + m) \, f_m(n - iM) \tag{3}$$

Considering the transmission through linear time-invariant (LTI) channel $c(l)$ with lT_b delay, the received signal is:

$$r(n) = \sum_{l=0}^{L-1} c(l) \, s(n - l) + n(n) \tag{4}$$

where $n(n)$ is additive Gaussian noise with a two-sided power spectral density of No/2.

Substituting equation (3) into (4)

$$r(n)=\sum_{l=0}^{L-1}c(l)\sum_{m=0}^{M-1}\sum_{i=-\infty}^{\infty}a(iM+m)\,f_m(n-iM-l)+n(n) \tag{5}$$

Finally the output signal can be written as

$$d(n)=\sum_{k=0}^{M-1}\sum_{j=-\infty}^{\infty}r(jM+k)\,h_k(n-jM) \tag{6}$$

where $\mathrm{h_m(n)}$ represents the receiver analysis filter of the m^th branch.

Substituting equation (5) into (6)

$$d(n)=\sum_{l=0}^{L-1}c(l)\sum_{k,m=0}^{M-1}\sum_{j,i=-\infty}^{\infty}a(iM+m)\,f_m(jM+k-iM-l)\,h_k(n-jM)+\sum_{k=0}^{M-1}\sum_{j=-\infty}^{\infty}n(jM+k)\,h_k(n-jM)$$

(7)

We can write the non-noise part in other form

$$d_s(n)=\sum_{l=0}^{L-1}c(l)\sum_{k,m=0}^{M-1}\sum_{j,i=-\infty}^{\infty}a(iM+m)\,I_{k,m,j,i} \tag{8}$$

where $I_{k,m,j,i}=\sum_{k,m=0}^{M-1}\sum_{j,i=-\infty}^{\infty}f_m(jM+k-iM-l)\,h_k(n-jM)$ (9)

ICI happens if $I_{k,m,j,i}\neq 0$ for $k\neq m$

ISI occurs when $I_{k,m,j,i}\neq 0$ for $j\neq i$

The total noise power in the noise term in equation (7) does not change (ideally), therefore the bit error rate can be written as

$$P_e=Q\left(\frac{E_b}{ICI+ISI+\frac{No}{2}}\right) \tag{10}$$

3.2 Gaussian Channel

In the AWGN channel, zero-mean White Gaussian noise is added to the transmitted signal *s(t)*, so that the received signal r(t) can be represented as

$$r(t)=s(t)+n(t)$$

Where *n(t)* is a zero-mean White Gaussian noise process with power $\frac{N_0}{2}$.

3.3 Slow Fading Channel

Rayleigh fading involves two independent characteristics: time spread introduced in the signal that is transmitted through the channel and the time variation behaviour of the channel [8]. In our modulation systems, we employed a Doppler shift for cellular communications, assuming the vehicle speed is 60*km/hr*, the carrier frequency f_c is 1800Hz, then the resulting Doppler shift is 100Hz.

In a slow fading channel the symbol duration of the signal is much smaller than the coherence time of the channel $T_s << T_c$ [8]. In our simulation systems, the symbol period is given by $T_s = \frac{1}{R_s} = \frac{1}{10^4} = 0.1ms$. This value is much smaller than the coherent time given by $T_c = \frac{0.423}{f_d} = 4.23ms$ [7]. In our modulation trials, we use a new improved Jakes' Model [5] to simulate the Slow Fading channel. If we consider a frequency-nonselective fading channel comprised of the propagation paths, the normalized low-pass fading process of a new statistical sum-of-sinusoids simulation model is defined by:

$$X(t) = X_c(t) + jX_s(t)$$

$$X_c(t) = \frac{2}{\sqrt{N}}\sum_{n=0}^{M} c_n \cos(\omega_n t + \phi_n) \qquad X_s(t) = \frac{2}{\sqrt{N}}\sum_{n=0}^{M} s_n \cos(\omega_n t + \phi_n)$$

$$c_n = \begin{cases} \sqrt{2}\cos\varphi_0\,, & n = 0 \\ 2\cos\varphi_n\,, & n = 1,2,\ldots,M \end{cases} \qquad s_n = \begin{cases} \sqrt{2}\sin\varphi_0\,, & n = 0 \\ 2\cos\varphi_n\,, & n = 1,2,\ldots,M \end{cases}$$

$$\varphi_n = \begin{cases} \pi/4, & n = 0 \\ \pi n/M, & n = 1,2,\ldots,M \end{cases} \qquad \omega_n = \begin{cases} \omega_d\,, & n = 0 \\ \omega_d \cos\frac{2\pi n}{N}, & n = 1,2,\ldots,M \end{cases}$$

Where φ_n and ϕ_n are statistically independent and uniformly distributed over $[-\pi,\pi)$ for all n. The Doppler shift ω_d is given by: $\omega_d = 2\pi f_d$. The coefficient $X(t)$ is a complex Gaussian random variable and the fading envelope is the absolute value of $X(t)$, that is $|X(t)| = \sqrt{X_c(t)^2 + jX_s(t)^2}$ [5]. $|X(t)|$ is a Rayleigh distributed random variable. The received signal is given by:

$$r(t) = s(t)|X(t)| + n(t)$$

As in the AWGN channel, $s(t)$ is the transmitted signal and $n(t)$ represents Gaussian noise.

4. Results

4.1 Comparison of the Daubechies, Root raised cosine filter and the PR filter

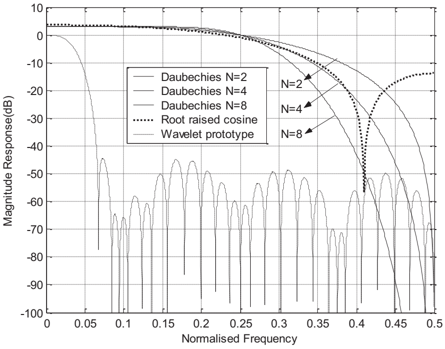


Figure 2 : Frequency responses of the Daubechies Wavelets, Root-raised cosine filter and the PR prototype filter

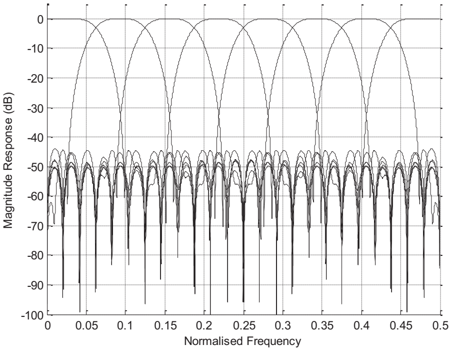
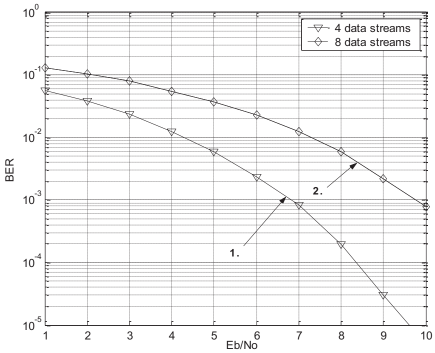


Figure 3 : Frequency response of 8-channel PR cosine-modulated filter bank

Figure 2 compares the frequency responses of the Daubechies N=2, N=4, N=8 wavelet filters, the Root raised cosine filter and the 8-band PR prototype filter. Figure 3 shows the frequency response of 8-channel PR cosine-modulated filter bank employed in our simulation system.

4.2 The performances of the two transmission schemes

For the wavelet modulation, we proposed two transmission schemes by using 8-channel PR filter bank, scheme 1 divides an input serial data into 4 streams and transmits these streams in every other consecutive channel. In this case, this means the signal data only transmit in 1, 3, 5, 7 channels, scheme 2 separates one serial data into 8 streams and transmits them in each channel (1~8 channel). Due to the biorthogonal filter bank characteristics, the whole system performs the perfect symbol recovery without the use of any channel. However, this is not true if the transmission passes through a noise channel. Figure 4 compares the BER performances Vs E_b / N_0 of the two transmission schemes. It is evident that the system performance of scheme 1 is much better than that of the scheme 2. This could be caused by the inter symbol interference and inter channel interference. As we can see in Figure 3 there are large overlapping areas between the adjacent channels and very little overlapping areas in every other channels. Thus in the following simulation, we only focus on scheme 1, and we refer to it as 4-band wavelet modulation (WM).



1. data transmitted in every other channel 2. data transmitted in each channel

Figure 4 : BER vs. E_b / N_0 performance comparison between the two transmission schemes in an AWGN channel

4.3 Comparison of 4-band WM & 16-QAM in a Gaussian channel

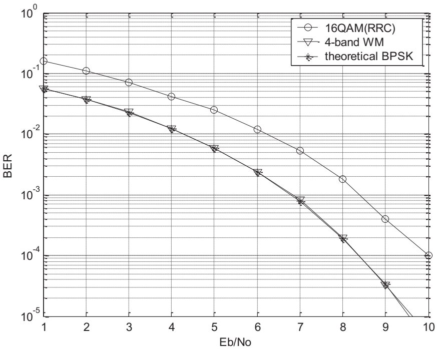


Figure 5 : BER vs. E_b / N_0 of 4-band WM and 16-QAM in an AWGN channel

In Figure 5 the comparison between the 4-band wavelet modulation and the 16-QAM are presented in terms of BER probability versus the ratio E_b / N_0 expressed in dB. We assume that the interference on the channel is caused solely by AWGN (additive white Gaussian noise). In this experiment, the 16QAM is simulated as the reference equivalent system to compare with the 4-band wavelet modulation. We use a pair of Root raised-cosine filters, before and after, the transmission to form the matched filter detection. The 4-band WM has a much better performance than 16QAM and the BER curves are almost identical to the theoretical BPSK in AWGN channel. The average improving gain is around 2dB, which means for a given error probability P_e, e.g. $P_e = 10^{-3}$, the signal-to-noise ratio required 8.5dB for 16QAM system while only 7dB for the 4-band WM system.

4.4 Comparison of 4-band WM & 16-QAM in a slow fading channel

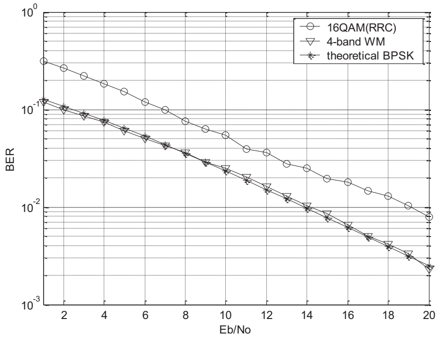


Figure 6 : BER vs. E_b / N_0 of 4-band WM and 16-QAM in a slow fading channel

Figure 6 compares the BER curves between the 4-band WM & the 16-QAM in a slow fading channel with a Doppler shift of 100Hz. However, the wavelet modulation still displays a prominent improvement over the 16QAM and the improving gain is varied from 3dB to 5dB. Also, we can see that the 4-band WM performance matches the theoretical BPSK in this slow fading channel.

5. Conclusion

In this paper, a new M band wavelet modulation has been proposed for wireless digital communication. The 4-band WM employs every other channel of the 8-channel PR synthesier and analyser filter banks as the transmitter and receiver respectively. There is quite obvious difference between the signals transmitting in every other channel and that in every channel. This is due to the large overlapping areas in the adjacent channels, and the biothogonormal PR filter bank can only perform the perfect symbol recovery in a channel free system.

However, we can improve our whole wavelet system by using the equaliser techniques, this can be dealt with in our future work.

Another main contribution of our work is that we found that M-band WM performances are equivalent to BPSK in an AWGN channel, as well as in a slow fading channel. The results are quite similar with Manish J.Manglani and Amy E. Bell's [3] investigation on the performances of the Fractal wavelet modulation.

6. References

[1] G.W. Wornell and A.V. Oppenheim. "Wavelet-Based Representations for a Class of Self-Similar Signals with Application to Fractal Modulation," IEEE Transactions on Information Theory, Vol. 38, No. 2, March 1992.

[2] Peter Steffen, Peter N. Heller, Ramesh A. Gopinath, and C. Sidney Burrus, "Theory of regular M-band wavelet bases," IEEE Trans. SP, vol. 41, no. 12, pp. 3497–3510, Dec. 1993.

[3] M. J. Manglani and A. E. Bell, Wavelet modulation performance in gaussian and rayleigh fading channels, Proceedings of MILCOM, 2001.

[4] M. Vetterli and C. Herley. "Wavelets and filter banks," IEEE Trans. Signal Processing, vol. 40, pp. 2207--2232, 1992.

[5] Yahong Rosa Zheng and Chengshan Xiao, "Simulation Models With Correct Statistical Properties for Rayleigh Fading Channels," IEEE Transactions on Communications, VOL. 51, NO. 6, June 2003.

[6] Truong Q.Nguyen, "A Tutorial on Filter Banks and Wavelets", International Conference on Digital Signal Processing, Cypress, June 1995.

[7] Bernard Sklar. Rayleigh fading channels in mobile digital communication systems. IEEE Communications Magazine, 35(7):90.100, 1997.

[8] Proakis, J. G. *Digital communications*, McGraw Hill 1995.

Author Index

Distributor:

School of Computing, Communications & Electronics
University of Plymouth
Drake Circus
Pymouth
PL4 8AA
United Kingdom

ISBN 978-1-84102-118-8
90000
9 781841 021188