

ITB Journal



Issue Number 9, May 2004

Contents

<i>Editorial</i>	3
<i>Table Of Contents</i>	4

**Special Edition for
ITB Research Conference 2004
Conference papers
22/23 April 2004**

The conference organisers are:

**Dr. Matt Smith
Dr. Brian Nolan
Ms. Mairead Murphy
Mr. Stephen Sheridan
Mr. Hugh McCabe**

The academic journal of the Institute of Technology Blanchardstown



Views expressed in articles are the writers only and do not necessarily represent those of the
ITB Journal Editorial Board.

ITB Journal reserves the right to edit manuscripts, as it deems necessary.

All articles are copyright © individual authors 2004.

Papers for submission to the next ITB Journal should be sent to the editor at the address below. Alternatively, papers can be submitted in MS-Word format via email to brian.nolan@itb.ie

*Dr. Brian Nolan
Editor
ITB Journal
Institute of Technology Blanchardstown
Blanchardstown Road North
Blanchardstown
Dublin 15*

This edition of the ITB Journal is sponsored by Macromedia Ireland



Editorial

I am delighted to introduce the ninth edition of the ITB Journal, the academic journal of the Institute of Technology Blanchardstown. The aim and purpose of the journal is to provide a forum whereby the members of ITB, visitors and guest contributors from other third level colleges can publish an article on their research in a multidisciplinary journal. The hope is that by offering the chance to bring their work out of their specialised area into a wider forum, they will share their work with the broader community at ITB and other academic institutions.

This issue is dedicated to the refereed academic papers presented at the ITB Research Conference in April 2004. The ITB Research Conference is a peer-reviewed, annual, multidisciplinary event intended as an opportunity for researchers from ITB and other third-level colleges to present their work in front of an informed audience. The Institute of Technology Blanchardstown (ITB) has now entered its fifth year of existence. As well as putting in place a wide range of innovating undergraduate programmes, the Institute has begun to establish a significant research profile. We currently have a number of funded research projects underway, and a growing cohort of postgraduate research students. These projects are being carried out in a range of research areas across computing, engineering, business and the humanities. Although there are many presentations from external researchers and research students, the focus of the conference is on dissemination of the range of research activity taking place at ITB. Details about the conference are available on the ITB web site at the following location: www.itb.ie/itb_conference_2004/. The conference organisers are:

Dr. Matt Smith
Dr. Brian Nolan
Ms. Mairead Murphy
Mr Stephen Sheridan
Mr Hugh McCabe

Once again, we hope that you enjoy the papers in this “chunky” issue of the ITB Journal devoted to the ITB Research Conference papers.

Dr. Brian Nolan
Editor
ITB Journal
Institute of Technology Blanchardstown
Blanchardstown Road North
Blanchardstown
Dublin 15

Table of Contents

<i>Developing a Distributed Java-based Speech Recognition Engine.....</i>	<i>6</i>
Mr. Tony Ayres, Dr. Brian Nolan.....	6
<i>Design of a Wear Test Machine for Diamond Saw Segment.....</i>	<i>19</i>
D. Nulty J. Dwan and Y. Blake.....	19
<i>An Adaptive eLearning framework – Design Issues and Considerations.....</i>	<i>28</i>
Marie Brennan, Larry McNutt.....	28
<i>Consistency of Academic Performance in Higher Education: A Study of an Irish Business Degree Programme.....</i>	<i>43</i>
Julie Byrne 1 and Conor Heagney 2.....	43
<i>The Value of Enterprise: Overcoming Cultural Barriers in Northern Ireland.....</i>	<i>54</i>
Julie Byrne 1 and Sharon McGreevy 2.....	54
<i>Speech Synthesis for PDA.....</i>	<i>63</i>
Peter Cahill and Fredrick Mtenzi.....	63
<i>E-Business - Making the Move from Traditional to Next Generation Marketing Strategies: An Exploratory Study of Irish Organisations.....</i>	<i>75</i>
Ethel Claffey.....	75
<i>An Evaluation of On-Screen Keyboards for First Time Single Switch Users.....</i>	<i>86</i>
Paul Ahern and Rupert Westrup.....	86
<i>Modelling a Mechatronic System using “Matlab/Simulink” and “Dyanst”.....</i>	<i>98</i>
Paul Dillon.....	98
<i>Profiling the International New Venture - A literature review of the empirical evidence</i>	<i>112</i>
Natasha Evers.....	112
<i>Justification of Investment in IT systems.....</i>	<i>128</i>
Aidan Farrell.....	128
<i>Architecture and development methodology for Location Based Services.....</i>	<i>145</i>
Aaron Hand1, Dr. John Cardiff2.....	145
<i>Camera Control through Cinematography in 3D Computer Games.....</i>	<i>160</i>
James Kneafsey & Hugh McCabe.....	160
<i>Novel Design of a Variable Speed Constant Frequency Wind Turbine Power Converter</i>	<i>170</i>
Aodhán MacAler1 & Joe Dunk2.....	170
<i>Strengthening the Practices of an Agile Methodology?.....</i>	<i>180</i>
Jimmy Doody1, Amanda O’Farrell2.....	180
<i>A Qualitative Method for Determining the Quality of BGA Solder Joints in a Lead-Free Process.....</i>	<i>188</i>
Shane O’Neill1, John Donovan1 & Claire Ryan2.....	188

<i>Application of the Hough Transform to Aid Raised Pavement Marker Detection on Marked Roadways.....</i>	<i>197</i>
Colin O'Rourke ¹ , Catherine Deegan ¹ , Simon McLoughlin ¹ & Charles Markham ² ..	197
<i>Investigation Into The Correct Statistical Distribution For Oxide Breakdown Versus The Oxide Thickness Used In Integrated Circuit Manufacture.....</i>	<i>205</i>
James Prendergast ¹ , Eoin O'Driscoll ¹ , Ed Mullen ²	205
<i>Emotion Authentication: A Method for Voice Integrity Checking.....</i>	<i>213</i>
C. Reynolds ¹ , L Vasiu ² and M. Smith ³	213
<i>Neural Networks for Real-time Pathfinding in Computer Games.....</i>	<i>223</i>
Ross Graham, Hugh McCabe & Stephen Sheridan.....	223
<i>A New Integrated Style to Teaching Engineering Mathematics at Third Level Engineering Courses.....</i>	<i>231</i>
Mohamad Saleh ¹ B.Sc. M.Eng., Ph.D., CEng, MIEE.....	231
Colm McGuinness ² B.Sc., Ph.D., CMath, MIMA.....	231
<i>Design Study of a Heavy Duty Load Cell Using Finite Element Analysis: A practical Introduction to Mechatronic Design Process.....</i>	<i>239</i>
Mohamad Saleh B.Sc. M.Eng., Ph.D., CEng, MIEE.....	239
<i>Measurement of the Frequency Response of Clinical Gas Analysers.....</i>	<i>247</i>
Kabita Shakya ¹ , Catherine Deegan ¹ ,	247
Fran Hegarty ² , Charles Markham ³	247
<i>Smart Growth and the Irish Land-use Stakeholder: From Rhetoric to Reality.....</i>	<i>258</i>
Dorothy Stewart,.....	258
<i>Soft, Vertical Handover of Streamed Multimedia in a 4G Network.....</i>	<i>270</i>
Ger Cunningham, Philip Perry and Liam Murphy.....	270
<i>Questions of Ethical Responsibility in the Research of Unaccompanied Minors.....</i>	<i>276</i>
Oonagh Charleton & Dr. Celesta McCann James.....	276
<i>Web Enabled Embedded Devices.....</i>	<i>286</i>
Brian Myler and Dr. Anthony Keane.....	286
<i>Developing Real-Time Multimedia Conferencing Services Using Java and SIP.....</i>	<i>293</i>
Gavin Byrne and Declan Barber.....	293
<i>Convergence Technologies for Sensor Systems in the Next Generation Networks.....</i>	<i>302</i>
Conor Gildea and Declan Barber.....	302
<i>Implementing Test Patterns to Dynamically Assess Internet Response for Potential VoIP Sessions between SIP Peers.....</i>	<i>313</i>
Declan Barber, Gavin Byrne & Conor Gildea.....	313
<i>Auto Generation of XLIFF Translation Documents from Proprietary File Formats.....</i>	<i>322</i>
Kieran O'Connor & Geraldine Gray.....	322

Developing a Distributed Java-based Speech Recognition Engine

Mr. Tony Ayres, Dr. Brian Nolan

Institute of Technology Blanchardstown, Dublin, Ireland

tony.ayres@itb.ie

brian.nolan@itb.ie

Abstract

The development of speech recognition engines has traditionally been the territory of low-level development languages such as C. Until recently Java may not have been considered a candidate language for the development of such a speech engine, due to its security restrictions which limited its sound processing features. The release of the Java Sound API as part of the Java Media Framework and the subsequent integration of the Sound API into the standard Java development kit provides the necessary sound processing tools to Java to perform speech recognition.

This paper documents our development of a speech recognition engine using the Java programming language. We discuss the theory of speech recognition engines using stochastic techniques such as Hidden Markov Models that we employ in our Java based implementation of speech signal processing algorithms like Fast Fourier Transform and Mel Frequency Cepstral Coefficients.

Furthermore we describe our design goal and implementation of a distributed speech engine component which provides a client server approach to speech recognition. The distributed architecture allows us to deliver speech recognition technology and applications to a range of low powered devices such as PDAs and mobile phones which otherwise may not have the requisite computing power onboard to perform speech recognition .

1. Introduction

In the past speech recognition engines have been developed using low level programming languages such as C or C++. Early versions of the Java programming language would not have been candidate languages for the development of such speech systems. The sandbox security model employed by the Java platform prevents rogue Java code from damaging system hardware but it also limits access to the hardware needed to perform speech recognition, namely the sound card. This problem could be overcome by using native code (i.e. C, C++, etc.) to implement functionality that Java could not and then use the Java Native Interface to link the Java and native code together. With this solution the platform independence that Java brings to software development is lost and the complexity is increased. When Sun released the Java Sound API [1] and the Java Media Framework [2] (JMF), these extensions to the Java platform allowed developers to create applications that took advantage of not only sound processing but also advanced video and 3D functionality also, moreover these extensions provide this added functionality without compromising the Java security model. With these APIs Java is not only a capable development platform for building speech recognition engines, but an engine developed in Java inherits its benefits of platform independence and object oriented development.

The proliferation of the Java platform from the desktop computer to handheld devices and set-top boxes provides part of the motivation for developing a Java based speech recognition engine, given that Java 2 Micro Edition is present on over 100 million mobile phones world wide, speech recognition applications could be delivered to a large number of people users. Other motivating factors include the project from which this development has stemmed, which is entitled Voice Activated Command and Control with Speech Recognition over WiFi. In this project we are developing speech recognition software for the command and control of a remote robotic device. Our requirements for this project included a Linux based distributed speech recognition engine and a speech recognition engine for an iPaq Pocket PC PDA. Java's proficiency for operating seamlessly across multiple platforms coupled with its networking capabilities through RMI and sockets made it an obvious choice for developing our software.

The remainder of this paper describes our approach to developing speech software in Java. We explain the theory behind speech recognition and map the theory to a Java based implementation and code. We also describe our design of a speech distributed recognition engine.

2. Speech Recognition and Java

As has been detailed already, pure Java based speech recognition software would not have been feasible prior to the release of the Java Sound and Media APIs. Despite this, speech processing capabilities have been available to Java based applications since the release of the Java Speech API 1.0 [4] in 1998. The API offers speech recognition and synthesis functionality leveraged through a speech engine provided on the host operating system. For example, on the Windows platform the Java Speech API can take advantage of the recognition and synthesis capabilities provided by the Microsoft SAPI [5] engine. Other engines supported by JSAPI include IBM Via Voice [6], Dragon Naturally Speaking [7] and Phillips Speech SDK 2.0 [8]. JSAPI provides no speech processing functionality of its own, therefore it is not a solution to the question of Java based speech processing tools.

The Sphinx [9] project taking place at Carnegie Mellon University, offers a large scale, set of speaker independent speech recognition libraries which can be used in various application scenarios. Sphinx II, is designed for fast real time speech applications and Sphinx III offers a slower more accurate speech engine. Both of these engines are written in the C programming language. Speech recognition continues to be an area of much research in the Java community, Sphinx 4 is being developed in Java, with support from Sun Microsystems. The source code for Sphinx project is open source, much of the implementation of our speech system is based on

this code and related Java code from an open source project entitled OCVolume [10]. OCVolume is a small Java based speaker dependent speech recognition engine which is also based on the front end of Sphinx III.

3. The Speech Recognition Process

Automatic Speech Recognition (ASR) is the ability of computers to recognise human speech. There are two distinct types of speech recognition engine namely, continuous and non-continuous. Continuous speech recognition engines allow the user to speak in a normal conversation style, in contrast non-continuous engines require speech to be input in a more constrained manner, for example some require longer than average pauses between words. Both continuous and non-continuous recognition engines can be classified as either speaker dependent or speaker independent.

Speaker dependent engines require a user to train a profile of their voice before the engine can recognise their voice; speaker dependent systems offer the greatest accuracy as the engine is tailored to the characteristics of the users voice. Speaker independent engines do not require any training, while a speaker dependent engine has a profile of the users voice, speaker independent systems use databases of recorded speech. The databases contain voice samples of many different users, words, phrases and pronunciations.

Hidden Markov Models to date represent the most effective and accurate method of performing speech recognition [11]. A Hidden Markov Model is specified by the states Q , the set of transition probabilities A , defined start and end states and a set of observation likelihood's B [12]. Section 3.1 describes Markov models in greater detail.

Figure 1 diagrams the speech recognition process from the speech input at the microphone to the recognised speech output.

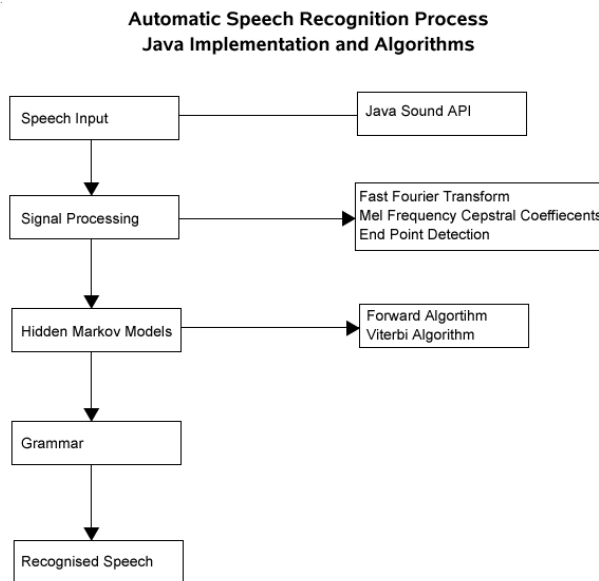


Figure 1 The Speech Recognition Process

3.1 Speech Input

Speech input is taken from a microphone attached to the system sound card, the sound card handles the conversion of the analogue speech signal into digital format. Depending on the recognition software type i.e. Continuous or non-continuous the engine may need to listen continuously for sound input, in this case a continuous audio stream will need to be opened and read from (Section 3 documents our implementation of this procedure using the Java Sound API).

3.2 Signal Processing

Speech input is taken from a microphone attached to the system sound card, the sound card handles the conversion of the analogue speech signal into digital format. At this point we have a digitised version of the speech signal. Speaking comes naturally to people, when we are spoken to we hear individual words, sentences and pauses in the speech, more so our understanding of language allows to interpret what was said. Consider what happens when we hear people speaking in a language which is foreign to us, we don't hear the individual words in the language and the speech sounds like one continuous stream of noise. The same scenario is true when we speak to computers for the purposes of speech recognition. The process of finding word boundaries is called segmentation.

Applying this knowledge to our digitised signal data, we need to process this signal in order to determine the word boundaries in the signal and also to extract relevant data which we use to

model the signal and eventually perform recognition through Hidden Markov Models [section 3.3]. We use the Fast Fourier Transform algorithm, an implementation of the discrete Fourier transform, to quickly calculate the Fourier equation in real time.

3.3 Hidden Markov Models

Hidden Markov Models (HMM) have proven to date to be the most accurate means of decoding a speech signal for recognition. HMM are stochastic in nature, that is, they generate a probability that an outcome will occur. In our speech system the input to the HMM will be a speech signal sampled at a particular moment in time, the output from the Markov Model will be a probability that the inputted signal is a good match for a particular phoneme. We can create numerous Hidden Markov Models to model speech signal input at time samples of the signal, each Markov Model can represent a particular phoneme. By combining the probabilities from each Markov model we can produce a probabilistic output that a given speech sequence is a representation of a particular word.

The fundamental concept in a HMM is the Markov assumption, this concept assumes that the state of the model depends only on the previous states. Given this concept, we can introduce the Markov process, a process which moves from state to state depending only on the previous n states. [11]

$$\text{HMM} = (\pi, A, B)$$

π = the vector of initial state probabilities

A = the state transition matrix

B = the confusion matrix

Figure 2 Definition of a Hidden Markov Model

Figure 2 shows a formal definition of a Hidden Markov Model, π represents the initial set of states. A represents the transition matrix or the probability of transiting from one state to another. B represents the confusion matrix or the set of observation likelihood's, which represent the probability of an observation being generated from a given state. [11].

Having modelled the speech input utterances using Hidden Markov Models we need to determine which observation sequence produces the best probability, we use the Viterbi algorithm to do this. The goal of the Viterbi algorithm is to find the best state sequence q given the set of observed phones o .

3.4 Grammar

Given a speech corpus which consists of many thousands of speech samples, during the recognition process the hidden markov model may have to search the entire corpus before finding a match. This searching problem is known as perplexity and can cause the speech engine to operate inefficiently, which translate into delays in recognition for the user. Grammars can be used to reduce perplexity. Grouping Markov Models of phonemes together we can form words. For example, a grammar can specify a particular word order, such that if we can recognise word A and we know that word B never follows word A we can eliminate searching the corpus for word B.

3.5 Recognised Speech

The output from these steps of the process is a String of text which represents the spoken input, at this point the recognised speech can be applied to some application domain such as in a dictation systems or, in our case, a command and control scenario.

4. Implementation

The implementation of our speech system is a speaker dependent system which requires the words to be recognised be recorded or sampled by the user prior to the use of the system.

The system accomplishes its goal by use of the process described in section 3, with one additional component. Given the need to distribute the speech process we developed a Java class to handle network communications between the client and the server. The client application is responsible for signal capture, while the server handles the process of recognition.

4.1 Signal Acquisition with the Java Sound API

Acquisition of the speech signal is achieved through a microphone using the Java Sound API. The audio signal is recorded as pulse code modulation (PCM) with a sample rate of 16KHz, this is implemented in Java using floating point numbers. The Java Sound API objects TargetDataLine and AudioFormat are used to create the input data line from the sound card, the method open() called on the TargetDataLine object opens the line for audio input. The AudioFormat object is used to create audio input of the specified type, in our case this is PCM signed with a frequency of 16KHz. Code fragment 1 shows the basic Java code used to open a line for audio input. The Java code in this class is implemented as a thread, which allows the system to do other work in the recognition process while continuously listening for audio input.

```

TargetDataLine SampleLine;
AudioFormat format = new AudioFormat(AudioFormat.Encoding.PCM_SIGNED, 16000.0F, 16, 1,
16000.0F, 2, false);
DataLine.Info info = new DataLine.Info(TargetDataLine.class, format);

if (AudioSystem.isLineSupported(info))
{
    int soundData = 0;
    try
    {
        SampleLine = (TargetDataLine) AudioSystem.getLine(info);
        SampleLine.open(format);
        SampleLine.start();
        soundData = SampleLine.read(audiostream, 0, BUFFER_SIZE);
        //do some work with soundData captured e.g. Call another method
        SampleLine.stop();
        SampleLine.close();
        sampleOutputStream.close();
    }
    //catch exceptions statement Here
}
//remainder of the program

```

Code 1 Capture Audio with JavaSound

4.2 Signal Processing Algorithms

Stage two of the speech process requires the signal to be processed in order to identify key features in the speech signal e.g. word boundaries etc. The acoustic signal data from the audio recording process is used as the input to our signal processing and analysis stage. Figure 3 shows a block diagram of the phases which make up signal processing. The mathematical formulae for calculating the stages in figure 3 and the corresponding Java implementation code is shown below. As space is limited only code relevant to each formula is included rather than entire methods and objects.

The Mel Scale [14] is a perceptual scale based around the characteristics of the human ear, it attempts to simulate the operation of the human ear in relation to the manner in which frequencies are sensed and resolved. Using the Mel scale in speech recognition can vastly improve the recognition accuracy. Calculation of the Mel scale is achieved using the Mel Filter bank, which involves applying triangular filters in the signals frequency power spectrum. The purpose of the Mel Filter Bank is to smooth out the pitch harmonics and noise in the speech signal and to emphasise the formant information.

The procedure for Mel cepstrum coefficients is:

1. Divide the signal into frames
2. Obtain the power spectrum

3. Convert to Mel Spectrum
4. Use Discrete Cosine Transform to get Cepstrum Coefficients

This procedure is integrated into the feature extraction process shown in figure 3

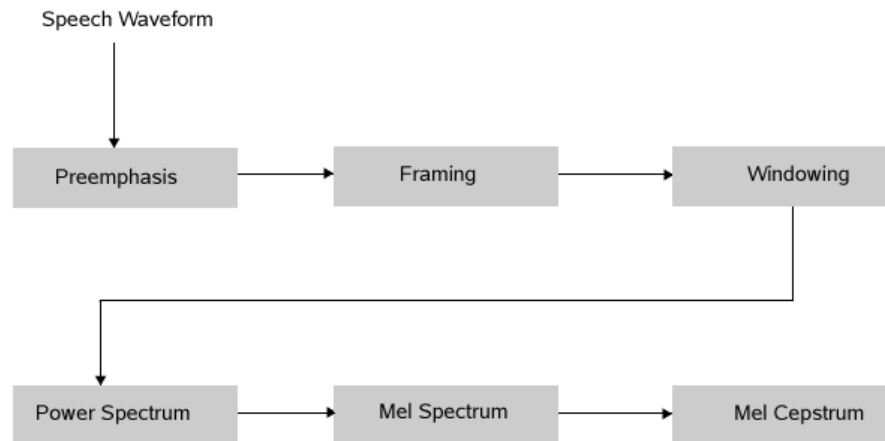


Figure 3 Signal Processing in Detail (Feature Extraction)

Pre-emphasis

In voice signals frequencies below 1KHz have greater energy than higher frequencies, this is due to the energy from the glottal waveform and the radiation load from the lips [16]. In order to remove the tilt we can use a High Pass Filter, applying this filter removes the glottal waveform and radiation load from the lower frequencies and distributes the energy equally in all frequency regions.

FIR Pre-Emphasis
 $y[n] = x[n] - \alpha x[n-1]$

```

Java
double outputSignal[] = new double[inputSignal.length];
for (int n = 1; n < inputSignal.length; n++)
{
    outputSignal[n] = inputSignal[n] - preEmphasisAlpha * inputSignal[n-1]; //preEmphasisAlpha = 0.95
}
return outputSignal;
  
```

Equation 1 Pre-Emphasis

Framing

The framing process involves breaking up the signal into frames with a shift interval in order to create a 50% overlap with a previous part of the signal.

Windowing

The frame is multiplied by a hamming window [13]:

$$w[n] = 0.54 - 0.46 \cos \left(\frac{2\pi n}{N-1} \right)$$

where N is the length of the frame

Java

```
double w[] = new double[frameLength];
for (int n = 0; n < frameLength; n++)
{
    w[n] = 0.54 - 0.46 * Math.cos( (2 * Math.PI * n) / (frameLength - 1) );
}
```

Equation 2 Windowing

Power Spectrum

The power spectrum is calculated by performing a discrete fourier transform through a fast Fourier Transform algorithm. The sum of the square of the resulting real and imaginary arrays from the fourier transform yields the power spectrum. [13]

```
s[k] = (real (X[k]))^2 + (imag (X[k]))^2

double pwrpectrum[] = new double[frame.length];
FFT.computeFFT( frame );
for (int k = 0; k < frame.length; k++){
    pwrpectrum[k] = Math.pow(FFT.real[k] * FFT.real[k] + FFT.imag[k] * FFT.imag[k], 0.5);
}
```

Equation 3 Power Spectrum

Mel Spectrum

The Mel spectrum of the power spectrum is computed by multiplying the power spectrum by each of the mel filters and integrating the result [13]. The corresponding Java implementation is not shown due to its size.

$$S[l] = \sum_{k=0}^{N/2} s[k] M_l[k] \quad l=0, 1, \dots, L-1$$

N is the length of DFT, L is the total number of mel filters.

Equation 4 Mel Spectrum

Mel Cepstrum

A discrete cosine transform is applied to the natural log of the mel spectrum to calculate the cepstrum. [13]

$$c[n] = \sum_{i=0}^{L-1} \ln(S[i]) \cos\left(\frac{\pi n (2i+1)}{2L}\right) \quad c=0,1 \dots C-1$$

C is the number of cepstral coefficients.

```

Java
double cepc[] = new double[numCepstra];

for (int i = 0; i < cepc.length; i++) {
    for (int j = 1; j <= numMelFilters; j++) {
        cepc[i] += f[j - 1] * Math.cos(Math.PI * i / numMelFilters * (j - 0.5));
    }
}
return cepc;

```

Equation 5 Mel Cepstrum

4.3 Hidden Markov Model Implementation

The Hidden Markov Model is implemented as a Java object, the constructor of the Markov takes two integers corresponding to the number of states and number of observation symbols. The initial state is set to 1 and the transition probabilities are initialised to random values. The Viterbi algorithm is implemented to find the best (most probable) path through the Markov trellis. The input to the algorithm is an observation sequence corresponding to an input signal (i.e. speech utterance). The signal has been pre-processed by the feature extraction stage; the Viterbi algorithm returns the probability that the input utterance is a recognised word. The training process for the Hidden Markov Model involves the use of the Baum Welch Algorithm [15]. This algorithm is designed to find the HMM parameters.

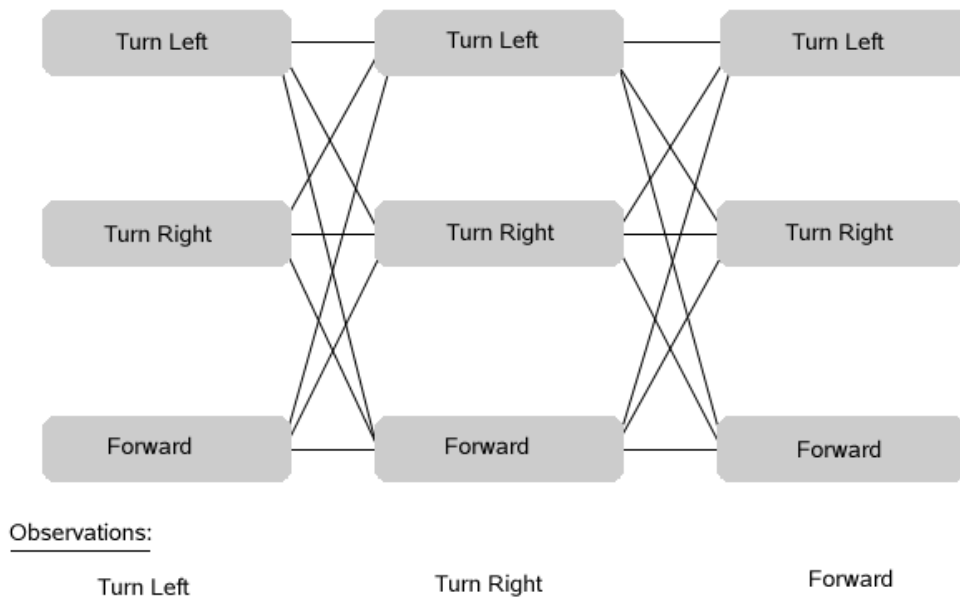


Figure 4 HMM Trellis

4.4 Distributed Engine Architecture

The speech engine we are developing is intended to be deployed across numerous disparate hardware/software systems, including desktop PCs running the Windows and Linux operating systems respectively, and an iPaq handheld PDA running Pocket PC 2003. Of these systems, the iPaq presents us with the greatest challenge, given its limited processing power and memory storage. With this in mind we devised a distributed speech recognition architecture, where the speech processing work could be performed by a more powerful server, with the Java based client on the iPaq only responsible for capturing the signal. This architecture is shown in Figure 5.

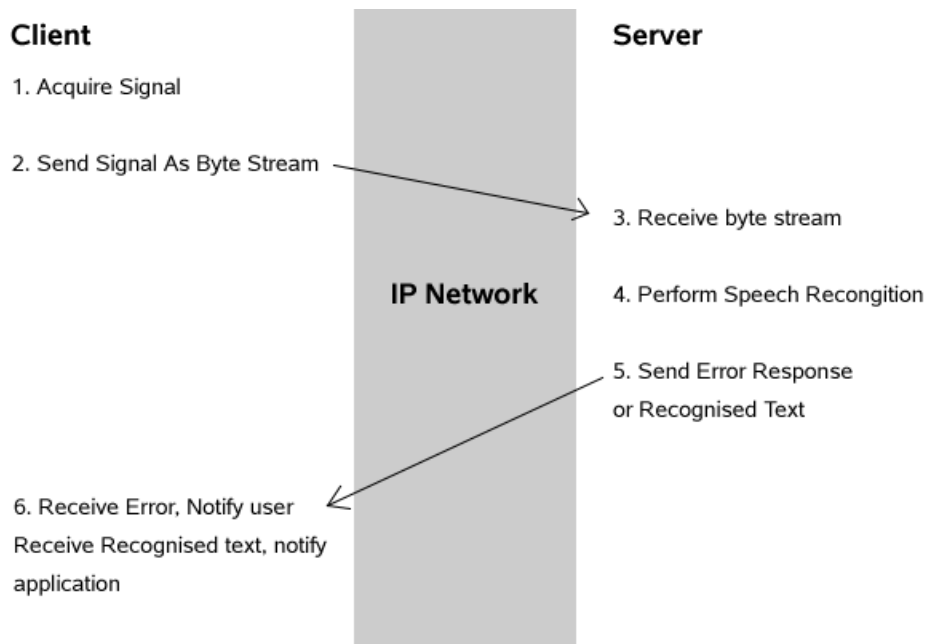


Figure 5 Distributed Speech Recognition Architecture

Using the Java Sound API, the Java Client can capture the signal as byte data which is then sent over an IP network to the server using a socket. The TCP/IP protocol is used to ensure reliability and the correct ordering of the packet data arriving at the server. The choice of TCP/IP as the protocol is not a trivial one, an alternative is to use the UDP protocol. UDP datagram packets can arrive in different i.e. unordered sequences (or not at all since UDP does not guarantee delivery), the correct order of the speech signal is vital, as recognition errors will occur otherwise. If the UDP protocol is employed the server will need to wait for all packets to arrive and order them accordingly before processing them for recognition, this could cause delays in the recognition process. Implementing fail safes for these factors can be totally avoided with TCP/IP.

5. Conclusions and Future Work

The Java programming language provides a suitable development platform with capable APIs for developing speech recognition technology. We have seen how the objects and methods provided through the Java sound API can be used to deliver continuous signal acquisition for a speech recognition process. The basis for a solid Java speech recognition engine is in place, the next phase is to investigate and make changes to the code such that it can be deployed on the iPaq and on the Linux operating system.

The current implementation of this system requires the user to train the words and phrases which will be recognised. The training process is quite intensive for the user, involving frequent manipulation of the user interface (Starting and stopping the record process), this process is not conducive to creating vocabularies of any any greater size than a few dozen words. While improving the usability of the training process could increase the potential vocabulary of the system, however a speaker independent approach would eliminate all training requirements. Speaker independent speech recognition systems do not require the user to train the system prior to use. Speaker independent systems rely on acoustic information provided to them by large speech corpus. Integrating a speaker independent component into this system through a speech corpus, would be the next logical step in the systems evolution.

6. References

1. Sun Microsystems, Java Sound API [online at] <http://java.sun.com/products/java-media/sound/>
2. Sun Microsystems, Java Media Framework [online at] <http://java.sun.com/products/java-media/jmf/>
3. A. Spanias, T. Thrasyvoulou, S. Benton, Speech parameterization using the Mel scale [online] <http://www.eas.asu.edu/~spanias/E506S04/mel-scale.pdf> (18/3/04)
4. Sun Microsystems Ltd, Java Speech API, [online at] <http://java.sun.com/products/java-media/speech/>
5. Microsoft Corporation, Microsoft Speech and SAPI 5, [online at] <http://www.microsoft.com/speech/>
6. IBM, Via Voice [online at] <http://www-306.ibm.com/software/voice/viavoice/> (15/1/2004)
7. ScanSoft, Dragon Naturally Speaking, [online at] <http://www.scansoft.com/naturallyspeaking/>
8. Phillips, Phillips Speech SDK 2.0, [online at] <http://www.speech.philips.com/> (20/1/2004)
9. CMU Sphinx Project, CMU Sphinx [online at] <http://www.speech.cs.cmu.edu/sphinx/index.html> (15/1/2004)
10. Orange Cow, OCVolume [online at] <http://ocvolume.sourceforge.net> (19/3/04)
11. Boyle RD, Introduction to Hidden Markov Models, University of Leeds, online at http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html
12. Jurafsky D. & Martin J.H. , Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2000, Prentice Hall, New Jersey.
13. Seltzer M., Sphinx III Signal Processing Front End Specification, CMU Speech Group, August 1999.
14. S. Molau, M. Pitz, R. Schluter, H. Ney, Computing MelFrequency Cepstral Coefficients on the Power Spectrum, Proc. Int. Conf. on Acoustic, Speech and Signal Processing, Salt Lake City, UT, June 2001
15. Boyle RD, Introduction to Hidden Markov Models, University of Leeds, online at http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/hmms/s2_pg3.html

16. Mustafa K, Robust Formant Tracking for Continuous Speech with Speaker Variability, Dept. of Electrical and Computer Engineering, McMaster University, [online]
<http://grads.ece.mcmaster.ca/~mkamran/Research/Thesis%20Chapters/3%20-%20The%20Formant%20Tracker%20-%20pp.%2041-70.pdf>

Design of a Wear Test Machine for Diamond Saw Segment

D. Nulty J. Dwan and Y. Blake

Department of Mechanical Engineering, Institute Technology Tallaght, Dublin 24.

Abstract

Diamond tools are used in a wide range of industrial areas such as construction industry, metal machining and exploration drilling. The diamond impregnated tools used in the stone and construction industry are metal matrix composites. Diamond saw blades are the most commonly used tools.

The optimum operation of diamond saw blades is determined by the cutting action of the diamond grit and the bounding of the metal matrix. However the wear behavior of the diamond saw has been less studied. Currently in the blade development, actual full blade tests often have to be conducted for optimization and the testing process is very slow and expensive to carry out. So the development of a testing machine that could reduce the blade development time would be very advantageous.

This paper describes the design and construction of a wear apparatus which simulates the wear conditions that a diamond impregnated saw blade experiences by using just a single segment. It is also our intention to present single segment wear tests on ceramic-based materials, which can be used for the testing and developing of a full blade diamond saw.

1. Introduction

Diamond tools are used in a wide variety of applications of which the most important include the drilling and sawing of rocks and concrete, grinding of glass or metals, polishing of stone. The most dramatic increase in the importance and growth of industrial diamonds have been the advent of synthetic diamonds and the use of diamond tools into the stone and construction industries [1]. Nowadays synthetic diamond holds the largest share at 90% of the industrial diamond tool market [2].

1.1 Diamond saw blades

Diamond impregnated tools consists of randomly dispersed and orientated diamond crystals in a metal matrix which bonds diamond together. As the tool cuts the matrix erodes away from the diamond. This exposes the diamond and gives the diamond crystals clearance for penetrating the material and also allows the cuttings to be flushed out. For proper operation the metal bond must wear away progressively so that new diamonds are exposed to continue the cutting action during the life of the diamond tool.

The key factors that determine the choice of matrix metal include (a) the wear resistance of the matrix; (b) the type and size of diamonds to be used and (c) the tool fabrication process.

For a particular application, currently the blade is selected by picking a stock item recommended by the tool manufacturer. However, often an off-the-shelf blade does not fully fulfill the requirements of blade life and cutting performance. At this stage, the diamond tool manufacturer has to develop a new matrix either by modifying the existing matrix or by designing a completely new matrix. Where full scale blade testing is carried out on different

types of stone or concrete materials, it can be very expensive as well as time consuming when one considers that on a typical 600 mm diameter blade there can be 40 segments. Substantial cost savings could be made if this type of testing could be carried out using single segments with different matrices and diamond combinations [3].

1.2 Design Requirements

The cutting performance of the saw blade can be related to segment wear and blade life. It is also important that a blade being capable of “free-cutting”. A “free-cutting” blade is a blade that cuts without too much difficulty. If the matrix is too soft, or excessively “free-cutting”, the blade life can be prematurely short. The opposite of “free-cutting” is a blade matrix that is too hard or has high wear resistance.

The design requirements of the machine are that it accommodates the normal range of blade diamond matrices, cutting speeds and feeds as closely as possible to that found in the field. The following are the typical test conditions which are normally encountered:

- (a) Peripheral blade speeds, ranging from 25 to 60 m/s, which are typical blade speeds used in industry.
- (b) Variable normal force applied on the saw segment.
- (c) Range of diamond concentrations, grit sizes and grades
- (d) Different matrix compositions
- (e) Range of different types of stone and concrete for testing.

2. Design analyses

The analysis initially examined the cutting action of a multi-segment blade and compared it to a single-segment ‘flywheel’ type blade. The analysis also examined the scenario of a segment-on-disc arrangement and compared it to the single-segment ‘flywheel’ and multi-segment blades. The investigation was approached from the perspective of examining the cutting action of the diamonds. The assumption was that the single-segment is identical to that of the multi-segment.

Bütner [4] developed an equation which described the cutting process that takes place in grinding wheels. These grinding wheels are similar to saw blades in that the diamond used has similar crystal shapes, and the matrix is a metal bond. Bütner proved for a grinding wheel that the ‘average area cut by each diamond, $\overline{A_{ps}}$ ’ was given by,

$$\overline{A_{sp}} = \frac{u_1}{v_1} \frac{1}{N_k} \sqrt{\frac{a_1}{D_1}} \dots\dots\dots(1)$$

where u_1 = feed rate

v_1 = blade speed (peripheral blade speed)

a_1 = depth of cut

D_1 = diameter of blade

N_k = No. of effective cutting grits per unit area

Normally though diamond impregnated sawblades are made up of segments. The spaces between each segment are called gullets. However a grinding wheel is a continuous rim which has diamond all around its circumference, but in a sawblade there is a reduction in the diamond cutting area because of these gullets. Bienert [5] modified Büttner's equation to take this into account by including a quotient ($\lambda_1 = l_2 / l_3$) which compared the length of the segment to the length of the segment plus the gullet length.

Bienert's equation for a segmented sawblade is

$$\overline{A}_{sp} = \frac{u_1}{\lambda_1} \frac{1}{v_1 \cdot N_k} \sqrt{\frac{a_1}{D_1}} \dots\dots\dots(2)$$

According to Bienert [5] the cutting action of each individual diamond grit is the same for both the multi-segment blade and the single segment blade. The only difference is the amount of material removed, the multisegment, having more segments, will remove more than the single segment wheel. The feed rate for the single segment blade is slower to take account of the reduced cutting surface.

The proposed wear testing is a pin-on-disc type wear machine, in which the diamond impregnated segment plays the role of 'pin' and the material to be tested forms the disc. In order to compare a segmented sawblade with the proposed testing machine, the cutting action of the diamonds is examined. The test segment used here is equivalent to that used in a multi or single segment blade. Using Bienert's[5] statement that the effective cutting performance of each cutting diamond is the same, then the following equation for calculating the feed rate ' u_2 ' of the segment into the stone disc was derived for the downward feed rate of the test segment on the test machine:

$$u_2 = \frac{u_1 \cdot l_4}{\lambda_1 \cdot \pi \cdot D_2} \cdot \sqrt{\frac{a_1}{D_1}} \dots\dots\dots(3)$$

where, ' u_2 ' = axial feed rate of the segment into the stone disc,

' u_1 ' = feed rate of multisegment blade,

' l_4 ' = POD segment length,

' D_2 ' = diameter of stone track on stone disc,

- $'D_1'$ = diameter of sawblade,
 $'a_1'$ = depth of cut using multisegment blade,
 $'\lambda_1'$ = segment quotient of multisegment blade.

The derived above equation (3) gives the downwards feed rate for the single segment into the tone disc which is equivalent to the cutting action of a multisegment sawblade.

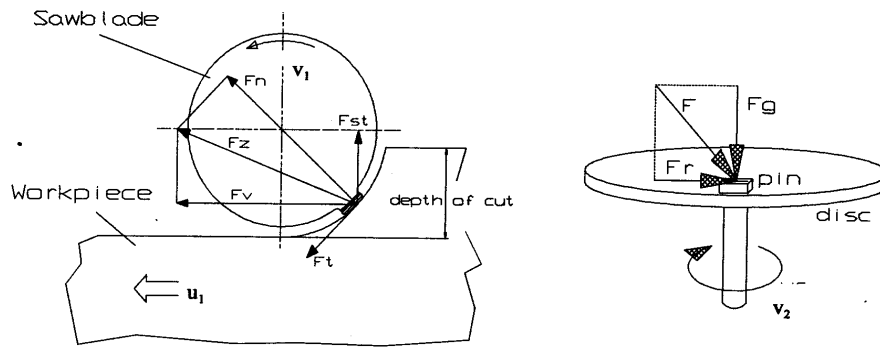


Figure 1 Single segment sawblade and pin-on-disc machine

3. Machine Design and Construction

After investigation on several of designs of pin-on-disc machines the following is the specific descriptions of the chosen design. The diameter of the stone disc was calculated from data determined from 'cutting arcs' which different sized blades would describe if cutting in a downward mode of cutting. From the initial analysis, a saw blade range was chosen from 500 mm to 850 mm in diameter. From these sawblade diameter ranges, a percentage range of depth of cut (a_1) was selected. It started at 10% of blade diameter, increasing by 5% to a maximum depth of cut of 45% of blade diameter.

From the various ranges for 'depth of cut' for each blade size, the different angles for α , the angle of engagement, were calculated. These values for the angle of engagement, α , gave the lengths (l_1) of the different cutting paths or trochoids which different blade sizes and 'depths of cut' would generate when sawing.

Translating these values for the cut path lengths (l_1) to the wear testing machine, the cut lengths ' l_1 ' are represented as half the circumference of different diameters on the disc. To separate each ' l_1 ' a slot is cut in the stone disc. From this, the diameters of the cutting tracks on the disc were calculated with each track representing twice the ' l_1 ' cut path for each depth of cut for each blade diameter combination,

The corresponding peripheral blade speeds for the different wear track diameters and the motor rpm were calculated. The motor (4 kW) had a capability of rotating from 1000- 3500 rpm through a speed inverter controller. A spreadsheet was developed which related the ‘% depth of cut’ with the peripheral blade speed. Two different blade sizes could have the same cutting track, but it would only differ by the % depth of cut in actual practice in the field.

Measurement of the input to the electric motor of a machine tool provides a convenient method of measuring the power consumed in cutting. An approximate method is given by Black et al. [6] where if the value of the power supplied when the machine is running idle is subtracted from the power reading taken under the cutting load, a reasonable estimate of the power consumed in cutting is obtained. A wattmeter is used so that the power consumed can be recorded as the different segment compositions are used and the different stone materials tested

As the cutting forces in action can be considered to be similar to those in a drilling machine, a hydraulic cylinder with a load cell to measure the vertical axial cutting force is used. It would therefore be possible to monitor the cutting action of the diamond saw segment. It is intended to measure the frictional force resulting from the cutting action of the test segment on the stone disc. Two methods were investigated, one using a load cell, the other using strain gauges. Strain gauges were chosen and are mounted on the cantilever segment holder. A multiplexer with a PC with a data acquisition card is connected to the circuit with signal capture at 1/1000 times a second. From the calculated forces resulting from the bending of the cantilever the frictional force can be measured. The final design is shown in Figure 2. A pin-on-disc type machine was designed to simulate the cutting action of the diamonds in a diamond impregnated tool.

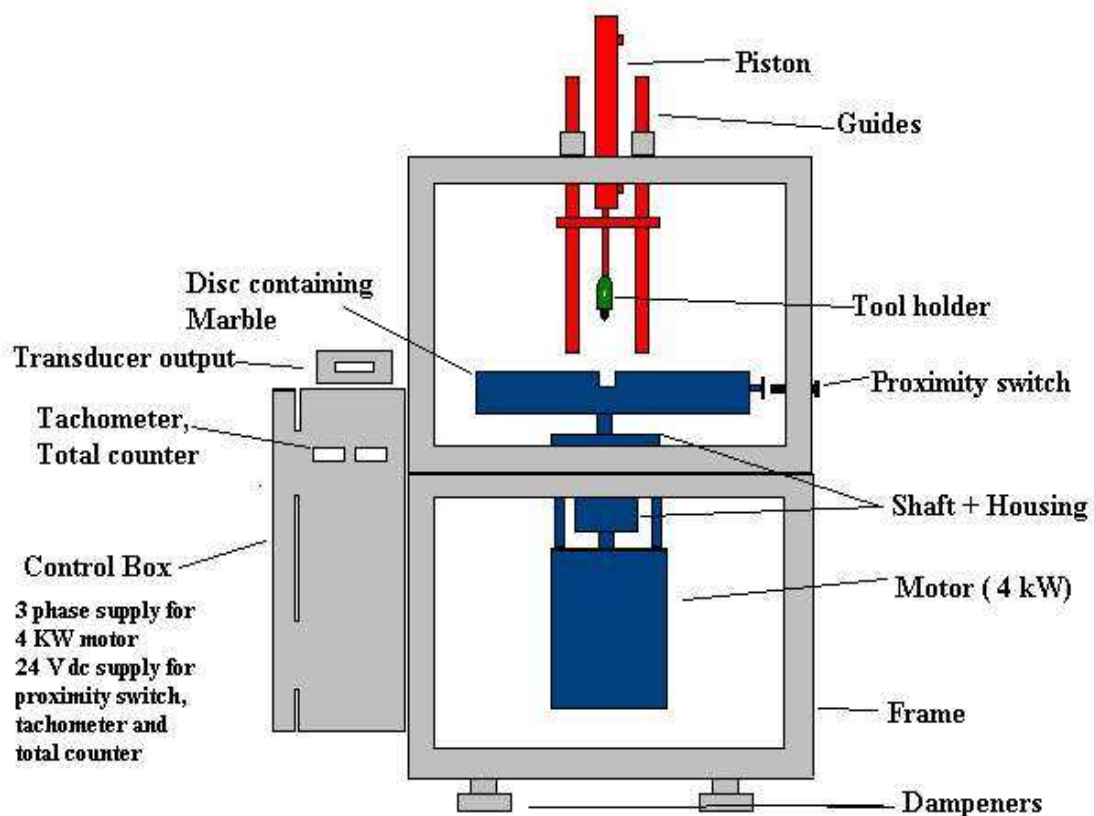


Figure 2. Pin-on-disc type testing Machine with dimensions base 500mm by 500mm and height 990mm.

The construction of the machine was carried out according to the design specification. The preliminary tests were conducted and the following modifications were provided to improve the testing performance:

- An inverter was added to alter the frequency to the motor thereby controlling the speed
- A hydraulic pump which is restricted through a pressure relief valve
- Pressure transducer was used to indicate the load applied to the disc
- Transducer digital output unit indicated the load applied
- A proximity switch was used to indicate speed and revolutions via a tachometer and total counter
- Safety measures were implemented.

4. Experimental results and discussions

On completion of the testing machine experimental work were conducted on two materials: marble and limestone. Diamond impregnated segments were used for various diamond sizes, concentrations and metal matrix compositions. In the table below listed the details of these diamond saw segments:

	Diamond Crystal Size	Diamond Concentration
Segment A	US Mesh 30/35	DC10
Segment B	US Mesh 30/35	DC40
Segment C	US Mesh 40/45	DC10
Segment D	US Mesh 50/60	DC40

4.1 Test results on Marble

Marble discs were tested by using two different diamond specimens. The testing was performed under the normal load 200 N and the speed of the disc was 2000 rpm. Figure3 shows the test results.

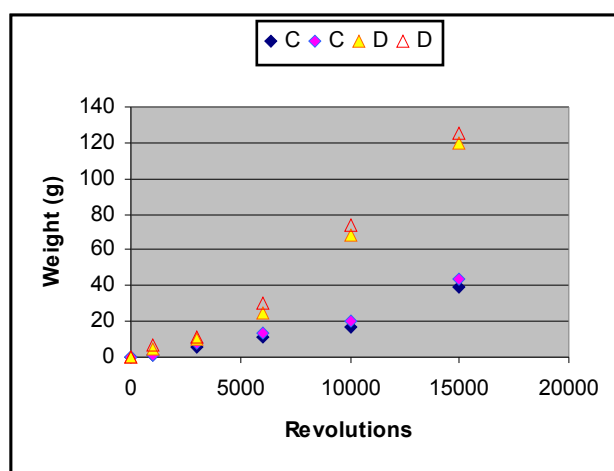


Figure3. The weight removed from marble sample against the revolutions of the marble disc.

Plotting the revolution of the tool segment travelled on the marble disc against the weight removed from the disc we can see an approximate linear relationship of the two parameters. The upper lines show the cutting performance of segment D, for which the size of the diamond is US mesh 50/60 and the diamond concentration is DC 40. It is obvious that this segment removes marble faster than the other type of segment, which is C. The test results show that under similar cutting condition the cutting performance is related to the diamond grit size and the concentration. The higher concentration and larger grit size are more efficient for removing of marble.

4.2 Test results on Limestone

Limestone discs were tested by using three different diamond specimens. The testing was performed under the normal load 200 N and the speed of the disc was 2000 rpm. The test results are showing in Figure4.

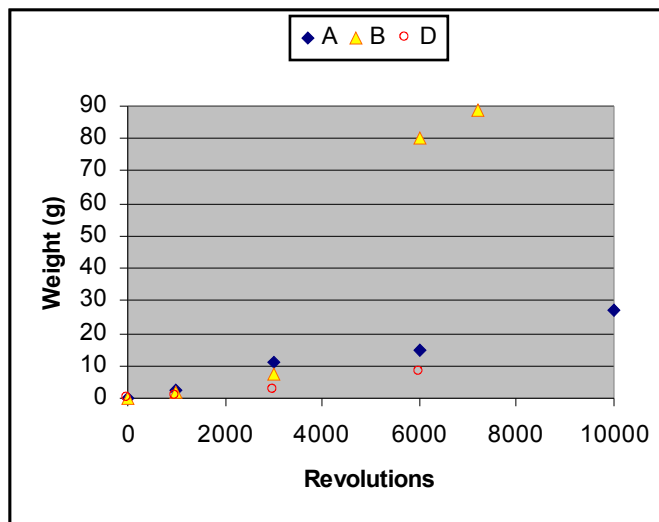


Figure 4. The weight removed from marble sample against the revolutions of the limestone disc.

The top line gives the cutting performance of segment B, for which the size of the diamond is US mesh 30/35 and the diamond concentration is DC 40. This segment removes marble most efficient in comparison with the other two segments. The middle line displays the results from segment A, and the bottom line are the results from segment D. Among three of them segment B has the highest concentration of diamond. The influence of the concentration on the cutting efficiency is most apparent. For the segments A and D, with the same concentration of diamond crystals, segment A, which has smaller grit size is more favourable for the cutting operation.

5. Conclusions

The pin-on-disc wear testing machine provides a possibility of simulating the cutting performance of a single saw segment, which can be used for the selection and development of diamond saw blade under variable conditions.

The method described enables different diamond segments to be tested at desired normal load at variable rotating speed.

The experimental results obtained provide valuable information on the performance of the single saw segment, which is related to the grit size and diamond concentration.

Further tests should be carried out on different types of diamond segment in order to predict their wear behaviors.

References

- [1] Jennings M., Wright. D., Guidelines for sawing stone, Industrial Diamond Review 2/89
- [2] Boothroyd G., Fundamentals of Metal Machining and Machine Tools, McGraw-Hill, Washington, 1982.
- [3] Dwan, J. D., A Design of a wear test machine for diamond impregnated tools. IMC-13, Cork, Ireland, Sept. 1996
- [4] Büttner, A., Das Schleifen sprödharter Werkstoffe mit Diamanttopfscheiben unter besonderer Berücksichtigung des Tiefschleifens, Diss. TU Hannover, 1975
- [5] Bienert, P. Dipl.-Ing., Kreissägen von Beton mit Diamantwerkzeugen, Hannover, Dissertation, Technische Universität Hannover, 1978.
- [6] Black S. C., Chiles V., Lissaman A.J., Martin S.J., Principles of engineering manufacture, 3rd Edit. Arnold. London, 1996

An Adaptive eLearning framework – Design Issues and Considerations

Marie Brennan, Larry McNutt

Institute of Technology Blanchardstown

Funded By: HETAC IT Investment

Contact email: maria.brennan@itb.ie

Abstract

This paper discusses the issues and motivations surrounding the design and development of an adaptive e-Learning facility. The problem facing developers is the deliverance of high quality educational opportunities via the web that are equivalent or even better than face-to-face classes. Because of rapid developments in the information and communications technologies with regard to on-line distance education it is possible to improve the quality of the system deliverance. This is where the concept of using individual learning styles is adhered to. If a system is designed where the individual learning style of the student is discovered, the system can then be designed to best suit them. By implementing such a design students can learn in a manner they prefer therefore leading to an increased willingness to learn. Learning styles will be determined through questionnaires. Once these styles are determined it is possible to design appropriate system modules for them. This paper discusses the relevance of learning styles and system design of computer education to prove the question “Is there a link between student learning styles and successful online learning” and “Is the design and development of an adaptive e-learning system an effective eLearning environment”. This is at present a work in progress.

Keywords

Adaptive System, learning Style, learning preference, eLearning, distance education

1 Introduction

eLearning can be described as the convergence of the Internet and learning, or Internet-enabled learning. Implementing eLearning frameworks today is not an uncommon occurrence. In Ireland for example many institutions have adopted eLearning such as University College Dublin, Trinity College Dublin and the Dublin Institute of Technology. The above third level colleges provide computer-based training through a wide range of specialized applications and processes and are only a few of many throughout the country that have introduced this form of Learning

Once implemented these eLearning frameworks are set up with a standard system for all users. All information is presented to the users in a pre-designed format, this format being text, graphics, audio, video etc. There are limitations however with having a standard system for all users. Using a standard system may be effective as face to face tuition for some but may pose problems for others. This paper will attempt to bridge this divide by presenting a solution in the form of an adaptive system for the eLearning framework.

This is where the area of learning styles is introduced and where different learning styles will be uncovered. An attempt will be made to identify and define the attributes of several types of learning styles. The system will be designed solely on what presentation format suits these particular learning styles and their suitability to the learner. The outcome of any teaching process is that the learner learns. What influences most how they learn is their individual learning style and so how an on-line course is designed, developed and delivered depends on that. (Kirkwood 1998).

2 Background

It has only been 10 years since the coding language for the World Wide Web was developed and Wide Area Information Servers became the tools for “surfing the net”. Since that educational institutions, research centers, libraries, government agencies, commercial enterprises and a multitude of individuals have rushed to log on to the internet (Johnson, 1999).

One of the consequences of this tremendous surge in online communication has been the rapid growth of technology-mediated learning at the higher education level. E-learning is the solution to the training challenges the Internet economy has created. E-learning refers to education that is enhanced by or delivered via the Internet. ELearning began in corporate training departments, schools, and universities as a supplement to standard teaching methods. Today, it encompasses a rich set of solutions that can be used throughout an organization from corporate communications and marketing to technical documentation to share information, experience, and ideas. E-learning can give learners the ability to turn change into an advantage by tapping existing knowledge resources and packaging them in a more accessible, customized, learner-centric format.

E-learning systems can enhance traditional teaching methods and materials, such as classroom discussion, textbooks, CD-ROMS, and non-Internet computer-based training. ELearning provides the added advantage for students in that they can develop online communities for providing mutual support and sharing information via discussion rooms and bulletin boards. Teachers can provide feedback and direction to learners, answer questions, and facilitate these discussions. ELearning can provide on-demand audio and video technologies can present material in a stimulating fashion to actively engage learners. Knowing a little bit about learning styles can help an individual in determining if online learning is for them. The interaction and delivery methods used in online classes are dramatically different from traditional classes, so understanding how one learns is a good part of the decision-making process.

The three predominant learning styles are visual, auditory, and tactile/kinesthetic. Broken down further, people learn by:

- Reading (visual)
- Listening (auditory)
- Seeing (visual)
- Speaking (auditory)
- Doing (Tactile/Kinesthetic)

The first three on the list are passive types of learning, while the last two are active types of learning. How much we tend to remember is a function of the type of learning we prefer and our level of involvement in the learning. People often learn through a combination of the ways described above. To a lesser degree, environment is a factor too. Students get only what they need and study at their own pace. Student information requirements vary. In addition, the knowledge and behavior of students with respect to the learning process can change both over time and at the same time (Riding et al 1995). Given this scenario, it is possible to suggest a need to develop interfaces and systems that help each student to reflect on, identify and develop their information needs.

This paper will provide some information into the background of distance education as well as the use of using technology to support eLearning. It will also look at a suitable adaptive architecture.

3 Distance Education

One of the first universities to deliver distance learning in an organized manner was Pennsylvania State University, establishing its first distance learning network in 1886. Penn State used the state of the art technology of the day, U.S. Mail, to communicate with its distributed students.

In the 1960's the UK Labour Government approved the setting up of 'The University of the Air'. This was later to become the Open University. The OU was originally set up to offer degree studies through broadcasts such as TV and Radio in partnership with the British Broadcasting Corporation and later computer mediated communication became a vital ingredient in distance delivery of under graduate taught programmes.

Distance education has walked through many of the problems now facing eLearning has much to offer eLearning. Distance education departments at colleges and universities have spent

decades addressing challenges of creating and designing learning resources to be used by students studying on their own. Many of the concerns currently facing eLearning are a high drop out rate, problems with creating interactivity and fostering community among learners. The outcomes from initial research into these issues suggest that the simple effects of technology on teaching style, learning style, grades obtained, and course satisfaction may not be very robust (Grasha and Hicks, 2000).

Gee (1990) studied the impact of learning style variables in a live teleconference distance education class. Students in the distance learning class who possessed a more independent and conceptual learning style, had the highest average scores in all of the student achievement areas. People with the lowest scores in student achievement in the distance learning course had a more social and conceptual learning style. Students with both a social and applied learning style performed much better in the on-campus class. The outcomes of the Gee study suggested that successful distance education students favored an independent learning environment while successful on-campus students showed a preference for working with others.

Students who study at a distance are separated both from their tutors and their peers. For some this can be a particular problem, and for all, some of the time the separation poses potential difficulties. Social interaction, such as the sharing of ideas, discoveries, successes and failures and general social support, are all to a certain extent, missing from the distance learning environment. Students may therefore feel isolated, start to lose motivation, experience frustration or anger, and a host of other unwelcome emotions.

When designing systems and materials for distance delivery, lecturers must consider not only learning outcomes, but also content requirements and technical constraints. Also to be considered are the needs, characteristics, and individual differences of both the students and the teachers.

The task of the distance educator is therefore to dispose of these problems as much as possible by mixing and matching techniques, creating and maintaining a stimulating environment, and offering opportunities for students to communicate with each other and with the teaching staff on a regular basis.

To mix and match techniques and make a teaching environment more stimulating for a student a good idea then is to teach each student exclusively according to his or her particular choice or style of learning thus making the whole experience of distance education for the student better. It is far more beneficial for the teacher to strive for a balance of the various styles of learning. By discovering the students learning style and designing and presenting information to them in that particular manner will reap considerable benefits.

4 Using Technology to Support Learning

Technology Based Training (TBT) is a computer based training methodology that includes web-based, intranet based, DVD and CD based training on any topic. The content is designed and developed with the same instructional objectives as classroom-based learning. TBT breaks course material into small bites by breaking large blocks of content into modules that can be searched and then completed in a short amount of time. Dissecting a skill into many segments allows users to gain competency quickly. Learning objects also serve as a helpful tool for users who need to brush up on a skill once they're back on the job. Students can quickly scan a course module list and find the lesson they need without wading through pages of unnecessary content. Benefits of TBT include:

1. Enhances retention rate by 25 – 60%. It provides for self-reinforcement. Interactivity improves the retention of the skills being taught and simulations help walk students through actual scenarios helps identify mistakes when they make them.
2. The ability to customize the learning material to students own needs, with more control over the learning process, leads to a 60% faster learning curve compared to instructor-led learning.
3. Saves time. A comprehensive skill assessment performed prior to taking the learning determines which topics you need to focus on. The delivery of content in smaller units, called “learning objects” contributes further to saving time and has a more lasting learning effect.
4. TBT interactivity accommodates different learning styles and fosters learning through audio, visual and testing.
5. Learn at your own pace: Don't feel constrained by an instructor-led class that is too fast or too slow for you. You can learn at comfortable pace, further increasing skill retention.
6. TBT is flexible – students can navigate through the learning to cover topics in whatever order is more beneficial in light of their specific learning needs. This allows students to select learning materials, or to be directed to content that meets their level of knowledge, interest and what they need to know to perform more effectively in their particular activity.

When evaluating TBT, the most important considerations are content and design. Good training requires a preliminary assessment of those needs the training must address. The assessment can be very detailed, focusing on learner characteristics, the learning environment and a variety of other issues. However, the single most important requirement involves the

identification of the standards a properly trained student must satisfy to effectively do the job. Each standard should yield a specific number of learning objectives. These learning objectives define what the trainee needs to know and how the trainee needs to apply that knowledge.

Properly constructed TBT reduces the amount of information the learner/trainee must retain during training to the lowest possible level. To accomplish this, certain characteristics must exist. First, information must be organized in a detailed, sequential fashion from the job, to the duties required to satisfy the job. Good TBT focuses on tasks and sub-tasks. Second, TBT chunks task and sub-task information within an architecture that supports retention and its quick transfer from short term to long term memory.

Technology-based training allows more room for individual differences in learning styles. TBT's interactivity accommodates different learning styles and promotes the growth and development of learning through audio, visual, testing, and by having learners "do what they are learning". TBT also provides a high level of simulation that can be tailored to the learner's level of proficiency.

People can learn at a pace that suits them and review course material as often as since they can customize the learning material to their own needs, students have more control over their learning process and can better understand the material, leading to a faster learning curve, compared to that of an instructor-led course.

Technology based training is changing the way corporations and individuals obtain skills in almost every single segment of the business process. Initially IT training dominated the TBT market. However front office, management, sales, customer service, and professional development training are increasing at a rapid rate.

5 Limitations of Existing Technology

When considering what exists in the area of eLearning today it is imperative to consider ones knowledge, skills, and abilities and what is the individual's academic and professional background. Is the individual comfortable with the media? Is the individual predisposed to self-learning? Is the content appropriate and is it suitable for the individual learning styles of each learner? Although classified as at your leisure and at your own pace it is important that the user of the system does not get bored or generally uninterested in continuing with the course for any reason . One of the problems that plague the eLearning industry today is a high dropout rate. There are no national statistics, but a recent report in the Chronicle of Higher Education (United States) found that institutions are seeing dropout rates that range from 20 to 50 percent for distance learners (Karen Frankola, 2001).

After research by Frankola it was found that some of the reasons for this high drop out rate included lack of management, lack of motivation, problems with technology, lack of student support. An area that also appears to have contributed greatly to this drop out is Individual learning preferences and poorly designed courses. Students with different preferences in how information is displayed to them have problems with how these courses are designed and presented to them. It is these students who then feel isolated from the course and lack the motivation to continue. With the overwhelming amount of information that must be streamlined, the most advantageous opportunity for eLearning will be getting what you want, when you want it, and how you want it .

Each individual learns differently. Some individuals require more direction and guidance from a trainer or teacher than others. The same idea applies to computer based learning. The way that information is displayed online appeals to some and may not to others. This depends on how different individuals perceive the information that is displayed in front of them. The solution to this problem reverts back to how the system is designed

There are few or no guidelines, for what constitutes effective human-computer interfaces for educational purposes. Due to a lack of proper system design guidelines, designers of educational software often use the styles that would have been used for lectures in a college environment. Here the student is handed the lecture notes as well as listening as the lecturer explains them. In an eLearning online environment these can be boring and do not keep the interest of the student. If the problem is the system design and how information is portrayed to the students then a simple solution is to adapt the system. And a better solution is to adapt the system to suit the learning style of the individual.

6 Adaptive Architecture

The major problem that exists is the design of the system used to display the information to these students. If this is so, then solutions to the problem is in determining the learning styles of the individuals taking the courses and understand the display methods that best suit them. Having knowledge of these methods can then allow the adaptation of the standard system engine building the online course. The content management system engine that will be used for this particular project will be WebCT. (6.1).

The next step will be determining the learning styles. There are many questionnaires that can determine individuals learning styles. There are many different styles of learning and these will be studied in detail to decipher what media content would best suit them. The next step then is to adapt the system for each type of style and present the given information in a variety of ways to the users. All of the above constitutes the adaptive architecture that will build this adaptive framework.

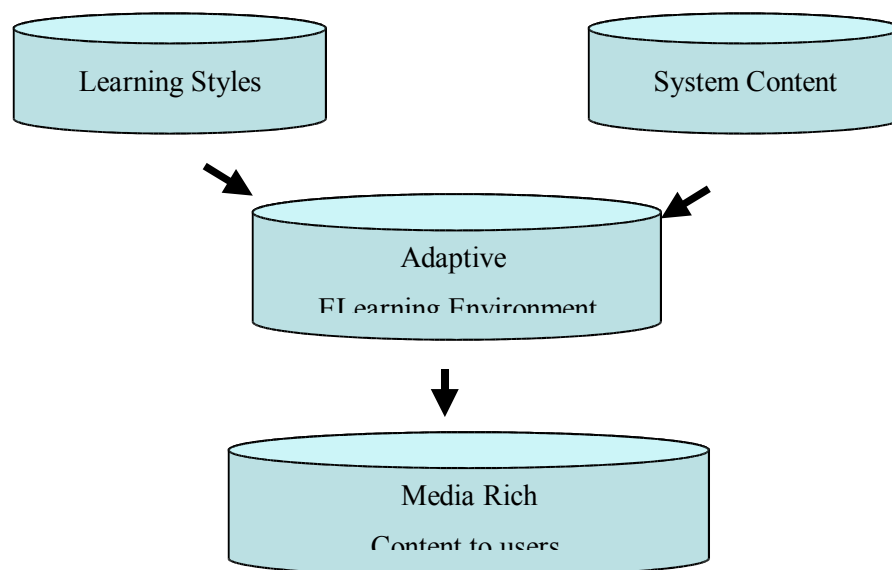


Figure 1: System Architecture

The focus of this project will be to analyze, develop, implement and evaluate an adaptive eLearning environment using chosen proprietary products. These will be course management systems that enable the efficient delivery of high quality online education. Course material will be presented to the students through the adaptive system mentioned above. The material that will be presented to them will include:

- Assignments
- lectures
- Content Module
- Course Map

The above will then be adapted to suit their learning styles.

6.2 Learning Styles

We are all different, and that applies to how we learn information, as well. Research has found that the two major categories of learners are those who learn best in visual ways and those who work better in auditory ways. Your learning style is determined primarily by your brain, whether it relies more on your eyes or your ears to comprehend new data. Those who respond better to what they see are visual learners. Those who respond better to what they hear are auditory learners. Those who are equally as good at interpreting data that they see and hear are known as “balanced” learners. Balanced learners will recognize aspects of what they're good at in both the visual and auditory learning style descriptions.

Of course, everyone relies on their eyes at some times and their ears at others. But when faced with new information, the majority of people fall back on their dominant learning style. And as more is being studied about learning styles, some sub-styles are being identified, such as kinesthetic, the learning style that relies on learning by doing.

Everyone uses both faculties, but most people tend to favor one over the other. In the 1940s Isabel Briggs Myers developed the Myers Briggs Type Indicator (**MBTI**), an instrument that measures, among other things, the degree to which an individual prefers sensing or intuition (Quenk and Wiley 1993). For succeeding decades the MBTI has been given to hundreds of thousands of people and the resulting profiles have been correlated with career preferences and aptitudes, management styles, learning styles, and various behavioral tendencies.

The Myers-Briggs Type Indicator is based on psychologist Carl Jung's theory of psychological types (Boeree, 1997). This indicator is thought to affect many of the behaviors and attitudes of a person including his or her approach to learning. Sensors are good at memorizing facts and intuitive learners are good at grasping new concepts. Sensors are careful but may be slow; intuitive learners are quick but may be careless. Knowing this information teachers and lecturers can present information to these learners in a way that they would understand a lot easier. The *Myers-Briggs Type Indicator* discovers the learning styles that best suits each individual's preferences.

Sensor

Creating web based training that appeals to this kind of learner means including details, well laid out procedures, verifiable facts, and practical applications of the information being presented.

Intuitors

Simulations and the opportunity to explore other web sites would probably be more appealing to this kind of learner. The implication to society or civilization as a whole of the practical application of the information being presented would make web based training of more interest to this kind of learner.

Extraverts

These learners would probably be better served by chat rooms, discussion forums and dialog databases included in their web based training. Interaction with a "virtual" teacher would probably also be useful.

Introverts

The impersonal, almost private, nature of web based training will probably make it very appealing to this kind of learner.

Thinkers

Presenting logical arguments and research results associated with the new material being presented is more likely to be the best kind of web based training for this kind of learner.

Feelers

Showing how the information affects people will make it more interesting to this kind of learner. Devices such as chat rooms that let them know how other learners and the "virtual" teacher respond emotionally to the information are useful.

Judgers

Web based training designed to go from beginning to end over a prescribed route would probably be most appealing to this kind of learner. Keeping them informed of their progress along the route might also be valuable.

Perceivers

Web based training that includes simulations and the opportunity to explore other web sites would probably be most effective with this kind of learner. They may respond best to an open ended learning agenda if one is possible.

Research using the Myers-Briggs Type Indicator, in general, shows that the majority of college students do not have well-developed independent or abstract-thinking learning styles. Their

interests and energy are centered on the world of "people, objects, and events" and not on the exploration of ideas (Grasha 1996; Lawrence 1982). That is particularly true of students attending large urban universities and less-elite, small liberal arts colleges. Thus, teachers employing technology need to understand the learning styles of their students when designing course activities. And those promoting technology in courses must recognize that not every student will easily benefit from its use.

6.3 Media Rich Content

These online educational systems can be improved by methods of dynamically adapting learning content to the knowledge level, and adapting presentation of content to the perception preferences of the learner. Once these preferences have been determined the task is then presenting the same information in a variety of different ways without losing any educational content in the process. It may be possible that if a student is more visual and a series of diagrams are displayed in front of him that he may be deprived of some valuable information that would have been initially presented to him in more textual manner.

The idea then would be to design the system where the content is presented in a way that information is predominantly centered towards a certain learning style and where valuable information is given in added text or otherwise so that valuable information is not lost in the process.

Figure 2 displays various learning styles and their suited corresponding presentation styles. For the sensory and perspective learner the emphasis is on the content of the material. It may be adapted so that they can understand it better maybe by giving better examples. The visual/auditory learning styles preference is that of how the information is presented to them. They would rather a more diagrammatic presentation of information with possible audio content. These include visual as in sights, pictures, diagram and symbols, auditory as in sounds and words.

Active learners tend to retain and understand information best by doing something active with it by discussing or applying it or explaining it to others. They also tend to like group work more than reflective learners. They should then be given the chance here to join discussion groups and provide feedback from them on a given topic; this will make the experience more enjoyable for them. They should also be encouraged to join a chat room to discuss various topics with other students. Reflective learners prefer to work alone so an avoidance of group work for these

learners would be more beneficial. Information presented to these learners should involve diagrams and text and assignments that involve research.

Most college courses are taught in a sequential manner. Sequential learners may have difficulty following and remembering if information jumps from one topic to another. To present information to a sequential learner then it may be necessary to fill in these steps for them or let them fill them in themselves or by presenting them with references. When they are studying, they need to take the time to outline the lecture material for themselves in logical order. In the long run doing so will save them time.

Inductive learners prefer to learn a body of material by seeing specific cases first for example observations, experimental results, numerical examples and then work up to governing principles and theories by inference. Inductive learners prefer less structure in their presentations. Information can be prepared for them in a less structured manner but allows them to work out solutions to given problems so that they do a lot of the work for themselves as they prefer.

Deductive learners prefer to begin with general principles and to deduce consequences and applications. Since deduction tends to be more concise and orderly than induction, students who prefer a highly structured presentation are likely to prefer a deductive approach. Research shows that of these two approaches to education, induction promotes deeper learning and longer retention of information but that most college science instruction is exclusively deductive probably because deductive presentations are easier to prepare and control and allow more rapid coverage of material.

9 Methodology

The proprietary products that will be used to undertake this project will be WebCT and Macromedia Breeze. Using Macromedia Breeze course content will be created in the familiar PowerPoint environment and automatically convert it into rich media experiences through Flash. Before the content is uploaded to WebCT for delivery to the students, each students learning style must be determined. To determine these learning styles a questionnaire similar to that of the Myers Briggs Type Indicator must be filled out by all students. The questionnaire will be in the format of around 12 questions that the student will be expected to answer prior to taking the online course. The questionnaire will be made available online for the student when they log on to the course.

Preferred Learning Style	Corresponding Preferred Presentation
Sensory } perception Intuitive	Concrete } content Abstract
Visual } input auditory	Visual } presentation verbal
Inductive } organization deductive	Inductive } organization deductive
Active } processing reflective	Active } student participation passive
Sequential } understanding global	Sequential } perspective global

Figure 2: Preferred learning styles and their corresponding preferred Presentation

The style of question will be in the following format:

Visual Modality

often seldom

I remember information better if I write it down

☒ ☐

Looking at the person helps keep me focused

☒ ☐

I need a quiet place to get my work done

☐ ☐

Auditory Modality

often seldom never

My papers and notebooks always seem messy.

☐ ☐ ☒

Pages with small print or poor quality copies are difficult for me to read

☐ ☐ ☒

Kinesthetic/Tactile Modality

often seldom never

I start a project before reading the directions.

☐ ☐ ☐

I prefer first to see something done and then to do it myself

☐ ☐ ☐

I have a difficult time giving step-by-step instructions.

☐ ☐ ☒

A score of 21 points or more in a modality indicates strength in that area. The highest of the 3 scores indicates the most efficient method of information intake. The second highest score indicates the modality which boosts the primary strength. For example, a score of 24 in the Visual Modality indicates a strong visual learner and such a learner would benefit from text, filmstrips, charts, graphs etc. If the second highest score is auditory, then the learner would benefit from audio tapes and lectures as well to supplement their learning. Furthermore, if your second highest score is kinesthetic/tactile, then taking notes and rewriting class notes will reinforce information.

Once these various styles have been determined then the content of the material can be adapted in a way that best suits the individual's style and in a way that will keep him/her interested in the material. The idea is to adapt the system to each learning style. For example if a student is a more visual type of learner the information can be presented to them in a more graphical format e.g. diagrams and charts. It is the designer and teacher that will predominantly be responsible for adapting the information in this case. WebCT does allow you to adapt the interface and contains a dyslexia screen if needed. As for the format of information that is presented to the student the responsibility lies with the teacher. Students will then be observed to view their progress as a result of this change in information presentation.

10 Summary and Evaluations

This paper has described how distance learning has become an integral part of education systems today. It has also discovered that there are certain problems that exist in this area of education. People have been learning at distance for centuries but not at the forefront of education as we know and are accustomed to. As a result of this there has been a high drop out rate within this form of education. One important reason that has been discussed is the area of preferred learning styles. Many learning styles have been covered and how information can be adapted to suit them. The outcome then is, to determine the learners preferred style and present information to them in an appropriate form. Learning styles will be discovered using a questionnaire designed specifically to determine a user's learning style. The system presented to the learners can then be altered as needed to suit them.

A message to communicate, the ability to write, the proper tools, and an infrastructure to deliver writings are the only requirements for distance education. By reinforcing the system, providing the students with a more adaptable system could prove invaluable for this type of education to expand.

The goal of this project then is to observe eLearning and distance education and to determine the relevance between the design of the system and the significantly successful outcome of presenting different system designs based on the learning styles of each of the users.

It is suggested that the identification and use of learning styles and learning strategies combined with adaptive systems can help facilitate the achievement of the goals of distance education and eLearning. These goals hope to include a lower than average drop out rate. By providing these systems adjusted according to the different needs of different users is hoped to do just that. Such differences present a profound challenge for instructional designers and it is hoped that through further research that the quality of learning material is enhanced if the material is designed to take into account learners' individual learning styles.

References

- Kirkwood, A. (1998).** New media mania: Can information and communication technologies enhance the quality of open and distance learning? *Distance Education*, 18(2), 228-241.
- Australian Journal of Educational Technology**1999, 15(3), 222-241.
- Riding, R. J., & Rayner, S. (1995).** The information superhighway and individualized learning. *Educational Psychology*, 15(4), 365-378.
- Rasmussen, K. L. (1998).** Hypermedia and learning styles: Can performance be influenced? *Journal of Multimedia and Hypermedia*, 7(4), 291-308.
- Johnson, James.1999.**"The Thread of a Great and Long Tradition" Vol 1, No.1, pp. 9-12.
- Grasha, A. F., & Yanguarber-Hicks, N. (2000).** Integrating teaching styles and learning styles with instructional technology. *College Teaching* 48 (1): 2-13.
- Sherry, L. (1996).** Issues in Distance Learning. *International Journal of Educational Telecommunications*, 1 (4), 337-365.
- Grasha, A. (1996)** *Teaching with Style*, Pittsburgh, Alliance.
- Lawrence, G. (1982).** People types and tiger stripes. [Second edition].
- Kolb, David(1984).** *Experiential Learning: Experience as the Source of Learning and Development*. Prentice-Hall, Englewood Cliffs, NJ, 1984.
- Wilson, R.C.(1986)** "Improving Faculty Teaching: Effective Use of Student Evaluations and Consultants." *Journal of Higher Education* 57:196-211; 1986.
- Frankola, K. (2001).** Why online learners drop out. *Workforce*, 80, 53-60.
- Jacquelyn Abromitis.** (2001) Literature Review: Learning styles of distant Educators
- Boeree, Dr. C. George (1997).** Carl Jung 1875-1961.
- Quenk, Naomi L. & John Wiley (1999)** Essentials of Myers – Briggs Type Indicator: Assessment.

Consistency of Academic Performance in Higher Education: A Study of an Irish Business Degree Programme.

Julie Byrne¹ and Conor Heagney²

¹Lecturer, School of Business and Humanities, National College of Ireland.

²Senior Lecturer, School of Business and Humanities, IADT Dun Laoghaire.

Abstract

This study examines the level of consistency of the academic performance of business students, both as a whole and within two fields of study – accounting and human resource management (HRM). The examination results of 177 students are ranked at different stages and compared with the rank of final year exam results. By utilising Spearman's (1904) coefficient of rank order correlations rather than absolute marks, this paper hopes to facilitate the process of comparison. The research found that the level of consistency increases substantially once students enter the degree irrespective of their field of study.

Introduction

This study aims to examine the level of consistency of the academic performance of business students, both as a whole and within two fields of study – accounting and human resource management (HRM). The research focuses on the consistency of that majority of students who have completed one business degree programme, on schedule, in the National College of Ireland. This paper examines the relationships between academic performance at entry qualification level and within the business degree. Assessment results in the final year of the degree are ranked and then correlated with the ranked assessment results of previous years in the degree as well as with entry qualifications. For the purposes of this paper the level of correlation between different stages indicates the level of consistency of academic performance. In the short-term it is hoped that this will provide students and those interested in academic performance with an overview of relative academic performance which may stimulate reflection. Further research establishing trends in consistency across courses, institutions and countries may provide a basis for subsequent targeted examination of the underlying causes of these relationships. By utilising rank order correlations rather than absolute marks, this paper hopes to facilitate comparison of academic consistency nationally and internationally. Thus, in time, it is hoped that a national or international picture of performance relationships will indicate fruitful avenues of investigation into the causes of such performance. Much of the literature suggests some degree of consistency between academic performance but that the level of consistency varies according to field of study. Although there is a body of research on academic performance in the field of accounting (Bouillon and Doran, 1991; Clark, 1995, Koh and Koh, 1999; Gammie et al, 2003), there is relatively little available on academic performance in the field of human resource management (HRM).

BACKGROUND

The National College of Ireland offers courses in Business, Computing, Informatics and Community Development. It is a state-funded institution with awards conferred by the Higher Education and Training Awards Council (HETAC). The BA in Accounting and HRM (the degree) commenced in 1991 and is a three year full-time course providing a general business background for a specialised education in either the management of the financial or human resources of organisations. The course is comprised primarily of students entering directly from second level education with up to 30% of places in total reserved for mature and transfer students as well as students from disadvantaged backgrounds. The degree is accredited by HETAC and has examination exemptions from both the HRM and accounting professional bodies. In year one students study six subjects in total. In years two and three students must select either the accounting or HRM stream. They then study two common subjects and three specialist subjects.

Literature Review

This section examines the nature of the relationship between earlier academic performance and degree performance. This relationship is further analysed by field of study.

Degree Performance and Earlier Academic Performance

Much of the research in this area demonstrates some degree of correlation between final year degree performance and earlier academic performance. The nature of these relationships seems to vary across fields of study and institutions. Early work by Sear (1983) shows a small but significantly positive correlation (0.3) between A level score and final year degree result. Peers and Johnston (1994) meta-analysis of previous studies in the area found that A level and university and polytechnic final year degree performance display a small but significantly positive relationship correlation of 0.28 overall. Peers and Johnston (1994) concluded that success in final year degree examinations is related to success in entry qualifications but is also influenced by other factors such as learning approach and environment. In reviewing literature from Thomas, Bol, and Warkentin, (1991) and Entwistle and Entwistle (1991) they suggested that a close match between entry qualifications and higher education performance is not to be expected or even desired given the expected development of conceptual understanding and more mature study habits at higher education. In American research, House (2000) found that high school class percentile rank and ACT (American College Testing) score were significant predictors of first year degree performance among science, engineering and maths students. In Irish research Moran and Crowley (1979) identified that the pass rate in first year increases monotonically with performance in Leaving Certificate with clear cut differences between students with low and high scores and between different fields of study. The Points Commission (Lynch et al, 1999), found a clear relationship between Leaving Certificate Grade

Point Average (LCGPA) and performance in higher education at first and final years of award. However, this relationship is not linear with the LCGPA of first class award students slightly below that of second class honours award students. This study also indicates that the relationship is indirect rather than direct with a number of factors mediating the relationship including institution type, field of study and gender.

Academic Performance Differences between Fields of Study

International research indicates that the consistency of a student's degree performance can vary widely across field of study and across institutions even within the same field of study. Most of this research centres on Arts, Humanities, and Science. These fields of study will be reviewed to establish a broad context for the results of this research in the accounting and HRM fields of study. Within the field of business management education, there has been some research on performance within the accounting field of study. However, there is little evidence of research on educational performance in HRM or Personnel Management.

Peers and Johnston (1994) identified that the relationship between entry qualifications and final year degree performance differs according to field of study. The correlation is higher for medicine, science, language and engineering (0.3 – 0.36) and lower for social sciences, architecture and arts (0.15 – 0.25). Chapman, K. (1996) also found that this relationship varied with the field of study with the strongest correlation for biology (.47) and the lowest for politics (0.23). The correlation for accounting in this study was 0.35. Irish research by Moran and Crowley (1979) on the link between Leaving Certificate results and first year degree performance identified engineering as the field of study with the highest correlation coefficient (0.714). In research by Lynch et al (1999) students in the university and college of education sectors with roughly identical LCGPA were more likely to be awarded a first or upper second class degree in the science field than in humanities. However, in the institutes of technology, students in the humanities were awarded the highest grades. While LCGPA was higher for students in the technological field than in science, a higher proportion of science graduates were awarded first or upper second degrees. Also within Business subjects in the University sector, students awarded a pass or third class degree had noticeably higher LCGPA than those awarded a middle or lower second award. Clarke (1995) found that prior study of accounting was the most significant determinant of first year accounting performance but of lesser importance in second year performance. CAO points were not important in terms of pass/fail classification but were important in explaining overall accounting scores in year one.

Description of Research and Methodology

This section of the paper specifies the research hypotheses, the method of data collection and describes the group and sub-groups used for the study. The section also explains how academic performance was measured and the statistical approach adopted.

Research Hypotheses

The research hypotheses are centred on the relationships between three *explanatory* variables and a *dependent* variable. Each explanatory variable is considered separately (univariate analysis). The null hypothesis is that there is no relationship between each of the explanatory variables and final year rank performance for the whole group or the accounting and HRM groups.

Hypothesis One - The Whole Group.

There is no relationship between final year degree performance and:

- (a) Leaving Certificate performance, or
- (b) Year One performance, or
- (c) Year Two performance.

Hypothesis Two - The Accounting Group.

There is no relationship between final year degree performance and:

- (a) Leaving Certificate performance, or
- (b) Year One performance, or
- (c) Year Two performance.

Hypothesis Three - The HRM Group.

There is no relationship between final year degree performance and:

- (a) Leaving Certificate performance, or
- (b) Year One performance, or
- (c) Year Two performance.

Data Collection

The data was collected from the CAO files (Leaving Certificate data) and from the College's examination files (years one, two and final year results data). Where the whole group has been sub-divided on a field of study basis, the performance rankings are based on rankings within each field of study sub-group. The data was processed using SPSS for Windows.

The Group

The group under review is comprised of students who entered the degree by means of the CAO system and who attempted the final year on schedule. Accordingly, mature, transfer and other

‘non-standard’ entry students and students who repeat a year are not included in this group. The range of CAO points for this group is 305 – 445. Of those students in the accounting sub-group, 83% studied accounting for the Leaving Certificate. Three cohorts who entered the degree in the years 1995, 1996 and 1997 and had completed a full cycle were selected for this research. As the degree encompasses two fields of study; accounting and HRM, the group is also split into two field of study sub-groups. Although students do not choose their field of study until the beginning of year two, the split of the whole group into field of study groups has also been applied to the entry qualifications and year one academic performance variables. A summary of the cohort and group sizes follows in **table 1**.

Table 1: Cohort Student Numbers by Group

Field of Study/Cohort	1995	1996	1997	Total	%
Accounting	34	32	26	92	52%
HRM	45	24	16	85	48%
Total	79	56	42	177	100%

Measurement of Academic Performance and Ranking

Academic performance in the Leaving Certificate is based on the CAO points total of the student. Academic performance in year one of the degree is based on the mean of the results achieved by the student in each of the six individual subjects for that year. Academic performances in the second and final years of the degree are based on the mean over five subjects. Where students have attempted the Leaving Certificate or any degree year exams more than once, only the first attempt has been taken. The academic performance measurements are converted into rank values within the whole cohort and within the field of study group (accounting / HRM) of that cohort. Accordingly, each student has two rank values i.e. a whole cohort ranking and a field of study cohort ranking. The CAO points totals of the students in each of the three cohorts are ranked with the highest total being ranked as one. The mean results of the students in years one, two and the final year in each of the three cohorts are ranked with the highest total being ranked as one. In ranking performance, students with the same performance level constitute ‘ties’. If two students had the same CAO points total, the rank value allocated to each of these students is the mean of both ranks applicable to that level of performance. For example, the ranks applicable to the top two students are one and two so the mean rank of 1.5 is allocated to both students.

The decision to translate academic performance into rank format is important in the context of this paper. The use of ranked data involves the loss of detail (in that the rank does not reflect the actual distance between each point on the ranking scale) and also limits the range of

statistical methods that can be applied to the data. There are two reasons for using ranked data. Firstly, the use of ranked data facilitates the comparison of performance across the College and Leaving Certificate systems. In addition, ranked performance data facilitates the comparison of the degree correlations with those of other higher education institutions either within or outside Ireland even though such institutions may have different assessment systems. Secondly, applying the principle of Adam's Equity Theory (Moorhead and Griffin, 1998: 145) we can conclude that students attach importance to relative performance. This theory suggests that people are motivated when they believe they are being treated fairly in relation to others. In an academic context, this means that students compare their inputs e.g. study time and their outcomes e.g. assessment results, with those of another student. If the ratio of input to outcome is perceived to be the same for both people, the individual perceives the situation to be fair and is motivated to continue inputting. The student perception of relative performance is often reinforced by ranking-based selection decisions of organisations recruiting graduates.

Statistical Approach

As the performance data is expressed in rank form, it is ordinal in nature. Accordingly non-parametric tests are applicable (Foster, 1998). The non-parametric test applied in this study is Spearman's (1904) coefficient of rank-order correlation subsequently referred to as 'Spearman'.

Descriptive Statistics

The CAO points mean and standard deviation for the accounting group is a little bigger than those of the HRM group. However the differences between the means, standard deviations and ranges of the whole group and of field of study sub-groups are not substantial. The accounting group performs quite a bit better than the HRM group at the CAO stage as the accounting group have, on average, seven more CAO points than their HRM counterparts. This difference reduces a little as students progress through the degree.

Table 2: *Group CAO Points Statistics*

Group	Group Size	Mean	Standard Deviation	Minimum	Maximum
Accounting	92	374	24	305	445
HRM	85	367	23	320	445
Whole Group	177	370	24	305	445

Results of the Research

Hypothesis One – The Whole Group.

The null hypothesis is rejected by Spearman. The correlation values for years one and two are particularly high.

Table 3: Whole Group: Spearman Coefficients

Explanatory Variables	Spearman Coefficient	Significance (2-tailed)	Number in Group
Leaving Certificate	.275**	.000	177
Year One	.681**	.000	177
Year Two	.744**	.000	177

** = Correlation significant at the .01 level (2-tailed).

Hypothesis Two – The Accounting Group.

As for the whole group, the null hypothesis is rejected in each case for the accounting group by Spearman. The accounting group shows a marginally stronger relationship for the Leaving Certificate variable than for the whole group. The relationships for years one and two for the accounting group are similar to those for the whole group.

Table 4: Accounting Group: Spearman Coefficients

Explanatory Variables	Spearman Coefficient	Significance (2-tailed)	Number in Group
Leaving Certificate	.340**	.001	92
Year One	.651**	.000	92
Year Two	.785**	.000	92

** = Correlation significant at the .01 level (2-tailed).

Hypothesis Three – The HRM Group.

As in the cases of the whole group and the accounting group, the null hypothesis is rejected for the HRM group by Spearman. However the Leaving Certificate correlation for the HRM group is weaker than for the accounting group. The correlations for years one and two for the HRM group are similar. The relationship values for these years are quite strong.

Table 5: HRM Group: Spearman Coefficients

Explanatory Variables	Spearman Coefficient	Significance (2-tailed)	Number in Group
Leaving Certificate	.259*	.017	85
Year One	.702**	.000	85
Year Two	.690**	.000	85

** = Correlation significant at the .01 level (2-tailed).

* = Correlation significant at the .05 level (2-tailed).

In summary, all null hypotheses are rejected. The relationship between each of the explanatory variables and final year ranking is statistically significant in all cases. The relationship strengthens when the student enters the College and continues to get stronger within the College

i.e. from year one to year two, with the exception of the HRM group. The relationships between the explanatory variables and final year ranking are stronger for the accounting group (compared to the HRM group) with the exception of year one.

Discussion and Implications of Research Findings

In this section the findings and implications of the research will be discussed firstly by whole group and then by field of study sub-group. Within each group the discussion is split between entry qualifications and interim degree performance. The findings should be considered within the context of the sample which was selected from one business degree course within one higher education institution. The sample size was 177 and the CAO points range was 300 to 445.

Whole Group - Final Year Degree Performance and Entry Qualifications Performance

As indicated by previous studies (Sear, 1983; Peers and Johnson, 1994; Lynch et al., 1999; House, 2000) this study also found a statistically significant positive relationship between entry qualifications and final year degree performance (based on ranking of assessment results). With a Spearman coefficient of .275 this relationship is similar to that reported by Sear (1983) and Peers and Johnson (1994). However, this level of correlation supports Lynch et al.'s view (1999) that entry qualifications are a far from perfect predictor of performance in higher education. In the light of Peers and Johnson's (1994) comments about the expected change in learning approach from secondary to higher education, this level of correlation seems appropriate. In other words as learning approaches change from one system to the next we would not expect a very strong relationship between performances in both systems. This finding may be of interest to Leaving Certificate students, their families and their teachers. Although Leaving Certificate results are used to gain admission to the degree, performance in the Leaving Certificate is not necessarily reflected in final year degree performance. On this basis Leaving Certificate students should not assume that their Leaving Certificate performance will be consistent with their final year degree performance.

Whole Group - Final Year Degree Performance and Interim Degree Performance

This study also found a strong and significantly positive relationship between interim and final year degree performance. The Spearman coefficient for year one is .681 and for year two is .744. Interim degree performance appears to have a stronger relationship with final year degree performance than entry qualifications do. In essence, the closer the student gets to final year performance the stronger this relationship gets. This increase in the strength of the relationship may be expected for a number of reasons. Firstly, there is a consistency in the measurement of academic performance as each of the stages occurs within the same higher education system.

Secondly, in line with Peers and Johnson's (1994) reasoning, both interim and final year degree performance are based on the development of similar approaches to learning. Thirdly, as students are pursuing a programme of study that they selected it is possible that a 'drive' or motivation factor is relevant.

Between years one and two, there is a number of changes e.g. the subjects, an increase in subject specialisation in the accounting/HRM areas, the number of subjects, class sizes (smaller as split into fields of study), age of student (one year older) etc. However, it is interesting to note that despite these many changes the level of correlation is broadly similar across years one and two. The strong relationship between performance in years one/two and the final year would appear to question the perception of some students that all that matters in years one/two is progressing to the following year. This may give students cause to reflect on their relative performance in years one and two, not just in final year.

Field of Study Sub-Groups - Final Year Degree Performance and Entry Qualifications Performance

Peers and Johnston (1994) identified the correlation between entry qualifications and final year degree result as relatively high in medicine, science, language and engineering (0.3 - 0.36). The correlation for the accounting group is within this range. It is noteworthy that the low correlations (0.15 – 0.25) in Peers and Johnston's study were observed in the social sciences, architecture and arts. It is interesting to note that the HRM correlation (.259) is just outside this range. Chapman (1996) computed a 0.35 correlation for the accounting field of study regarding entry qualifications and final year degree performance. The Spearman correlation coefficient for the accounting group in this study is 0.340 which is similar to Chapman's findings. The level of correlation between entry qualifications and final year performance is higher for the accounting group than for the HRM group i.e. .340 (accounting) as against .259 (HRM). As 83% of the accounting group studied accounting for entry qualification purposes, it is possible that the presence of accounting and the absence of HRM in the entry qualifications curriculum may contribute to part of this difference. As in the case of the whole group, Leaving Certificate students, irrespective of their field of study, should not assume that their Leaving Certificate performance will be consistent with their final year degree performance.

Field of Study Sub-Groups - Final Year Degree Performance and Interim Degree Performance

As in the case of the whole group, each field of study group has much stronger correlations for interim degree performance than for entry qualifications. It is interesting to note that while the

interim degree relationship strengthens for the accounting group (.651 to .785), it remains similar for the HRM group (.702 to .690). Although the levels of correlation differ as between the fields of study, these differences are not substantial. Thus it appears that the choice of field of study, within this degree, does not have a major impact on the relative performance relationships.

CONCLUSION

In conclusion, this study examined the level of consistency of the academic performance of business students within one degree course in one Irish higher education institution. The level of consistency was examined for the whole group and within two fields of study sub-groups – accounting and HRM using Spearman's coefficient of rank-order correlation.

Using rank order correlation, it was found that a weak but statistically significant relationship exists between Leaving Certificate performance and final year degree performance. The relationships between interim degree performance and final year degree performance are much stronger. The above relationships are broadly similar for the accounting and HRM groups within the degree. Thus, it appears that the level of consistency increases substantially when students enter the degree irrespective of their field of study. It is hoped that this research develops into a longitudinal study incorporating a new cohort each year. In this way the results should become more robust with time. In addition, as the sample size increases, the possibility of analysing the data by reference to age, gender and mode of entry may become feasible. It is also hoped to develop the research further by fostering cross-institutional studies in this area through collaboration with colleagues in other higher educational institutions. The use of rank order correlation should facilitate this process.

REFERENCES

- Bouillon, M.L. and Doran, B.M. (1991)**, 'Determinants of Student Performance in Accounting Principles I and II', *Issues in Accounting Education*, Vol. 6, No. 1, pp 74-85.
- Chapman, K. (1996)**, 'Entry Qualifications, Degree Results and Value-Added in UK Universities,' *Oxford Review of Education*, Vol. 22, No. 3, pp. 251-264.
- Clarke, P.J. (1995)**, 'Some Determinants of Student Performance in University Accounting Examinations', *The Irish Accounting Review*, Vol. 2, No. 1, pp. 49-68.
- Entwistle, N.J. and Entwistle, A. (1991)**, 'Contrasting Forms of Understanding for Degree Examinations: the student experience and its implications', *Higher Education*, Vol. 22, No. 3, pp. 205-228.
- Foster, J.J. (1998)**, *Data Analysis Using SPSS for Windows, Versions 8 to 10 (1st edition)*, London : Sage Publications
- Gammie, E., Jones, P.L. and Robertson-Millar, C. (2003)**, 'Accountancy undergraduate performance: a statistical model', *Accounting Education*, Vol. 12, No. 1, pp. 63-78.
- House, J.D. (2000)**, 'Academic background and self beliefs as predictors of student grade performance in science, engineering and mathematics', *International Journal of Instructional Media*, Vol. 27, No. 2, pp. 207-21.

- Koh, M.Y. and Koh, H.C. (1999)**, 'The determinants of performance in an accountancy degree programme', *Accounting Education*; Vol. 8, No. 1, pp 13-29.
- Lynch, K., Brannick, T., Clancy, P. and Drudy, S. (1999)**, *Points and Performance in Higher Education: A study of the Predictive Validity of the Points System, Research paper number 4 for the Commission on Points System*, Dublin:Stationery Office.
- Moorhead, G. and Griffin, R.W. (1998)**, *Organisational Behaviour – Managing People and Organisations*, New York: Houghton Mifflin.
- Moran, M.A., and Crowley, M.J. (1979)**, 'The Leaving Certificate and First Year University Performance', *Journal of the Statistical and Social Inquiry Society of Ireland*, Vol. 24, No. 1, pp. 231-266.
- Peers, I. and Johnston, M. (1994)**, 'Influence of Learning Context on the Relationship Between A-level Attainment and Final Degree Performance: A Meta-analytic Review,' *British Journal of Educational Psychology*, Vol. 64, pp.1-18.
- Sear, K. (1983)**, 'The Correlation Between A Level Grades and Degree Results in England and Wales,' *Higher Education*, Vol. 12, pp. 609-619.
- Spearman, C. (1904)** 'The Proof and Measurement of the Association Between Two Things'; *American Journal of Psychology*; Vol. 15, pp 72-101.
- Thomas, J., Bol, L. and Warkentin, R. W. (1991)**, 'Antecedents of college students' study deficiencies: the relationship between course features and students' study activities', *Higher Education*, Vol. 22, No.3, pp. 227-250.

The Value of Enterprise: Overcoming Cultural Barriers in Northern Ireland

Julie Byrne¹ and Sharon McGreevy²

¹ Lecturer, School of Business and Humanities, National College of Ireland

² Head of Department, School of Business and Humanities, IADT Dun Laoghaire

Abstract

This paper presents some of the evaluation results of a learning intervention which took place in Northern Ireland from 1999-2002. The intervention called the KEY project was funded by the International Fund for Ireland and aimed to bring together young people from conflicting cultural and political backgrounds living in Northern Ireland and the border counties of the Republic of Ireland. The objective was to develop the interpersonal and enterprise skills of the participants by involving them in a range of learning activities. In this way the project hoped to redress the disadvantages of children born into marginalised communities and also to help sustain peace and reconciliation by bringing together young people from different political and cultural traditions. The project involved partnership with selected schools which sent their pupils to four residential enterprise camps. These camps combined traditional classroom methods with outdoor adventure activities and the real-life creation of a business. Participants from different political and cultural backgrounds were grouped for outdoor activities and business creation. The empirical research was gathered using a combination of quantitative and qualitative methods. The findings suggest that the teamwork and focus required in outdoor activities and business creation can successfully overcome the political and cultural barriers that can impede classroom based learning. The effect of the programme on attitudes to those from different backgrounds will be explored as will the difference in attitude between the genders. Finally, the durable effects of the programme will be explored.

Keywords: enterprise education, cultural barriers, peace and reconciliation, Northern Ireland education

Introduction

This paper provides an overview of the content and impact of an intervention that took place in Northern Ireland in the period 1999-2002. This programme, called the KEY project, aimed to redress social disadvantage and overcome religious and cultural barriers through enterprise education. This paper presents the findings of questionnaires and interviews conducted with the participants on the programme, their parents and teachers.

The KEY Project

The KEY project was a joint initiative between two philanthropic organisations, Young Enterprise Northern Ireland and Junior Achievement Ireland. Junior Achievement which was founded in the USA in 1953 commenced operations in the Republic of Ireland in 1995. It currently provides enterprise education to over 30,000 primary and secondary school children many of whom are living in areas of disadvantage. The sister organisation Young Enterprise Northern Ireland, shares many common characteristics with Junior Achievement. This year alone, Young Enterprise Northern Ireland expects to deliver programmes to in excess of 40,000 school children in Northern Ireland covering over 250 Primary Schools and 80% of all post

primary schools in Northern Ireland. In contrast to Junior Achievement, Young Enterprise Northern Ireland secures 75% of its funding through the public sector. It has been operating in Northern Ireland since 1986. The KEY project, designed and delivered by these two organisations, was envisaged as a supplement to the national educational system for young people aged between 14 and 16 years. Over a three year period 900 young people from partner schools sent their pupils on four residential sessions. They were drawn equally from three communities, Northern Ireland Catholic, Northern Ireland Protestant and Republic of Ireland. The objective was to redress the disadvantages of children born into marginalised communities by raising their aspirations and self-esteem and by teaching them enterprise skills. The project aimed to help sustain peace and reconciliation by bringing together young people from different traditions and breaking the cycle of hostility to those of a different political and cultural tradition. The project combined traditional classroom methods with outdoor adventure activities and the real-life creation of a business. The course was delivered by both KEY project staff and business volunteers.

Locating the KEY intervention

The KEY project can be seen as an intervention targeting economic disadvantage but also bearing some of the characteristics of a peace education intervention. It could be argued at this point that the peace education dimension in the program is implicit rather than explicit. In Northern Ireland the pursuit of social and economic equity is frequently considered in conjunction with religious and cultural differences. Finnegan (1998:1367) reports that policy targeting resources on Northern Ireland's most disadvantaged areas and peoples frequently has the objective of 'reducing community differentials between Catholics and Protestants'.

Although the project has separate objectives as identified above, there is some evidence to suggest that what an individual does when he or she comes in contact with a second culture has an effect on such factors as self esteem and academic performance (Phinney, 1991 and Coleman, 1995 in Coleman, 2001). Thus, the objectives of the programme can be seen as being highly interrelated. There are a number of strategies that target economic disadvantage including efforts to increase the quality and quantity of participation in the labour market. One of the main program initiatives in this area is programmes aiming to develop employability for youth. According to Anderson (1998) such programs tend to include summer programs and out of school youth programs. KEY is an example of such a program and it addresses both general employability as well as enterprise education. According to Hynes (1996: 11) enterprise education may be described as the process or series of activities which aims to enable an individual to assimilate and develop knowledge, skills, values and understanding that are not simply related to a narrow field of activity..

The benefits of camp adventure programs in helping participants to overcome cultural barriers have been documented by Edginton and Martin (1995). The camp provides a laboratory for diversity giving participants the opportunity to explore the customs, commitments history and language of different cultural groups. It can also help combat myths and stereotypes by exposing participants to real interaction and providing the opportunity to gather accurate information on those from different cultural groups. It can also provide diverse role models for participants interacting as they do with counsellors, mentors and tutors from varied backgrounds.

The KEY programme although not explicitly described as peace education by its organiser appears to utilise some of the seven principles guiding a pedagogy of peace suggest by Shapiro (2002). The emphasis in the KEY project on team working and co-operative enterprise building incorporates many of these principles and helps participants to find a way to live together often by focusing on shared tasks in a neutral environment. In this way the KEY programme shares many characteristics with the peace workshops for Jews and Palestinians run by the School for Peace (Feuerverger, 1998:1) which aims to break down the ‘barriers of fear, hate and mistrust that have saturated their daily existence.’

Evaluation Methodology

The KEY project was independently evaluated annually using a combination of qualitative and quantitative methodologies. This paper presents results from three years annual evaluation reports and reflects upon the one of the key objectives of the programme – to sustain peace and reconciliation between the participants from different religions and cultures on the programme.

Due to the nature of the programme, three stakeholder groups were identified for the purposes of the evaluation. These groups consist of the participants, their parents and teachers. Questionnaires were the main method used each year to gather information from participants. In line with the agreement with the funding body, the International Fund for Ireland, the programme intake is divided broadly evenly among Northern Ireland Catholic schools, Northern Ireland Protestant school and Republic of Ireland schools. It is also divided evenly among male and female participants. Each year a c.33% representative sample was taken from the population of participants. This sample completed a pre-test questionnaire on the first morning of their first day of the programme and post-test questionnaire on the last day of their last day on the programme. KEY staff were responsible for distributing and collected these questionnaire and response rates were extremely high varying between 80 -100%.

Questionnaires were designed using relevant indicators from the Life and Times Survey and OECD research. The questionnaires were designed specifically for analysis by SPSS (Statistical Package for the Social Sciences) Version 11, which facilitates large scale data sets and multi-dimensional analysis. Some elements of the pre and post test questionnaires were designed to track attitude change and therefore consisted of similar questions to facilitate comparative analysis. In the first two years of the programme, group interviews were also used to explore the opinions of participants. In the case of the parents, information was sought from a representative sample of parents attending the graduation ceremonies for completing participants held one month after the end of the programme. Short individual and small group interviews were used to elicit their opinions. In the case of the teachers, information was sought from a representative sample of teachers who had attended the programme with the participants. In depth individual and small group interview were used with this stakeholder group. With the permission of parents and teachers, interviews were audio-recorded. These audio tapes were then transcribed and analysed using the outline view function of Microsoft Word. Themes were identified from the transcripts and representative quotes selected.

Results from the KEY Intervention

This section of the paper will present a selection of the quantitative and qualitative results from three key stakeholder sources: parents, teachers and the participants themselves.

Participant Feedback

Quantitative findings show that over the first three years of the programme, on average 76% of participants agreed that the programme gave them a better understanding of those from other religions. Participants also appeared less hostile to connections with those from other religions as shown in tables 1 and 2 below.

Table 1: I wouldn't mind being taught by a teacher of a different religion.

Year	2000	2001	2002
Pre Test	71%	74%	70%
Post Test	86%	89%	70%

Source: Participant Pre and Post Test Questionnaires KEY Project 2000, 2001, 2002

When asked whether they would mind being taught by a teacher of a different religion, there were large shifts in opinion in 2000 and 2001 with a c.15% increase in the number of participants saying that they would not mind such an occurrence. In 2002 there was no movement on this indicator.

Table 2: I wouldn't mind if a relative were to marry someone of a different religion.

Year	2000	2001	2002
Pre Test	71%	71%	67%
Post Test	80%	84%	70%

Source: Participant Pre and Post Test Questionnaires KEY Project 2000, 2001, 2002

Similarly when asked if they would mind if a relative were to marry someone of a different religion there were quite large movements in 2000 and 2001 with a smaller shift in opinion in 2002. One of the reasons for the smaller movements in 2002 appears to be the schools' selection of participants for the programme. In the first two years of the programme, 2000 and 2001, the schools tended to select their more able students and the gender distribution was broadly even. However, as their confidence in the programme grew and their knowledge of the benefits increased, the schools started to select some of their more troubled students who tended to be male. In 2002 the gender distribution was 70% male and 30% female. As the following cross-tabulations reveal, this over representation of males affects the overall averages as the males on this programme consistently appeared less open to attitudinal change in terms of those from different backgrounds.

Gender Cross-tabulations

The relationship between gender and peace and reconciliation was explored by using cross tabulations. Generally the findings from each year suggest that male attitudes are significantly more entrenched than those expressed by female participants and that the females participants seemed more open to living together with those from a different religion.

In general the males started the programme less open than the females to connections with those from a different religion. In addition, at the end of the programme they males also seemed to display less movement in opinion than females with regard to these connections. The following statistics highlight the polarity of attitudes and levels of hostility demonstrated by each group according to their gender. For example in 2001 after completing the programme;

- 73% of males compared with 86% of females strongly agree that they now have more understanding of other religions.
- 43% of males compared with 64% of females would prefer to live in a mixed religion neighbourhood.
- 11.1% of males would mind a little or a lot if taught by a teacher of another religion. Only 2% females who responded stated they would mind a little.

During interviews in 2000 and 2001 participants spoke openly about religion and their perception of other religions. It was clear that many barriers had been broken down during the programme and that many participants had made friendships with people they would not otherwise have spoken to. The findings suggest that the teamwork and focus required in outdoor activities and business creation can facilitate the process of just learning to be together.

Parent and Teacher Feedback

During interviews with parents and teachers a number of common themes emerged which will be illustrated below using quotes. Firstly, the parents interviewed were very positively disposed to the concept of their children meeting and mixing with people from other religions and backgrounds. This was the case even when they admitted that they had not been exposed to the same sort of interaction. In other words many parents who had not had the opportunity to learn to live together were happy that their children were engaged in that process.

'We live in what you might call a ghetto with the peace line at the top of our street. The kids fight and throw stones at each other and I think its great for them to meet each other away from all that and see what that they don't have horns on their head, they're just normal people trying to get on with their lives, find a job, bring up a family. My generation thinks too much about the past, it's right to do this with the kids because they're the future.' (Parent)

'We grew up in a more fearful environment; I still wouldn't go into the city in the evening. I think they're freer and meeting each other like this helps a lot.' (Parent)

'I think (the peace and reconciliation dimension) is great, she had a chance to meet other people and that's where we all need to move on.' (Parent)

'The most obvious thing is that she's more open, more broad minded in terms of dealing with different sides. Now when she sees fighting on telly she says 'I know these people and they're not bad.' She's much more relaxed.' (Parent)

'Students from different areas and different religions normally wouldn't get a chance to come together and experience the good points of working together. It gives them a chance to break down the barriers and see the benefits of teamwork.' (Teacher)

The second theme to emerge from interview with parents and teachers was the development of intercultural friendships and romances on the programme which are being sustained over large geographical distances. For some of these new friends however, meetings can only take place on 'neutral ground'. Just because they as individuals have learned to be together, their immediate communities have not necessarily made the same leap.

'They made tremendous friends and developed their own social and interaction skills among their own peer group and I think they've made friends for life.' (Teacher)

'You can see the effect in some boy/girl relationships and with friends. They phone each other and some of my kids arranged to meet others in Belfast and go to an exhibition.' (Teacher)

'We're a Catholic school and although the other Protestant school is only 200 yards away it's a case of 'never the twain shall meet'. The kids you will hear are rioting are separated by one street. Now they've made friends, they meet on common ground, they still won't go into each other's areas so they go somewhere neutral and they still keep in touch with people they met.' (Teacher)

The third theme points to the softening in hostility as the participants learned literally to live together over the four residential sessions. The teachers were present during these residentials and thus were in an ideal position to observe how the participants responded to meeting people from a different background. They spoke of the role of symbols at the start of the programme and how interaction based around tasks overcame cultural symbols and barriers.

'At the beginning they held back, you could see by the last residential they were very confident. They spoke about religion very openly. We were very pleased.' (Teacher)

'They weren't sitting in their school groups for each meal and they even changed groups between meals and were comfortable. They even told me they'd had discussions about religion ... so there is more understanding there.' (Teacher)

'After the first weekend there was maybe ambivalence towards the group they had met, our kids reckoned that the other kids were different. By the end of the programme, that had turned around completely. The quality of their value judgements about the kids from other schools had matured so much.' (Teacher)

'The wearing of [Celtic and Rangers] football shirts was very important to begin with but then they gave them up.' (Teacher)

The final theme to emerge was concerns about the influence of family and community attitudes. In many cases the participants had learned how to live together on a short term basis but some teachers felt that they were returning to communities which had not necessarily learned the same lesson.

'How far they take it into their communities is another matter. In North Belfast, feelings run very high and some of the kids will say 'Well the ones I know are alright but not that lot.' It's mainly to do with the family and community messages. I hope it will transfer but they need to be strong to stand up to what's going on around them.' (Teacher)

'The children got on really well but the problem is beyond the children. Because of the areas they live in it's difficult for them to utilise it. The programme gets them out of that and gives them an opportunity to interact and be friendly but it can still be hard for them when they go back.' (Teacher)

Interviews reveal that participants clearly leave this programme with a greater understanding of and openness toward friendship with people from different religions and backgrounds. They

discuss their differences on the programme and work together on shared projects and activities. The statistics tracking attitudes in this context however, do not display the radical movement that is evident from the interviews with teachers and parents.

Durability of Programme Benefits

Quantitative data gathered 18 months after the 2000 programme reveal that 77% of participants either agreed or strongly agreed that they had greater understanding of those from other religions and backgrounds. However, participants were less willing 18 months after the programme to accept a teacher of a different religion than immediately after the programme. They were also more inclined to live in a single religion neighbourhood. However, 33 % of participants rated the peace and reconciliation dimension as the most important aspect of the KEY project.

Discussion

Drawing on research from Coleman, Casali and Wampold (2001) these results can be interpreted in the context of a non-linear process describing how the participants in the KEY project dealt with individuals from a second culture. Traditionally it would have been assumed that when faced with having to deal with someone from a different culture, an individual would move in a linear way starting with separation from his/her own culture then onto an acculturation stage and finally ending at assimilation of the other culture. However Coleman et al (2001) suggest that there is a possibility of maintaining involvement with one's culture of origin and developing competence in a second culture. They offer three additional descriptions of second culture acquisition – alternation, integration and fusion. For some KEY participants, discussion and collaboration on the programme with those from a second culture is contrasted with a day to day existence where cultural separation is the strategy supported by the home and peer group. In these circumstances, some participants are clearly utilising the alternation strategy – where he or she associates with two cultural groups but not at the same time. Although the participants clearly have learned how to learn to live together during the programme there is limited evidence to suggest that the integration strategy or fusion strategies are being used to help them live together in their communities. In other words there is limited evidence that they are choosing to have their culture of origin co-exist with the second culture (the integration strategy) or to blend both cultures (the fusion strategy).

Summary and Conclusions

Whilst the KEY project has the stated objective of targeting disadvantage it also has a spin off effect of overcoming cultural barriers through the medium of enterprise education. The

evaluation of this intervention highlights a softening of hostility, an increase in communication and the building of friendships across the traditions promoting greater understanding of people from other backgrounds. The quantitative results in this context however, do not display the radical movement that is evident from the interviews. Throughout the evaluation, male attitudes appeared more entrenched than those expressed by females. Longer-term evaluation suggests that once back in their own cultures there is some hardening of attitudes again perhaps due to the impact of family and community. The neutral and safe environment created by the KEY programme and the focus on specific activities and tasks facilitates participants being together. Over time they learn how to live together in the residential setting provided by the programme with the support of very positive teachers and KEY staff. However, when they return to the more complex home environment they must make extra effort to meet each other and deal with a greater range of negative forces urging them to remain separate. These findings may be suggestive that inter-cultural interventions must be delivered in an incremental on-going basis perhaps in the community if those from different religions and traditions are to learn to live together in a sustainable way.

Bibliography

- Anderson, B. (1998)** Employment and training solutions for the economically disadvantaged: an essay, *The Review of Black Political Economy*, 25(4), 77-84.
- Ashford, R. and Pratten, J. (1999)** Developing Business and Enterprise Skills Through Vocational Training in Schools and FE Institutions: the European model? *Journal of Further and Higher Education*, 23(3), 339-49.
- Cadwallader, A. (2002)** In Northern Ireland, a setback for peace efforts, *Christian Science Monitor*, 94 (225), 7-10.
- Coleman, H., Casali, S. and Wampold, B. (2001)** Adolescent Strategies for Coping with Cultural Diversity, *Journal of Counselling and Development*, 79 (3), 356-64.
- Edginton, C. and Martin, E. (1995)** Camp adventure: promoting cultural diversity, *The Journal of Physical Education Recreation and Dance*, 66(4), 31-2
- Feuerverger, G. (1998)** Neve Shalom/Wahat Al-Salam: A Jewish-Arab school for peace, *Teachers College Record*, 99(4), 692-731.
- Finnegan, M. (1998)** Equity as a policy objective: the case of Northern Ireland, *International Journal of Social Economics*, 25(9), 1367-79.
- Harris, I. (1990)** Principles of peace pedagogy, *Peace and Change*, 15(3), 254-72.
- Hynes, B. (1996)** Entrepreneurship education and training, *Journal of European Industrial Training*, 20 (8), 10-18.
- Shapiro, S. (2002)** Educating against violence, *Tikkun*, 17(1), 44-9.
- Todd, R. (1996)** How are we Irish, how are we British, *Civilisation*, 3(1), 13-5.

Speech Synthesis for PDA

Peter Cahill and Fredrick Mtenzi

*Computer Science Department, School of Computing, Dublin Institute of Technology, DIT
Kevin Street, Dublin 8, Ireland*

Contact email: peter.cahill@student.dit.ie

Abstract

A Text-To-Speech (TTS) synthesiser is a computer-based system that should be able to read any text aloud. This paper presents the design and implementation of a speech synthesiser for a PDA. Our implementation is based on FreeTTS by Sun Microsystems Inc. This paper focuses on the issues that arise during the development and the differences with the desktop synthesiser. We carry out detailed experiments on different platforms to show how the quality and speed of conversion varies. Our TTS implementation on a PDA, apart from being platform independent produces the same sound quality with a far less powerful processor than the desktop synthesiser on which it was based.

Keywords

PDA, Text to Speech (TTS), Speech Synthesis, Mobile devices, J2ME

1. Introduction

Speech Synthesis is a simple idea; a synthesiser inputs text and outputs audio. The implementation of a synthesiser is far more complex than at first glance. Even the theory of the process is quite complex. Ideally, a speech synthesiser should sound like a real person, defining what a real person actually sounds like is difficult as people from different areas have different language and dialect. The program to input the text can even be quite complex as some languages are written in a very different way than others, an example would be a comparison between English, Japanese and Arabic. Today's software speech synthesisers do produce intelligible speech; however, they still do sound very artificial, and can be difficult to understand (Lemmetty, 1999).

Speech synthesis software has existed for about twenty years, as it was then when computer hardware was advanced enough for a real time Text To Speech (TTS) conversion. Hardware limitations that existed were in areas such as processing power, memory and audio hardware. Until about 1990 the cost of hardware that performed speech synthesis was far too expensive for most labs and universities. Recent advancements in computer hardware has lead to a significant increase in speech technology, many universities are now participating in research in this area.

In recent years portable computer hardware has advanced significantly for example three years ago a ~5Mhz CPU would have been seen to be more than enough processing power, where as now the new Nokia N-Gage comes with a 104Mhz processor (Nokia Press Release, 2003).

This advancement in portable hardware has led to the advancement of the applications for portable hardware. Thus bringing forward the possibility that these portable devices are now advanced enough to perform real time speech synthesis.

There are many applications of speech synthesis for portable devices such as a PDA. Some potential applications follow: Email has always been a large problem on portable devices, as the idea behind a portable device is that it should be very small, meaning it will have a very small display. Trying to read emails that were always intended for large displays on a screen that is significantly smaller is extremely difficult. Speech synthesis would allow the portable device to read the email to the user. Another application would be for visually impaired people, who would not normally be able to read emails and SMS messages; speech synthesis could read the text to them. Also, people who have a speech disorder could communicate with people over a phone by entering the text into the speech synthesis system, and the speech synthesis system could output the audio through the phone line. Audio books are becoming increasingly popular as more portable audio players become available. The current disadvantage to audio books is the large amounts of data required to store a book in audio. An example of this is the Lord of the Rings audio book, which is 17 CDs. If consumers could obtain the book as computer text, and allow the computer to read the text to them, it would solve this problem.

Our aim was to develop a speech synthesiser for a PDA. We modified the desktop speech synthesiser project by Sun Microsystems Inc. called FreeTTS (FreeTTS, 2004) so that it can be used in a PDA. This involved rewriting core components of FreeTTS. We tested our application on a PDA with a 400MHz Intel Xscale chip, running Windows CE and the CrEme Java virtual machine. The remainder of the paper is organised in the following fashion. Section two describes the design of the Speech Synthesiser for the PDA. Implementation details and testing are discussed in section three. Summary and results of our experimentation are presented in section four. And section five discusses conclusions and future work.

2. Design of Speech Synthesiser

We originally aimed to develop a speech synthesiser in J2ME. After researching modern speech synthesisers and also from discussing the project with the head of the FreeTTS group, Mr. Walker (Walker, 2004), it became clear that it would be best to convert the existing FreeTTS engine from being J2SE 1.4.1 dependent to be J2ME compatible. The process of converting an existing engine involves a number of steps, rather than the common software design model. Before any changes can be done to the FreeTTS engine, it was necessary to study it so that any modifications made will be done correctly, to avoid introducing bugs into the engine.

J2ME development is done differently depending which configuration is being used. There is a toolkit for the connected limited device configuration (CLDC), including an emulator for the Windows platform. Since the CLDC is far too restricted for a speech synthesiser, it was

necessary to use the connected device configuration (CDC). The CDC does not have the same support as the CLDC (Lamsal, 2003). It is possible to use a PDA emulator, and install a CDC virtual machine on it. This still results in a lengthy process when wanting to run a Java program. The CDC is somewhat similar to the JDK 1.3, although some differences do exist. The design and development of the project can be done in two steps. The first step is to convert FreeTTS to be compatible with JDK 1.3, and the second step is to convert FreeTTS to be J2ME CDC compatible. One of the immediate advantages of using FreeTTS was that we were building on top of a very well designed program. The core elements of the engine are in the 'com.sun.speech.freetts' namespace. Other packages used by the engine go in sub-packages. There are eight sub-packages, and each one manages a single area or type of data. There is also the 'com.sun.speech.engine' package, which is used to provide compatibility with the Java Speech API. Another package used is 'de.dfki.mbrola', which gives support for the MBROLA voices (this package does only work on MacOS X java). A text to speech researcher in Germany, Marc Schroder, added the MBROLA support (Schröder, 2003).

Figure 1 shows the classes, interfaces and their relations in the main package. The relations shown are to show the main relations between the classes, however more relations have been omitted around the packages to keep the diagram legible (Cahill, 2004). The FreeTTS class is the control class, and the main method will output a list of possible arguments if ran without any. The FreeTTS class is used to start the synthesis engine. This can be done in a number of ways, but will generally involve loading text from a file to be synthesised. Text may also be entered from the command line. Different voices can be selected, or it is possible to get FreeTTS to output a list of available voices. As seen in Figure 1, the FreeTTSTime class inherits the FreeTTS class. This class is a standalone class that is used when using the 'alan' voice for telling the time. The InputMode class that the FreeTTS class uses are used to specify different input modes, which can be a file, lines in a file, from terminal input or given text. The only other class used by the FreeTTS class is the Voice class. Voice is used to represent a voice. When initialising an instance of the Voice class, the lexicons and the AudioPlayer class must be set. The AudioPlayer class is the class used to output audio. The Age and Gender classes are used to apply small changes to the voice. The Voice class uses the OutputQueue class to manage the processing queue for an utterance. The other classes used by the Voice class are used to perform the actual speech synthesis itself. The DynamicClassLoader and the VoiceManager classes are used to manage all of the available voices.

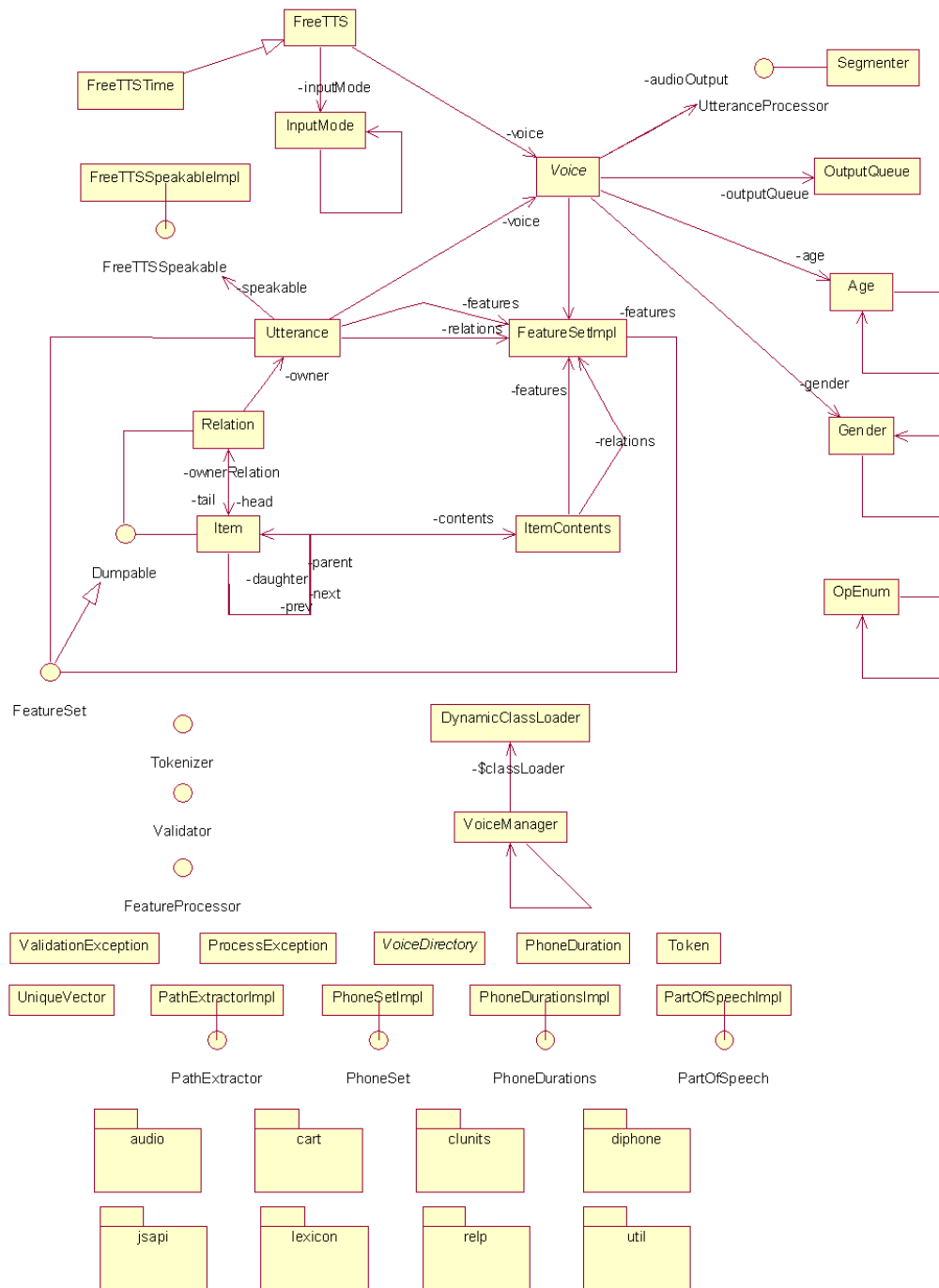


Figure 1: FreeTTS Engine Architecture

The FreeTTS class uses the VoiceManager to enumerate through all of the available voices in the current classpath. The voices in the classpath must be in jar files, where the 'MANIFEST.MF' file identifies them. Other interfaces used include the Tokenizer, Validator and the FeatureProcessor interface. Some interfaces are indirectly used by the FreeTTS package. These interfaces are used and implemented by other classes, which are also indirectly accessed. Due to the amount of classes in the FreeTTS engine, the other classes not mentioned yet exist in sub-packages. There are even more classes than mentioned here, these classes are

nested classes and their functionality is only relative to their respective class. There are eight packages: audio, cart, clunits, diphone, jsapi, lexicon, relp and util. All the package names are straightforward, possibly with the exception of 'relp', which is abbreviated from Residual Excited Linear Predictive decoding (Black, Taylor, Caley & King 1999).

2.1 Stage one

The FreeTTS program requires JDK 1.4.1, but CDC is very similar to JDK 1.3 therefore it was necessary to convert the FreeTTS program to JDK 1.3. Converting the FreeTTS program from JDK 1.4.1 to JDK 1.3 is a complex task. The program will not be functional until the conversion is complete and hence, a single error in the process will probably not be noticed until the late development stages. This is even more sensitive when dealing with audio as a media format; a single error anywhere in the conversion process would most likely result in the synthesiser outputting noise. This makes design and unit testing both very crucial stages. To make the conversion process easier to develop and test, we split the conversion into two stages: the first stage is to convert from JDK 1.4.1 to JDK 1.4.0 and the second stage is to convert from JDK 1.4.0 to JDK 1.3.

2.1.1 Converting to JDK 1.4.0

The main difference between JDK 1.4.1 and JDK 1.4.0 is support for regular expressions. Regular expressions are supported in the 'java.util.regex' package. The package consists of two classes that are not encapsulated within the package, they are: Pattern and Matcher. The Pattern class is used to compile a regular expression pattern, similar to the regcomp() function in the POSIX C library. After compiling a regular expression pattern, then the Pattern class is used to get an instance of the Matcher class, which will match the pattern against a given character sequence. The Matcher class does have some additional features for replacing matches in the source character sequence and to group matches. The regular expression package uses classes that implement the CharSequence interface.

Since regular expression support only became part of J2SE in version 1.4.1, different regular expression projects had existed before this time. Open source regular expression packages have been made by the following organisations: Apache (have made 3 different ones), IBM, GNU and the independently made JRegEx (Samokhodkin, 2002). These engines are not drop in replacements for the J2SE 1.4.1 regular expression engine, but they are similar as they are all based on the same theory of compiling a regular expression pattern and matching it with a search buffer.

2.1.2 Converting to JDK 1.3

The conversion to JDK 1.3 is one of the larger development stages. The differences between JDK 1.4 and JDK 1.3 are somewhat overwhelming at first, and care must be taken when developing at this stage. The regular expression support previously done for it to be JDK 1.4.0

compatible does need modifications, as does all classes that use any of the following functionality:

- Debug Assertions
- Strings and Character Sequences
- Hash Mapping
- Input
- File Input
- W3C Document Model
- Timers

While there are only seven items on this list, what it means is that almost every class in the synthesiser required modifications at some point. Even some very basic datatypes such as Strings and Character handling is done differently in JDK 1.3. Many of the modifications can be done by changing the actual syntax, rather than doing significant changes to the design of the synthesiser. Wrapper classes can be used for the string and character sequence classes to allow core parts of the engine to remain unchanged. JDK 1.3 does support hash mapping with the HashMap class, it is not as efficient as the JDK 1.4 class but does work. After all these parts were been changed, the engine was JDK 1.3 compatible.

2.2 Stage two – converting to J2ME

The process of converting to JDK 1.3 involves changing the majority of the classes in the program. The program was tested before progressing to the J2ME conversion, as debugging on a desktop for a desktop is far quicker and easier than on a desktop for a different platform. The conversion for FreeTTS to be JDK 1.3 compatible is interesting, and does cover a lot about the structure and the development of the Java platform for desktops. Developing for portable devices is significantly different than for a desktop.

The CDC was the target platform for the project. Using the CLDC was ruled out as it does not support enough memory for a voice database, and it does not natively support floating points. The CDC is similar to JDK 1.3 and some programs are directly compatible, however differences do exist between the platforms. A point worth noting about java virtual machines is the virtual machine terminology. The term JVM is used to describe desktop Java virtual machines, CDC Java virtual machines are referred to as CVMs, and CLDC Java virtual machines are referred to as KVMs.

While J2ME CDC itself does not have any memory or processing limitations, most of the operating systems on portable devices do. An example of this is that Windows CE has a memory limit of 32Mb per process (Grattan, 2000). However, this does not mean that a program can use 32Mb of RAM as Windows CE does not have the advanced memory management capabilities that exist for desktop operating systems. This results in unusable memory blocks wasting memory and taking up part of the 32Mb.

The timers in FreeTTS were modified for them to work on J2ME. We expanded on this step so that we added additional features to the timers. Being able to monitor the amount of processing time required by the synthesiser at different tasks is essential. The timing results can give ideas as to what processing steps are requiring the longest time so that they can be optimised further. The timers can also be modified to output formatted time information. The information can be printed to the standard output, and can be formatted with extensive use of the tab character. This allows the outputted time data to be then imported into most spreadsheet programs, resulting in further flexibility in the analysis of the data.

2.3 Design of the Java Sound API

The biggest barrier in developing for the CDC is audio support. Like Java on desktops, J2ME does not natively support sound. J2ME does support predefined beeps, but standard Java does not include any sound support. Sound support does exist for Java on desktops, however it is implemented as an extension (in the 'javax.sound' namespace) and is optional to the standard. This is interesting, as what this means, is that a JVM developer could develop a JVM that would be fully compatible with JDK 1.4 without any sound support, yet it would be capable of getting Sun JVM certification.

There are currently no implementations of the Java Sound API for J2ME. This is mostly because the Java Sound API defines sound output functionality, which is handled differently on all operating systems resulting in platform dependency. This finding resulted in writing a partial implementation of the Java Sound API for J2ME. As the FreeTTS program will be tested on J2ME on Windows CE and the implementation would have to be platform dependant, the Java Sound API implementation would be for Windows CE. Of course a full implementation of the Java Sound API would be an entire project in itself, so the aim was to make a partial implementation that supports sampled audio output. The specification does allow for audio to be written into various file types depending on the implementation. The partial implementation of the Java Sound API was made to support writing to audio files in a platform independent manner. This means that for audio output the implementation must be running on Windows CE, however, if its not being run on Windows CE it will still be possible to output the audio into a formatted 'wav' file. This feature is not included in the implementation of the Java sound API for desktops.

The use of the Java Native Interface (JNI) and a library written in a platform dependant language such as C is required for audio output. Most of the API can be implemented in J2ME, just when it comes to the stage of outputting the sampled audio it is necessary to use a native library to initialise the audio hardware and to stream the audio data from Java into it. The native library needs to use the Windows CE Audio API, which is written in C. The standard way of using libraries with JNI is to develop a dynamic linker library (DLL) that will provide

the platform dependent functionality that originally could not be achieved from Java. Developing with JNI allows the Windows library to have functions that are callable from a Java class. Also, the DLL library can access any of the classes, instances of classes, variable types and variables in the JVM. Accessing the JVM at this level must be done with caution, as there is no error checking. Any methods called presume that all parameters are valid and there is no memory bounds checking.

3. Implementation and Testing

We took an approach of reducing development time by spending time on constructing a customised development environment. The modern integrated development environments (IDEs) are very flexible and can be modified to suit a program like this.

The development environment is the suite of all the tools required for a project. The tools used are going to differ for different projects, and in this project there is a very wide range of tools used. While during the implementation a collection of open source development tools were used, the most important tool was Apache Ant (Atherton, 2004).

Apache Ant is a build tool for Java that was first released in 2000. Ant is often compared to the UNIX make tool, as both programs are the same type of application. Since Ant was developed so many years after make, Ant was developed with the flaws that exist in make in mind. The main difference between Ant and make is that Ant is designed to be platform independent, which is an advantage when using it for Java programs. The other common build tools, such as the different varieties of make are all shell based. They depend on the flexibility of shell tools. Using shell scripts and tools can be very useful, however it does result in the build tools being UNIX specific. Ant is different; it is not based on shell scripts or shell tools, so it is fully platform independent.

The use of Ant does dramatically speed up the build process. The FreeTTS engine consists of approximately 250 class files, in different folders in the package hierarchy. A single Ant 'build.xml' file can be used to compile the entire project, or just any changed files since the last build. After the compiling stage Ant can be used to generate Java archives (JAR files) containing the classes. The use of JAR files does reduce the program size. This size reduction is achieved by it removing the file system file block overhead and also the files are compressed. The same Ant file can also be used to build additional JAR files from the binary databases used for speech synthesis. Files such as the voice databases can be put in a JAR file, as can the lexicon letter to sound (LTS) rules.

Doing most of the testing on a desktop can increase development speed of the project. The time required to download the program to a portable device, load the CVM on the portable device and execute the synthesiser on the portable device is considerable. While if it were possible to perform some testing on a desktop the development time could be decreased. The only desktop

platform that does have a CVM is Linux. However, Java is unsupported on Linux, and does contain bugs when being used for audio related applications. The solution that we found to this problem was to modify an existing JVM to resemble the CVM standard.

All modern synthesisers have been in some way derived from the Festival Speech Synthesis System (Lemmetty, 1999). While the internals of commercial synthesisers is kept within their respective companies, the market leaders that do develop them do partially fund Festival and its related projects (e.g. IBM and AT&T), suggesting that even the commercial synthesisers are related to Festival. The Flite and FreeTTS projects are based on Festival, and are both far more efficient at speech synthesis than Festival (Black & Lenzo, 2003). The difference between these programs and our program is the target platform. This brings forth the approach to change part of the synthesiser so that it can perform better on portable devices.

Flite was always meant to be a lightweight, efficient synthesiser. Considering that Flite has been in development for a number of years, with many authors, one can assume that Flite is reasonably efficient. Memory usage and processing are both minimal in Flite. This means that Flite is already optimised, and with FreeTTS being based on Flite, it does inherit this efficiency.

The synthesis process involves synthesising utterance by utterance. All of the synthesisers we are aware of use the following process: Read next utterance into memory, process utterance, output audio. This process would work fine for real time synthesis on a desktop. On a slower device (e.g. PDA), this will result in there being a long pause at the start of each utterance while the utterance is being processed. While the synthesis would still work, it would not be in real time.

We modified this technique, so that when outputting audio, the raw audio can be copied into a small buffer, where a process thread will play the audio from it. This means that the synthesiser does not need to wait for the audio to be played before processing the next utterance, but instead when one utterance is being played the synthesiser is using the spare system resources to process the next utterance. The result is that the processing wait is made invisible to the user. For example, it takes about 8 seconds to say "Festival was primarily written by Alan W Black, Paul Taylor and Richard Caley". During these 8 seconds the system is almost idle, as playing back raw audio does not require much processing. If the proceeding utterance required seven seconds to process, there would be no need for any pause if the utterance were processed during the 8 seconds the hardware was almost idle. The synthesiser does still process an utterance at a time, resulting in other factors such as prosody and pitch remaining unaffected.

Unit testing is the standard testing approach for any large object orientated program. The theory behind it is that most objects are used to represent a single type of entity in the system. Unit testing involves establishing test cases to be run on the individual entities in the system.

The test cases do generally involve checking that the entity itself is working, and then it will do extensive error checking. Error checking will involve calling methods with invalid parameters, such as null object references and negative integers. We used unit testing at a number of stages during development to ensure that the changes we made worked for the purpose intended. The use of unit testing does not guarantee that the program would fully work, but it did give us insight into bugs we discovered at an early stage.

Much of the conversion process is focused on porting the actual synthesiser engine. The engine will work at this stage, but further modifications should be done to make it more suitable for a PDA. The engine was modified for it to support loading binary FreeTTS lexicon databases. ASCII databases are unsupported due to the large disk space required for an ASCII database. In the case that a user had an ASCII database they wanted to use, the desktop version of FreeTTS does have a program to convert an ASCII database to a binary one that would work on the J2ME synthesiser. The synthesiser supports FreeTTS voices. In addition, FreeTTS does contain a program to convert Flite voices to the FreeTTS format.

While it would be ideal to support Multi-Band Resynthesis OverLap Add (MBROLA) voices, there are currently no PDA's with enough processing power to handle MBROLA voices in real time. This is due to the large processing and memory requirements for them (Bozkurt, 2002). The Java Speech API does not yet exist for J2ME. The Java Speech API version 2 (JSR113) is currently being developed by Sun Microsystems and it is aimed for both J2SE and J2ME.

Memory is a serious problem when using a speech synthesiser on a PDA. Windows CE has a per-process memory limit of 32Mb (Grattan, 2000), while the synthesiser itself would not use this much, the memory is shared with the CVM process. Memory fragmentation and the use of large voice databases can result in Windows CE refusing to allocate more memory, even if the system does have it. We carried out tests on this and Windows CE refuses to allocate more memory when approx 22Mb has already been allocated.

4. Summary and Results

After developing a platform independent speech synthesizer for J2ME, we took carried out tests on the time it spent doing the individual tasks. Specifically we measured the time spent on audio processing. Our improvements in the audio output technique allows for real time speech synthesis on much slower processors than the desktop synthesisers require.

Figure 2 and Figure 3 are comparing the time spent by a 400Mhz PDA when performing the two most complex audio tasks. It can be seen that our approach results in the PDA (which is using our audio classes) being more efficient than the desktop (which is using the very same synthesiser but with Suns audio classes). We also performed further testing to analyze the outputted sound quality. We used FreeTTS v1.2 to write a given utterance to a formatted 'wav' file. We then used our J2ME synthesiser on the PDA to output the very same utterance. After

recording the two samples, we used the UNIX 'diff' command to verify that the files were identical. This result showed that the audio quality outputted by our J2ME speech synthesiser was identical to that outputted by the FreeTTS program on a desktop computer.

5. Conclusions and future work

A design and implementation of a TTS for a PDA was carried out using J2ME. Testing clearly demonstrates that the quality of the sound produced by a PDA is identical to that produced on a desktop. However, the PDA is much more efficient in real time speech synthesis. We intend to carry out more testing on different types of PDA and Java virtual machines. We have noticed that because of there being different virtual machines for PDA's there are some incompatibility problems between them. Further testing will also give insight to further optimisation possibilities.

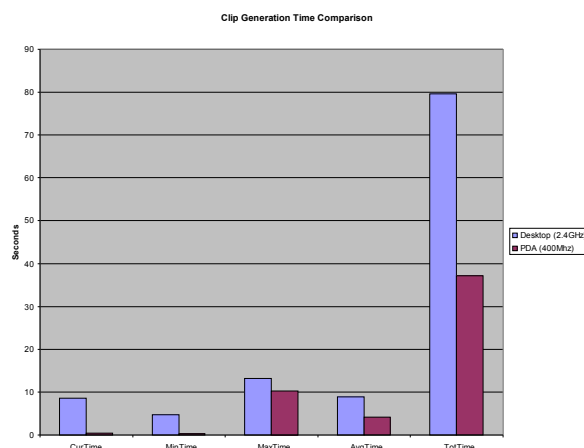


Figure 2: Clip generation time comparison

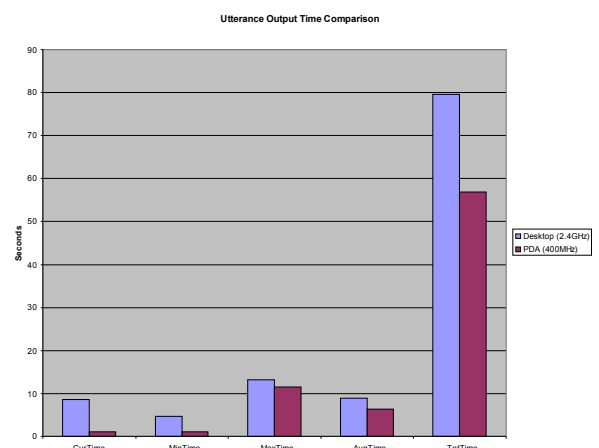


Figure 3: Utterance output time comparison

It will also be interesting to perform the same tests on other platforms such as Sun and Macintosh computers and see how the results differ, especially with the audio output packages. Other future work will include the use of the Java Speech API version 2 and possibly support for MBROLA voices when it becomes feasible. Portable hardware is developing rapidly, and it is most likely that it would be possible to support MBROLA in the near future.

References

- Atherton, B. (2004).** Apache Ant Project, The Apache Software Foundation.
- Black, A. & Lenzo, K. (2003).** Building Synthetic Voices. Languages Technologies Institute, Carnegie Mellon University.
- Black, A. Taylor, P. Caley, R. & King, S. (1999)** Edinburgh Speech Tools Library – System Documentation. Technical report. University of Edinburgh.
- Bozkurt, B. Dutoit, T. Prudon, R. D'Alessandro & C. Pagel, V. (2002).** Improving Quality of Mbrola Synthesis for Non-Uniform Units Synthesis, Proc.IEEE TTS 2002 Workshop, Santa Monica, September 2002.
- Cahill, P. (2004).** Speech Synthesis for Portable Devices. Technical Report. Dublin Institute of Technology.
- Grattan, M. & Brain, M. (2004).** Windows CE 3.0 Application Programming, Prentice Hall, ISBN: 0130255920
- FreeTTS (2004).** Speech Synthesiser written in Java. Speech Integration Group. Sun Microsystems Inc.

- Lamsal, P. (2002).** J2ME architecture and related embedded technologies. Technical Report. Helsinki University of Technology.
- Lemmetty, S. (1999).** Review of Speech Synthesis Technology. PhD Thesis. Helsinki University of Technology.
- Nokia Press Release (2003).** Nokia N-Gage technical specification.
- Samokhodkin, A. (2002).** The JregEx project. Available from: <http://jregex.sourceforge.net/>
- Schröder, M. (2003).** MBROLA Voice support for the FreeTTS engine on Mac OS. DFKI, Saarbrücken, Germany.
- Sun Microsystems (1995).** Sun Announces Three Editions of Java 2 Platform. JavaOne conference, San Francisco.
- Walker, W. (2004).** FreeTTS Programmers Guide. Technical Report. Speech Integration Group, Sun Microsystems Inc.

E-Business - Making the Move from Traditional to Next Generation Marketing Strategies: An Exploratory Study of Irish Organisations

Ethel Claffey

School of Business, Athlone Institute of Technology, Athlone

Contact email: eclaffey@ait.ie

Abstract

In an era in which Irish firms face an increasingly competitive environment, it is essential that they explore the e-business option. Experience worldwide shows that this is now one of the most effective ways to increase sales, productivity and profits, if incorporated effectively into the strategic planning process. While much has been written about the potential impact of the Internet on marketing strategies, very little of the debate has been focused on Irish conditions and practice and on how Irish companies are adapting to the e-business environment.

This paper reviews the diffusion of Internet access and use, with a view to appreciating the current Internet penetration rates around the world and highlighting the importance of this medium for Irish marketing managers. The current state of e-business strategies and the strategic marketing implications are assessed. Also examined is the extent to which 30 successful Irish firms, across a range of sectors, have utilised the Internet and innovative web technologies as part of their marketing activities.

Keywords

E-business/E-marketing, Strategic Marketing.

Introduction

Since 1994 the Internet has evolved from a scientific network into a platform that is enabling a new generation of business and business activities. The e-business approach has broadened the types of business conducted on the web. E-business is about convenience, availability and world-wide reach to enhance existing performance or to create new virtual business ventures (Amor, 2002). Application of the concept offers efficiencies in terms of increased market access and information, as well as decreased operating and procurement costs (Rosen and Howard, 2000). The introduction of the online environment has made real-time, customised, one-to-one advertising, marketing and commerce possible (Chaffey et al., 2000).

E-marketing is relatively new in the business world. It can be defined as 'achieving marketing objectives through use of electronic communications technology' (Stauss & Frost, 1999). Some of the new marketing practices identified in the literature, arising from the ability of e-business technologies, include improved customer benefits such as faster processing of orders, personalisation of consumer data, one-to-one customer interaction and many-to-many communication. Biswas and Krishnan (2003) describe the impact of the Internet on marketing through three dimensions. Firstly, the Internet enables companies to create value through one-to-one marketing. Secondly, it allows cost-

effective integration of electronic trading into a company's core business. Thirdly, the Internet provides sellers with the opportunity to learn more about their best customers, one at a time, thereby making possible the provision of a customised service.

To survive and succeed in today's complex business world all companies – from established industry leaders to feisty upstarts – must develop a strategy that embraces and takes maximum advantage of the latest trends in technology. Yet, many organisations have been slow to exploit e-business technology. Some have implemented focused e-business strategies to build cutting-edge enterprises that serve and retain customers. Others, unfortunately, are lured into ill-fated ventures by the ever-changing roster of 'buzzwords', 'fads' and analogies.

Internet Usage

That the world has been changed by the Internet is beyond doubt. The diffusion of Internet access and use has had tremendous implications for marketing strategy and practice.

The penetration of the Internet varies widely across countries. Market research consultants, Nua, reported that there were 605.60 million Internet users at the beginning of 2003, approximately 10% of world population. According to Nua's figures this shows an increase of 25 per cent (approx) from the beginning of 2001. Table 1 breaks down the 605.60 million figure by region. In 2003, Europe had 31 per cent of the online market, Asia/Pacific was a close second with almost 31 per cent, and Canada & the USA ranked third with 30 per cent of the online market.

Table 1 Internet Users Online

Region	Number of Users, Million
World Total	605.60
Africa	6.31
Asia/Pacific	187.24
Europe	190.91
Middle East	5.12
Canada & USA	182.67
Latin America	33.35
Ireland	1.31

Source: Nua.com, 2004.

More recent data suggest that Internet usage rates in all regions is climbing steadily, particularly in Estonia and Slovenia where penetration levels are on par with those in Western Europe. By 2006, it is predicated that around 27 per cent of Internet users in Central and Eastern Europe will go online at least once a month. In 2003, the Internet penetration rate in Ireland was 40 per cent, up from 36 per cent in September 2002. Data has revealed that four out of every ten Internet users in this country made online purchases during the last three months of 2002. Ireland's mobile usage rate also increased by two per cent from September to December 2002, up to 79 per cent. According to the latest figures, 72 text messages were sent per month per subscriber in Ireland from September to December 2002.

IDC consultants predict a rise in ecommerce spending in Central and Eastern Europe. They estimate that B2B will account for 90 per cent and will continue to constitute the bulk of ecommerce spending over the next five years. In 2006, the total ecommerce market should reach USD17.6 billion. The key ecommerce markets in the region are the Czech Republic, Hungary, and Poland, which together constitute 90 per cent of the market's total value. According to a new study by [RoperASW](#) and [AOL Time Warner](#), 45 per cent of European online consumers are expected to buy more products over the Internet in the coming years, as compared to 41 per cent of American consumers. In addition, nearly three out of four Europeans surveyed said they regularly or occasionally use the Internet to research purchases before buying products offline.

In 2002, Enterprise Ireland reported that 84 per cent of Irish organisations had Internet access and that 67 per cent of this group had their own web site. The adoption rate per sector is highlighted in Table 2.

Table 2 Adoption Rate of Internet Access in Ireland per Sector

Sector	% Web site Ownership	% of Sales via Internet
Pharmaceuticals	68	2.3
Engineering	55	0.2
Electronics	76	0.3
Consumer Food	44	0.3
Consumer Products	61	0.2
InfoComms	94	1.5
Finance/Healthcare/Software	95	5.6
Digital Media/eCommerce/Training	89	0.4
Total	67	0.7

Source: Enterprise Ireland, 2004.

The extent of basic Internet access, usage and literacy suggests the possibility of radical change in society and business. Individuals can now acquire previously inconceivable and detailed information from the web and enter dialogue with businesses in a manner that revolutionises what was traditionally an asymmetrical power relationship. Murray and Ko (2002) argue that the Internet and Information Technology are having a profound impact on society and on the practice of management and marketing globally. In the field of marketing, their effect has been to create a new digital environment for markets, sometimes referred to as the Cybermarket, in which buyers, sellers, their intermediaries and agents interact in novel ways.

E-business Strategies - Background Literature

Online channels offer marketers a unique combination of capabilities, including detailed customer data, real-time customer targeting, rapid market feedback, and on-the-fly campaign adjustments. Deighton and Barwise (2001) claim that being first to master and deploy the new technology and media will be critical to the creation of a new cohort of consumer companies for the twenty-first century. The Internet has been described as being a significant global medium for communications,

content and commerce. It is a powerful tool for building relationships with all of a company's communication targets (Fahy & O'Callaghan, 2002). ONS UK (2002) in their study of pioneering companies discovered that active e-business users, on average, report a doubling of sales when entering a new market.

Yet, online marketing performance remains poor. By employing traditional direct marketing practices, many marketers are simply not tapping into the full potential of this online channel. Research conducted by Day and Hubbard (2002) shows that only 30 per cent of senior managers see the impact of the Internet on their ability to manage customer relationships as a major opportunity, while 52 per cent regard it as a minor opportunity, and 13 per cent are of the opinion that it has little impact.

The first phase of e-business was one of rapid growth and change. Business fortunes were made and lost. We are now entering into the second phase where a huge variety of businesses have become more accessible via electronic media. Online goods and services currently available include home banking, holidays and public administration. The e-business revolution has been described as being twofold. Technology has revolutionised the way we can do business, but business itself is only slowly adapting to the new possibilities. Amor (2002) describes this emerging economy as needing a new paradigm, but the process of conversion will take some time to complete. It is his contention that the best IT infrastructure will be useless if the correct business plan is not in place. He divides the Internet presence of an enterprise into six phases:

Phase 1: The company has set up a web page. However, no real structure is provided and there is no search engine or means of communicating with the consumer. Only some company information is provided.

Phase 2: The web site has some structure, including a search engine to seek key words and information about a company, and a method to exchange messages within the company.

Phase 3: Here the company is trying to sell information, goods etc. online, but the system is not connected to the real databases on the company intranet. It is slow, costly, and lacks security.

Phase 4: The web site has a direct link into the legacy systems of the company's intranet, allows retrieval of information from internal databases, and uses secure protocols to transmit data between the company and the customer or another business. Cost saving and profit making should prevail.

Phase 5: Using any device that contains a chip (cellular phone, car, etc.), people are able to connect to your data and transmit or receive the desired information to conduct e-business.

Phase 6: All chip-based devices are interconnected and create one huge information resource. The devices are able to interchange any type of information on an object-oriented level. Applications are transparent to these devices. Users won't know the source of solutions to their problems.

Amor (2002) believes that most companies are somewhere between phases 2 and 3 but are moving towards phase 4. He believes that to succeed in the future it seems inevitable that companies must progress to phases 5 and 6.

Over the last decade, companies have spent substantial amounts of money building web sites and online communication strategies with a lack of customer-centred vision. This has resulted in consumer frustration and the problem of poor customer retention for the company (Philips, 2003). Deighton and Barwise (2001) identify three properties of digital communication which must be mastered: fragmented attention, radical interactivity, and instrumentality. The marketer must struggle with integration across consumers' fragmented vision and attention. Internet conversation has difficulty in generating emotional involvement; it is fundamentally an instrumental medium.

Rayport and Jaworski (2001) postulate that senior management must complete three sequential tasks when considering an online offering:

1. Identify the scope of the offering
2. Identify the customer decision process
3. Map the offering to the consumer decision process.

Smith and Chaffey (2001) note that a key success factor in e-marketing is achieving customer satisfaction through the electronic channel. This raises issues such as whether the site is easy to use, does it perform adequately, what is the standard of associated customer service, and how are physical products dispatched. The new millennium has seen the rise and fall of dot.com firms, coupled with hype and disappointment over what the Internet will bring to business. Problems of embracing or responding to the Internet remain a challenge for managers of established firms. For today's e-business organisation, it may be simple to set up a web presence, but it has proved difficult to create a sustainable online business. Chen (2001) concludes that, gripped by a misunderstanding of competitive dynamics, dot.coms squandered their ample but ultimately limited resources. The dot-com model was inherently flawed: a vast number of companies all had the same business plan of monopolising their respective sectors through network effects. It was clear that even if the plan was sound, there could only be at most one network-effects winner in each sector, and therefore most companies with this business plan would fail. In fact, many sectors could not support even one company powered entirely by network effects (Cassidy, 2002).

However, according to Porter (2001), the greatest impact of the Internet has been its ability to reconfigure existing industries. Those who recognised the potential of the technology at an early stage used it to create new business models. Many entrepreneurs spotted voids in the market that could be filled by using the Internet. Pioneers included:

- Jeff Bezos who created the world's largest Internet-based bookstore – Amazon.com
- Pierre Omidyar who successfully established a virtual auction house – ebay.com
- Michael Yang who set up one of the world's largest search engines on the WWW - Yahoo.com.

Research Objectives and Methodology

E-business, in all its forms, is projected to continue growing at double-digit rates, thereby becoming the fastest growing form of commerce in the world. As Irish organisations enter the era of e-

marketing, the crucial question becomes not whether to incorporate the Internet into the marketing plans, but rather how to build effectively on the success of traditional strategies by capitalising on the strength of the Internet.

This research seeks to make a contribution by examining the e-strategies being employed by a range of successful Irish firms across the financial, pharmaceuticals, engineering, services, healthcare, and retailing sectors. Given the limited literature and complex nature of the issues under consideration, an exploratory approach was adopted for the study. This approach was based on 30 in-depth interviews with senior managers/policy makers from a variety of industry types and sizes. The objective analysis and all data can be referenced; however, due to the confidentiality of certain issues discussed, findings are not accredited to individual companies. Table 3 outlines the business sectors targeted in this research.

The interviewees were selected on the basis of their reputations and their willingness to co-operate. Both B2B and B2C sectors were represented, including those from a range of different sectors. Small-scale qualitative studies suffer from problems in the reflexivity of the design and the generalisability of the research findings. However, this approach was considered necessary in order to gain a broad insight into the current e-business strategies employed by Irish firms, as well as serving to develop directions for further research. The interviews were unstructured, though 10 core themes were used to guide the discussion:

- What is the role of marketing in the organisation?
- What e-business activities does your company carry out?
- What do you see as the greatest benefit(s) of e-business to your company?
- Do you consider e-business as having any limitations?
- How do you rate how effective your web site is?
- Do you have an e-business strategy incorporated into your strategic plan/marketing plan?
- Would you consider using m-business (mobile business)?
- What challenges/threats do you foresee in the e-business environment?

Table 3 Business Sectors Surveyed

Industry Type	Size of Company	Number of Companies
		Surveyed
Financial	Large	2
Financial	Small	1
Pharmaceuticals	Large	2
Pharmaceuticals	Small/Medium	1
Retailing	Large	2
Retailing	Small/Medium	4
Manufacturing	Large	3
Manufacturing	Small/ Medium	4
Insurance	Large	2
Healthcare	Large	2
Services	Small	3
Electronics/IT	Small/Medium	2
Education	Large	2

Findings

The Role of Marketing

Findings from the surveys indicated that only 40 per cent of the participating companies had a marketing person employed at a strategic level. This figure included managing directors involved in conducting strategic marketing planning. However, all of these companies recognised the need for change, with 70% citing that they were planning, or in the process of planning, a strategic marketing plan whereby the recruitment of a marketing person at senior level would be considered. The most effective marketing activity identified by all respondents was the use of personal selling and demonstrations at conferences, trade shows or other similar events.

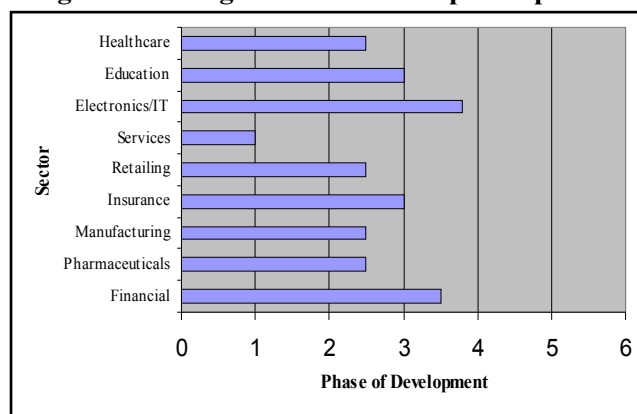
The Role of E-business in Marketing Practices

The findings reveal that the Internet is slowly diffusing into marketing practice in Ireland. In total, 95% of companies surveyed had established a presence on the web through the development of a web site and increased use of email. Interestingly, only three of the companies studied had an e-business plan and related budget in place. Overall, e-business strategies and planning were at the embryonic stage, with related activities being used in an *ad hoc* way to achieve company objectives.

The majority of companies surveyed had a basic, static web site containing company and product material or a simple interactive site where users were able to search and make queries to obtain information, such as product availability. Berton et al., (1998) have discredited these methods, because they feel the company does not take into account the variety of cultures and socio-economic backgrounds of purchasers that may visit the site, or the fact that the market may be fragmented.

In the context of Amor's six-phase theory referred to earlier, figure 1 outlines the average stage of development for each sector surveyed as part of the study. It is evident that findings concur with Amor's theory in that most companies are somewhere between phase 2 and phase 3 but are moving towards stage 4 in the e-business development process. Moreover, it is evident that none of the companies surveyed had yet progressed to a high level of advancement in relation to e-business activities.

Figure 1 Average Phase of Development per Sector



However, Amor fails to allude to the fact that e-business progression can depend on industry type. In summary, the research findings revealed that some e-business activities seemed to be more suited to certain sectors, for example:

- The electronics/IT sector and the financial sector, which included a leading Irish bank, seemed to be more advanced in terms of their e-business strategy. The majority of interviewees stated that their company had an e-business plan and related budget in place. One respondent commented on the fact that the transition to the e-business environment had not been difficult due to the sector's historical suitability to electronic transactions.
- The services sector, which included the legal, accountancy and auctioneering professions, was at the early stages in the e-business adoption process. Two companies had brochureware sites, whereas one company did not even have a web site. Amongst reasons cited for this level of e-business activity included client confidentiality, the need for personal contact with clients, and the type of the product/service offered. Due to the nature of real estate transactions, for example, customers require physical site/property inspections and, therefore, the Internet is used to build awareness and attract customer interest.
- The pharmaceuticals/manufacturing sectors mainly used the Internet for e-procurement activities, for example sourcing suppliers. Customer-facing web sites were used to provide corporate overview information. On a positive note, the majority of respondents anticipated running supply-chain management projects in the future.
- The retail sector varied in progression, with a range from static to transactional management sites. E-business activities differed depending on the nature of the products being sold. Giftware, holidays, books and music, for example, seemed suitable for online selling. One interviewee, representing a large travel agency, commented on the fact that the tourism sector had been profoundly transformed by e-business technology. Historically, the tourism industry has been an early adopter of new technologies, for instance Computer Reservation Systems (CRS); therefore, the transition to the e-business environment was considered a natural progression. On the contrary, clothing and footwear, for example, were perceived to be unsuitable for online trading. Surprisingly, a major bookstore manager indicated that their company wouldn't benefit from an online sales facility and that online companies such as Amazon were not a major threat to its business. In fact, this company indicated that customers use sites such as Amazon to source books and music which they then order or purchase locally.

In general, web site presence or sophistication did not change with organisation size, based on the profile of companies surveyed. Not surprisingly, given the number of brochureware sites, only one quarter of the companies visited were gaining any significant benefit from their sites. Four main reasons were cited for this perception:

- web sites were out of date and needed to be revised (33%)
- web sites were up to date but needed enhanced marketing effort to create awareness of their existence (50%)

- search engines needed to perform more efficiently (10 %)
- due to the nature of the business, it was unlikely to gain significant benefits via the web (10%).
(In this case, the majority of companies perceived benefits only in terms of additional sales).

In many cases, sales over the web were not regarded as a realistic option. However, it was evident that there were still significant opportunities for improved customer service through innovative use of the technology.

Only one interviewee claimed to rate the effectiveness of the web site presence by calculating the increase in sales for that company as a percentage of 'hits'. Others used the number of enquiry emails as a method of evaluation. However, two thirds of interviewees stated that they had no way of rating the efficiency of their web sites.

Not more than five per cent of companies surveyed had supply-chain management projects or were running customer-relationship management applications. It is important to note that these companies were in the financial services or electronics/IT sectors which seemed to be most progressive when adapting to the e-business environment. Only five companies were using mobile business (the use of wireless devices, such as cell phones or handheld devices to conduct transactions over the Internet) in their marketing activities, and this was mainly to keep in touch with colleagues when out of the office. Twenty per cent of companies were embracing and integrating electronic communication into their operations in an effort to strengthen databases and enhance relationships. Two thirds of those surveyed indicated that they used the Internet for business purposes and believed that the technology can reduce costs and enable easier and cheaper access to foreign markets.

All respondents felt that the Internet had the potential of generating a larger customer base. Moreover, findings showed that the Internet has proved to be an essential information source for Irish business organisations. This was identified as being one of the greatest benefits offered by the technology. All interviewees stated that the Internet as a marketing/communications tool lacked the personal touch with customers which they felt could be achieved only through face-to-face contact.

Challenges/threats anticipated in the e-business environment

All interviewees felt the Internet offered opportunities that were previously unavailable with traditional marketing methods. Two respondents indicated that the greatest challenge of the e-business environment was to build a relationship with customers that would lead to a perception of brand loyalty on their part, and consequently, repeat sales. Creating value proposition that would be readily embraced by the target market was another challenge cited. One respondent indicated that to implement e-business on a wide scale in that organisation would require a huge process of learning and cultural change. The growing importance of security and privacy in e-business transactions was

cited as another major challenge. However, all those surveyed were confident that they could adopt the new technology in the future with no major difficulty. Interestingly, 60 per cent of the interviewees stated that their e-business activities had been successful and that, on average, they expected to increase e-business spending by 15 per cent over the next year. Approximately one third anticipate that they will be pursuing significant opportunities in mobile commerce within three years.

Conclusion

For today's e-business organisation it may be simple to set up a web presence. However, it is difficult to create a sustainable online business, despite the continuing growth of online users. While the Internet has presented a variety of new strategic options to organisations, it has not changed the fundamentals of marketing management. The essence of e-business is still the generation of value by offerings to customers (Mathur and Kenyon, 2001). Many argue that the pace of innovation will continue to quicken in the next few years as companies exploit the still largely untapped potential for e-business, especially in the business-to-business arena where most observers expect the fastest growth.

In summary, the interviews with Irish managers from a variety of different sectors revealed interesting attitudes to the Internet which was seen as providing an additional challenge for those surveyed. A large number of companies involved in the study used web sites as an information source but they are not currently utilising them in a trading capacity. Consequently, their views on the impact of e-business on their operations varied. It was evident that the majority of companies surveyed had the potential to utilise their web sites more efficiently. Key issues that should not be overlooked include keeping web sites up-to-date and improving site performance on the main search engines through a judicious use of keywords and links. It would seem essential that companies perceive their site as a vehicle for customer service as well as a potential source of direct revenue. Moreover, the demands of allowing partners, customers, and sometimes, even competitors inside the e-business infrastructure will multiply security and privacy challenges.

However, many managers are now re-examining how best to integrate e-business with existing marketing channels and practices. These observations are encouraging and there seems to be a growing sophistication and realism amongst Irish managers about e-business. The right e-business solution represents an immediate challenge. Successful organisations will need e-business solutions that provide robust functionality, thereby enabling them to carry out marketing activities quickly while also meeting the specific and continually changing needs of their channel partners. Chen (2001) argues that an organisation's ability to succeed with its e-business initiative correlates directly to the management team's capacity to first document and articulate its corporate strategy, and then adapt to a specific e-business model which supports that strategy.

E-business enables a significant array of business opportunities and activities, and it would seem that, in time, it will become embedded in all aspects of companies' operations. It can be argued that there is a better strategic path for Irish enterprises to take in the future. If Irish

managers wish to reinvent their organisations, they will have to take a long and hard strategic look at how to satisfy the consumer in a rapidly changing environment. E-business mandates rethinking and repositioning and changes the nature of communication and interaction between almost all parties involved in the value chain. Traditional strategy formulation is no longer apt for the digital age.

References

- Amor, D. (2002), *The E-business Revolution, Living and Working in an Interconnected World*, Second Edition, Prentice Hall.
- Biswas, A. & Krishnan R. (2003), *The Internet's Impact on Marketing*, Introduction to the JBR special issues on 'marketing on the web – behavioural, strategy and practices and public policy', *Journal of Business Research*.
- Brady, M., Saren M., Tzokas, N. (2002), *The Assimilation of IT into Marketing Practice*, *Irish Marketing Review*, Volume 15, Number 2, P17.
- Cassidy, J. (2002), *Dot.com, The Greatest Story Ever Sold*, Penguin Books, Ltd.
- Chaffey, D., Mayer R., Johnston K. (2000), *Internet Marketing*, Pearson Education Limited, England.
- Chen, S. (2001), *Strategic Management of e-Business*, Wiley & Sons, England.
- Day, G.S. & Hubbard, K.J. (2002), 'Customer Relationships Go Digital', Working Paper, Wharton School, University of Pennsylvania, February.
- Deighton, J. & Barwise, P. (2001), 'Digital Marketing Communication', in Wind, J. & V. Mahajan (2001), *Digital Marketing*, Wiley, New York.
- Enterprise Ireland. (2004), *E-business and Information Technology Resource*, <http://www.enterprise-ireland.com/ebusiness>.
- Fahy, J. & O'Callaghan, D. (2002), *Is the Internet Dumbing Down Marketing?* *Irish Marketing Review*, Volume 15, Number 2, P59.
- IDC. (2004), <http://www.idc.com/analysts>.
- Nua Internet Surveys. (2004), http://www.nua.ie/surveys/how_many_online/index.html
- Mathur, S. & Kenyon, S. (2001), *Creating Value: Successful Business Strategies*, Butterworth-Heinemann, Oxford, UK.
- Murray, J.A. & Ko, E. (2002), *Researching the Cybermarket*, *Irish Marketing Review*, Volume 15, Number 2, P5.
- Online National Statistics, UK. (2002), <http://www.statistics.gov.uk>.
- Phillips, P. (2003), *E-Business Strategy, Text and Cases*, McGraw-Hill Education, UK.
- Porter, M.E. (2001), *Strategy and the Internet*, *Harvard Business Review*, March Edition.
- Rayport, J.F. & B.J. (2001), *e-commerce*, McGraw-Hill, Irwin, New York.
- Smith, P.R. & Chaffey, D. (2001), *eMarketing eXcellence: at the heart of eBusiness*, Butterworth Heinemann, Oxford, UK.
- Stauss, J. & Frost, R. (1999), *Marketing on the Internet: Principles of Online Marketing*, Prentice-Hall.

An Evaluation of On-Screen Keyboards for First Time Single Switch Users

Paul Ahern and Rupert Westrup

School of Science & Technology
Institute of Art, Design and Technology,
Dun Laoghaire, Dublin, Ireland
paul.ahern@iadt-dl.ie
rupert.westrup@iadt-dl.ie

Abstract

This paper presents an evaluation of three designs for on-screen keyboard layouts for use by Single Switch Users (SSUs). SSUs are those users who have a disability which means that they are able to activate a switch into an “on” or “off” position but are not able to locate or position a particular switch. The first keyboard layout design was alphabetically based, the second design was based upon the most frequently occurring letters in the English alphabet and the third design was the standard QWERTY layout found on most keyboards. The research shows that there is no significant difference in words per minute (WPM) for first time users between the three layouts. The researchers believe there was a significant learning effect for users going from one layout to the next and are of the opinion that further research on the layouts with experienced rather than first time users will yield interesting results. The overall aim of the research is to investigate different screen layouts with the goal of finding a screen layout best suited to SSUs.

1 Introduction

Augmentative and Alternate Communication (AAC) refers to any method of communicating that supplements the ordinary methods of speech and hand writing, where these methods are impaired. AAC can be unaided, for example where a person uses gestures or signing to aid communication, or aided, where some equipment such as symbol charts or computers are used. With aided communication, there is a range of technological options available ranging from low technological devices such as photo boards to high technology options such as computers. (Millar & Scott, 1998, pp. 3-5) The main difficulty in the area of AAC is that the rate of communication is between one-half to five WPM (Words Per Minute) (Vanderheiden, 1998), which is no where near the rate of normal conversation where the WPM rates are between 100 and 200. Gunderson (1985) has pointed out that the layout of the letters on the communication device (in the case of this experiment, an on-screen keyboard) and the method used to access it are relevant to the communication rate achieved.

One aspect of AAC is that of Single Switch User (SSU) communication. SSUs are those users who are able to activate a switch into an “on” or “off” position but are not able to locate or position a particular switch. In the field of AAC, there is a wide range of switches available. One way in which switches are classified is into contact and non-contact categories. With a contact switch, a user hits a button, while non-contact switches work by detecting movements of parts of the body. Non-contact switches can detect gross movement such as a flailing arm or very small movements such as an eye blink. (Nisbet & Poon, 1998)

For the purposes of this experiment to find the most efficient on screen keyboard layout, a contact switch (the return key on a keyboard) was used. The authors have also successfully used an eye blink switch based on the electromyograph (EMG) signals picked up from the facial muscles with the on screen keyboards.

2 Methods of Selection from an On Screen Keyboard

The problem for entering text as a SSU is speed. When an able-bodied user wants to type it is possible to choose a symbol and then select it, so there are two actions available to the user which are locate and select. A single switch user only has one action available. Therefore it is not possible for the SSU to actively locate a letter before selecting it to the same degree as an able-bodied user. A solution to this problem is to have an on-screen keyboard that iterates across the letters and when the user sees the wanted symbol they can use a switch to select that symbol.

In order to select an on-screen symbol there are two options available - direct selection or scanned selection. The choice of selection depends on the disability level and the switch being used. If the user can move their head or any limb in a controlled way, then direct selection may be possible. Direct selection is when the user can make a selection without having to pass through other symbols. If the user is unable to scan through the letters of the alphabet, it is necessary for the application to iterate through the letters allowing the use to select the wanted letter.

In scanned selection the user can choose to accept the current symbol by employing the switch, or not to accept the current symbol by either taking an action or by not taking action for a period of time. The problem with this method is that the SSU must wait until the application locates the correct letter and so there will be periods of inaction, which reduces WPM rates.

Scanned selection can be in one of two forms – linear or grouped. Linear scan selection scans through each available symbol in some order. Grouped scan selection divides the symbols into groups. The user selects a group and then the group is scanned (through linear scanning) and the user then selects the desired symbol. Since the user is selecting from the group and then from within the group, the user needs to make twice as many actions in grouped scan selection as compared to linear scan selection. The positive aspect for the grouped scan selection is that it is faster than the linear scan selection method. As shown in Venkatagiri (2003) the row-column scanning method (used in the layouts in this research) is nearly twice as fast as a linear scanning method.

3 Description of the on-screen keyboard layouts

Three on-screen keyboard layouts were designed and implemented. The first layout organised the letters alphabetically, the second layout ordered the letters according to their frequency in the English language. The third layout organised the letters in accordance with the standard QWERTY keyboard layout.

All three layouts had the same basic structure. There were two arrows: one horizontal arrow and one vertical arrow. The horizontal arrow iterated through the rows of the on-screen keyboard. When the row that contained the desired letter was indicated by the horizontal arrow, the SSU could then activate their switch. This action caused the horizontal arrow to stop on the current row and activated the vertical arrow. The vertical arrow kept iterating through the selected row until the SSU selected a letter by activating the switch when the vertical arrow indicated the correct letter. Once a letter had been selected, both the arrows ‘jump’ (return) to the first row/first column and then the horizontal arrow began to iterate again. For the purpose of the experiment the iteration occurred at the rate of 1 row or column per second. This rate was customisable so that users could increase their speed as their level of experience increased.

Predictive text was incorporated into the three layouts. When a letter was chosen by the user, the predictive function displayed the most frequently occurring words that begin with that letter. A maximum of five predictive words were displayed on-screen at a time. The predictive words became more refined as the SSU selected more letters. Predictive words were selected in the same way as letters i.e., the SSU stopped the horizontal arrow on the row and then stopped the vertical arrow on the column. After a predictive word was selected a space was provided automatically, so as to increase communication speeds.

3.1 Alphabetical Layout

This layout ordered the letters of the alphabet in alphabetical order – A, B, C etc. The interface had a 7 * 5 layout (Figure 3.1). As well as the 26 letters of the alphabet there were also four basic text editing commands – SPACE, RETURN, CLEAR and DELETE. The SPACE and RETURN commands added a space character and started a new line respectively. The DELETE command acted like the backspace key on a keyboard, deleting the character to the left. The CLEAR command deleted any text typed up to that point.

3.2 “Frequently Occurring” Layout

The “frequently occurring” layout (Figure 3.2) was in essence the same as the Alphabetical layout described in 3.1 except that the alphabet was arranged in a layout based on the most frequently occurring letters in the English language (Appendix 1). The idea behind the most frequently occurring layout is based on Morse’s principle. The most frequently occurring letter should be quickest to select. This layout also contained the four text editing options that were available in the Alphabetical layout.

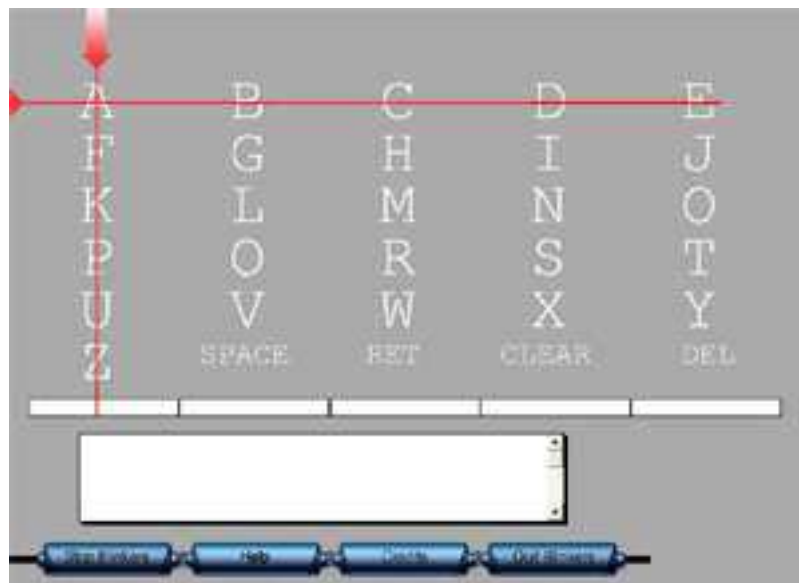


Figure 3.1 – Alphabet Interface

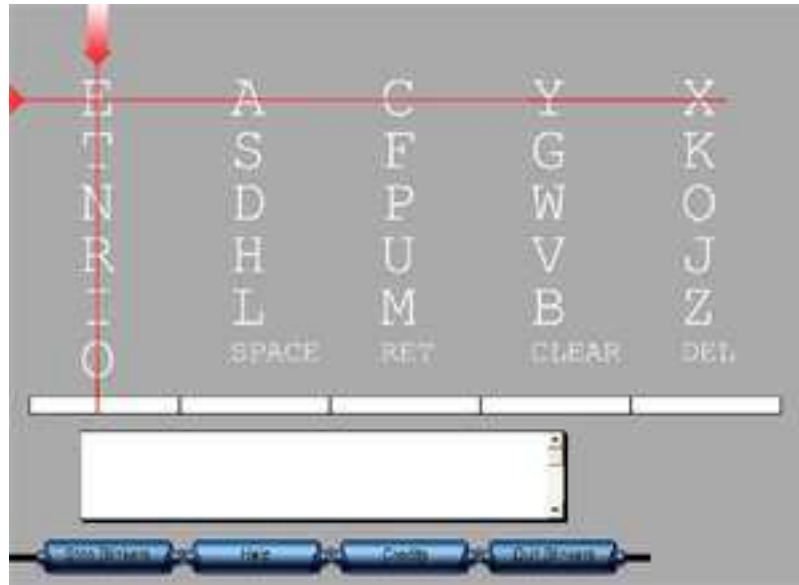


Figure 3.2 – Frequently Occurring Interface

3.3 QWERTY Layout

The QWERTY layout (Figure 3.3) had a different row-column layout to the Alphabetical or Frequently Occurring layouts. This interface had a 4 * 10 design with the formatting options laid out in different locations. The QWERTY layout was slightly different from an ordinary hardware keyboard. There were no numbers present and the positioning of the RETURN and DELETE keys was different. Another difference was that this interface gave the SSU two chances to select a predictive word. This was due to the fact that the field containing the predicted word was spread over two columns.

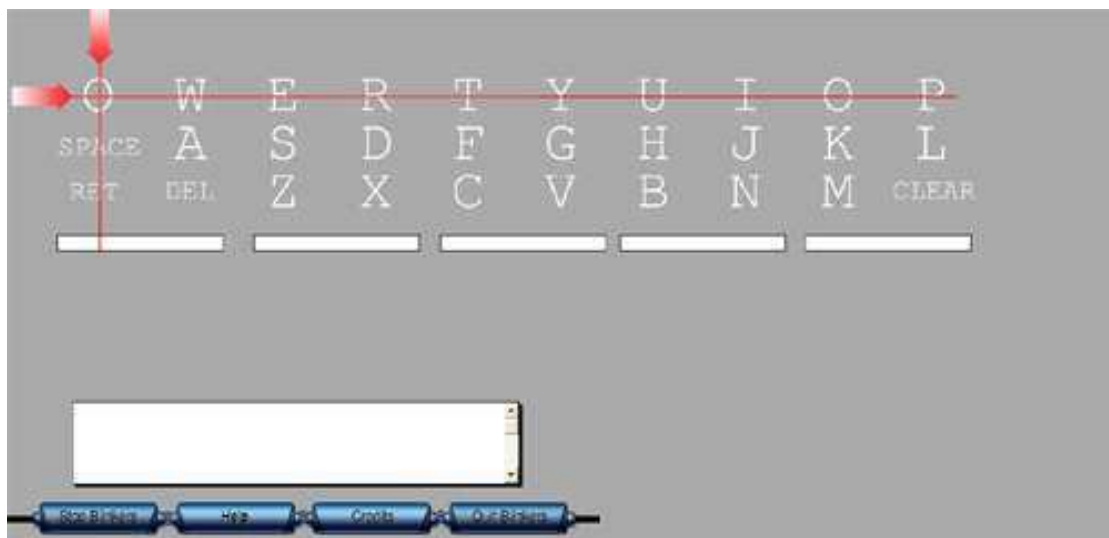


Figure 3.3 – QWERTY Interface

4 Experimental Design

The object of the experiment was to determine if there were significant differences between the efficiency of the three on-screen keyboard layouts described in Section 3. The experimental design was a one-factor analysis of variance (ANOVA). The factor in question was the interface layout with three on-screen keyboard layouts: Alphabetical, Frequently Occurring and QWERTY.

4.1 Subjects

Seven participants were used in this study. Two were female and the remaining five were male. Ages varied from 21 to 56. None of the subjects had any prior experience with using on-screen keyboards. Also, all of the subjects were able-bodied. Each subject was trained in the use of each keyboard prior to testing.

4.2 Procedure

Each of the seven subjects entered 4 sentences using each of the three layouts. The three layouts were all programmed in Macromedia Director. The users used the RETURN key as their switch. The input from the subject appeared on the screen in the field located in the lower half of the screen (Figures 3.1, 3.2, and 3.3).

Once trained in the use of a layout subjects practiced with the interface so that they were comfortable in its use before the testing began. Most subjects said that they felt comfortable with less than one minute of practice. The order of interfaces used by the subjects was varied to minimise the 'learning effect'. The practice period for the first interface was usually longer – lasting up to two minutes.

Each of the three interfaces recorded the data in the same fashion. The time from the first keystroke until the 'RETURN' function was selected (denoting the end of a sentence) was recorded. The number of actions performed was also recorded for further analysis. Time between sentences was not recorded. The arrows scrolled at a rate of 1 second per row or column which was not adjustable by the participants.

The sentences used for this study were those used by James & Reischel (2001) to evaluate speeds of mobile phone text entry. For that study, sentences had been derived from newspapers and general conversation. The sentences are shown in Table 1. Conversational sentences were chosen as the layouts were intended for everyday, conversational use. All words in the sentences were in the predictive text database but the user was not made aware of this prior to the test. In order to judge words per minute scores, the value of 5.98 characters per word was assumed (Dunlop & Crossan 2000). After each subject had completed all three tests they were

asked to complete a questionnaire regarding the speed, ease of use and interface preference. The experiment took between 40 minutes and 80 minutes for each subject.

Number.	Sentence	Characters
1	HI JOE HOW ARE YOU WANT TO MEET TONIGHT	39
2	WANT TO GO TO THE MOVIES WITH SUE AND ME	40
3	WE ARE MEETING IN FRONT OF THE THEATRE AT EIGHT	47
4	LET ME KNOW IF WE SHOULD WAIT	29

Table 1 – The four conversational sentences used in the tests.

5 Results

In general the subjects found the act of entering text very tedious and three of the subjects commented on being “exhausted” after completing all three interfaces. Most subjects complained about the amount of focus required to use the interfaces at speed. The frustration at selecting the incorrect letter was usually displayed vocally by all test subjects.

5.1 Performance Results

The mean speeds (in WPM) were 1.6 (standard deviation (s.d.) 0.3), 1.5 (s.d. 0.4) and 1.7 (s.d. 0.4) for the Alphabet, “Frequently occurring” and QWERTY layouts respectively. The ANOVA results showed an experimental F ratio for layout of less than one which implies that there was no significant difference between the three layouts.

5.2 General Observations

No matter what order the subjects tested the interface all (bar one subject) recorded faster WPM rates for the remaining interfaces.

All subjects suggested how the interfaces could be improved. These suggestions varied from using a larger font for the predictive words to reorganising the entire interface.

Most of the subjects continually failed to use the predicted words. It seems that they were focussed on entering the letters. Most subjects (5 of the 7) commented on the positioning of the predictive words being awkward – in scanning the predictive words for the wanted word they would often miss the next letter.

All of the subjects were frustrated by the inability to leave a row without selecting a letter. Subjects 4, 6 and 7 commented that if they were typing there own sentences that they would have been faster.

All subjects agreed that with more practice their times would improve especially with the “Frequently occurring” interface.

5.3 Confounding Factors

The use of the RETURN key as a switch.

This experiment used the RETURN key as a switch. The switch for actual users may range from a signal from a user blink to hitting a large switch with the hand. The difficulty of using a switch by as actual user may reduce WPM rates.

Low number of subjects.

With a greater number of subjects, it is more likely that a significant difference between layouts may have been found.

Prescribed Sentences

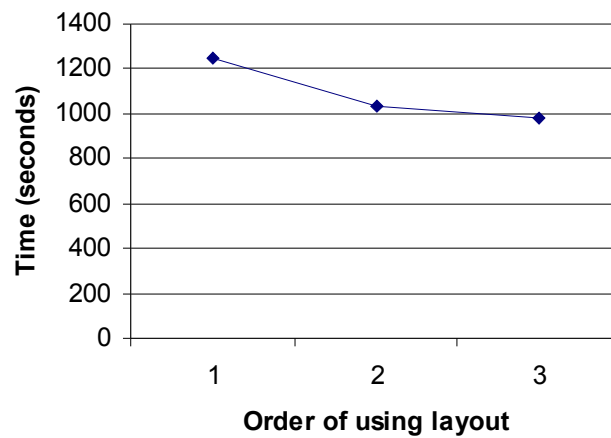
The fact that subjects were using sentences that were not their own meant that the subjects frequently had to move their focus from the interface to the required sentences.

Unrealistic Sentences

Subject 6 noted that in a real world situation a language more similar to the abbreviated text found in text messages might be used. For example using “c u tmrw” as opposed to “see you tomorrow”.

Learning Effect

The researchers noticed that there was a learning effect of going from one layout to another, which is illustrated in Graph 1. The points in the graph represent the mean values for all layouts for all participants for those layouts which were completed first, second and third. The graph shows that the layouts which were completed first took an average of 1250 seconds to complete, the layouts which were completed second took an average of 1038 seconds and the layouts which were completed third took an average of 979 seconds.



Graph 1 Mean time to complete the 4 sentences based on order that the layout was taken

6 Comparison with Related Work

A comparison between the work carried out in this study and the work being carried out by Venkatagiri (2003) is shown in Table 2. The Venkatagiri layouts are different to the layouts used in this experiment. The Venkatagiri layouts (43 – key layouts) contain numbers, and some punctuation options – full stop, comma and question mark. None of the Venkatagiri layouts contain a “CLEAR” function.

Venkatagiri’s results are simulated and appear much higher than the empirical results discovered in this study. The differences shown in Table 2 could be due to the differences in layout – Venkatagiri’s layouts have 13 more symbols to iterate through. Also the layouts in this study have predictive text, which in conjunction with fewer symbols should have shown the layouts proposed in this study to be faster.

It is proposed that the reason for the large gap between the results of this study and the results of Venkatagiri’s is the lack of experience in using the layouts for the subjects in this experiment.

	Alphabetical (in WPM)	Frequently occurring (in WPM)	QWERTY (in WPM)
Venkatagiri(2003) – 43 Key Row-Column	2.14	2.53	1.77
This Study – 30 Key Row-Column	1.60	1.54	1.66

Table 2 - Results of this study compared with Venkatairi’s 43-key, row-column access layouts

	Alphabetical (in WPM)	Frequently occurring (in WPM)	QWERTY (in WPM)
Venkatagiri(2003) – 43 Key Linear	0.98	1.34	0.77
This Study – 30 Key Row-Column	1.60	1.54	1.66

Table 3 – Venkatairi’s 43-key, linear access layouts compared with the results from this study.

7 Further Work

An area of further research would be to test the layouts of the interfaces repeatedly with the same subjects to evaluate how experience affects the words per minute rates. This work has already begun and the results for one subject are shown in Table 4 compared with Venkatagiri’s results.

	Alphabetical (in WPM)	Frequently occurring (in WPM)	QWERTY (in WPM)
Venkatagiri(2003) – 43 Key Row-Column	2.14	2.53	1.77
Experienced User of this study – 30 Key Row-Column	2.90	3.18	2.92

Table 4 – Experienced user compared with related work.

Smith & Zhai (2001) have used mathematical algorithms to create very productive on-screen keyboards for use with a touch screen. Their “Metropolis” keyboard reported a WPM rate of 43.1 based on Fitts law. Their research incorporates statistical transition probabilities between symbols to improve the keyboard performance. Similar techniques have the potential to improve keyboard layouts for SSUs.

8 Conclusion

This study compared three different interfaces for allowing test input for first time single switch users. The results did not show a significant variance between interfaces. All the results were very close, with the QWERTY interface proving to be the fastest with a rate of 1.7 words per minute. The Alphabet interface was next with 1.6 words per minute and the “Frequently occurring” interface was slowest with a rate of 1.5 words per minute. The researchers felt that further work with experienced users could show a significant difference between the three layouts. The reason for testing experienced users is that the learning effect apparent in this research should be eliminated. In turn the effects of the different layouts should become more apparent.

9 Acknowledgements

This research was funded by the Institute of Art Design and Technology, Dun Laoghaire Seed Fund. The authors would like to thank Mr. Cyril Connolly and Dr. Mark Riordan, both of the Institute of Art Design and Technology, Dun Laoghaire, for their advice and comments.

References

- Dunlop, M. & Crossan, A. (2000),** *Predictive text entry methods for mobile phones*, Personal Technologies, pp. 134-143.
- Gunderson, J. R. (1985),** Interfacing the Motor-Impaired for Control and Communication. In J.G. Webster, A.M. Cook, W.J. Tomkins, & G.C. Vanderheiden (Eds.), *Electronic Devices for Rehabilitation*. New York, New York: John Wiley.
- James, C. L. & Reischel, K. M. (2001),** *Text input for mobile devices: comparing model prediction to actual performance*, Proc. of CHI2001, ACM, New York, pp.365-371.
- Millar S., & Scott J., (1998),** *Augmentative Communication in Practice: Scotland – An Introduction*, University of Edinburgh, CALL Centre
- Nisbet P., & Poon P., (1998),** *Special Access Technology*, University of Edinburgh, CALL Centre
- Smith B.A., & Zhai S., (2001),** *Optimised Virtual Keyboards with and without Alphabetical Ordering*, Proc. Of INTERACT 2001, International Conference of Human-Computer Interaction, Tokyo, Japan, pp, 92-99
- Vanderheiden, G. C. (1998),** Overview of the Basic Selection Techniques for Augmentative Communications: Past and Future. In L. E. Bernstein (Ed.), *The Vocally Impaired: Clinical Practice and Research*, 40-83. Philadelphia, Pennsylvania: Grune and Stratton.
- Venkatagiri, H (2003).** *Efficient Keyboard Layout for Sequential Access in Augmentative and Alternate Communication*. Unpublished, Iowa State University.

Appendix 1 – English letter frequencies used in “Frequently occurring” interface.

English Letter Frequencies – Per 1000 letters			
Sorted by Letter		Sorted by Frequency	
A	73	E	130
B	9	T	93
C	30	N	78
D	44	R	77
E	130	I	74
F	28	O	74
G	16	A	73
H	35	S	63
I	74	D	44
J	2	H	35
K	3	L	35
L	35	C	30
M	25	F	28
N	78	P	27
O	74	U	27
P	27	M	25
Q	3	Y	19
R	77	G	16
S	63	W	16
T	93	V	13
U	27	B	9
V	13	X	5
W	16	K	3
X	5	Q	3
Y	19	J	2
Z	1	Z	1

Source: <http://library.thinkquest.org/28005/flashed/thelab/cryptograms/frequency.shtml>

Modelling a Mechatronic System using “Matlab/Simulink” and “Dyanst”

Paul Dillon

Institute of Technology Tallaght, Tallaght, Dublin.

Paul.Dillon@it-tallaght.ie

Abstract

This paper presents the process of modelling a mechatronic system using two modelling methods. One the well known Matlab/Simulation package the other, Dynast, uses a novel energy based approach. The system modelled is a single axis positioning system driven via an armature controlled DC motor. The model consists of the main items in the system; pulse width modulated amplifier, DC motor, pulley and belt drive, and ball screw thread. Effects of inertia of parts and also frictional effects are also accounted for in the models. The problems of combing different disciplines within each method are shown. Simulation results are presented from the two methods. Dynast was found to have significant advantages over Matlab.

Introduction

Computer simulation has reached a certain maturity with a wide range of software packages available to simulate different disciplines e.g. pSpice in electrical electronics, Matlab-Simulink (Mathworks 2004) for control systems. There still remains the problem of integrating different engineering disciplines within one software package. Such a package would have distinct advantages for the technician or engineer who works with Mechatronic Systems as it would negate any transfer of data from separate simulation packages or having to learn the process of modelling within different packages. Also possibility of using the computer as a tool to formulate the equations underlying systems is becoming more wide spread with “Dynast” (Dynast 2004) and “Mathworks-Power Systems Blockset” (Mathworks 2004), being just two examples amongst others, (Virtual Action Group 2004, Job van Amerongen et al 2003). By investigating two separate approaches to modelling of a mechatronic system, and comparing the process and results achieved in each, a designer would appreciate any differences and advantages one system has over the other.

Mechatronic System

Mechatronics has been defined as the “synergetic integration of physical systems with information technology and complex decision making in the design, manufacture and operation of industrial processes and products”, (Tomizuka 2002).

In this paper we are concerned with the modelling, simulation and operation of an industrial process namely a positioning system in a single axis, that would be typical as found in

computer numerical control (CNC) machines. The system consists of single axis drive based on a linear ball screw, driven by an armature controlled DC motor supplied by a “Servopack”, using velocity feedback, therefore incorporating many aspects of a mechatronic system. The schematic layout of the system is given Figure 1 while a photograph of the pulley end is shown in Figure 2. The table mass consists of a plate, which is mounted on the ball screw nut. The plate also is supported on two slide bearings. The ball screw nut is driven via the screw thread, by a large pulley, which in turn is driven via a toothed belt from a small pulley, mounted on the motor shaft. The motor is slung underneath the support table, thus allowing for tensioning of the belt by the weight of the motor. The motor has an integral tachogenerator and encoder for use as velocity feedback and position control in a digital controlled system respectively. The control box consists of the Yaskawa “Servopack”, main power supply and protection circuits. The physical systems that will be considered are the “Servopack”, the DC Armature Controlled Motor, the pulley and belt drive, the linear ball screw and the mass of the table, the position of which is the end measured variable.

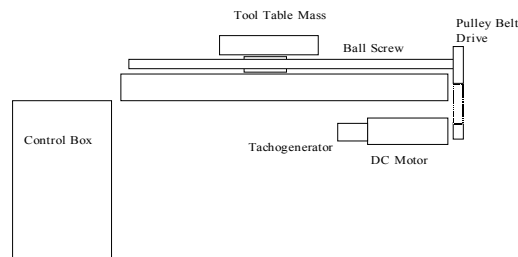


Figure 1 Schematic Layout

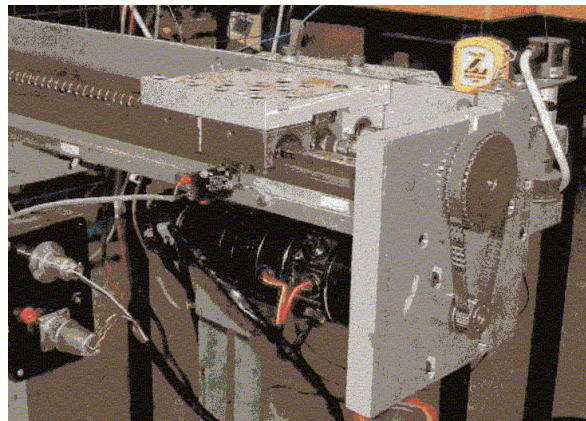


Figure 2 Pulley end of the Positioning System

Motor

The motor is a Yaskawa DC motor; the complete specifications are given in the motor technical bulletin (Yaskawa Electric Miniertia 1985). In order to understand and simulate the motor, the process followed would be to start with a sketch of the schematic arrangement of the motor main parts, as in Figure 3. Then work towards a mathematical model of the system in as much

detail as possible. The model here neglects higher order effects such as commutator contact resistance and windage effects. It would be usual to use differential equations for the relationships and then process these using Laplace Transforms.

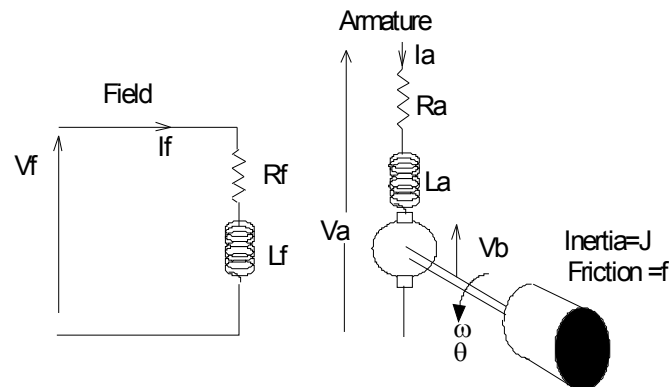


Figure 3 Schematic of a DC Motor (Dorf 1989)

One such model derived is as follows

$$G(s) = \frac{\omega(s)}{V_a(s)} = \frac{K_m}{(R_a + L_a s)(Js + f) + K_b K_m} \quad \text{Eq. 1 (Dorf 1989)}$$

The transfer function expresses the output $\omega(s)$ as a function of the input armature voltage $V_a(s)$ and the transfer function consisting of motor parameters. The values of the motor parameters as required for Eq. 1 are sourced from the motor specification sheet (Yaskawa 1985). $K_m=0.119$ [NmA⁻¹], $R_a=0.41$ [Ω], $J_M=0.000168$ [kgm²], $L_a=0.0006$ [H], $f=0.0000253$ [Nm (rad/sec)⁻¹], $K_b=0.0124$ [V/rpm]

Note the units used are inconsistent (but typically expressed in such a manner) and would have to be altered for use in a modelling package.

Servopack (Yaskawa Electric DC Servomotor Controllers 1987)

The Servopack allows the motor to develop approximately 190W and operate with a maximum speed of 3000rpm either in forward or reverse mode. The electronic controls supplied with the unit include over travel protection and a proportional and integral (PI) control of speed, based on a tachogenerator feedback signal from the DC motor. Separate to the actual Servopack, are a line filter, mains transformer, and thermal relay. The schematic of the circuit within the Servopack is shown in Figure 4.

Type CPR - FR 01 B
- FR 02 B

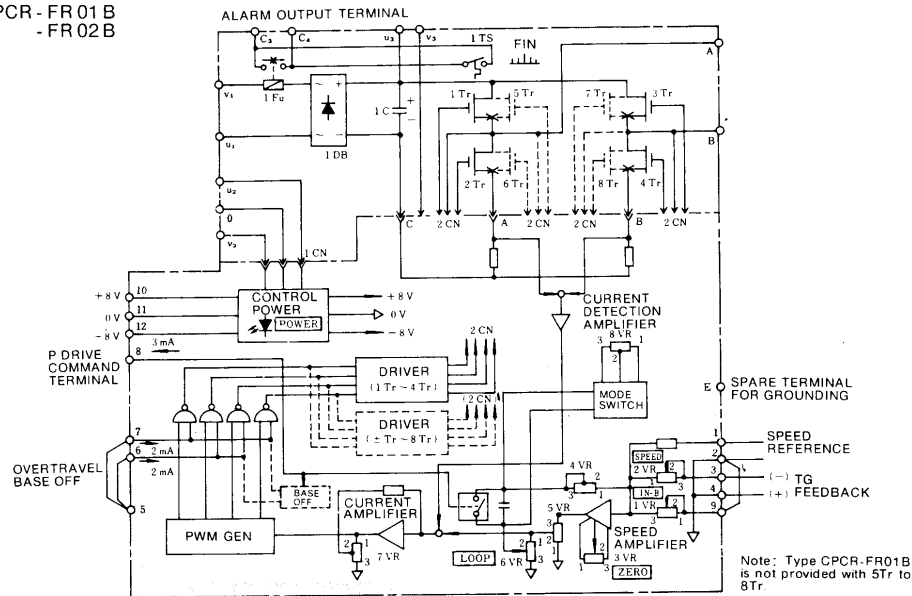


Figure 4 Schematic of Internal Connections of Yaskawa Servopack CPR-02B (Yaskawa 1987)

A designer would start with a circuit design and work on sizing the parts for the application, possibly using a different software package. At the time that this work was carried out (Brehens 1997, Price 2001) access was kept to Matlab/ Simulink which had no direct way of modelling such an electronic circuit within the modelling package. As can be seen from the schematic, the main circuit consists of a H-Bridge circuit based on power FETs, which are driven by a pulse width modulated (PWM) signal. The schematic giving qualitative but not quantitative information, part values and other circuit information could be obtained by back engineering the unit. For this reason the behaviour of the circuit was only considered.

To obtain a system model, it was therefore decided to take a decision to assume that the dynamic behaviour of the PWM drive circuit was faster than that of the mechanical motor and so a simple look up table was produced mimicking the voltage output of the drive circuit. This look up table was arrived at by testing the circuit to derive the control relationship between the inputs, which were a voltage signal representing a desired speed, a feedback voltage from the tachogenerator and the output drive voltage which was pulse width modulated Figure 5. The values of the potentiometer were varied in order to look at the motor output voltage. The test was conducted with a 7V feedback signal, from the simulated tachogenerator.

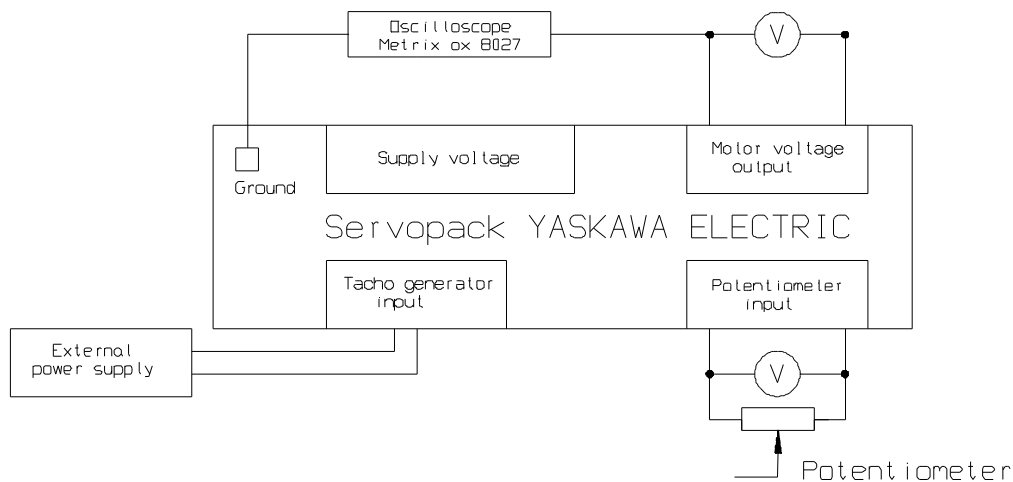


Figure 5 Test Setup for deriving the Control and Drive Relationships for Servopack. (Brehrens 1997)

The results, in Figure 6, showed that the drive voltage has a narrow proportional band based on the difference between the desired input voltage and the feedback voltage and also there is a proportional relationship between the desired input voltage and feedback voltage from the tachogenerator. The tachogenerator voltage being 3.5 times that of the input voltage. This plot is representative for the motor output voltage behaviour for the other tests at 14V and 20, merely showing that the curve is moved along the input voltage axis.

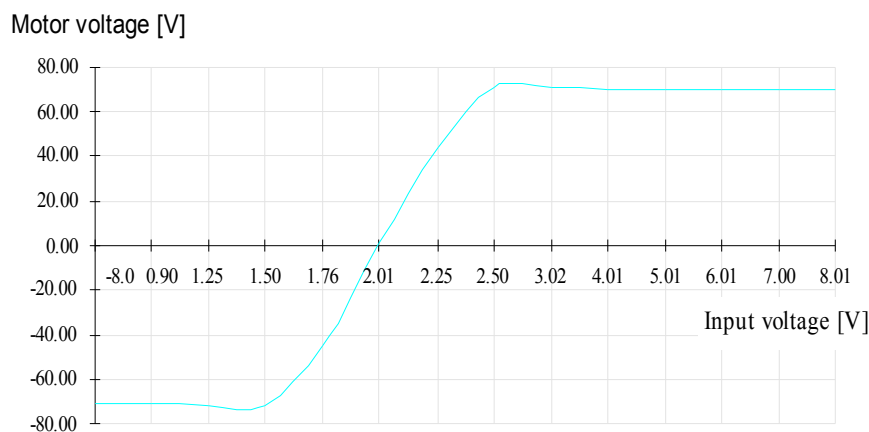


Figure 6: Control behaviour of the Servopack

Transmission System

The transmission system consists of a pulley drive, which has a ratio of 2.892. The belt is toothed in order to minimise the possibility of backlash. The pulleys, ball screw and tool table mass are considered to have a certain reflected inertia to the motor. The calculation of which is covered by Gross 1983. The total reflected inertia excluding the motor inertia J_{ext} can be then calculated by the following

$$J_{ext} = J_{sp} + \frac{J_{lp} + J_{bs} + J_{lin}}{i^2} \quad \text{Eq 2}$$

$$J_{sp} = \text{Inertia of small pulley wheel} = 7.482 \times 10^{-6} \text{ kgm}^2$$

$$J_{lp} = \text{Inertia of large pulley wheel} = 6.7991 \times 10^{-4} \text{ kgm}^2$$

$$J_{bs} = \text{Inertia of ballscrew} = 3.2 \times 10^{-5} \text{ kgm}^2$$

$$J_{lin} = \text{Inertia of the driven linear element} = 5.75 \times 10^{-6} \text{ kgm}^2$$

$$i = \text{Pulley ratio} = 2.892$$

$$J_{ext} = 9.329 \times 10^{-5} \text{ kgm}^2$$

The motor inertia as given by Yaskawa (Yaskawa 1985) is $J_M = 0.000168 \text{ [kgm}^2]$. The motor is directly attached to the small pulley wheel.

The total inertia reflected back to the motor can then be included in the block diagram of Figure 3 by

$$J = J_m + J_{ext} \quad \text{Eq 3}$$

$$J = 1.68 \times 10^{-4} + 9.329 \times 10^{-5} \text{ kgm}^2$$

$$J = 2.610 \times 10^{-4} \text{ kgm}^2$$

Apart from the inertia loading on the motor, we have to consider the frictional load exerted by the transmission elements. Friction is generated in a number of locations, friction in slide guides, frictional losses in the feed screw bearing, frictional losses in the feed screw nut and finally pulley and belt drive losses.

The influences of the friction torque and losses from the feed screw drive can be described as the sum of $T_f \text{ [Nm]}$ the friction torques. We obtain the friction torques reflected on the motor by equation 4.

$$\sum T_f = \frac{\left(\frac{T_{fsg}}{\eta_{fsn}} \right) + T_{fsb}}{\eta_p \cdot i} \quad \text{Eq. 4. (Gross 1983)}$$

Where, T_{fsg} = friction torque in the slide guides

η_{fsn} = efficiency of feed screw nut

η_p = efficiency of pulley drive

T_{fsb} = friction torque of feed screw bearing

i = ratio of pulley drive

Friction in slide guides. Gross (1983) states that the friction in the slide guides can be taken as

$$T_{fsg} = \mu_f \cdot \frac{P_{bs}}{2\pi} \left[(m_T + m_{wp}) \cdot g + F_c \right] \quad \text{Eq. 5 (Gross 1983)}$$

Where, μ_f = the speed dependent friction factor

m_t = mass of tool table [kg]

- m_{wp} = mass of workpiece [kg], which is zero because it does not exist
 p_{bs} = pitch of ball screw [m]
 g = $9.81[\text{ms}^{-2}]$, gravitational acceleration
 F_c = force of cutting [N], which is not present

The speed dependent friction factor μ_f , of the actual friction guide, is difficult to decide on. Therefore we assume a mean value a factor between 0.05 and 0.1.

$$T_{fsg} = 0.075 \cdot \frac{0.01}{2\pi} \cdot 2.268 \cdot 9.81 = 2.656 \cdot 10^{-3} [\text{Nm}]$$

Frictional losses in the feed screw bearing are again accounted by referring to Gross 1983, page 127. The feed screw is supported with two ball thrust bearings. If there was a force working against the tool table, in the direction of the bearings, the load on the bearings become larger, due to the force. This load causes a greater friction in the bearings and therefore a friction torque which has an effect on the motor. According to the INA company this friction torque is approximately

$$T_{fsb} = \mu_{sb} \cdot 0.5 \cdot d_m \cdot F_{al} \quad \text{Eq. 6 (Gross 1983)}$$

- where, μ_{sb} = speed dependent frictional factor of the feed screw bearing
 d_m = mean value of the bearing diameter [m]
 F_{al} = axial feed screw load [N]

For the speed dependent frictional factor of the feed bearings μ_{sb} a mean value can be taken and is chosen to be between 0.003 and 0.005, according to Gross, page 228. The mean value of the diameter is $0.019[\text{m}]$. The axial feed screw load is described, according to Gross, as a machining force. For the modelled system no machining force exists and therefore the equation would reduce to zero, the force F_{al} is not present. Therefore we assume a force caused by *Newton's law of motion*. We get

$$F_{al} = m_t a_t \quad \text{Eq. 7}$$

- where, m_t = mass of tool table [kg]
 a_t = acceleration of table $[\text{ms}^{-2}]$

If we substitute the above equation for F_{al} into the T_{fsb} equation we obtain the friction torque of the feed screw bearings T_{fsb} [Nm] as follows

$$T_{fsb} = 8.618 \times 10^{-5} \cdot a_t [\text{Nm}] \quad \text{Eq. 8}$$

Taking the frictional losses in the feed screw nut next. The feed screw nut in the original model, is a ball screw nut, these type of nuts are commonly used, because less friction results.

Nevertheless, we will consider the efficiency of the ball screw nut. This coefficient η_{fsn} can be calculated approximately as below

$$\eta_{fsn} \approx \frac{1}{1 + 0.02 \cdot \frac{d_{bs}}{p_{bs}}} \quad \text{Eq. 9 (Gross 1983)}$$

where, d_{fs} = diameter of ball screw [m]
 p_{bs} = pitch of the ball screw [m]

The pitch p_{bs} is given with 0.01[m]. The ball screw has a diameter of 0.015 [m]. By applying eq. 9 we get

$$\eta_{fsn} \approx \frac{1}{1 + 0.02 \cdot \frac{0.015}{0.01}} \approx 0.971$$

Finally taking pulley and belt drive losses, the efficiency of the belt and pulley system can be taken to be 95 to 98% under correct operating conditions.

Applying Equation 4 we obtain the sum of the friction torques T_f [Nm] reflected on the motor, shown below.

$$\sum T_f = \frac{0.00273532 + 0.00008162 \cdot a_t}{2.83416} \quad \text{Eq. 10}$$

Modelling Packages

Matlab Simulink

If we now consider the problem of modelling the complete system within the chosen software package. Taking the Matlab/Simulink system first, the complete system is shown in Figure 7.

A number of submodels are shown within the block diagram. The motor submodel is shown in Figure 9. This is a direct application of the block diagram Figure 8, resulting from the transfer function for the DC motor, equation 3.

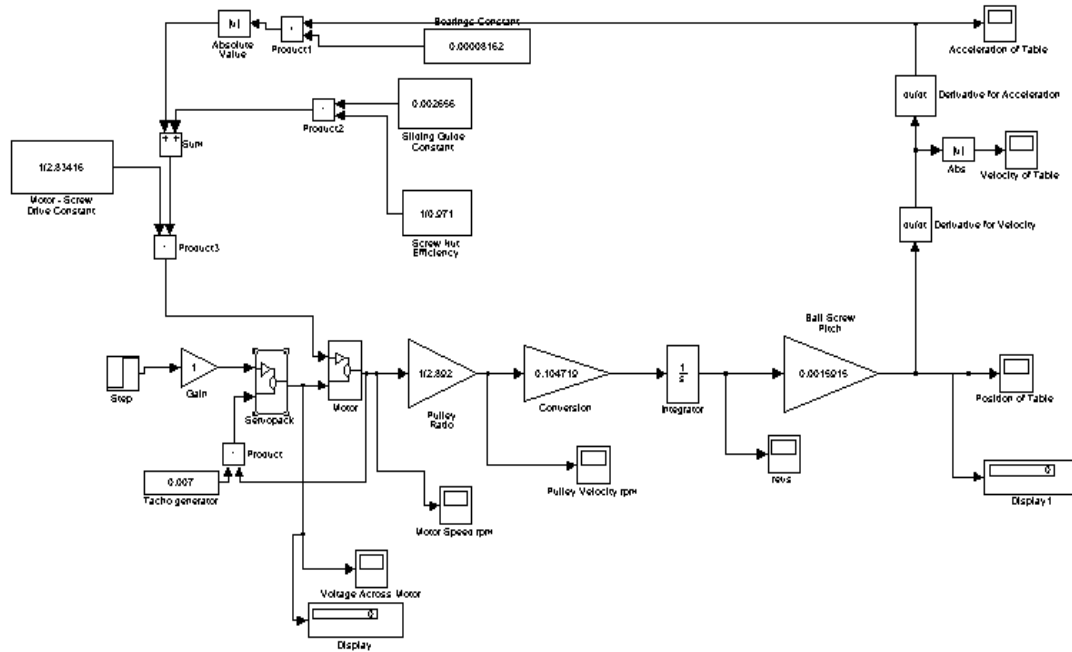


Figure 7 Complete Matlab Simulink Model of the Motor Pulley Screw Drive

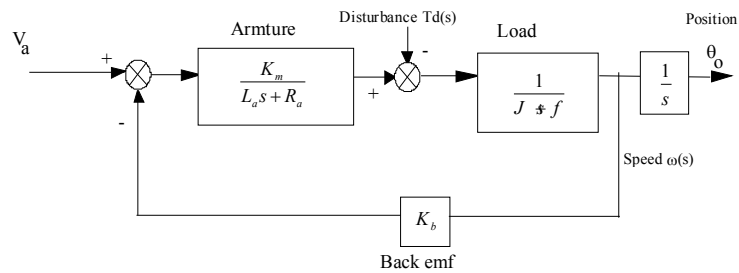


Figure 8 Block Diagram of Motor (Dorf 1989)

The block diagram allows for the influence of an external disturbance torque $T_d(s)$ to be included, which in our case will simulate the frictional torques. By substituting in the motor parameters the resulting simulation block diagram is shown in Figure 9.

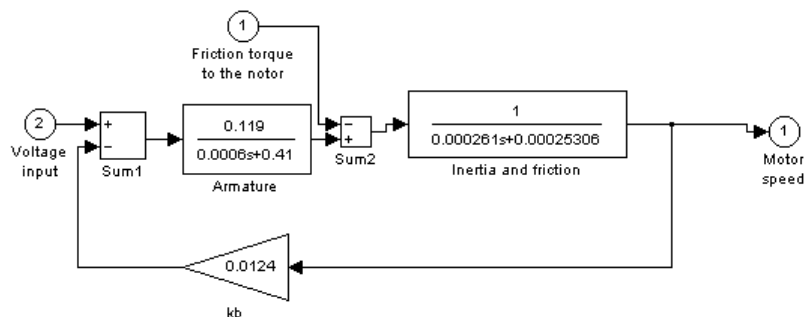


Figure 9 Simulink Diagram for DC Motor Transfer Function

The motor “Voltage Input” point, derives from the model of the Servopack which is set up as a look up table. Figure 10. The “Friction torque to the motor” point depends on from the acceleration of the tool table mass. This is accounted for in the simulation by taking a signal of the position of the table and completing a double derivative to generate an acceleration signal. This is then used to implement equation 10 as derived. The other submodel is the servopack. Figure 10. This is quite simple, the two inputs are the desired motor speed” and the feedback voltage from the tachogenerator. These are summed to provide an input to the look up table which generates a voltage output, such as given in Figure 6.

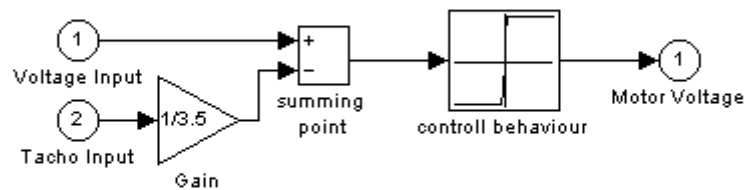


Figure 10 Simulink Block Diagram representing Servopack operation.

Dynast (Dynast 2004)

To contrast the above process. Dynast (Dynast 2004) uses a novel method of modelling, the concept of “multipole” modeling. This allows the integration of different engineering disciplines within the one modelling package. Multipole modelling is similar to the nodal analysis approach in electrical systems. The basic procedure is to break up the proposed model into disjointed subsystems. Similar to forming free body diagrams in mechanics. The disjointed subsystem can then be analysed under the assumptions that the mutual energy interactions between subsystems take place at a limited number sites such as electrical connections, pipe connections (fluid systems), or mechanical contacts. The energy flow through each contact point can be expressed by a product of two complementary power variables such as voltage and current, force and velocity. The energy entry point is modeled as a pole associated with a pair of power variables. Individual physical elements such as an electrical resistor can then be expressed as a two pole model, with the power consumed being governed by

$$P_e(t) = i_e(t) \cdot v_e(t)$$

where i_e is the through variable flowing through the element and v_e is the across variable between the element poles.

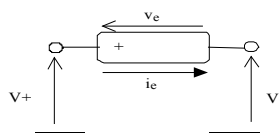


Figure 11 Variables associated with a generic physical element (Dynast 2004, Mann 1999)

Once a multipole model for an element has been developed, it can be contained within a library for further use, allowing the user to build up a resource of increasing complexity and sophistication. Simulation models of real systems can then be built by including these multipoles within a complete diagram. The user can start with a simple model based on simple elements then add in more elements to represent behaviours of higher order or smaller effects. Figure 12 shows just one variation of multipole diagram as crude representation of the mechatronic system. This includes the PWM circuit, a complex multipole model of the DC motor, the multipole model of the transmission system. Included is a PID block with the I and D turned off. This block gives feedback control of the PWM circuit. As can be seen the Dynast representation allows the use of symbols that schematically look like the physical element concerned. The PWM circuit is built up from resistors, capacitors and op amps. The Transmission section is constructed of multipole elements representing inertias, and damping elements for rotational and linear motions. The transmission structure also has transformer elements to represent the pulley and belt ratio and the transformation from rotary motion to linear translation motion.

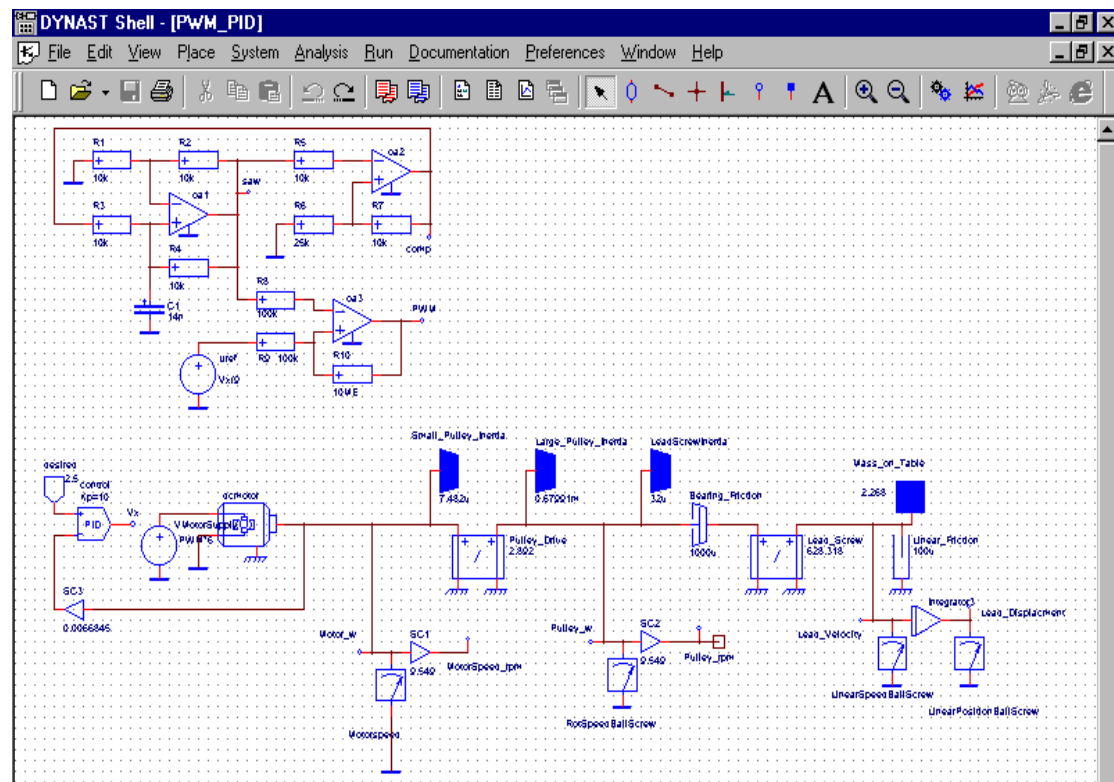


Figure 12 Dynast Representation of the System Including PWM Circuit, DC motor and Transmission System, along with Proportional Control of Speed.

Results

Figures 13 and 14 show the results for simulations run using the Matlab/Simulink model and the Dynast model, for a time of 0.5 seconds. The aim being to compare the resulting calculations for the motor speed (rpm) and the lead screw displacement (m).

We are interested in the lead displacement output in this paper for comparison. (Other parameters within Dynast would be available such as the PWM voltage output wave). From Figure 13 one can see that the simulations compare well in the end result. The slight difference in lead displacement is due to the difference in motor speed as shown in Figure 14. One can see that the Matlab model achieves a steady state speed of 1200rpm by 0.02 seconds while the Dynast model responds faster and has a higher final speed. This difference is due to tuning of the parameters on the PWM and the proportional control values. By re-tuning of values and investigation of the PWM circuit as well as incorporation of an H Bridge circuits it is hoped to approach a closer model to the real system. Variations on circuits have not been shown here.

Future Work and Conclusion

Further work will have to be completed on the design of the PWM circuit to allow for direct application of the velocity feedback on the pulse width, with an electronic circuit rather than a block diagram. Consideration will be given to modelling a commercial PWM driver such as the HIP4081 or the UC3637 and changing the BJTs to MOSFETs in the H Bridge. Considering the approach used for each model, derivations of physical values of mechanical parts were common to both, although Dynast applied them singly in direct relation to where the inertias were in the real system and also the frictional effects are handled in the same manner. Also Dynast was able to bring both the electronic and mechanical aspects in the one environment, the PWM circuit, the multipole model of the DC motor, the mechanical transmission system. A designer of electronic circuits would no doubt make more appropriate circuit for the PWM than the author. The Matlab model was assembled from equations representing the behaviour of the system. Dynast has some advantages over Matlab, one being able to incorporate the different disciplines the most obvious. The modelling method Dynast uses of directly generating the model rather than generating equations is also significant. Dynast can also be used as a toolbox for Matlab for modelling and control (Mann et al 2000). Dynast is freely available to use across the Internet thereby making it useful from a teaching/student perspective. A course on Modelling and Control with toolboxes for designing "Virtual Experiments" is currently being developed for support of teaching in the area using Dynast, (DynLAB 2004). Investigations of other multidiscipline simulation tools such as 20 Sim (van Amerongen et al 2003), which uses Bond Graphs (Karnopp 2000), along with extensions to the Matlab toolbox/blockset will be carried out, for future support of Mechatronic courses.

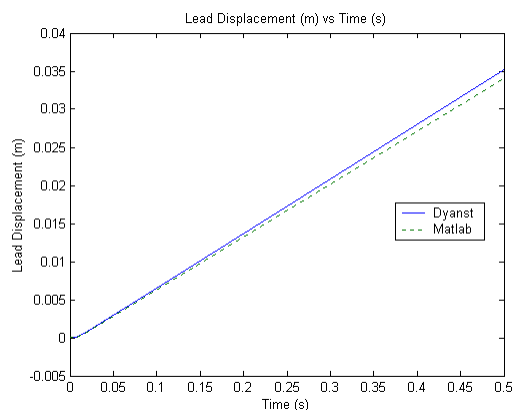


Figure 13 Lead Displacement (m) V's Time (s)

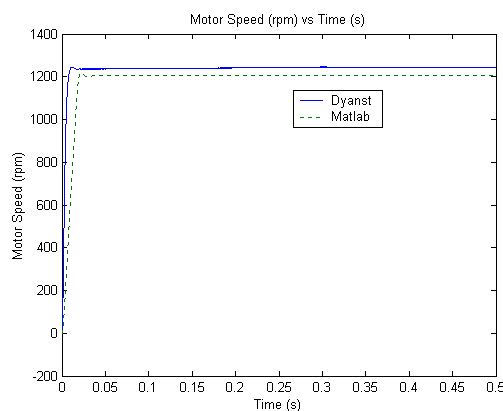


Figure 14 Motor Speed (rpm) V's Time (s)

Acknowledgements

The author would like to gratefully acknowledge the support of the Leonardo da Vinci Program of the European Union for the “DynLAB” project.

References

- Brehrens, M.**, *Simulation and Construction of a Machine Tool Feed Drive*, 1997, Project Thesis, IT Tallaght.
- Dorf R. C.**, *Modern Control Systems*, 5th Edition 1989 pp 49 to 52
- Dynast**, Website for Modelling System <http://virtual.cvut.cz/dyn/> 2004
- DynLAB**, Website of the DynLAB the Leonardo da Vinci Vocational Training Programme Pilot Project at <http://icosym.cvut.cz/dynlab/> 2004
- Gross, H.**, *Electrical Feed Drives for Machine Tools*, John Wiley and Sons 1983.
- Karnopp D., Margolis D., Rosenberg R.**, *System Dynamics*, Wiley Interscience, 2000
- Mann H.**, *Support for Efficient Modelling of Multidisciplinary Systems*, Proceedings of the ASME Vol. 67. Nasville, 1999.
- Mann H., M. Ševčenko, J. Pavlík.** *Dynast – modeling toolbox for Matlab formulating equations automatically* Proc. 8th Conf. on Matlab 2000, Prague 2000, pp. 416-420.
- Mathworks** <http://mathworks.com/> 2004
- Price, A.**, *Instrumentation of a Motor Drive Rig and Comparison with a Simulink*, 2001, Project Thesis Degree, IT Tallaght
- Tomizuka, Masayoshi**, *Mechatronics: from the 20th to the 21st Century*, Control Engineering Practice 10 (2002) pp877-886.

van Amerongen Job, Breedveld Peter, *Modelling of Physical Systems for the design and Control of mechatronic systems*, Annual Reviews in Control 27, 2003 pp87-117

Virtual Action Group for Multidisciplinary System Simulation Technical Committee on Computer Aided Control System Design of the IEEE Control Systems Society at <http://icosym.cvut.cz/cacsd/dyn/> 2004

Yaskawa Electric DC Servomotor Controllers “Servopack™” Transistor PWM Reversible Type CPR-FR01B to FR05C TSE-C717-11G Bulletin January 1987

Yaskawa Electric Miniertia® Motor RM Series Type UGRMEM TSEC 253-10C Bulletin August 1985

Profiling the International New Venture - A literature review of the empirical evidence

Natasha Evers

School of Business & Humanities
Institute of Technology, Blanchardstown
Natasha.evers@itb.ie

Abstract

International New Ventures (INVs) are defined as “business organisations that, from inception, seek to derive significant competitive advantage from the use of resources and the sale of outputs in multiple countries” (McDougall et al., 1994: 470). Globalisation, the dismantling of trade barriers, and the pervasive impact of new technologies have created unprecedented opportunities for young firms to go international early in their life cycle. This is also evidenced by the large number of long established firms that have recently internationalised. In light of these global trends, size of the firm is no longer a key determinant of internationalisation as it once was - the INVs are clear evidence of this.

This paper presents a critical review of the empirical findings popular in the current INV literature, with the objective of developing a profile of INVs which includes the basic characteristics, such as the competitive strategies, products and choice of market entry modes. There are certainly enough similarities in the empirical studies to generate a more comprehensive profile of the INV than exists in the literature already. A review of theoretical contributions on the INV is equally of interest but is outside the scope of this paper. The first section of the paper describes the emergence of the INV firm, which has challenged traditional theories of internationalization. The second section identifies some of the key driving forces behind the firms. Then, drawing on empirical findings, a profile of the INV is drawn up. Finally, noting the consensus in the literature, a conclusion is reached. This is followed by observations from the author and the recognition of research avenues, which may warrant further investigation.

Introduction

There exists a limited amount of theory building and empirical investigations on the early internationalisation of young entrepreneurial firms. The phenomenon of the International New Venture (INV) as a form of international entrepreneurship is still an unexplored area of research in entrepreneurship and international business literature. International business literature has tended to focus on large multinationals and long-established SMEs which have already gained maturity in their home market prior to internationalisation, whereas research on entrepreneurship has primarily focused on venture creation and on the management of SMEs within a domestic context.

Zahra & George (2000:11) broadly define International Entrepreneurship as “the process of creatively discovering and exploiting opportunities that lie outside a firm’s domestic markets in the pursuit of competitive advantage”. McDougall & Oviatt (2000: 903) go further and add risk to the definition: “(International Entrepreneurship is) a combination of innovative, proactive, and risk-seeking behaviour that crosses national borders and is intended to create value in organisations”. INVs constitute a new form of entrepreneurship, as they are

entrepreneurial from the outset with a strong international orientation. This combination of International Business and Entrepreneurship seems to naturally encompass the INVs. McDougall and Oviatt's (1994) theory on *New Venture Internationalisation* has been the closest yet to theorising on this concept. They describe INVs as firms having "an international vision (...) from inception, an innovative product or service marketed through a strong network, and a tightly managed organisation focused on international sales growth" (Oviatt and McDougall, 1997:47).

From the empirical standpoint, it has been only in the last decade that INVs have received a lot of attention in research. An abundance of empirical evidence has emerged revealing that an increasing number of firms are following this pattern of internationalisation in the US and western European economies (McDougall & Oviatt, 1994;1997; Bell, 1997; McKinsey & Co, 1993; Lindmark et al 1994; Preece et al 1999; Christensen and Jacobsen 1996; Waago et al 1993; Madsen & Per Servais, 1997; Larimo 2002; Bell 1995). However, according to Knight & Cavusgil (1996), INVs have been around since the late 1970s. This claim is based on documented examples of internationalisation patterns similar to such INVs in different countries.

On titling the INV, Rialp et al (2002) argue that the "denomination given to this specific phenomenon has been rather diverse and somewhat confusing, because not always synonymous labels have been used to describe those firms that decide to internationalize from inception" (pp20). Researchers have given such firms many titles but in essence they refer to the same thing. The term "Born Globals" has been the most popular (Rennie, 1993; Knight and Cavusgil, 1996; Madsen and Servais, 1997; Bell and McNaughton, 2000; Madsen et al., 2000; Rasmussen et al., 2001; Larimo, 2001; Aspelund and Moen, 2001; Moen, 2002), other names such as Global Start-ups (Oviatt and McDougall, 1995), High Technology Start-ups (Burgel and Murray, 2000), Global High-Tech Firms (Roberts and Senturia, 1996), Instant Internationals (Fillis, 2001), and International New Ventures -INVs- (Oviatt and McDougall, 1994; McDougall et al., 1994; Oviatt and McDougall, 1997; Servais and Rasmussen, 2000) have also been used for designating these firms. The name INV has been adopted in this review.

INVs are "a business organization that, from inception, seeks to derive significant competitive advantage from the use of resources and the sale of outputs in multiple countries" (Oviatt and McDougall, 1994, p. 49; McDougall et al., 1994, p. 470). Similarly, Knight and Cavusgil (1996, p. 11) conceptualize INV firms as being "small, technology-oriented companies that operate in international markets from the earliest days of their establishment". Knight (1997:1) defined INV firms as "a company which, from or near its founding, seeks to derive a

substantial proportion of its revenue from the sale of its products in international markets". The various definitions in the literature agree that these firms acquire significant export involvement early in their life cycle. An increasing number of firms can be classified as INVs according to the Knight & Cavusgil (1996) definition. The accelerated process of internationalisation constitutes an increasingly distinctive pattern of the internationalisation process of some SMEs when seen in comparison to other types of businesses (Rennie, 1993; Madsen et al., 2000; Servais and Rasmussen, 2000; Aspelund and Moen, 2001; Larimo, 2001; Moen, 2002; Lindmark et al; Munro & Coviello 1994; Bell, 1995; Moen 2001; Waago et al 1993).

Many authors (McDougall & Oviatt 1994; Knight & Cavusgil 1996; Madsen & Servais 1997; Larimo 2001; Moen 2002) have discussed the difficulties of explaining the development of INVs through the traditional "stages" or "Uppsala" model (Johanson & Vahlne, 1977, Johanson & Vahlne 1990; Johanson & Wiedersheim-Paul 1975), and the Innovation-Related Internationalisation Model (Cavusgil 1980). In their scholarly publication, Knight & Cavusgil (1996) discuss the term "INV" at length for the first time. They identified an increasing number of firms that do not follow a slow gradualist approach when internationalising. Findings from their case study research in developed economies have revealed that some firms 'leapfrog' stages found in the traditional process models and go international immediately or within two years of their inception. Welch & Loustarinen (1988) discussed reports of small English firms, Australian start-ups, and established Swedish firms that skipped important stages and quickly engaged in foreign direct investment (FDI). In another cross-country comparative study, Bell (1995) looked at the internationalisation of small computer firms in Ireland, the UK and Finland. He discovered that these firms leapfrogged the stages model of internationalisation, thus reconfirming the inadequacy of traditional theory explaining the process of INV internationalisation. McKinsey & Co (1993) identified many INVs whose management viewed the world as their marketplace right from birth.

In spite of the growing evidence to suggest that the traditional approach does not fully explain the internationalisation process of INVs, there has been a paucity of theoretical contributions attempting to explain the growing phenomenon (Madsen & Servais 1997). McDougall et al (1994) state that these traditional theories are invalid for INVs, for a number of reasons. Firstly, these theories assume that firms become incrementally international long after they have been formed, as the bias within international literature is to study mature large firms and long established SMEs. That the formation process of INVs appears largely inconsistent with traditional stages theories has been well documented by many authors (McDougall et al., 1994, 1999; Knight and Cavusgil, 1996; Roberts and Senturia, 1996; Oviatt and McDougall, 1997; Madsen et al., 2000; Moen, 2002). Secondly, these theories, they argue, focus too much on the firm level and ignore the individual and small group level of analysis (i.e. the

entrepreneur and their network of business alliances and contacts). They propose that Entrepreneurship theory (Kirzner, 1973) and the Strategic Management theories can better explain the phenomenon.

Driving Forces behind INVs

A starting point in the literature on INVs has been to propose reasons which may explain the emergence of this type of firm. Knight & Cavusgil (1996) present several recent trends, these are as follows:

1. there is the increasing role of niche markets, compelling small entrepreneurial firms to be competitive by delivering to niche markets across multiple foreign markets;
2. there are the advances in process technology;
3. there are the advances in communication technology;
4. there are the inherent advantages of the small company - quicker response time, flexibility, adaptability, and so on.
5. A fifth factor is the means of internationalisation - knowledge, technology, tools, facilitating institutions and so forth.
6. then is the growing importance of global networks and alliances.

According to many authors, (Knight and Cavusgil, 1996; Madsen and Servais, 1997; Servais and Rasmussen, 2000; Moen, 2002) four of the trends proposed by Knight & Cavusgil (1996), are of most importance: 1) new market conditions in many sectors of economic activity (including the increasing importance of niche markets for SMEs worldwide); 2) technological developments in the areas of production, transportation and communication (IT); 3) the increased importance of global networks and alliances; and 4) more elaborate capabilities of people, including those of the founder/entrepreneur who starts the born global/INV. These factors are interrelated. So far, these different driving forces have been only superficially explored, and they are not easily integrated in most of the theoretical frameworks of reference (Rialp et al 2002). Furthermore, it can be expected that such trends will be even stronger in the next years, thus making the phenomenon of INVs more widespread in the future.

Profiling the INV based on empirical findings

Many authors have attempted to characterise these firms based on empirical evidence and past literature (McKinsey & Co 1993; Knight, 1997 and Rialp et al 2002). Now, at this stage in the literature, a profile of the attributes of INVs may be established with little controversy.

The first piece of pioneering empirical work was carried out by McKinsey & Co (1993) on a group of small Australian exporters. Based on their findings, the report outlined the key characteristics of INVs. As will be discussed below, the results of several later studies have supported most of these characteristics:

- Management views the world as its marketplace from the outset of the firm's founding. Unlike traditional companies, they do not see foreign markets as simple adjuncts to the domestic market.
- INVs begin exporting one or several products within two years of their establishment and tend to export at least 25% of total production.
- They tend to be small manufacturers, with average annual sales usually not exceeding \$100 million.
- The majority of INVs are formed by active entrepreneurs and tend to emerge as a result of some significant breakthrough in process or technology.
- They may apply cutting edge technology to developing a unique product idea or to a new way of doing business.
- The products that INVs sell typically involve substantial value adding; the majority of such products may be intended for industrial uses.

So, the INV firm often possesses a knowledge-based competitive advantage that enables them to offer value-added products and services (McKinsey & Co., 1993). These characteristics have been widely accepted amongst scholars as these are quite generic and can be easily applied to most INVs. Larimo (2001) sums up the nature of these firms following his research on Finnish INVs, which are focused on niche, high-tech fields and base their competitiveness on high quality, technological innovation, close relationships with customers, and networking. Equally, in Jolly et al (1992), case study research on four high tech start-ups also emphasised the importance of having niche focus in foreign markets, global vision from the outset, and special emphasis on customer value and product quality at the heart of their strategy.

Empirical research carried out in the Northern European countries also demonstrates the prevalence of the INV phenomenon. Building upon a definition along the criteria suggested by Knight (1997), Madsen, Rasmussen, and Servais (2000), suggest that Danish INVs have a quite distinct profile when compared with other type of exporters in terms of product and market characteristics, geographical markets served, entry modes, and control of marketing activities. According to the study, the INVs do not follow the stages approach and do not set up sales or production subsidiaries abroad. Instead, they appear to be comfortable with operating at arms length through different types of collaborative arrangements with foreign partners (agents, distributors, suppliers, etc). They do build on experiential knowledge but they use this to gradually internationalise. Also, INVs have a unique profile compared with all other groups of exporters. For instance, they do not focus on a specific geographical region in their most important export market: rather they seem to target a narrow customer segment which may be allocated in different geographical places: this niche strategy is consistent with other studies of INVs (Zuchealla, 2002; Moen, 2002)

In relation to export channels, INVs rely much more on foreign distributors and less on direct sales, compared to other firms who engage in more FDI via established subsidiaries. This export strategy is consistent with firms with the same low resources of INVs given their small size and age at internationalisation. Thus network patterns, collaborative arrangements, and hybrid structures have appeared to be a more productive and cost effective route to take for the INVs. There are, however, similarities between large International firms and INVs in terms of using specialised production as part of their competitive strategy and strong global orientation. (Rasmussen, Madsen & Servais 2000)

McDougall et al (1994), from extensive case study research, found that the success of INVs was primarily due to an international vision from inception, an innovative product or service marketed through a strong network, and a tightly managed organisation focused on international sales growth (Jolly et al 1992; McDougall, Shane and Oviatt, 1994). Based on the literature, the most important and recurring attributes and characteristics of INVs are identified and expounded upon below. They are:

- Global orientation and entrepreneurial vision
- Experiential knowledge prior to start-up initiation
- Niche-focused and customer orientation
- High levels of product customisation
- Networks and business relationships
- Technological superiority as the source of competitive advantage
- Focus on high technology sector
- Export strategy

Global orientation and entrepreneurial vision

“The most distinguishing feature of INVs is that they tend to be managed by entrepreneurial visionaries who view the world as a single, borderless marketplace from the time of the firm’s founding” (Knight & Cavusgil, 1996:12). Unsurprisingly, a global orientation and entrepreneurial vision have been the two most prominent characteristics of INVs based on empirical findings, especially when compared to locally-based and traditional exporting firms, which have been long established prior to exporting. Prior to McKinsey’s report, a number of case studies had emerged in the 1980s documenting the existence of firms that were internationally oriented from birth. Ganitsky (1989) referred to these as innate exporters, which were more agile and more internationally focused from a management perspective. Jolly (1992) used a number of case studies of high-tech firms labelled technological start-ups. One of the characteristics of these firms was a founding entrepreneur who possessed a wealth of international experience and followed a strategy towards international niche markets, serving them with standardised products. Their findings concluded that these firms had a global vision from the start, and confirmed from their cases that the “the founders were able to inculcate their global vision throughout their organisations” (1992:74). Individual traits of the founding entrepreneur were important in shaping the future orientation of the firm. These attributes tie in with McDougall & Oviatt’s definition of International Entrepreneurship (2000).

The global orientation of the manager has been a key factor related to the export performance of the firm (Moen 2001; Zou & Stan 1998). A strong global orientation has been more readily associated with INVs. Moen’s research (2001; 2002) on Norwegian and French INVs showed that these firms possessed characteristics of global orientation from the start at founding, and that decisions made at founding influence the future path dependency of the firm which would ultimately play a key determinant in whether a firm would be an INV or a Local New Venture. Moen (2001) interprets global orientation to include international vision, proactiveness and customer orientation. This definition also extends to knowledge of customers’ markets and the use of information and communication technologies (ICT) to facilitate the business activities. He found that a distinguishing factor between global and locally based non-exporting firms was the former had a stronger global orientation (customer, vision and commitment /knowledge) and this was suggested as a key factor explaining why some firms remain in their home market ... the absence of a global orientation. Moen (2002) also found that both Norwegian and French INVs view the world as their marketplace, communicate the importance of their export activities to all employees, and focus on developing resources for their export activities.

The study concludes that a firm is either an INV or Local New Venture based on the vision of the firm at start-up.

Similarly, based on their case studies of INVs, McDougall & Oviatt (1994) have highlighted the importance of international vision from inception, an innovative product or service marketed through a strong network, and a tightly managed organisation focused on international sales growth. Moreover, the INVs are led by alert entrepreneurs, who are able to link resources from multiple countries to meet different international markets. These people are individuals who see opportunities from establishing ventures that operate internationally from inception.

Experiential knowledge prior to start-up initiation

Other studies have shown that the founders career background and industry knowledge are key enablers for early internationalisation (Madsen & Servais, 1997; Moen, 2001; Larimo 2001; Larimo & Pulkkinen, 2002). These assets allow them to combine a particular set of resources across national borders rather than just on their home markets, thereby forming an INV.

Christensen and Jacobsen's (1996) study of Danish INVs report that the firms use "established contacts and knowledge acquired prior to the initiated new business"(p 7). So, the entrepreneur's market knowledge, his personal network of international contacts and the experience transmitted from former occupation, are important factors. Harveston et al (2001) pay strong attention to entrepreneurial orientation and international entrepreneurial ventures. Based on their findings on UK high-tech entrepreneurial firms, international entrepreneurial orientation was highly important in terms of risk taking, and the proactiveness of the founder. Wickramasekera & Bamberry's (2001) study of Australian wineries found that accelerated internationalisation is brought about by management experience in the industry, international market knowledge, and overseas contacts (networks). Thus, whereas a firm may be new in terms of legal foundation, it may be much older in terms of the length and variety of the management experiences and the access to external networks embodied in their managers. (Welch & Loustarinen, 1988; Madsen & Servais, 1997).

Niche-focused and customer orientation

The literature suggests that most INV strategies have tended to adopt an international niche-focused strategy, serving narrow global segments with highly specialised products combined with a strong customer orientation (Moen 2002; Larimo 2001; Aspelund and Moen, 2001; Bell 1995;). Moen 2002 showed that, based on a sample of Norwegian firms, small firms were more likely to follow niche-focused strategies than larger firms. In Bell's (1995) transnational

study of small computer firms, sectoral targeting and specialised niche-focused strategies featured as highly characteristic of firms in Finland, Ireland and Norway. More recent research on Finnish INVs (Larimo 2001) also found these firms to be focused on niche markets or high-tech fields and the firms had based their competitiveness on quality, technological innovativeness, and close relationship (customer orientation) networking. Online technologies have made closer customer relationships easier to manage worldwide.

Product customisation

There appears some disagreement in the literature in relation to the degree of product customisation. Mainstream research states INVs are customer orientated and engage in product specialisation according to client requirements. Similarly, McKinsey & Rennie (1993) suggested that INVs tend to customise their products according to customer requirements in international markets. This view has been challenged by findings which suggest that high-tech start-ups choose a business area with homogenous customers which means minimal adaptation in the marketing mix as a means to lower costs and achieve economies of scale globally. Further, Jolly et al (1992) found that their firms offered standard products to global niche markets where a minimal amount of customisation occurred. This is further supported by a study on Italian INVs which identified firms that adopted this global niche strategy by offering standard but luxury goods without incurring costs of localisation. Although this warrants further investigation, it is true that advancements in technology have enabled small entrepreneurial firms to engage in product customisation through flexible manufacturing.

Networks and business relationships

McDougall & Oviatt, following extensive case study research (1994), found their firms preferred to use hybrid governance structures for their international activities to conserve resources during the cash-draining formation process. They found obvious differences between established firms and INVs in relation to resource base, and claimed that the INV entrepreneur must rely on hybrid structures for controlling sales, distribution and marketing activities abroad. Rasmussen and Servais's (2000) findings elaborate further on network, identifying three types of networks used by INVs. The first is the personal network, established by the founder prior to firm founding; the second is the network established through sales cooperation with distributors, agents, etc; and third is the network in development, both for production and sales, created through canvassing, participating in trade fairs, etc.

Research on INVs in Italy (Zuchella, 2002) revealed that one of the critical success factors for firm survival is the ability of the firm to build up international network alliances of a predominantly non-equity nature. Networking was perceived to be a strategic option as a way

to expand geographically. Networking was shown to be vital for firms of small size with related resource constraints to compete internationally. Further, Burgel and Murray (1999) found that their firms were inclined to engage in collaborative relationships with intermediaries to conduct sales abroad as a means to gain access to resources and capabilities that they do not possess. Zuchella (2002) also noticed that the location of the INVs in industrial clusters created a lever for internationalisation not present for firms in non-industrial districts. A further feature that has emerged in the literature is the role of *Industrial Clusters* as a growth environment for INVs. Moen (2001) concludes that INVs often operate in highly international industrial sectors, with firms that were likewise born in an industrial cluster located in the same geographic area (i.e. Italian case) with a tradition of servicing customers worldwide (pg 170). Rasmussen & Servais (2000) found that their Danish INVs relied on resources in a industrial district during their founding process. Some of the other firms were tied to an industrial district abroad. Evidence appears to be growing about the role industrial districts play in supporting INVs, particularly in terms of access to resources during the founding process. The importance of industrial districts have been a recent finding in INV research and an interesting addition to the profile of INVs (Andersson & Wictor, 2000).

Technological superiority as the source of competitive advantage

Competitive advantage literature has mainly focused on product or technology and price marketing advantages (Porter 1980). Most studies have favoured technology as the main source of competitive advantage for INVs (Rennie 1993; Moen 2002; Bell 1995; Knight, 1997; Larimo 2001). Rennie (1993) describes INVs as competing on quality and value created through innovative technology and product design. Moen (1999; 2002) also found that INVs were stronger in product technology than local (non-INV) firms. Aspelund and Moen (2001) recently identified three generations of small exporters according to their year of establishment: the traditional exporters, the flexible specialists, and the INV generation (comprised of the 36 firms in the sample which were established in or post 1989). Building upon a sample of 213 Norwegian exporters, these three generations of exporting firms were then compared using competitive advantage, manager orientation, market/environment features, and export strategy. The results showed that the various generations of exporting firms had different export performance antecedents, the INVs being those which were found to have technological advantage and niche focus combined with strong customer orientation as the key factors in determining export performance.

More recent evidence on Finnish INVs (Larimo & Pulkkinen, 2002) came from a survey of 470 Finnish firms which investigated the concepts of global orientation, competitive advantage and

export strategies, and their links with international intensity and age of SMEs. The study attempted to identify the differences between established and newly exporting firms. Competitive advantage was not only stronger in INVs than old and new locally established firms in manufacturing sectors, but technology advantage was the core competency of the INVs and was used to build up a position that enables them to rapidly expand to foreign markets. Similarly, Bloodgood et al's (2000) research findings on 61 high-potential ventures in the US across different industries also found that internationalisation was directly related to the use of product differentiation as a source of competitive advantage, mainly through high technology. Other important factors included the international work experience of the board of directors and the firm size at the time of Initial Public Offering (IPO) (Bloodgood et al, 2000).

Sectoral context

High-tech firms have become the favourite empirical contextual model for research on INVs. Much of the literature has focused on high-technology sectors (Burrill & Almassy, 1993; Burgel & Murray, 1995; Autio et al, 2000; Zahra & al, 2001; Crick & Jones, 2000, Jolly et al, 1992; Knight & Cavusgil 1996; Jones, 1999; Larimo 2001). This may have created the false impression that INVs are solely of a high-tech nature, rendering the nature of product and industry a qualifying feature of the INV profile. Following a limited number of empirical studies, INVs do not necessarily operate solely in high-tech fields but also exist across various low-tech industries, such as Services, the crafts industry (McAuley, 1999; Fillis, 2000), the Seafood sector (Bell et al 2001 and McDougall & Oviatt, 1994), and the Wine industry (Wickeraksama & Bambera 2001).

Based on research on the New Zealand seafood industry, Bell et al (2001) show that most seafood firms can embark on a rapid and dedicated internationalisation path from inception. They offer value added products. The study concluded that firms in the New Zealand seafood sector display the attributes of INVs (McDougall & Oviatt) despite the traditional nature of the sector. However, it has been argued that the nature of the industry has been influential in accelerating the international process of firms. In Coveilly and McAuley's (1999) study on New Zealand INVs, they found that patterns of firm internationalisation did not follow the traditional stages model. The authors claimed this behaviour was influenced by the nature of the industry the firms were operating in (Bell 1995) which was associated with relatively short product life cycles, high levels of competition, and a small domestic market, thus driving these firms to go abroad. This view has contributed to the sectoral bias in the literature towards hi-tech sector, as mentioned above.

Many studies have discussed the prevalence of firms that fit the INV definition which come from non hi-tech sectors, but which can also be classified as knowledge-intensive and as adding value to their products. Also, research conducted on Danish firms (Madsen & Servais 1997; Madsen, Rasmussen & Servais 2000) found that INVs were identified in all types of industries in both low and high-tech sectors. Studies on Italian industry sectors have revealed that INVs have emerged in non-high-tech industries (Zucchella, Marcarinni, 1999; Luostrairn & Makeisson, 2002). Thus, the INV does not only operate in high technology sectors. These firms also operate in traditional sectors as pointed out above. This context warrants further empirical investigation.

Export strategy

Despite the little empirical research has been conducted on the foreign entry modes and their determinants for INVs, important findings can still be found from those empirical studies by Linquist (1991), Bell (1995) and Schrader, Oviatt & McDougall (1997), and Rasmussen & Servais (2000). Based on Lindquist's research on Swedish firms (1991), the entry modes preferred were direct exporting and foreign sales through intermediaries. Similarly, in Bell's study (1995) on small computer firms, 70% of sales transactions were carried out by direct exporting or indirect exporting via agent and distributors. Firms with highly customised products relied on exporting directly.

In a more detailed examination on how INVs compared to other small established exporters, Rasmussen & Servais (2000) confirmed that the majority of INVs - young and small international firms - used a distributor in foreign markets. The second most common mode was to sell directly to the end user of the product in foreign markets. Establishing an office abroad appeared a 'no-go' option for small firms. Across a range of different industries, most Danish INVs relied on direct sales or relied heavily on intermediate modes of entry via agents. They did not engage in FDI by opening up a sales office or subsidiary. Rasmussen & Servais (2000) concluded that export agents and direct sales were dominant xport mechanisms amongst INVs and other small established exporters operating across many sectors.

This trend ties in with findings from Burgel and Murray's (2000) empirical analysis of high tech start-ups modes of entry in the US. Direct exporting appeared to be the most attractive amongst start-ups as it consumed less resources than using a distributor. Equally, Schrader, Oviatt & McDougall (1997) argue that different entry modes represent different degrees of resource commitment and their attendant foreign risk to the firm. Many start-ups experience cash flow problems in the early years and may lack the necessary human and financial

resources for effective commercialisation of their products. A pattern of entry mode across a wide range of sectors concludes that INVs would opt for the mode using the least resources, mainly direct exporting or indirect using an agent and possibly a distributor.

Country of origin

It is also worth noting the country context of these firms. The Scandinavian nations and the US are well represented in INV literature. The more recent empirical studies have been conducted in the Scandinavian countries of Norway, Finland and Denmark, being as they are the home countries of the researchers. A survey showed, however, that INVs are not very common in Sweden (Anderson & Wictor 2000). Country specific studies are based in the UK (Burgel & Murray, 2000), Finland, Denmark, Italy, and New Zealand. Two studies are based in Australia: first, the pioneering work of McKinsey & Co (1993); and second a sector-specific study on INVs in the Australian wine industry. A number of comparative studies have emerged in recent times: Bell's cross-country study on software INVs in Ireland, Finland and Norway; and Moen's (2002) study comparing exporting firms in France and Norway. The empirical research reflects the greater interest in the small, open, export-oriented economies of Finland, Norway and Denmark, traditionally dominated by SMEs with limited domestic market opportunities. The earlier studies on INVs were based in the US, led by the pioneering work of McDougall & Oviatt global case study research and by Gary Knight (1997) on US INVs. The large economies of Australia and the US remain an important research context for INVs.

Conclusion

This review has been limited to profiling INVs based on empirical evidence popular in the current INV literature. Several observations emerge from this review. There is sufficient consensus in the empirical findings to allow us to draw up a profile of the INV, notwithstanding the slight exceptions stemming from the degree of product customisation. Almost every author has attempted to elaborate on their own list of key success factors characterising INVs. Findings across studies are quite consistent. However, current empirical research is highly context-specific, primarily concentrating on the high-tech sectors. However, although research is limited, we know that INVs do not necessarily operate solely in high-tech fields but also exist across various low-tech industries, such as Services, the crafts industry (McAuley, 1999; Fillis, 2000), the Seafood sector (Bell et al 2001; McDougall & Oviatt, 1994), and the Wine industry (Wickeraksama & Bambara 2001). Further empirical investigation is required in low technology sectors and those that are traditionally low knowledge intensive that have internationalized early in their life cycle.

There is agreement about the rapid pattern of internationalisation which distinguishes these firms from others. They are characterised by their small size and flexibility. They possess an entrepreneurial drive and global vision from the outset. Much of this drive and vision comes from the proactive, experienced and committed managers/entrepreneurs. They tend to follow a niche strategy, enabled by closer customer relationships. They create unique, intangible knowledge-intensive offerings facilitated by technological innovation, usually associated with a greater use of IT. Their limited resource base has forced them to use different governance structures, such as hybrid forms through the strong use of personal and business networks (networking). Also, studies conclude that INVs would opt for the mode using the least resources, mainly direct exporting or indirect exporting: using an agent and possibly a distributor. These appear quite generic characteristics and could well be applied to firms regardless of size or life cycle.

However, some additional comments can be made regarding the profile of the INV. We can add a further characteristic based on the empirical findings of Zuchella (2002) and Rasmussen & Servais (2000). It is this: that the role of industrial clusters may be important breeding grounds for the INV, in terms of providing facilitating local and international network support enabling access to foreign markets. The majority of studies on INVs have been conducted in the high technology, knowledge intensive industries, with a limited amount in traditional industries. The latter context warrants further investigation. Also, concentration of research on hi-tech sectors limits the ability of theorists to apply findings to low-knowledge-intensive industries. Scant attention has been paid to service industries.

REFERENCES

- Aaby, N.E., & Slater, S.F. (1989). Management influence on export performance: a review of the empirical literature 1978-1988. *International Marketing Review*, 6/4, 7-22.
- Andersson, S. & I. Wictor (2001). Innovative Internationalisation strategies in new firms - Born Globals the Swedish case. Paper published in proceedings at the 4th MCGill conference on International Entrepreneurship, University of Strathclyde, Glasgow, Scotland, sept. 2001.
- Aspelund, A., & Moen, O. (2001) A generation perspective on small firms' internationalization- from traditional exporters and flexible specialists to born globals. In C.N. Axinn and P. Matthyssens, (eds.), *Reassessing the internationalization of the firm*, (Advances in International Marketing, 11) (pp. 197-225). Amsterdam: JAI/Elsevier Inc.
- Autio, E., Sapienza, H.J., & Almeida, J.G. (2000). Effects of age at entry, knowledge intensity, and imitability on international growth. *Academy of Management Journal*, 43/5, 909-924.
- Bell, J. (1995). The internationalization of small computer software firms: a further challenge to "stage" theories. *European Journal of Marketing*, 29/8, 60-75.
- Bloodgood, J., Sapienza, H.J., & Almeida, J.G. (1996). The internationalization of new high-potential U.S. ventures: antecedents and outcomes. *Entrepreneurship Theory and Practice*, 20/4, 61-76.
- Burgel, O. & Murray, G.C. (2000). The international market entry choices of start-up companies in high-technology industries. *Journal of International Marketing*, 8/2, 33-62.
- Cavusgil, S.T. (1980). On the internationalization process of firms. *European Research*, 8/6, 273-281.
- Chetty, S.K. (1996). The case study method for research in small-and medium-sized firms. *International Small Business Journal*, 15/1, 73-86.

- Christensen, P.R. (1991). The small and medium-sized exporters' squeeze: empirical evidence and model reflections. *Entrepreneurship & Regional Development*, 3, 49-65.
- Coviello, N.E. & Munro, H.J. (1995). Growing the entrepreneurial firm: networking for international market development. *European Journal of Marketing*, 29/7, 49-61.
- Coviello, N.E. & McAuley, A. (1999). Internationalisation and the smaller firm: a review of contemporary empirical research. *Management International Review*, 39/3, 223-256.
- Fillis, I. (2001). Small firm internationalisation: an investigative survey and future research directions. *Management Decision*, 39/9, 767-783.
- Granitsky, J. (1989). Strategies for innate and adoptive exporters: lessons from Israel's case. *International Marketing Review*, 6/5, 50-65.
- Johanson, J. & Vahlne, J-E. (1977). The internationalization process of the firm: a model of knowledge development and increasing foreign market commitment. *Journal of International Business Studies*, 8/1, 23-32.
- Johanson, J. & Vahlne, J-E. (1990). The mechanism of internationalization. *International Marketing Review*, 7/4, 11-24.
- Jolly, V., Alahuhta, M. & Jeannet, J-P. (1992). Challenging the incumbents: how high-technology start-ups compete globally. *Journal of Strategic Change*, 1, 71-82.
- Knight, G.A. & Cavusgil, S.T. (1996) The born global firm: a challenge to traditional internationalization theory. In S.T. Cavusgil & T.K. Madsen, (eds.) *Export internationalizing research - enrichment and challenges*, (Advances in International Marketing, 8) (pp. 11-26). NY: JAI Press Inc.
- Larimo, J. (2001). Internationalization of SMEs - two case studies of Finnish born global firms. Paper presented at the *CIMaR Annual Conference* in Sydney, Australia (November 20th), 1-21.
- Leonidou L. & Katsikeas, C. (1996). The export development process: an integrative review of empirical models. *Journal of International Business Studies*, 27/3, 517-551.
- Lu, J.W. & Beamish, P.W. (2001). The internationalization and performance of SMEs. *Strategic Management Journal*, 22, 565-586.
- Madsen, T.K. & Servais, P. (1997). The internationalization of born globals: an evolutionary process?. *International Business Review*, 6/6, 561-583.
- Madsen, T.K., Rasmussen, E.S., & Servais, P. (2000) Differences and similarities between born globals and other types of exporters. In A. Yaprak & J. Tutek, (eds.) *Globalization, the multinational firm, and emerging economies*, (Advances in International Marketing, 10) (pp. 247-265). Amsterdam: JAI/Elsevier Inc.
- McDougall, P.P., Shane, S., & Oviatt, B.M. (1994). Explaining the formation of international new ventures: the limits of theories from international business research. *Journal of Business Venturing*, 9/6, 469-487.
- McDougall, P.P. & Oviatt, B.M. (2000). International entrepreneurship: the intersection of two research paths. *Academy of Management Journal*, 43/5, 902-906.
- Moen, O. (2002). The born globals: a new generation of small European exporters. *International Marketing Review*, 19/2, 156-175.
- Moen, Ø. (1999). "The Relationship between Firm Size, Competitive Advantages and Export Performance Revisited." *International Small Business Journal* 18(1): 53-72.
- Wickramasekera, R. & Bamberry, G. (2001). Born globals within the Australian wine industry: an exploratory study. *Working-Paper* No. 1/01, Charles Sturt University, Wagga.
- Oviatt, B.M. & McDougall, P.P. (1994). Toward a theory of international new ventures. *Journal of International Business Studies*, 25/1, 45-64.
- Oviatt, B.M. & McDougall, P.P. (1997). Challenges for internationalization process theory: the case of international new ventures. *Management International Review*, 37/2 (Special Issue), 85-99.
- Oviatt, B.M. & McDougall, P.P. (1999) A framework for understanding accelerated international entrepreneurship. In A.M. Rugman, & R.W. Wright, (eds.) *Research in global strategic management: international entrepreneurship* (pp. 23-40). Stamford, CT: JAI Press Inc.
- Pulkkinen, J. & J. Larimo (2002). Global orientation, competitive advantages and export strategies of different types of SMEs: Empirical evidence from Finland. *Paper presented at the European International Business Academy (EIBA) Annual Conference, December, Athens, Greece*.
- Rennie, M. (1993). Global competitiveness: born global. *McKinsey Quarterly*, 4, 45-52.
- Rialp, A., Rialp, J. & G. Knight (2002) The Phenomenon Of International New Ventures, Global Start-Ups, And Born-Globals: What Do We Know After A Decade (1993-2002) Of Exhaustive Scientific Inquiry. *Paper presented at the European International Business Academy (EIBA) Annual 28th EIBA Conference, December, Athens, Greece*.

- Roberts, E.B. & Senturia, T.A. (1996). Globalizing the emerging high-technology company. *Industrial Marketing Management*, 25, 491-506.
- Servais, P. & Rasmussen, E.S. (2000). Different types of international new ventures. Paper presented at the *Academy of International Business (AIB) Annual Meeting* (november), Phoenix, AZ., USA, 1-27.
- Welch, L. & R. Luostarinen (1988) Internationalisation: Evolution of a concept. *Journal of General Management*, 14/2, 36-64
- Zahra, S.A. & George, G. (2002) International entrepreneurship: the current status of the field and future research agenda. In M. Hitt, R. Ireland, M. Camp, & D. Sexton, (eds) *Strategic leadership: creating a new mindset* (pp. 255-288). London, UK: Blackwell.
- Zahra, S.A., R. D. Ireland & M. Hitt (2000). International expansion by new venture firms: international diversity, mode of market entry, technological learning, and performance. *Academy of Management Journal*, 43/5; 925-950.
- Zou, S. & Stan, S. (1998). The determinants of export performance: a review of the empirical literature between 1987 and 1997. *International Marketing Review*, 15/5, 333-356.
- Zucchella, A. (2002). Born Globals Versus Gradually internationalising firms : an analysis based on Italian Case studies. *Paper presented at the European International Business Academy (EIBA) Annual 28th EIBA Conference, December, Athens, Greece*

Justification of Investment in IT systems

Aidan Farrell

School of Computing, Dublin Institute of Technology, Kevin Street, Dublin 8

Contact email: aidan.farrell@dit.ie

Abstract

For a company, capital investment of any sort is weighed up before a decision is made to invest. It is true that the vast majority of investments for companies can be quantified financially. Investment in Information Technology (IT) and Information Systems (IS) however has proved more complex than other investments as there are a large amount of intangible and non-financial benefits associated with this area of expenditure. Investments are traditionally rationalised by outweighing the costs and the benefits. The indirect costs associated with the deployment of IT/IS are equally difficult to put a measure on and hence the traditional methods of appraising IT/IS investments have proved to be inappropriate in this capacity. This paper details the lack of commitment by companies to fully justify their investment in IT/IS due to the apparent problems associated with measuring costs and benefits. After researching the areas of costs, benefits, risks, valuation and evaluation techniques, this paper provides a new framework for justifying investment in IT/IS. The framework proposes extensions to the current processes used by decision makers when justifying investment in IT/IS and may provide an additional tool to justify investment more accurately in this area.

Keywords: justification, investment, expenditure, benefits, decisions, information systems, information technology.

1 Introduction

Changes in technology have affected the environment in which organisations operate dramatically over the last few years. Technology has also changed the way organisations use IT/IS, (Irani, 1999). A globalised market and the influence of the Internet change the way organisations do business. This is changing the way people access information, communicate, shop and entertain themselves. This also changes the way businesses compete and operate, (Wessels, 2003). Decisions have to be made between modifying current systems and replacing existing systems with newer, more up to date technologies (Irani, 1999). Users in organisations are demanding that their information systems be more efficient and effective. With this, organisations are forced to invest heavily in information technology deployment and maintenance in order to obtain value and benefit and to stay competitive in this new fast paced, global environment. Therefore investment decision makers must acknowledge this change.

Although many IT/IS expenditure is regarded as costly and risky, many information systems investments go ahead without the use of formal investment appraisals or risk management techniques (Ward *et al.* 1996). Some organisations justify heavy investment in IT systems as an 'act of faith', where they have 'a gut feel' or assume that IT always pays off (Wessels, 2003). This is based on estimation and assumptions. The aim of this paper is to introduce and explain why traditional appraisal techniques used to justify investments in information technology and information systems are not suitable for this type of investment appraisal. Secondly, it aims to provide possible solutions to improve investment decisions by providing better information on how to identify and measure, more precisely, the costs and the benefits of information systems

to the decision-makers on IT/IS investment panels. Section 2 introduces the associated problems with justifying investment in IT/IS and its effects on the decision-making process and decision-makers. Section 3 provides an in depth look at the shortcomings of traditional methods when identifying and quantifying costs. It analyses methods of appraising the true costs and benefits of an IT/IS investment and looks at emerging valuation measures for justifying investment in IT systems. Section 4 outlines a new justification investment model or framework suited to information technology and information system expenditure, which may aid decision-makers justifying an investment in an information system. Section 5 offers some concluding comments.

2 Investing in information systems

Traditionally in business, capital expenditure has always been formally justified by means of appraising the benefits of the investment to the organisation. Capital investment can be purchasing a new fleet of vehicles for a delivery company or installing a new information system into a recruitment company. A traditional capital appraisal involves a statement of initial cost of the investment, the ongoing costs, the anticipated benefits and a calculation of suitable key performance indicators (KPI) or statistics (Wessels, 2003). Prior to the early 1990's, investment appraisal techniques were applied with relative ease, however with the changing nature of investment in information systems and the changing nature of the economic situation, management were finding it more difficult to apply these techniques to investing in IT/IS, (Wessels, 2003).

2.1 Associated problems

Organisations are currently finding it more difficult to justify the present levels of expenditure on information systems with many organisations not performing evaluations or cost benefit analysis (CBA) on their information systems at all. The fact that the benefits of IT systems are hard to quantify and are considered non-financial, those who do perform evaluations or (CBAs) sometimes report mixed or confused results (Remenyi, *et al.*, 1995). To justify an information system and show that it is appropriate for a particular business context, it must first be evaluated and then justified (Remenyi *et al.* 1995). This simple form of justification however cannot be applied to justifying an investment in information systems (Ward *et al.*, 1996). Investment in information systems is much more complex. Both the benefits and the costs of IT/IS are too complex to simply put down on paper due to their intangible and non-financial nature. It is hard to put customer satisfaction or improved workforce effectiveness on paper or in figures. The impact of IT/IS cannot easily be quantified in terms of benefits financially, which is the traditional medium to record both benefits and costs. Benefits are both tangible and intangible. Costs are both direct and indirect. As a direct result of this, most companies do not formally evaluate their investment in information systems. (Hochstrasser, 1992) reported that only 16% of companies used '*rigorous methods to evaluate and prioritise their IS investment*'. Further research from Hochstrasser found that, where investment appraisal of IS did take

place, it was usually based on financial techniques specifically designed to assess financial impact in terms of cost. In recent research about the justification of investing in information technology and information systems four issues have been identified as the core issues that prevent the practical application of appraisal techniques identified (Wessels, 2003). They are:

- The inability to clearly quantify the value from investing in a new information system.
- The complexity of models and methods suggested for solutions. They are too difficult to apply in practice.
- The process of human decision-making.
- The reach and range of information systems in the organisation.

There have always been problems to date over the ability to clearly articulate and quantify value. Volume and spending are easy to identify, unlike measuring value, which is much harder to define. Investment decisions are based on the human perception of value, however measured. The specification and implementation of information systems is often left to IT professionals themselves. There tends to be little or no involvement from managers or the users, which (Sauer, 1993) believes is the cause of many of the ineffective or failed information systems. (Earl, 1996) goes one step further and suggests that if information systems implementations are left to IT professionals and users alone, the investment is rarely recouped. This emphasises that management should be the ones who implement a new investment as they do so more efficiently and effectively.

The complexity of models and methods suggested in the past, have proved too difficult to apply in practice. Research by (Fitzgerald, 1998) shows that many companies define the costs associated with information technology in a narrow way. His study shows that they only include hardware and software. Organisations appear hesitant to include other costs in order to avoid putting further responsibility on a department. Costs on IT consumables such as paper, ink cartridges and disks are considered department costs and not IT costs. There were no Return on Investment (ROI) calculations on the IT investments because people believed them too difficult to carry out.

Decision-making process

As with all capital and large investments, management must make the decision to invest in an area of information technology or in an information system. (Simon, 1960) breaks the decision making process into different stages:

- Problem recognition
- Problem definition and structuring
- Identifying alternative courses of action

- Making and communicating the decision
- Monitoring the effects of the decision.

This model can be applied to making the decision on whether to invest in the information system or not. The decision point illustrated in the center of Figure 1 is the point at which management will decide to invest or not. Up to this point, the decision maker or decision makers will have to obtain as much information about the specific investment as possible in order to make the correct decision (Simon, 1960). The advantage of using this model is, it can illustrate the potential effects of an investment on the future of the organisation before a decision is made and hence reduce the risks of investing heavily in a system that may have little or no benefits to the company. The future of the organisation is influenced by the decision to invest or not to invest. Investment will either have a positive effect, no effect or a negative effect (Wessels, 2003). In order to aid the decision makers, the risk of investing and not investing are quantified. Different types of investments such as capital, revenue and strategic are analysed, evaluated and justified using different techniques, developed by professionals such as business analysts (Wessels, 2003). It is these techniques that have proved problematic for the area of information technology and information systems.

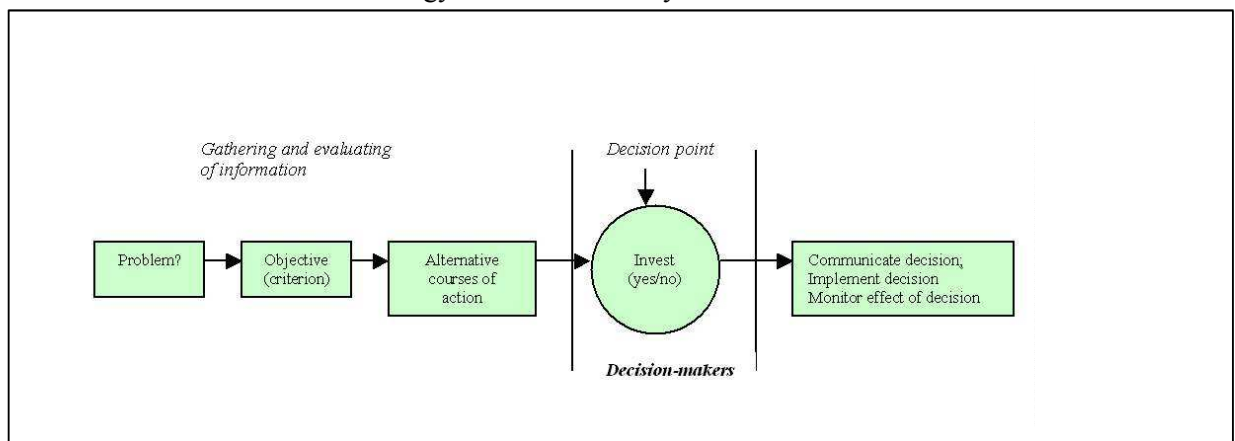


Figure 1: Decision-making process (Wessels, 2003).

The decision point, illustrated in figure 1 is the single most important moment of the decision making process (Wessels, 2003). The outcome of this decision point is affected by the amount of investigation and evaluation of the potential outcome of the investment by the organisation. The amount of research carried out has proved to differ vastly between the organisations, investments and decision makers. The investment techniques adopted by the decision makers are a primary tool used by decision makers to reach an investment decision. The decision to invest has to be made and it's effects monitored on a continuous basis. This makes it possible for the decision-maker to judge if their decision resulted in the expected and desired effect (Kelly, 2003). In the next section, the techniques that have proved successful for decision

makers are analysed for their value to decision makers investing in information technology and information systems.

3 Solutions to justifying an investment

The use of traditional appraisal techniques to justify investments in information technology and information systems area has received a lot of attention (Irani, 1999). The reason for this growing interest is due to the vast sums of money being spent on information systems and on the increasing need to justify this significant capital expenditure (Irani, 1999). Traditional appraisal techniques are a mixture of financial ones and strategic management ones. In the list of techniques below there are six, which are financial and have a monetary value. The last three are intangible but have been accepted in business management as techniques to appraise information systems investment. For this reason I have included them as traditional techniques.

- Cost Benefit Analysis (CBA)
- Return on Investment (ROI)
- Return on Capital Employed (ROCE)
- Payback
- Net Present Value (NPV)
- Internal Rate of Return (IRR)
- Compliance to Standards
- Risk Avoidance
- Scare Tactics

Over reliance on these traditional methods may lead to an excessively conservative IT portfolio and in turn a loss of competitiveness (Wessels, 2003). There are many reasons why these techniques are inadequate for appraising investment in complex information systems. Considering most IT/IS investments offer long-term strategic benefits, the focus of these techniques on the short term is one of the reasons for their inadequacy (Lefley, 1994). Organisations investing in information systems are usually replacing or adding to current systems, which are no longer appropriate for the tasks, the current business environment (described in section 1) demands. Some common reasons why organisations invest heavily in information technology and information systems include: (Defensive strategy)

- Changes in the business environment
- Risk avoidance e.g. keeping up with competing organisations
- Functionality problems with the current information system
- Technical limitations with systems
- Compliance with standards e.g. website accessibility for the visually impaired

- Scare tactics e.g. “if we don’t upgrade we’ll have serious problems in the future”

The other side of investing is more of an (Offensive Strategy), where an organisation aims to improve its efficiency, effectiveness and competitiveness. The justification in investing in a new information system would therefore not only include cost issues but also functionality, alignment with its particular business processes, opinions of users and compatibility with current technology (Wessels, 2003). All of these must be right before investing in a new information system. The following sections analyse the various areas, which are core to justifying an investment. These sections offer solutions in the form of reference tables to better appraise true costs, strategies to identify possible benefits and merging techniques to evaluate these more accurate costs against benefits. The result of this analysis is brought together in the form of a justification investment framework in section 4.

Appraising the true costs

IT/IS deployment can often be divided into direct and indirect costs (Hochstrasser, 1992). The following section will address both types and identify some important tangible and intangible costs, which are often overlooked. Tables 1, 2 and 3 offer solutions to these overlooked costs and can be used as reference lists when identifying costs in the future. They are represented in the investment justification framework under ‘identify costs’.

Direct project costs

Direct costs are those costs, which can easily be associated with the implementation or operation of information technology or information systems. Senior management often make decisions on what the projects budget should be and the ultimate investment decision based on what they think the project costs are. The reality is that direct project costs are often underestimated (Hochstrasser, 1992). The costs go far beyond the typical hardware, software and installation costs. Installation and configuration costs, which take in consultancy fees, installation engineers and networking hardware/software, are also classified as direct costs (Irani *et al.*, 1998). Direct costs can include unexpected additional hardware accessories, which typically happen when implementing new information systems on old hardware. Increases in processing power, memory and storage are not uncommon. Table 1 shows the direct project costs and some examples associated with IT/IS implementation, which should be used to improve the accuracy of identifying direct costs in the investment justification framework.

Direct project costs associated with IT/IS implementation	Examples of direct project costs associated with IT/IS implementation
Environmental operating costs	Air-conditioning facilities Uninterruptable power supply Computer furniture
Initial hardware costs	File server Terminals and Network printer
Initial software costs	Software packages and Networking software Operating system
Installation and configuration costs	Management consultancy support Installation engineers Network wiring, junctions and connectors
System development costs	External customising time, In-house customising time
Project overheads	Running costs: electricity, space Networking costs: telecommunication time, Rises in insurance premiums
Training costs	Vendor software familiarisation courses Software upgrade training courses
Maintenance costs	Yearly service contracts
Unexpected hardware costs	Secondary data and storage devices Upgrades in processing power
Unexpected software costs	Vendor module software upgrades Operating systems upgrades
Security costs	Protection against viruses and abuse
Consumables	Print cartridges/ribbons, disks and paper

Table 1: Direct project costs (Irani et al., 1998).

After looking at the issue of direct costs and the solutions reference table for direct costs, which may be overlooked when justifying investment, the next section looks at even more significant costs. Those costs, which are easier to overlook, are indirect costs associated with an IT/IS investment. (Hochstrasser, 1992) suggests that indirect cost may be up to four times greater than direct costs of the same project. These indirect costs can be divided into human and organisational.

Indirect human project costs

The single largest human cost is management time. (Irani *et al.*, 1998) This is specifically associated with integrating new systems into the organisations current work practices. One of the results of an organisation investing in a new information system will be the time management spend revising, approving and amending IT/IS related strategies (Irani *et al.*, 1998). Management will also investigate the potential of the new information system by experimenting with information flows and information reports. Example: The Human Resources department of an organisation implements a new information system. The administrative staff need time to absorb new operational work practices. The management of the administrative staff will require time to absorb new managerial work practices. During this period both parties will have developed new skills. Employees may seek salary increases or

benefits as a result of making themselves flexible to this new system and for learning new skills in the process. Any pay awards associated with the implementation of this new HR system, cost implications of increases of staff turnover will therefore be an indirect human project cost and should be into the justification criterion under costs. Table 2 below provides a solution to identifying some commonly underestimated or overlooked indirect human project costs, associated with the adoption of IT/IS.

Indirect human costs associated with IT/IS implementation	Examples of indirect human costs associated with IT/IS implementation
Management/staff resources	Integrating new systems into new/revised work practices
Management time	Devising, approving and amending IT and manufacturing strategies
Management effort and dedication	Exploring the potential of the system
Employee time	Absorbing the transition from traditional to new work practices
Employee training	Being trained and training others
Employee motivation	Interest in IT/IS reduces as time passes
Changes in salaries and structures	Promotion and pay increases based on improved employee flexibility
Staff turnover	Increases in recruitment costs: interview, induction and training costs

Table 2: Indirect human costs (Irani et al., 1998).

Indirect organisational project costs

Indirect project costs also effect the organisation, with new work practices emerging with the introduction of the new information system. At first, a temporary loss in productivity may be experienced, due to the employees going through a learning curve (Hochstrasser, 1992). Management may attempt to exploit the new system to its full potential in a strategic level and in turn additional organisational costs will incur. Example: An organisation set up an Electronic Data Interchange (EDI) link between a customer and a supplier. The implementation of this system will have knock on technology and cost factors. The cost factors will be both direct and indirect. The most important aspect of indirect organisational costs occurs when such a knock on project reduces the number of management levels it has. This is typical of companies with extensive IT/IS installations, which often leads to a changing of the corporate shape (Hochstrasser, 1992). The cost of restructuring an organisation is considered very high particularly where groups within resist change and are unwilling to make the transition. These indirect costs should therefore be built into the justification for any new IT/IS investment. Table 3 provides a list of indirect organisational cost associated with an IT/IS investment.

Indirect organisational costs associated with IT/IS implementation	Examples of indirect organisational costs associated with IT/IS implementation
Losses in organisational productivity	Developing and adapting to new systems, procedures and guidelines
Strains on organisational resource	Maximising the potential of the new technology through integrating information flows and increasing information availability
Business process re-engineering (BPR)	The redesign of organisational functions
Organisational restructuring	Covert resistance to change

Table 3: Indirect organisational costs (Irani et al., 1998).

Taking into account all three areas of costs, these three tables provide a comprehensive reference point for organisations that wish to eliminate underestimated and overlooked costs, associated with their IT/IS investment.

Appraising the true benefits

The benefits of information systems are a portfolio of tangible and intangible benefits. Tangible benefits are those, which can be quantified or assigned a financial value while intangible are much more complex. Intangible benefits cannot be assigned a monetary value. An example of common benefits from an investment, which offers intangible benefits would include a more efficient customer service or enhanced decision-making (Laudon & Laudon, 2000). This section represents possible solutions to the problem of identifying true benefits of investing in IT/IS systems. It analyses competitive advantage and the discipline of benefits realisation management in a bid to help identify benefits of IS investments more accurately. This section of the paper is identified in the investment justification framework in section 4 under ‘identify benefits’.

Competitive advantage as a benefit

Measuring the benefits of IT alone has been one of the major areas of research over the last 15 years (Davis *et al.*, 2003). It is accepted that IT does not pay off in the same way that traditional investment do. It is too difficult for management of an organisation to measure the return of an information systems investment. However management have always believed that IT/IS has the potential to provide them with a competitive advantage (Davis *et al.*, 2003). A competitive advantage is a business terminology used to define the position of one company performing better than rival companies. Competitive advantages are found by benchmarking one company’s performance to a competitor’s in the same industry. A competitive advantage provided by an investment in IT is yet another intangible benefit. It is difficult to financially

calculate having superior performance over competitors in any industry. When an organisation takes strategic actions, the performance relative to competitors is a measure of whether a competitive advantage was achieved or not. Traditional accounting techniques for measuring the performance of a competitor were popular in that accounting measures were publicly available for many companies (Davis *et al.*, 2003). This proved to be a useful tool in capturing information on competing companies operations.

Benefits realisation/management

Benefits management is the process of organising and managing such that the potential benefits arising from the use of IT are actually realised (Bennington, 2003). It is a management process not to be confused with cost/benefit methodologies. It is important to note that IT/IS alone do not create business benefits for an organisation. Benefits flow from using IT/IS. People in turn are enabled to work more effectively and efficiently and it is these actions that actually produce the benefits (Kelly, 2003). One of the problems with justifying investment in IT/IS is that the benefits are intangible and non-financial therefore they are hard to document. For this reason, a management process has to be put in place to identify the benefits with an investment otherwise they will not appear (Bennington, 2003). The result of IT/IS investment can be turned into a benefit using strategic management. This benefit can then be used in the decision-making process when justifying the investment. The benefits management process has five sub-processes and derives from the business strategies and plans.

How to identify and structure benefits

Step one from the benefits management process is identifying and structuring the benefits associated with investing in this new IT/IS venture. (Bennington, 2003) describes the typical key benefits of IT investments as enhanced employee productivity/efficiency, saved money or reduced costs, improved accuracy/presentation of information and compliance with legislation/regulations. But one of the problems faced by decision makers is in itself identifying the benefits of IT/IS investment. Benefits are complex, they evolve overtime, they are intangible and hard to measure. So how are benefits measured in benefits management? (Bennington 2003) identifies interviewing key stakeholders in workshop environments and examining the project objectives and deliverables as two ways.

Workshops are used to bring together all the key stakeholders to agree on the benefits that are required from the IT investment. Depending on the level of investment, it may be necessary to have a break down of different benefits and designate a separate workshop for each benefit. As the benefits are broken down, individuals are assigned benefits to strategically manage and

produce. Benefits management focuses on who is responsible and when will the benefits be achieved (Kelly, 2003). Key performance indicators (KPIs) are necessary to benchmark the process of whether an organisation has benefited in some way from the investment. They identify benefits to be measured, determine if benefits have been delivered, assign accountability to individuals or individual departments and allow management to make strategic decisions based on these key performance indicators.

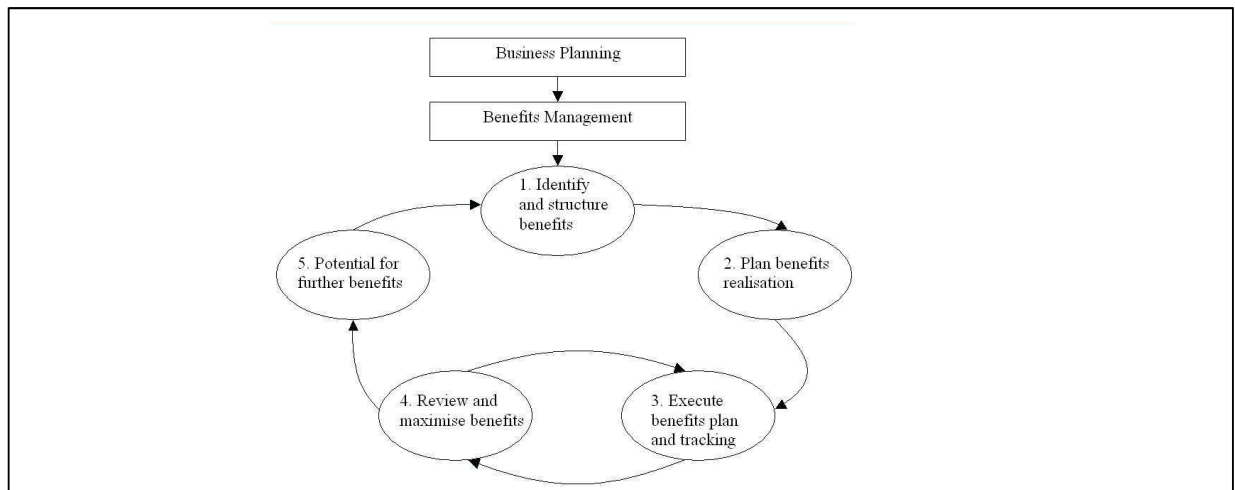


Figure 2: The benefits management process (Kelly, 2003).

Plan benefits realisation

The problem with benefits in relation to investing in IT is that they are difficult to identify and measure when outweighing with costs. To plan benefits realisation, structures must be put in place to help identify categories of benefits to help bring all benefits to the decision-maker's attention. There are three categories that benefits can be broken down into (Kelly, 2003). They are as follows:

- Financial – where financial value can be measured.
- Objective – where a percentage or number is most appropriate or a certain
- Criteria has been met which previously was not e.g. legislation compliance.
- Observable – where a financial or objective benefit cannot be found but where an individual may interpret a benefit has occurred e.g. customer satisfaction, image, good will and good name.

Execute benefits plan and tracking

The benefits being planned for must be executed and tracked. The important questions to answer in step three are where the benefits will occur, who will be responsible for them, who

receives the benefit, when will the benefit be realised, what action will the organisation have to take to deliver the benefit and how is it linked to output? (Kelly, 2003).

Review and maximise benefits

Project managers do not monitor benefits because they focus on managing the deliverables (Bennington 2003). In step three, persons within the organisation are assigned a benefit or benefits to plan and track. They must review these benefits with a view to maximising the benefit's value to the project.

Why realise benefits and the potential for further benefits?

It is valuable to an organisation to know what benefits they have achieved and not achieved. That way they can weigh up the benefits in reality against costs as opposed to comparing potential benefits and costs. A more realistic picture is painted for the decision-makers with actual benefits realised. These benefits will also be documented for future investment projects. It is recognised that those responsible for the realisation of a benefit remain accountable for that benefit. Benefits management is an appropriate process to use when justifying investment in IT/IS which can bring benefits to the attention of the decision-makers which otherwise have not been identified. Reasons why organisations do not identify benefits include pressure to deliver other projects, many of the IT/IS benefits are intangible or it is either too difficult or too costly to do so (Bennington, 2003). With this in mind here are some final points, which identify why benefits management could help decision-makers trying to justify investment in IT/IS:

- Senior management has doubts IT delivers real business benefits
- Benefits are often overstated to gain project approval (Kelly, 2003)
- It's difficult to predict benefits in advance.

Emerging IT/IS valuation measures

As discussed above, by far the most popular way of measuring performance is traditional accounting measures, easily applied to investments which have definite costs and whose benefits can be measured accurately. However, there are emerging IT/IS valuation measures, which can help to justify investment in IT/IS more easily and more accurately (Davis *et al.*, 2003). These valuation measures are not dependant on just tangible costs or benefits like the traditional techniques of Return on Investment (ROI) and Payback. Below is a list of the valuation techniques and a brief description of how they overcome the shortcomings of the more traditional valuation techniques. These merging valuation techniques are solutions to the problem of justifying the investment where the benefits of the investment appear to be

intangible and hard to quantify. The following section is represented in the investment justification framework in section four under 'justify – emerging valuation techniques'.

- **Balanced scorecard**

Integrates traditional financial measures described above with three key performance indicators (kpi's), (Davis *et al.*, 2003). They are customer view, internal business processes and organisational growth, learning and innovation.

- **Portfolio investment management**

By calculating risks, yields and benefits, this measure manages the IT assets from an investment perspective (Kelly, 2003).

- **Total Cost of Ownership (TCO) and Benchmarking**

The purpose of benchmarking is to gain sustainable competitive advantage. IT benchmarking is more of a concept. What is more recognisable may be the concept of Total Cost of Ownership (TCO), which compares the fully loaded cost per PC of an organisation to the same measure in other organisations (Kelly, 2003). TCO is just one component within the context of IT benchmarking.

- **Applied information economics**

Uses scientific and mathematical methods to evaluate the IT/IS investment process (Davis *et al.*, 2003).

- **Economic value added**

Is concerned with the true economic profit of an investment (Davis *et al.*, 2003). It is calculated by measuring the net operating profit and deducting the cost of all capital invested in the enterprise including technology.

- **Economic value sourced**

Calculates risk and time in monetary form and in turn adds these to the overall valuation equation (Davis *et al.*, 2003).

- **Real option valuation**

Values the flexibility of the organisation, tracking assets in place and growth options. This presents a wide variety of possibilities for the organisation in the future (Davis *et al.*, 2003).

- **Benefits Management/Realisation**

The process of organising and managing such that the potential benefits arising from the investment in IT/IS, is actually realised (Bennington, 2003).

4 **Investment justification framework**

The following section outlines some ways to improve the traditional techniques of justifying investment in information technology and information systems. This investment justification

framework does not supplement or replace the past techniques but is more of an extra measure to improve the accuracy of traditional techniques which have proved ineffective for justifying IT/IS investments. The framework was derived from the research carried out in this paper, looking into ways to identify, measure and track the benefits and the costs of IT investments. Section 3.1 looked at solutions to appraising true costs of an investment through reference tables of commonly overlooked costs and provided examples of such costs. They are represented in the first phase of the framework under 'identify costs'. Section 3.2 looked at benefits realisation management and competitive advantage as a solution to identifying true benefits of an IT investment. These solutions are also represented in the first phase of the framework under 'identify benefits'.

Moving from left to right, after the cost and benefits have been ascertained, the second phase of the framework justifies the investment through traditional financial techniques discussed in section 2 where suitable. It also justifies the investment through emerging valuation techniques discussed in section 3 as solutions to the problems arising from the more traditional financial techniques. If investing in an IT system is justified, the third phase of the framework is used. This section of the model recommends the use of management to implement the investment, as opposed to ICT professionals. This will help ensure a more successful project.

- **Identify all three costs brackets**

In order to fully identify all costs, direct costs, indirect human costs and indirect organisational costs should be identified and measured. The tables identifying these costs in section 3.1 provide a solution in the form of reference tables to identify areas overlooked by management previously.

- **Carry out benefits management/realisation**

Section 3.2 analyses competitive advantage and benefits realisation management as solutions to appraising true benefits of an IT investment. Realising the potential benefits and not just financial benefits, available in organisational figures, is an important way to bring out intangible benefits, which may otherwise go undetected.

- **Justify the costs versus the benefits using traditional financial methods**

The more traditional financial methods identified in section 2 should still be used in conjunction with the emerging valuation techniques, particularly to outweigh the more tangible costs and benefits.

- **Justify the costs versus the benefits using emerging IT valuation measures**

The new and emerging IT valuation measures analysed in section 3.4 should be used for intangible costs and benefits. The intangible nature of IT costs and benefits has proved to be the primary problem in justifying IT/IS investments with any great accuracy.

- **Ensure management in the implement the IT/IS investment**

Management must specify and implement any investment in IT/IS, as ICT professionals can be the cause of many ineffective or failed investments. (Earl, 1996) goes as far as saying that in a lot of cases the investment is not recouped.

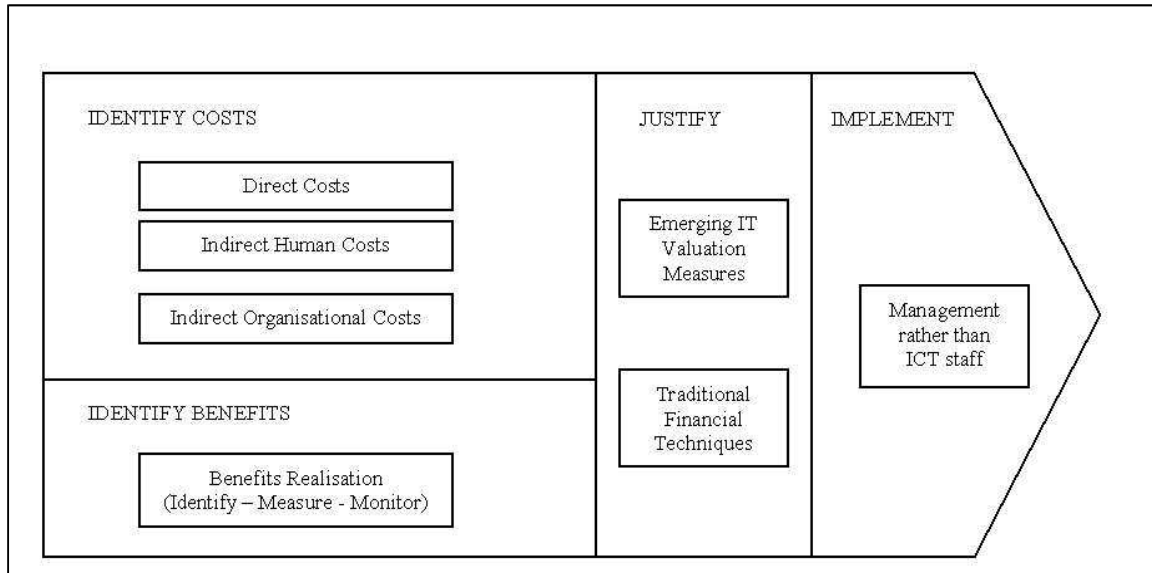


Figure 3: An investment justification framework (Author, 2003).

This model improves greatly on past techniques and other models, as it is not as complex as past models created to justify IT investment. It is developed as a left to right, linear model, which does not support traversing back once a phase is reached. This indicates that if an investment is not justified and will be not implemented, the decision-makers can not return to try and identify further benefits in order to force the justification of the investment to management. This would defeat the purpose of determining if an IT project investment offers a return. This framework is a useful aid to decision-makers justifying an investment in the area of IT systems. It offers a framework that is well-structured and not as complex as more traditional investment justification models detailed in section 2.1.

5 Conclusion

This paper has showed that the changes in technology in the last decade have affected the way organisations invest in IT/IS. In a global market driven by the Internet, organisations are forced to invest heavily in information technology deployment. If they are to obtain value and to stay competitive in this new global environment, they must begin to invest wisely. Section two has shown that the traditional justification techniques are not suitable for evaluating the intangible and non-financial costs and benefits, which are associated with IT/IS investments. These approaches have been shown to focus on short-term, non-strategic and tangible benefits. In a time when management are under mounting pressure to produce financial savings, there is a

possibility that IT/IS projects with a long-term focus could be excluded due to their intangible, non-financial benefits. The organisations, which reject the 'act of faith' policy and decide to justify the investment in IT/IS solely on financial/accounting strategies, will be faced with unrealistic costs and benefits. (Irani, 1998) provides numerous examples of the direct costs, indirect human costs and indirect organisational costs often overlooked by decision-makers. The benefits of IT/IS, which have always proved the hardest aspect of justifying IT expenditure, are realised more easily with the use of benefits management/realisation. Realising the potential benefits of investing in an information system will become an essential part of the justification process as it becomes more accepted at strategic management level. As it becomes clear that the traditional justification techniques are ineffective for IT/IS, new, emerging valuation measures will be used in justifying benefits and costs. The suggested justification framework in figure 3 incorporates these three aspects and concludes by identifying the problem with IT professionals specifying and implementing the investment. This should be left up to management as professionals rarely recoup the investment.

Many companies investing in IT/IS do so at great risk of failure. Organisations fail to see that IT is a long-term investment with intangible and non-financial costs and benefits. Companies fail to see that because of this, their accounting and financial techniques for justifying investment such as Cost Benefit Analysis (CBA), are not appropriate. As a result, there is an assumption that decisions are still being made on purely technical rational grounds. The investment justification model or framework could be used to improve an organisation's IT/IS investment justification strategy, identifying costs and benefits more accurately than the more complex models and formulae. It could also bring some issues like who should implement the information system if they are to invest, to the attention of decision-makers. The emerging IT valuation measures and the idea of benefits management/realisation are prominently put forward in the framework modeling it to be an up to date and useful tool in justifying investment in the IT area. This framework could easily be interpreted for investment in new web based media and evolving technologies such as mobile solutions. There is also scope to build upon this framework using more complicated justification techniques, incorporating formulae and statistics. Future research is necessary to show the success or failure of this newly proposed framework. A field study should be carried out, in order to obtain evidence to show its success or failure. This field study would involve applying an organisations IT/IS investment to this framework in order to come up with results for the different phases of the framework and hence prove how useful the framework is to this field of research.

6 References

- Bennington, P., (2003)**, “Benefits Management in IT Projects The Perth Perspective”, Government Employees Superannuation Board [Course Notes].
- Davis, L., Dehning, B., Stratopoulos, T. (2003)**, “Does the market recognize IT-enabled competitive advantage?”, *Information & Management*, no. 40, pp. 705-716.
- Earl, M.J. (1996)**, “Putting information technology in its place: a polemic for the nineties”, *Journal of Information Technology*, no. 7, pp. 100-108.
- Fitzgerald, G. (1998)**, “Evaluating information systems projects: a multidimensional approach”, *Journal of Information Systems*, no. 30, pp. 15-27.
- Hochstrasser, B. (1992)**, *Justifying IT investments. Advanced information systems: the new technology in today's business environment*, London: Chapman and Hall, London.
- Irani, Z., Ezingard, J-N., Grieve, R.J. (1998)**, “Costing the true costs of IT/IS investments in manufacturing: a focus during management decision making”, *Logistics Information Management*, vol. 11, no. 1, pp. 38-43.
- Irani, Z. (1999)**, “IT/IS Investment Justification: An Interpretive Case Study”, Proceedings of the 32nd Hawaii International Conference on System Sciences.
- Kelly, D. (2003)**, “Benefits Realisation Management”, Dublin: DIT. [Course Notes].
- Laudon, K.C. & Laudon, J.P., (2000)**, *Management information systems: organisation and technology in the networked enterprise*, Prentice Hall, New Jersey, pp: 354.
- Lefley, F. (1994)**, “Capital investment appraisal of manufacturing technology”, *International Journal of Production Research*, vol. 32, no. 12, pp. 2751-2756.
- Remenyi, D., Money, A., Twice, A. (1995)**, *Effective measurement and management of IT investments*, Butterworths, London, pp: 55-56.
- Sauer, C. (1993)**, “*Why information systems fail: A case study approach*”, Alfred Walter, Oxford.
- Simon, H.A. (1960)**, “*The new science of management decision*”, New York: Harper and Row, New York.
- Ward, J., Taylor, P., Bond, P. (1996)**, “Evaluation and realisation of IS/IT benefits: an empirical study of current practices”, *European Journal of Information Systems*, vol.5, no. 4, pp. 218-232.
- Wessels, P. (2003)**, “Justifying the investment in information systems”, *South African Journal of Information Management*, vol. 5, no. 2.

Architecture and development methodology for Location Based Services

Aaron Hand¹, Dr. John Cardiff²

¹ School of Science, Institute of Technology at Tallaght, Dublin 24

² School of Science, Institute of Technology at Tallaght, Dublin 24

Contact email: aaron.hand@itnet.ie

Abstract

This paper presents a LBS (Location Based Service) architecture, development methodology and a development tool to assist LBS developers in building new open standard LBS applications. The approach adopted has been to define new LBS systems based on open standards rather than proprietary driven technologies. SAGESS (Spatial Application Generic Environment System Standards) is an architecture platform for the development and deployment of wireless LBS applications. SAGE (Spatial Application Generic Environment) is a development methodology that outlines a step-by-step development approach to LBS application development. A prototype LBS application was deployed on the author's SAGESS architecture platform, developed using the author's SAGE development guidelines and the SAGE development toolkit. SAGESS and SAGE will decrease LBS development difficulties and provide an open standard approach to LBS application development.

Keywords: Mobile computing, LBS (Location Base Services) applications, LBS Architecture, LBS methodology.

Introduction

The emergence of location technology as potential new market revenue, the next “killer applications” (Albena Mihovska & Jorge M.Pereira, 2001) has caused existing GIS (Geographic Information Systems) application software and telecommunication technologies to try and combine this new technology in an unified mix and match style of architectures. Therefore an open standard architecture framework is required for future market growth and stability of the LBS/GIS market place. The SAGESS LBS architecture and SAGE development methodology goals are to solve existing wireless LBS problems and establish an open frame base for the development and deployment of wireless LBS applications. The research described in this paper outlines an LBS deployment architecture and LBS development methodology, to decrease the difficulty in LBS application development and deployment. This project also defines a LBS development toolkit in order to assist first time and existing LBS developers to develop flexible LBS applications for the present (2004) wireless devices and future wireless devices. The paper is organised as follows: SAGESS and SAGE overview followed by a prototype demonstration.

SAGESS Architecture

The SAGESS platform is a three-tier architecture. The first tier is the client communication tier, second tier is the application tier and the third tier is the GIS information tier. The SAGESS architecture incorporates findings of Rui José, Filipe Meneses & Geoff Coulson, (2001) and Bharghavan, V. & GUPTA, V., (1997) and Jin Jing, Abdelsalam Sumi Helal & Ahmed Elmagarmid, (1999).

Client Communication tier

The client communication tier is a protocol independent tier where the users location is established and communication with the application tier occurs. The user launches an Internet browser and makes a wireless Internet connection to the application tier. The user invokes the LBS application by entering the application URL (Uniform Resource Locator). Any wireless location method that returns the geographical location of the user can be used to locate the user's current location. The wireless communication between the client communication tier and the application tier can be any form of wireless Internet protocol communication or technology.

Application tier

The application tier performs all result-set mark-up (voice, text, image). The result-set mark-up is the output displayed from the LBS solution. The application tier has two internal components, an *XSLT* (eXtensible Stylesheet Transformations) *processor* and an *SQL/XML translator*. The application tier receives over a wireless Internet connection the user's current location, if the user's current location cannot be obtained at the client communication tier then it can be obtain at the application tier by means of either third party software or a MLC (Mobile Location Center). The location can also be determined at the GIS tier with location sharing between the telecommunications database and the GIS database. The application tier *SQL/XML translator* component communicates to the GIS tier the current user's location and the LBS application query. The *SQL/XML translator* transforms the result from the LBS application query into an XML formatted document. The *XSLT processor* transforms the XML document into a specific or a range of different mark-up formats (SVG, HTML, PDF, WML, VRML). The resulting outcome is then displayed in the user's Internet browser.

The application tier in the SAGESS architecture will transform dynamically generated XML into any desired mark-up output. The XSL applied to the XML document decides the output of the data in the XML data page. One XML document can have many different XSL depending on screen size, browser language (English, French) and preferred mark-up Language

content (HTML, WML, SVG, VRML). The application server can automatically discover the browser type of a calling application and automatically assign the appropriate XSL.

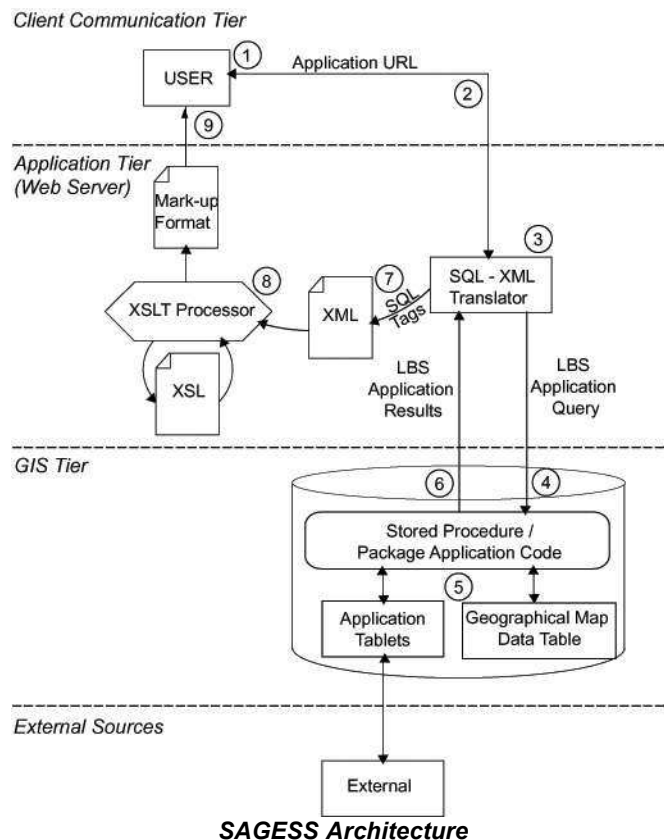
GIS (Geographic Information System) tier

The GIS information tier will perform all application query processing. The GIS tier stores, modifies and retrieves spatial data sets based on the LBS application query from the application tier. The application tier evokes an *application specific database stored procedure or package* which executes the core application logic. The *application specific database stored procedure or package* interacts with *geographical map data tables* (digital maps in relational formats) to perform LBS spatial related queries. The *application specific database stored procedure or package* interacts with the *geographical map data tables* and with the *application tables*. Application tables contain data that is utilised by LBS applications to perform more user specific and accurate LBS applications. *Application tables* can be a range of tables each table obtaining information from many different outside sources. *Application tables* could contain data on traffic information, road works, amenities location (ATM's, shopping centres, car parks) and services (restaurants, bars, cafés, hotels). These tables can be automatically and dynamically updated from existing external sources. The tight coupling of core application code and dynamic *application tables* allows for the creation of highly adaptive LBS applications. The GIS tier requires a database that can manipulate spatial data.

below displays an overview of the SAGESS architecture and demonstrates the message flow through the system. The steps of communication are:

1. User enters the HTTP address of the application. The user's location can be determined at this point or at a later point. **(Client Communication tier)**
2. The Web server accepts the user request and offers LBS application options if present. **(Application Tier)**
3. The *SQL/XML translator* makes a JDBC/ODBC connection to the database on the GIS tier with the LBS application query and user location. **(Application Tier)**
4. The core application logic (stored procedure/package) is evoked by the LBS application query sent from the *SQL/XML translator*. **(GIS Tier)**
5. The stored procedure/package utilises many application specific tables and manipulates spatial data to produce the application result set. **(GIS Tier)**
6. The application SQL result set is returned to the calling *SQL/XML translator* component. **(GIS Tier)**
7. The *SQL/XML translator* converts the SQL result set into an XML document with appropriate name tags. **(Application tier)**
8. The XSLT processor then transforms the XML document into the mark-up page specific by the XSL. **(Application Tier)**
9. The user's browser can then interpret the output from the mark-up page. **(Client Communication tier)**

below the numbers represent the steps above.



The SAGESS three-tier architecture design provides clear coupling of resources and a more flexible and scalable architecture than a two-tier approach. This style of architecture allows for development changes to occur on the server side and for thin accessing clients. Similar style of architectures are used in many commercial LBS systems and in the PARAMOUNT (Poslad S, Laamanen H., Malaka R., Nick A., Buckle P. & Zipf, A, 2001) and CRUMPET (Loehnert E., Wittmann E., Pielmeier J., & Sayda F., 2001) research system. The architecture is based on the use of standard protocols (HTTP, TCP) and open standard database connection interfaces (JDBC, ODBC) to provide a more flexible and adaptive platform for future wireless LBS development. SAGESS is made up of software components that can be swapped and updated but the overall message parameters remain the same. The SAGESS architecture combines two different areas of computing GIS and LBS underneath a coherent flexible architecture.

SAGE Development Methodology

The SAGE development methodology is intended to assist LBS developers to prosper from the SAGESS architecture and deliver simple to develop, flexible, device independent attractive LBS applications. The SAGE development methodology deals with issues of user data modelling (interface layout and language representation), complex functionality, geo-data processing, generic application development and LBS development environment design.

SAGE deals with one of the key LBS development problems of third party technology and not methodology being the driving force behind LBS application development. SAGE deals with LBS development of wireless database structure issues (Dunham M. H. & Helal A, 1995) and nomadic computing difficulties (Alonso R. & Korth H. F, 1998). The SAGE development methodology's principal factor is a database data centric approach to LBS application development. This data centric approach is necessary, as LBS applications are data intensive applications, which are accessed from limited wireless devices. The database data centric approach provides the developer with more flexibility, security and a stably controlled development environment.

SAGE defines clear development steps for LBS application development:

1. Database geographical tables.
2. Define application specific tables.
3. Develop core application functionality.
4. Define eXtensible StyleSheet (XSL) for result display.
5. Develop Internet based (GUI) Graphical User Interface to evoke LBS application

The SAGE development methodology couples the mandatory spatial map query with the core application logic. This enables all application function processing to be performed inside the database. The calling LBS application will receive the entire data result for the LBS application query. This style of LBS application function processing removes the application process cost, from the limited wireless devices to the large process GIS server, coupled with the large processing application tier.

The third step of SAGE is to develop core application functionality. The core LBS application functionality (application specific functions and location query function) are developed using database store procedures. Database procedures are implemented in a form of SQL (Structured Query Language). SQL is a standard interactive and programming language for manipulating data stored in a database. SQL is both an ANSI and an ISO standard. Database procedures define for application development should have:

- Parameters names related to their function.
- Common LBS functions implemented in dynamic SQL.
- A store procedure comments table set up.
- Transport parameter called transport.

A store procedure comments table should be set up inside the database that stores a LBS developer description of the store procedure. The description table will assist new developers in building LBS applications based on integrating existing procedure capabilities.

Dynamic SQL enables a procedure to assign dynamic values at run time. This feature enables stored procedure to adjust to varying databases conditions and one stored procedure can execute differently based on parameters. Dynamic SQL is ideally suited for repetitive task with changing parameters. Database procedures can now perform similar tasks to external programming languages. The advantages of database procedures languages over external languages are processing time and complex data type matches; two major factors in LBS applications.

Database store procedures can increase LBS processing speeds as store procedures are stored internally to a database in compiled code and procedure execution plans can become cached in memory vastly increasing execution time. LBS applications relay on a mandatory spatial data query. Database procedures perform this task quicker and with greater flexibility then any external procedure languages. Database stored procedure will execute identically in the same proprietary database, running on different operating systems (OS).

SAGE Toolkit

The SAGE toolkit is a wizard application to develop quick LBS applications based on the SAGE development methodology and deployed on the SAGESS architecture. The SAGE toolkit incorporates database spatial functions options and allows new LBS application to be developed based on existing development templates. The SAGE wizard toolkit will:

- Provide a wizard style GUI (Graphical User Interface).
- Incorporate XSL templates.
- Enable LBS query testing.
- Display database LBS query functions.
- List Procedure templates.
- Generate an automatic LBS application interface.

The SAGE development wizard will be initial set up with the default spatial settings for the proprietary database connection. LBS users can change these settings and all connection settings in the program. The SAGE development toolkit requires LBS developers to enter in the application specific LBS queries and the name of the appropriate XSL. The XSL used can be one of the existing templates or a new template based on the existing template.

SRF (Service Route Finder) Prototype Application

The SRF application is a prototype application that was developed using the author's toolkit on the SAGESS platform using the SAGE development methodology. The SRF application combines both LBS/GIS into a dynamic mobile solution for locating services (Restaurants, ATM's, Hotels, Cinema, etc.) on low resource devices over standard communication protocols. This chapter will detail the implementation structure and features of the SRF application.

SRF Application

The SRF is an LBS prototype application that was implemented on industrial technologies based on the SAGESS architecture. The SRF application performs a wide range of user and location based services.

The SRF application:

- Finds services that are geographically local to the user.
- Finds the shortest route to the service location.
- Calculates the route based on method of transport and on up to the minute traffic information, and road maintenance information.
- Open standard content, non-application specific.
- Dynamic user specific generated map images.
- Estimates arrival time, and distance along the route path.
- Provides a rerouting service back to the users original location.
- Provides detailed information on the selected service.
- Provides additional LBS services like nearest ATM, car park, monument on route.
- Incorporate external information systems to improve solution accuracy.

The SRF application delivers user specific maps in SVG format. The SVG maps are automatically generated and can be displayed on any Internet browser device that has SVG support. SVG can be dynamically generated by any of the proposed methods of John McKeown, & Jane Grimson (2000). SVG is a vector-based image therefore it can be scaled and altered without losing image quality. This feature will resolve the wireless development issues of delivering a solution to multiple wireless devices with different screen sizes. The SRF application generates maps with information that is specific to the user's request. The generated maps only show information (roads, ATMS, Car Parks, Monuments) situated on the users route (Dublin City Walking Nearest Restaurant). This feature allows for more tailored made service information and SVG images of lower size. The generated maps contain Internet-based intelligence of restaurant Internet site hyperlinks, dynamic travel time, voice description and name enlargement. The generated user adaptive maps will improve tourism and service location as just the information required is delivered in a clear and concise manner as described by Alexander Zipf, 2002. The SRF application can also deliver hypertext based information for devices that not SVG capable.

SRF Architecture

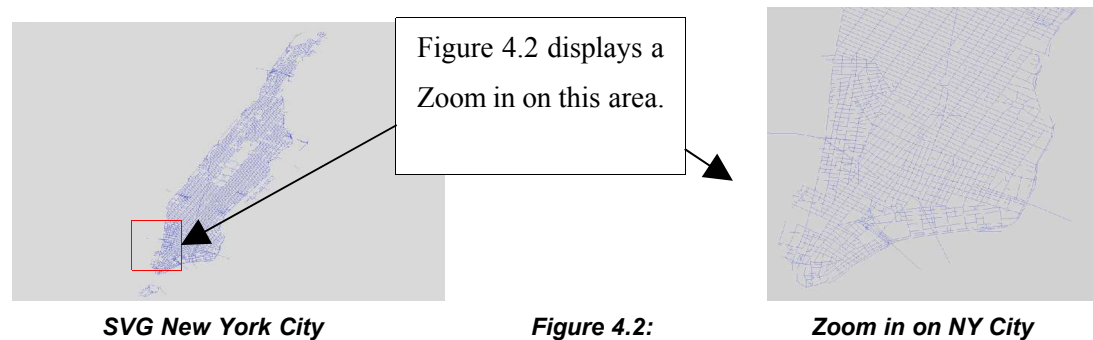
The developed SRF application was based on the SAGESS architecture. This style of architecture is a three-tier architecture. The first tier, the client communication tier, is an independent tier that allows any user with a wireless device and Internet browser capabilities to connect to the application. The application tier runs the Apache Web server on Linux Red Hat 9 operating system. Apache was selected because it is a free server that can run on many different platforms. The *SQL/XML translator* is implemented using the Oracle XDK for Java component.

In the GIS tier the SRF application access geographical data types from an Oracle 9i database with a spatial cartridge running on Linux Red Hat 9 based platform. Oracle was selected because it is the world leading database in Spatial, XML technologies and has been used in various other commercial and research systems. The SRF core application code was implemented in Oracle PL/SQL and is stored internally as a database stored procedure. The database procedure accepts parameters of the users location coordinates and service criteria. The SRF stored procedure uses a shortest route algorithm based on a link algorithm. The SRF stored procedure will utilise dynamically updated traffic, restaurant and road works tables to dynamically produce the shortest route. The SRF stored procedure will examine roads inside a service region area. The service area is based on the distance from the users current location to the service location multiplied by a geographic factor. This procedure will dramatically speed up route calculations because roads not within the service area are excluded from the search. If no route can be found then the service area is increased, the algorithm works on entire roads so if the start of a motorway is found, the entire motor junction inside the area will be included. The shortest route algorithm takes into consideration road speeds and other outside factors. If the user is very far away from the service, (for example the length of an entire country) then a link based algorithm search can be timely. To resolve this issue knowledge based routes on mid-range points are used. These routes are updated when appropriate traffic information source is received. The execution results of the stored procedure are stored in a user specific object-relational table. This facility allows for added security and base application tables to be made read only, preventing multiple application read/write locks. Database stored procedures add an extra layer of security because users can be granted only execute access on the procedure and not the underlying tables.

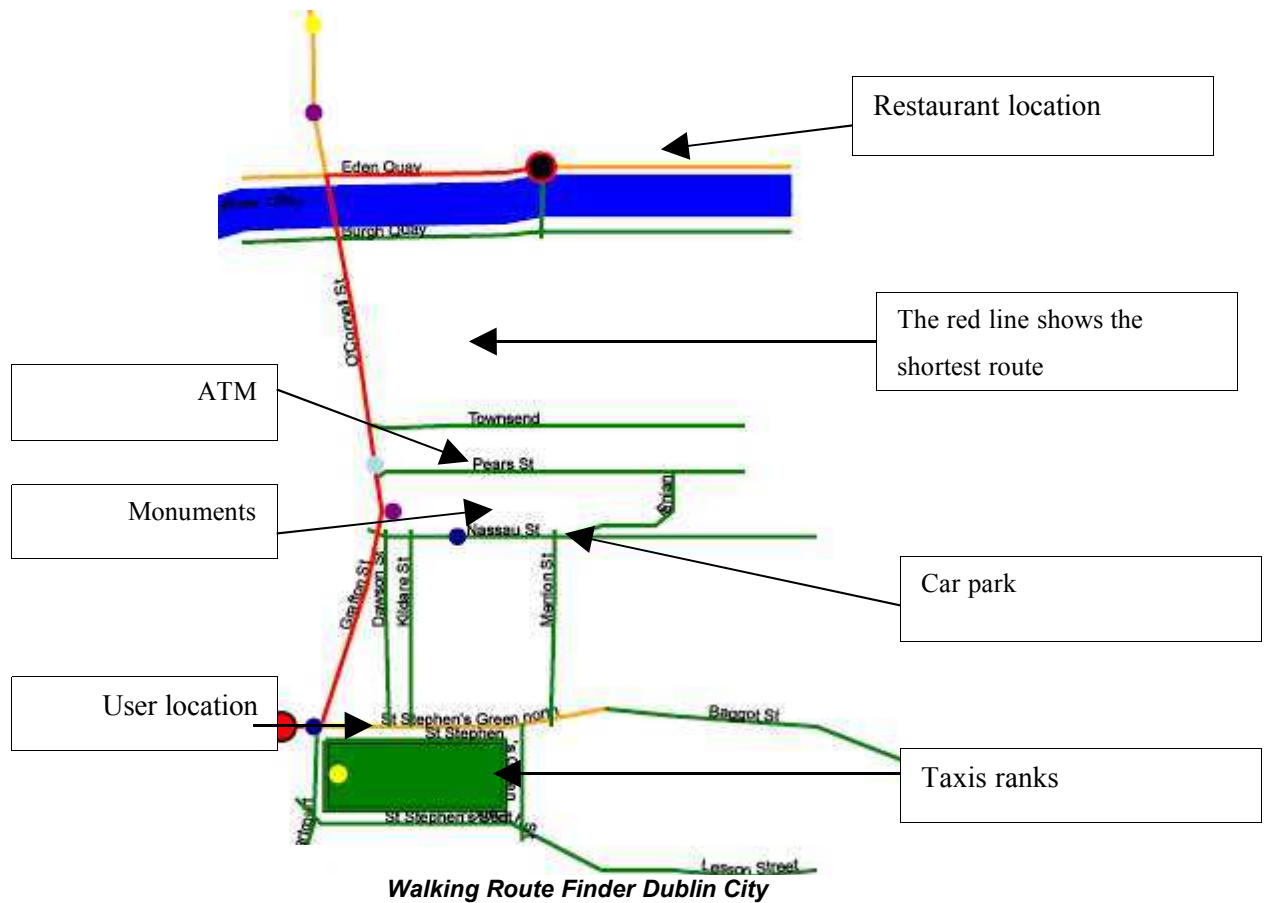
SRF Prototype System

The SRF prototype system was tested on two areas: Dublin City and New York City. The commercial GIS company MapInfo provided the map for New York City and an Oracle

supplied conversion tool was used to convert the map into an Oracle geometry type. The map of Dublin city was created by the author and is based on a non-electronic commercial map of Dublin city. The author created the map of Dublin to demonstrate a solution to the problem of obtaining digitalised map data. The map creation follows the guidelines of cartography map development and database geometric types. These two cities have two completely different city layouts. New York is grind oriented while Dublin like a typical European city with many side streets in no fixed pattern. Figure 4.1 and Figure 4.2 below demonstrate the dynamic SVG map generated from the SRF application of the entire city of New York. The image is generated from 80000 rows stored inside the database. To display the image takes a bit over five seconds on a typical GPRS connection.



The SRF application implements the most popular of LBS applications, the services guide. The SRF application implements a restaurant guide as it prototype service guide. The user can select a restaurant based on restaurant type and price. below displays the result of a user selecting a “French” type restaurant at the price of €40 and walking as the method of transport.

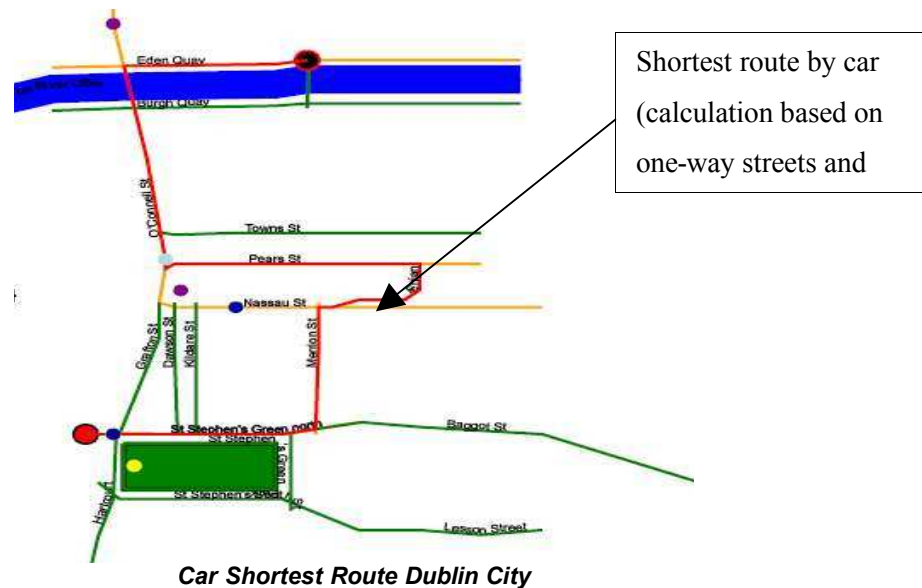


The dynamically generated SVG maps offer intelligent map features, not representative in non-electronic maps. These intelligent features include a road text enhancement, route travel details and restaurant Internet address hyperlinks. Road text enhancement displays the name of the road in larger text when the user clicks on a road. This feature is designed for better visibility and to resolve issue relating to text on line support in mobile SVG.

The user can select the nearest restaurant regardless of type. below demonstrates the SRF application ability to deliver user centric maps. The map area displayed in below is relevant to the users' specific queries.

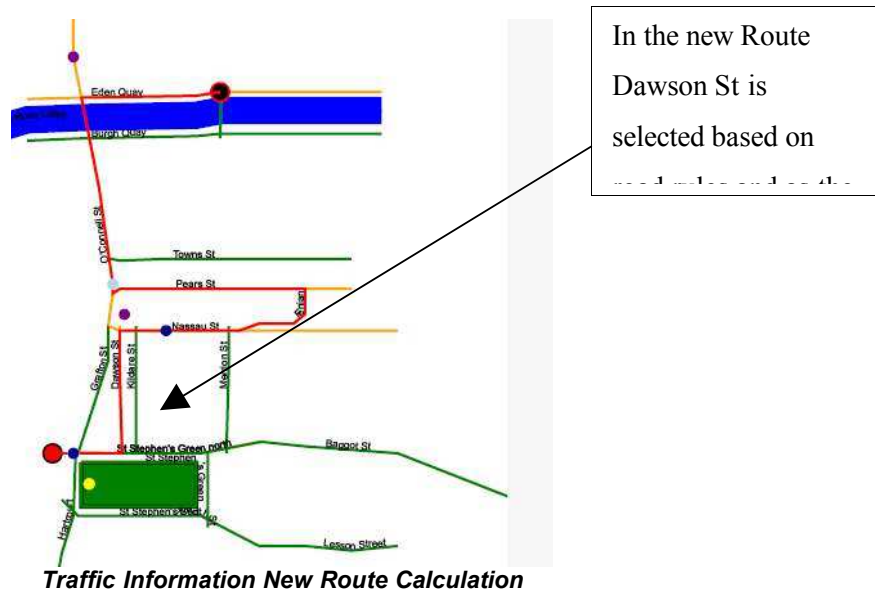


demonstrates the results when a user selects driving as their desired method of transport and “French” as their restaurant type with a menu price of €40. The users current location is the same as the two previous queries. The same XSL for the previous query is used to dynamically generate the map.



The dynamically generated map offers all the same capabilities as the Walking XSL in . The SRF application also displays the routing information specific to transport means and route selected. The development time and complexity of the car shortest route module into the SRF application was decreased because it is based on the Walking shortest route template and the core application logic. This demonstrates the power of the SAGESS architecture and the SAGE development strategy, as new application areas and product development is possible based on generic templates.

The SRF application incorporates dynamic outside traffic information to improve on the LBS application result accuracy. below the traffic information feed was received from an outside source (AA road watch, traffic light statistics, News information feeds). The dynamic outside traffic source informs of a 20-minute delay on the Merrion Street. below demonstrates the SRF application ability to adjust this dynamic traffic information to recalculate the shortest route.



The SRF application was designed with two test areas New York and Dublin. below demonstrates the SRF application on New York City.

The New York City development was based on the previous Dublin City development. The core application LBS Dublin City code required only minor alterations of a New York map table and New York information tables. The New York XSL was developed based on the Dublin City template and requires only to change the ratio number scale. There was only a small requirement of change because the XML dynamically generated document was based on generic table names. The dynamically generated New York SVG maps offer all the same interactive capabilities as the Dublin maps because they are based on the same XSL template. The use of templates and generic XML tags means that a first time LBS developer could have deployed the above LBS application with only two minor changes.

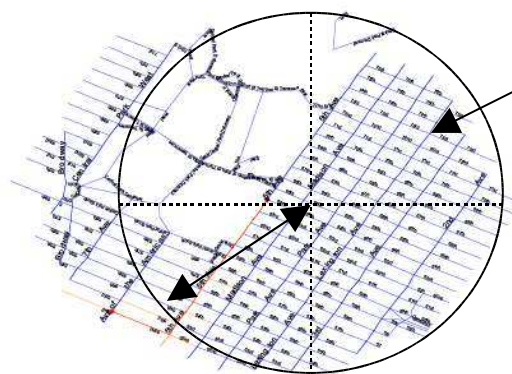
The user service area is the search area to be used to calculate the shortest route. below the entire map of New York is displayed. To search all of New York to find every possible route could take days or weeks and the map delivered would be difficult for the user to search.



Entire New York Search Area

The New York user and the restaurants are located in the given area. below shows the results.

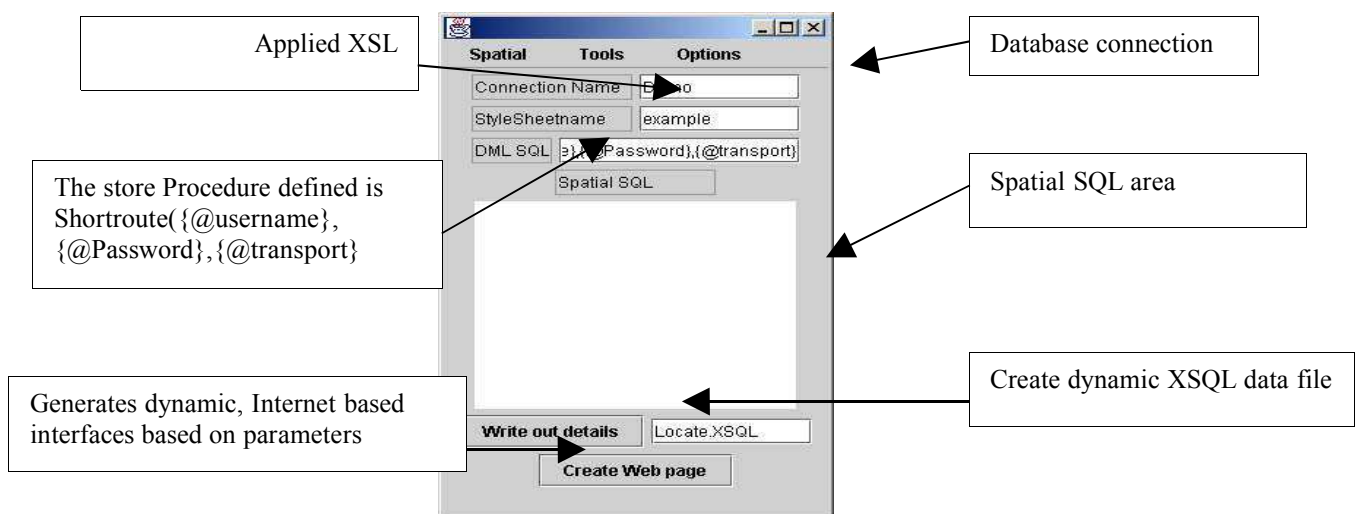
The SRF application delivers an SVG image of just the relevant service area. below the service area of the query is displayed. The use of a service area significantly decreases the shortest route calculation time.



New York Service Area

The service area size is the length from the users location to the service location multiplied by a location size factor.

below demonstrates the SAGE development toolkit.



SAGE Toolkit LBS Interface Generation

Three Parameters username, password and transport will be included as input boxes in the new dynamically created Internet page, displayed in below. The key word “transport” generates a drop down list box of common transport means.

Dynamically Generated Internet Page

below demonstrates the spatial LBS query options.

SAGE Toolkit Query Options

Conclusion

An open standard approach to LBS application development will enable LBS developers to deliver LBS applications that can be flexible and scalable to an ever-changing wireless telecommunications market. The SAGESS architecture and SAGE development methodology provides a framework from which LBS developers can deliver the next generation of dynamically generated user specific LBS applications based on industry standards.

REFERENCES

- Loehnert E., Wittmann E., Pielmeier J. Sayda F. (2001).** PARAMOUNT- Public Safety & Commercial Info-Mobility Applications & Services in the Mountains. 14th International *Technical Meeting of the Satellite Division of The Institute of Navigation ION GPS*.
- Forman G & Zahorjan J. (1994).** The Challenges of Mobile Computing, *IEEE Computer*, Vol. 27, No. 4 pp 38 – 47.
- Rui José, & Adriano Moreira, & Filipe Meneses, & Geoff Coulson. (2001).** An Open Architecture for Developing Mobile Location-Based Applications over the Internet. *6th IEEE Symposium on Computers and Communications, Hammamet, Tunisia*.
- Jin Jing, & Abdelsalam Sumi Helal, & Ahmed Elmagarmid. (1999).** Wireless Client/Server Computing for Personal Information Services and Applications, *ACM Computing Surveys*, Vol. 31, No.2.
- Dunham M. H. & Helal A. (1995).** Mobile Computing and Databases: Anything New? *SIGMON Record*, Vol. 24, No. 4, pp 5 – 9.
- Albena Mihovska & Jorge M.Pereira. (2001).** Location-Based VAS:Killer. *Applications for the Next-Generation Mobile Internet Personal, Indoor and Mobile Radio Communication*, 12th IEEE International Symposium .
- John McKeown, & Jane Grimson. (2000).** SVG: putting XML in the picture: *Proceedings of XML Europe Paris, France, Graphic Communications Association (GCA)*.
- Alonso R. & Korth H. F. (1998).** Database System Issues in Nomadic Computing, *SIGMON Record*, 1998, pp 388 – 392.
- Poslad S, Laamanen H., Malaka R., Nick A. Buckle P. and Zipf, A. (2001).** CRUMPET: Creation of User-friendly Mobile Services Personalised for Tourism. *Proceedings of: 3G 2001 - Second International Conference on 3G Mobile Communication Technologies*.
- Bharghavan, V. & GUPTA, V. (1997).** A framework for application adaptation in mobile computing environments. In *Proceedings of the 21st International Computer Software and Applications Conference (COMPSAC '97)*. IEEE Computer Society, New York, NY, 573- 579.
- Alexander Zipf. (2002).** User-Adaptive Maps for Location-Based Services (LBS) for Tourism. In: K. Woeber, A. Frew, M. Hitz (eds.), *Proc. of the 9th Int. Conf. for Information and Communication Technologies in Tourism*, Innsbruck, Austria.

Camera Control through Cinematography in 3D Computer Games

James Kneafsey & Hugh McCabe

School of Informatics & Engineering, Institute of Technology at Blanchardstown, Dublin 15

Contact email: james.kneafsey@itb.ie, hugh.mccabe@itb.ie

Abstract

Modern 3D computer games have the potential to employ principles from cinematography in rendering the action in the game. Using principles of cinematography would take advantage of techniques that have been used to render action in cinematic films for more than a century. This paper outlines our proposal to develop a camera control system that uses principles of cinematography for 3D computer games and provides a critical review of related research.

Keywords: Virtual camera, virtual environments, 3D computer games, cinematography.

1. Introduction

Interactive virtual environments present a view of a three-dimensional scene to a user by means of a virtual camera. The user interacts with the scene generally by controlling an *avatar*, i.e. a 3D representation of a character in the scene usually in human form. Throughout the interaction, the virtual camera must continually provide views of the scene that enable the user to carry out their assigned task. We propose that it should be possible for the virtual camera in 3D computer games in particular to present camera angles to the user that do not only show the user what they need to see to carry out their task but also draw the user into the story, add drama to the scene and invoke emotions in the user in the same way that cinema does. We propose that the principles of cinematography can be incorporated into a framework for controlling the virtual camera.

Cinematography is an art-form that has evolved over more than 100 years of film-making. Each revolutionary approach to camera operation which has been generally accepted by audiences, such as when D.W. Griffith first employed a moving camera, represents a new addition to the language of cinematography (Brown, 2002). Cinematography presents a number of principles that can be used to ensure that the viewer does not become disoriented and the presentation of the action is consistent throughout each scene. Particular camera treatments can make the viewer feel as though they are part of the story and help them to identify with the characters on the screen. The application of principles and techniques of cinematography to the camera can invoke emotions in the viewer in response to the action presented on the screen (Mascelli, 1965).

We propose that 3D computer games have a greater potential for the employment of cinematographic principles than other virtual environments due to the similarity in content to

equivalent real-world contexts. A horror game could borrow from the techniques used in horror films. A game about gangsters could use principles employed in gangster films. Despite these similarities, a considerable difference to note between computer games and contexts in the real-world is that computer games are interactive. Whereas there is substantial determinism in how the action will play out in a film because of its adherence to a script, the same cannot be said of computer games. Autonomous intelligent entities, i.e. one or more human players and a number of *non-player characters* (NPCs) which are characters driven by software algorithms, provide the action for the game in real-time and so the level of determinism in the future state of the 3D scene at any moment is decreased significantly. This presents an additional problem not associated with the filming of action for a cinematic film: The camera control module in the game must attempt to continually predict the game's future state in order to film the relevant elements of the scene and use these predictions to present the game player with a cinematographic view of the scene.

This paper presents a critical review of related research in the area of virtual camera control and outlines our aim to implement a virtual camera control system for 3D computer games through cinematography and to implement a framework for the quantitative evaluation of our implementation and that of other research projects. The remainder of this paper is structured as follows: Section 2 discusses current approaches to camera control in 3D computer games. Section 3 outlines the principles of cinematography that are relevant to the control of the virtual camera in a 3D computer game. Section 4 presents a review of research related to camera control in virtual environments. Finally, section 5 presents some conclusions and future work.

2. Camera Control in Computer Games

There are a number of different camera views used in 3D computer games suited to the particular game genre. Some games offer more than one type of camera. A *first-person camera* (Sánchez-Crespo Dalmau, 2004) depicts the action through the avatar's eyes (figure 1). This is often the default view for games that require accuracy of movement such as shooter games, e.g. *Quake III Arena* (id Software, 1999), *Half-Life* (Valve Software, 1998). A *third-person camera* films the action from above and behind the avatar and so the avatar is displayed at the bottom of the screen. This is the default view for *Hitman 2* (Eidos Interactive, 2002) and *Tomb Raider III* (Eidos Interactive, 1998), for example.



Figure 1: Types of camera views in 3D computer games. From left to right: First-person camera in *Quake III Arena* (id Software, 1999), third-person camera in *Tomb Raider III* (Eidos Interactive, 1998).

Inertial cameras are used in a number of games to ensure that the movements of the camera are not rigidly responsive to the movements of the avatar. The cameras movements are related to the avatar's movements via a spring model (Sánchez-Crespo Dalmau, 2004) so that as the avatar moves, the camera moves smoothly as if it were attached to the avatar by a spring.

Common approaches to camera control in a 3D computer game consider only practical issues such as the placement of the camera in the 3D world, fixing camera movements to a curve to ensure smoothness of movement and, as already mentioned, using a spring model for realistic inertial movement (DeLoura, 2000). Other issues addressed include how to prevent occlusion of the view by the avatar when using a third-person camera and also how to prevent collision of the camera with the geometry of the scene (Sánchez-Crespo Dalmau, 2004). Sánchez-Crespo Dalmau (2004) also discusses setting up a number of different cameras to film action from different angles and with different movements. These cameras are set up before the interaction and a real-time algorithm then selects the best camera to film each situation during the interaction. None of these approaches attempt to apply principles from cinematography to an intelligent camera control system in a game in real-time.

3. Principles of Cinematography

Cinematography is defined as the art of making films. Through cinematography extra meaning and emotional subtext can be communicated to the viewer visually. The camera can hide elements of the scene to build suspense or focus on particular events to elucidate the plot (Brown, 2002). In this section we outline the principles of cinematography we propose to implement. More complete references on cinematography can be found in Mascelli (1965) and Brown (2002).

3.1 Types of Camera Angles

The type of camera angle used to film a particular scene determines how much the viewer will be drawn into the scene and whether they will be well oriented or not. An *objective angle* (figure 2) tends to make the viewer more detached from the scene whereas a *subjective angle* is more engaging but has the potential to disorient the viewer if maintained for a long period of time.

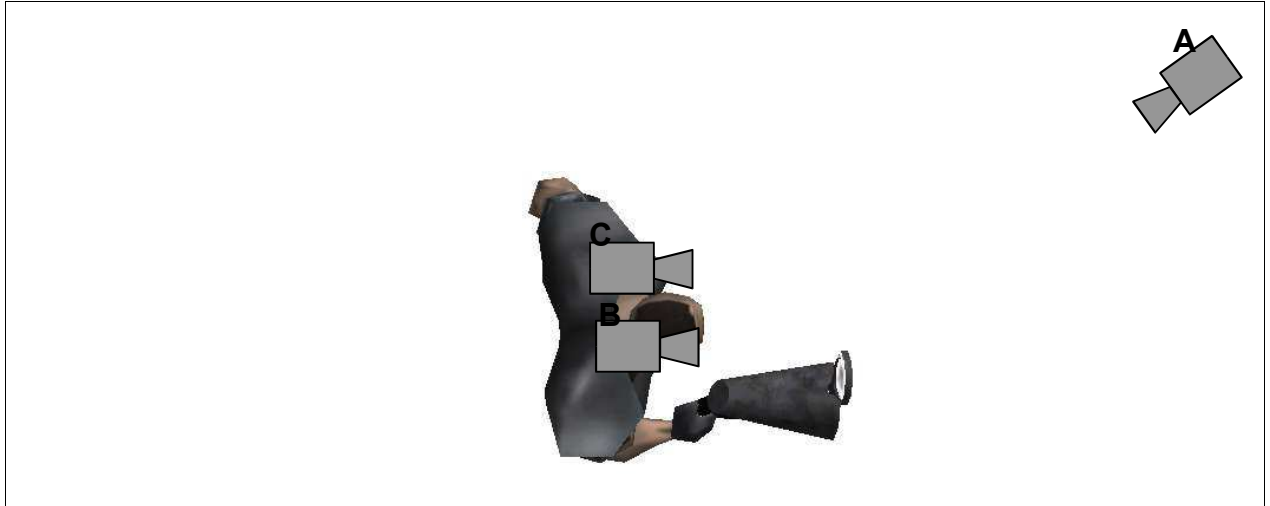


Figure 2: Types of camera angles. Camera A films the subject objectively, camera B is a subjective treatment and camera C is a point-of-view angle.

An objective angle presents the scene from the point-of-view of a neutral observer rather than that of any character in particular. A subjective angle can be used to make the viewer feel part of the scene and identify with the characters. The camera would be positioned closer to the action, for example it may move along with a speeding car. The camera may also replace one of the characters in the scene by filming what they would see. A *point-of-view angle* presents the scene from the point-of-view of one of the characters in the scene but not subjectively. It is as if the audience is cheek-to-cheek with the character. A point-of-view angle is as close to a subjective angle as the camera can get while maintaining objectivity.

3.2 Subject Size

Different subject sizes have different functions as a scene and characters are introduced and the narrative develops. A setting is often introduced with an establishing shot (a shot is a continuous unit of filmed action from the time the camera is turned on until it is turned off), usually a *long shot*. A long shot includes the entire setting, such as a room or a street, and the characters in it. Following this a *medium shot* (figure 3) may be used to introduce the principal characters. A medium shot depicts the characters from just below the waist or just above the knees up. A *close-up* may then be used to isolate one character or event in particular. There are a number of different close-ups. For example, a medium close-up depicts a character from

between the waist and shoulders to just above the head and a head close up shows just the head. The scene may be re-established with a long shot to prevent the viewer from becoming disoriented due to the subjective nature of closer shots, such as the close-up and medium shot. It may also be re-established when new developments occur.



Figure 3: Examples of different subject sizes.
From left to right: Medium shot, medium close-up, head close-up.

3.3 Camera Height

The height of the camera in relation to the subjects can have a psychological effect on the viewer and influence the element of drama in the scene. A *level angle* (figure 4) is used for eye-level shots. Objective level angle shots are filmed from the eye level of a person of average height except in the case of close-ups which are filmed from the height of the subject. Point-of-view level angle close-ups are filmed from the height of the opposing character and subjective level angle close-ups are filmed from the height of the subject.

A *high angle* shot is when the camera is tilted downwards to film the subject. These shots are used to make the subject seem small or make the audience feel superior to the subject. A *low angle* shot is when the camera is tilted upwards to film the subject. This has the effect of exaggerating the speed or height of a moving subject, or giving a character prominence in the shot.



Figure 4: Examples of different camera heights. From left to right: level angle, high angle, low angle.

3.4 Screen Direction

Screen direction ensures that positions, movements and looks (a character looking at something) are consistent from shot to shot. It guarantees that the action in a scene is presented in a uniform manner so the viewer is aware of the locations of elements of the scene and so is does not become disoriented. If the subject is looking or moving in a certain direction just before a cut (the ending of a particular shot) the following shot should depict them looking or moving in the same direction unless the two shots are separated by a different scene, a time lapse or an optical effect such as a fade.

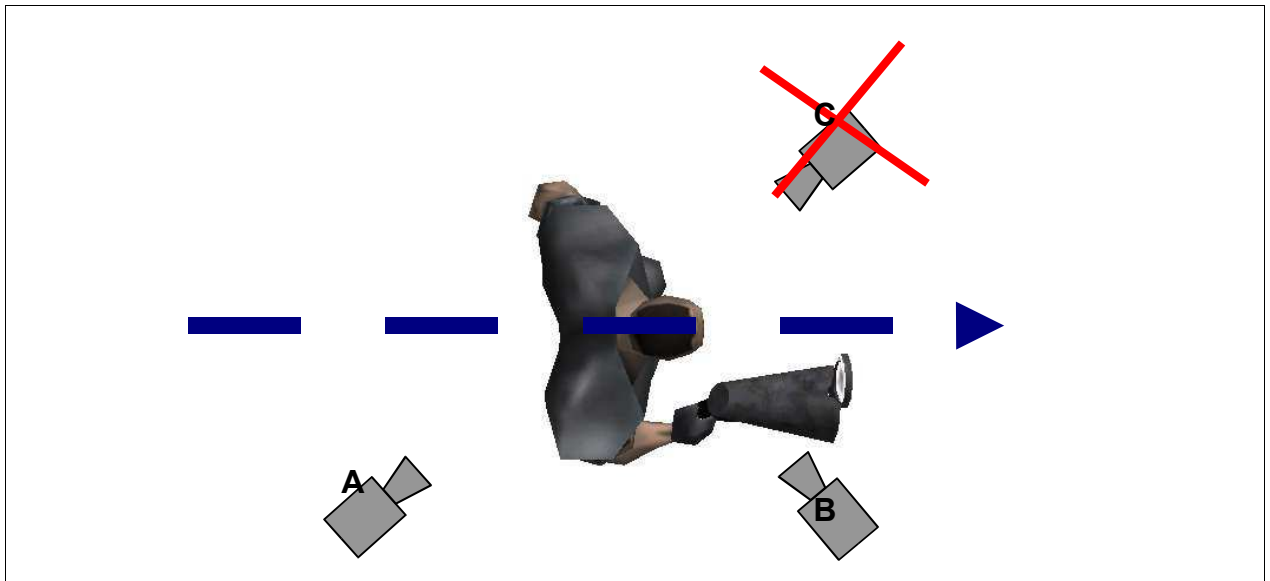


Figure 5: The action axis: If the subject is filmed with camera A, subsequent shots must keep to this side of the action axis dictated by the subject's motion. Therefore position B is valid while C is not.

Consistent screen direction is maintained by the adherence to the *action axis* (figure 5). The action axis is an imaginary line (it may be considered to be an imaginary plane in a 3D scene) established by certain factors, such as:

- A look: a character looking at something
- Movement, e.g., a car moving down a street
- A specific action, e.g., a character moving towards a door or two characters talking
- Physical geography: the layout of a room for example

When a shot ends the action axis must be noted so that screen direction will be maintained for the next shot. At the beginning of the next shot the camera must be placed on the same side of the action axis. This ensures that movement will continue in the same direction, a character will look in the same direction and characters' relative positions will be consistent with the previous shot. It is not necessary to adhere to the action axis if a time lapse has occurred or if the two shots were separated with a different scene. If the aim of a particular shot is to disorient the

viewer rather than keep them oriented, inconsistent screen direction can be used by not adhering to the action axis.

3.5 Cinematography in 3D Computer Games

Examples of some of the principles of cinematography discussed above can be seen in 3D computer games.

3.5.1 Types of Camera Angles

Many 3D computer games employ first- and third-person cameras. Both of these film the scene subjectively since the camera moves along with the avatar. First-person cameras are more subjective than third-person cameras because the user sees what the avatar sees.

3.5.2 Subject Size

3D computer games employing third-person cameras depict the avatar in long shot, medium shot or close-up. Subject size is not relevant to first-person cameras as the avatar is not visible.

3.5.3 Camera Height

Recently released 3D computer games allow the user to vary the height of the camera relative to the avatar in order to vary their view.

3.5.4 Screen Direction

There are generally no cuts within the same scene in recently released 3D computer games so screen direction is not an issue. The camera continually shoots the scene in first-person or third-person mode and never cuts.

3.5.5 Action Prediction

When filming a cinematic production in the real world, the director is at liberty to position the actors and props such that the filming of the scene conveys the required meaning or emotional subtext. The director can try different arrangements of the elements in the scene and attempt to employ different principles of cinematography until they achieve the shot they want. 3D computer games afford the camera control system only one attempt to film the action since it is supplied by human players and NPCs. Nothing in the scene can be repositioned and the system must attempt to predict the future locations of characters and props in the scene at each instant of time. These factors make the filming of action in a 3D computer game more similar to that in a documentary where the director has no influence over the action than a cinematic film.

4. Related Work

Previous work in the area of virtual camera control includes *CamDroid* (Drucker & Zeltzer, 1995). This work provides a general framework on top of which camera control mechanisms can be built for a number of domains within the area of virtual environments. Constraints can be applied to the camera to achieve the required shot treatment. For example a designer may require that the camera remain a certain distance from a character or that two particular characters are kept in the shot. This is a lower level approach than ours but it highlights some

of the principles upon which our research will build. Marchand & Courty (2000) discuss the positioning of the camera in a virtual environment and its reaction to changes in the environment. They address the issues of occlusion of the subject by scene geometry and the amplitude with which the camera responds to changes in order to avoid occlusion but do not consider the application of principles of cinematography to the virtual camera.

Christian et al. (1996) propose a Declarative Camera Control Language for the application of cinematographic constraints to the virtual camera. A number of *film idioms* are encoded in the camera control system; film idioms are the stereotypical ways of shooting particular scenes, for example, a scene of two people talking. User input determines the characters that are to be filmed over a period of time. The appropriate film idiom is selected and the action is filmed. The system has knowledge of factors relevant to cinematography such as subject size and shots. The system takes an animation trace from action that has already occurred so it is not suitable for use in a real-time environment. Tomlinson et al. (2000) present a camera system that attempts to reflect the calculated emotional state of the virtual actors in the scene while employing principles from cinematography. Each character in the scene has sensors, emotions, motivations, and actions. The entity controlling the camera, called the CameraCreature, also has these attributes. The CameraCreature evaluates the emotions of the actors in the scene and its own emotions to decide how to film the scene. This approach is similar to ours but it is limited in that the camera control system is purely reactive, i.e. it does not attempt to plan shots. Halper et al. (2001) present *A Camera Engine for Computer Games*. The camera control system attempts to continually predict the future state of the scene and applies constraints on the camera. A principal aim is to implement the constraints in such a way that they are not adhered to rigidly since this could produce erratic camera movements. A compromise is drawn between constraint adherence and frame coherence, i.e. continuity of movement of the camera. This work is more in line with ours in that it considers only computer games but it does not highlight the various principles of cinematography that can be used; however, it does present some methods upon which our research will build.

5. Conclusions and Future Work

We aim to implement a virtual camera control system for 3D computer games through cinematography. We plan to first implement features already present in recently released 3D computer games such as:

- Prevention of collision of the camera with scene geometry
- Prevention of occlusion of the view by the avatar or scene geometry
- Smooth camera movements

We propose to establish a number of typical scenarios in popular 3D computer game genres and apply the principles of cinematography outlined above to the virtual camera:

- The use of different types of camera angles, i.e. objective, subjective or point-of-view, depending on the context of the action
- The filming of the characters such that the appropriate subject sizes, e.g. medium shot, close-up, etc., are applied to each situation
- The filming of the characters with the appropriate camera height, i.e. level angle, high angle or low angle, for each situation depending on the dramatic or emotional subtext to be conveyed
- Adherence to the action axis when cuts are used during a scene if user orientation must be maintained

The major issue to address is the interactive nature of a computer game and its implications for the design of a system for camera control through cinematography. Another issue we propose to examine is the design of such a system such that a user will become easily accustomed to the camera work resulting from the principles employed. We will evaluate our implementation and that of other researchers by testing the systems on a number of typical scenes and compare the resulting camera work with the opinions of experts in the area. We will also evaluate our system from the point of view of playability, i.e. we will evaluate the effect of the cinematographic camera work on the user's ability to carry out the task required by the game.

References

- Brown, B. (2002).** Cinematography: Image Making for Cinematographers, Directors and Videographers. Oxford: Focal.
- Christian, D. B., Anderson, S. E., He, L., Salesin, D. H., Weld, D. S. & Cohen, M. F. (1996).** Declarative Camera Control for Automatic Cinematography. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, 148-155.
- Eidos Interactive. (1998).** Tomb Raider III, computer game for IBM compatible PCs, released by Eidos Interactive, San Francisco, CA.
- Eidos Interactive. (2002).** Hitman 2, computer game for IBM compatible PCs, released by Eidos Interactive, San Francisco, CA.
- Courty, N. & Marchand, E. (2001).** Computer Animation: A New Application for Image-Based Visual Servoing. ICRA 2001: 223-228.
- DeLoura, M. (2000).** Game programming gems. Rockland, Mass.: Charles River Media.
- Drucker, S., and Zelter, D. (1995).** Camdroid: A system for implementing intelligent camera control. In proceedings of 1995 Symposium on Interactive 3D Graphics, 139-144.
- Halper, N., Helbing, R. & Strothotte, T. (2001).** A Camera Engine for Computer Games: Managing the Trade-Off Between Constraint Satisfaction and Frame Coherence. In Proceedings of Eurographics 2001, 174-183.
- Halper, N. & Olivier, P. (2000).** CAMPLAN: A Camera Planning Agent. In Smart Graphics, Papers from the 2000 AAAI Spring Symposium, 92-100.
- id Software. (1999).** Quake III Arena computer game for IBM compatible PCs, released by id Software, Mesquite, TX.
- Marchand, E. & Courty N. (2000).** Image-Based Virtual Camera Motion Strategies. In Graphics Interface Conference, GI2000, 69-76.

- Mascelli, J. V. (1965).** The Five C's of Cinematography. Los Angeles: Silman-James Press.
- Sánchez-Crespo Dalmau, D. (2004).** Core Techniques and Algorithms in Game Programming. Indianapolis, IN : New Riders Publishing.
- Tomlinson, B., Blumberg, B. & Nain, D. (2000).** Expressive Autonomous Cinematography for Interactive Virtual Environments. In Proceedings of the Fourth International Conference on Autonomous Agents, 317-324.
- Valve Software. (1998).** Half-Life, computer game for IBM compatible PCs, released by Valve Software, Bellevue, WA.

Novel Design of a Variable Speed Constant Frequency Wind Turbine Power Converter

Aodhán MacAleer¹ & Joe Dunk²

¹ Department of Electrical Engineering, Limerick Institute of Technology, Limerick, Ireland.

² Department of Electrical Engineering, Limerick Institute of Technology, Limerick, Ireland

[¹Aodhan.MacAleer@lit.ie](mailto:Aodhan.MacAleer@lit.ie)

[²Joe.Dunk@lit.ie](mailto:Joe.Dunk@lit.ie)

Abstract

The operation and efficiency of wind turbines at present are hampered by the variable speed nature of wind, yet the constant speed requirements of electrical generators, hence wind turbines speeds are generally held down to a constant value regardless of wind conditions. This paper presents a novel design for a power converter that can produce a fixed output suitable for grid connection, while operating at variable speed governed by the available wind power. This novel design utilizes advances in high power, high frequency solid-state switches (IGBT's) based on an ADSP-21990 fixed-point mixed-signal digital signal processor (DSP). At present the system has been designed and simulated using Matlab and is currently in the build and test stage.

Keywords: Wind turbines, Variable Speed Constant Frequency, Power Converters, Pulse Width Modulation, DSP.

1. Introduction

Datta, Rajib and Ranganathan (2002) have stated that grid-integrated Wind Energy Conversion Systems (WECS) should generate at constant electrical voltage/frequency, determined by the grid. However, it is also advantages to vary the mechanical speed of the turbine/generator to maximise power capture with fluctuating wind velocities.

In the case of weak grid networks like Ireland's to successfully install wind turbines, a knowledge of the wind turbine impact on the grids is essential, this area is currently receiving intense focus in Ireland and European from wind turbine manufactures and grid controllers. The importance of this has currently led to a wind turbine moratorium [19] by the Electricity Supply Board (ESB) national grid preventing any addition wind turbines connecting to the grid until problem issues are solved.

2. Variable Speed Fundamentals

Albert Bertz (1947) proved that the amount of energy that can be successfully taken from the wind as the "power coefficient (C_p)", and calculated that the maximum physically value to be 59.3%, however in practise due to drag losses and losses in the mechanically components, power coefficients are often much less. The power coefficient is defined by Hau (2000) as the ratio of the extractable mechanical power to the power contained in the wind stream.

Another term principal in understanding wind turbine technology is the Tip-speed-ratio (TSR), which is defined as the ratio between the rectilinear speed of the blade tip and the wind speed. The advantages of variable speed WECS over fixed speed systems can be more easily understood from the C_p -TSR relationship (figure 1), which is indicative of all Horizontal Axis Wind Turbines (HAWT).

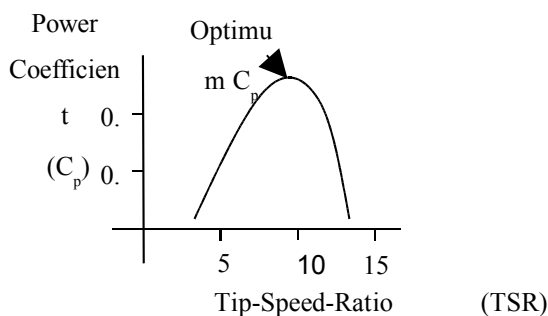


Figure 1 C_p -TSR relationship

Constants speed wind turbines are designed to operate near the optimum power coefficient at wind speeds that occur most often at the particular site. However from an understanding of the TSR (Ratio of the Tip Speed to Wind Speed), for a constant speed machine having a fixed speed rotor, a change in wind speed will cause a change in the tip-speed-ratio, hence fixed speed machines will often be operating at a non-optimal power coefficient, i.e. when the wind speed is above or below operating wind speeds for the specific wind turbine.

A variable speed system differs in that a changing wind speed will create a change in the rotor speed i.e a change in the rotor tip-speed, hence the TSR stays constant because both wind speed and tip speed change simultaneously. Hence a variable speed machine operates at near optimum power coefficient regardless of the varying wind velocity, excluding over speed operation where the rotor speed is restricted to prevent self-damage.

The added power capture from variable speed wind turbines, has been proved by Carlin (1995) to be on average up to 10% more annual energy. Additional advantages to variable speed operation include reduced loads on the drive train, and reduced turbine rotor fatigue loads, which reduce the overall turbine cost considerably.

In addition to improved energy capture possible with variable speed operation, ease of synchronisation to the grid is also an added benefit possible since the current can be controlled from zero to rated value via inverter control, unlike fixed speed operation where large inrush currents can pose problems for the grid.

3. Topical Issues

The Irish Electricity Supply Board National Grid (ESBNG) has called for a wind energy moratorium pending the resolution of the power systems reliability issues. This topic is especially specific to the Irish Network due to its largely isolated status, however this problem is currently being examined by other European network providers, due to the widespread increase in wind energy installation due partly to the Kyoto agreement fallout.

In a recent presentation [9] to the Joint Oireachtas Committee the Commission for Energy Regulation (CER) made the point that future & present wind turbines planned installations do not meet the requirements of the grid code, failing on several technical issues. The main provisions that wind farms struggle to comply with are “Fault Ride Through, Voltage and Frequency variations” among others.

Fault ride through refers to the reaction of power stations to faults on the transmission line, wind farms tend to trip of the system unlike conventional generator which continue generating regardless, this isolation from the grid could be significant when it can be argued the grid needs the extra generation the most, this is most prevalent to older manufactured wind turbines that are directly connected to the grid for speed stability etc.

This power converter designed is only peripheral connected to the grid, mainly for synchronisation to the grid, and does not require external signal to control generation, excitation or speeds, hence the serious technical issues presented by the ESB national grid will be solved in the most part.

4. Wind Turbine Configurations

Analysis and design of a power converter for wind turbine applications cannot be done without reference to the generator configuration used for power generation. Since in particular the rectifier and generator configuration are inextricably linked, hence in order to design or analyse the rectification process it is first necessary to consider the orientation and design of the power generator.

In wind turbine design the two main electrical machine synchronous or asynchronous can be utilised, synchronous machines are commonly encountered in the generating mode, and tend to be more in demand for large scale fuel burning power generation among other things due to their ability to generate reactive power and so improve the power factor of the installation. Although more expensive initially than an equivalent asynchronous generator slightly higher efficiency are possible, to offset the initial cost.

Synchronous machines are almost always brush-less, since the slip rings have been dispensed with in favour of static excitation hence brush-less systems, this development is mainly due to the progress and availability in power electronics components (diodes & thyristors for example) in recent years. Brush-less systems in the past have been constructed using a stator with a distributed winding and a rotor with salient pole field windings. These are supplied through rotating diode rectifiers assembly, fed by a shaft-mounted exciter.

Asynchronous generators or induction generators (IG) as they are commonly known tend to be either squirrel cage or wound rotor type, both referring to the physical properties of the rotor. IG may be designed to operate at more than one speed fixed speed by pole changing, where the stator winding can be switched between pre-determined number of poles, this is however not a widely used method. Slip rings IG or wound rotor (also termed doubly fed IG) involves removing the squirrel cage rotor and replacing it with a rotor with windings connected via slip rings to external elements (sometimes power converters). Principle disadvantages of wound rotor IG include higher initial cost compared to an equivalent squirrel cage IG, and maintenance due to the slip rings brushes.

However at present the most common electrical system utilised in commercial wind power plants is the induction generator (IG) directly connected to the grid. A major drawback is that the reactive power flow and thus the grid voltage level cannot be controlled. Another drawback associated with fixed speed systems is that the blade rotation causes voltage fluctuation of a frequency of 1 to 2 Hz on the grid. This fluctuation problem is not solved by using several turbines; on the contrary, Svensson (1996) states that if several identical wind turbines are installed in a wind park, the rotors can synchronize with each other and the power fluctuations are superimposed in phase.

The choice of synchronous versus asynchronous machine is dependant very much on the requirement of the wind turbine designer. In [8] Svensson states that the standard electrical system for fixed speed wind turbines tend to be squirrel cage asynchronous (induction) generator directly connected to the grid. This form of power generation is first-rate in reducing the reactive power demand and a capacitor bank is installed to compensate for the no-load current of the generator. In addition a thyristors equipped soft starter can be used to reduce the inrush currents present when the wind turbine is first connected to the grid.

For wide speed ranges i.e. variable speed operation, Larsson (1995) states that the synchronous generator with a rectifier and a thyristors inverter is the most common orientation. It should be noted that this combination is for wide speed ranges, it is possible to use induction motors with slip control to provide some limited variable speed operation, although for a smaller scope of wind speeds. The major advantages of synchronous generators compared with the induction generator is that it can be directly connected to the simple diode rectifier, however the synchronous generator is mechanically more complicated compared to the induction generator.

As previously stated synchronous generators have separate excitation and therefore do not require a power supply line like induction machines, however the output voltage of a SG is lower at lower speeds therefore the author presents a combination involving a boost converter between the rectifier and the DC link capacitors. The design makes use of a synchronous generator coupled to an uncontrolled diode rectifier, to reduce losses presented by more traditional use of induction generators with controlled rectifier bridges. At lower speed the boost chopper pumps the rectified generator voltage up to the DC link value necessary for the line side converter operation, therefore:

$$V_{DC} > V_{Line-Peak}$$

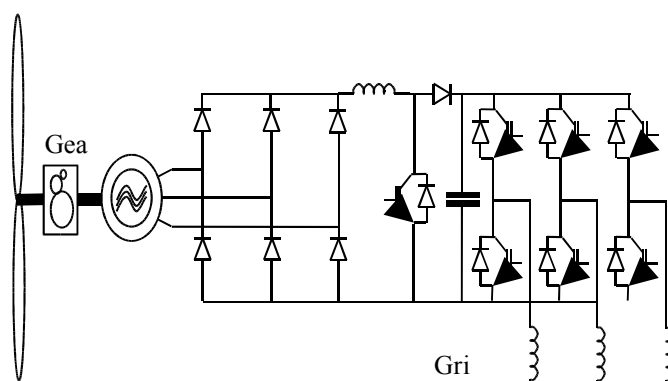


Figure 2 Synchronous Generator with diode rectifier, boost converter and PWM inverter

An additional advantage of using a synchronous generator based power converter is that it is also possible to use multipolar permanent magnet generators providing gearless operation hence increasing efficiency and reducing overall costs.

5. Basic Design Specification

The basis of this novel method of power conversion is constructed around a 16-bit Analog Devices ADSP-21990 fixed-point mixed-signal digital signal processor (DSP). The ADSP-21990 has a dedicated interface for Pulse Width Modulation (PWM) generation, and an on-

board 8-channel 12-bit analogue to digital converter (ADC), which can be controlled via software by the Analog Devices Visual DSP++ 3.0 development system.

The author achieves variable speed constant frequency power operation converter based on DSP control of fast switching IGBT's, it is also possible to measure the input and output and control a *buck*-boost converter using the same DSP board functionality.

With the aid of Matlab's powerful Simulink simulator, using a specialised power based add-on called SimPower it was possible to design the entire power converter with the aid of true-to-life waveforms for the main variables such as current and voltage.

The SimPower Model of the rectifier coupled to a programmable AC source can be seen from figure 3, although it is possible to represent the synchronous generator via SimPower, the final design will be tested via a computer controlled programmable AC source that can follow a practical real life speed pattern, hence the source used for simulation. In addition the voltage waveform generated by the diode rectifier input and output can be seen in figure 4, unfortunately DC outputs under variable AC rectifier inputs cannot possibly be displayed in the limited scope of this paper.

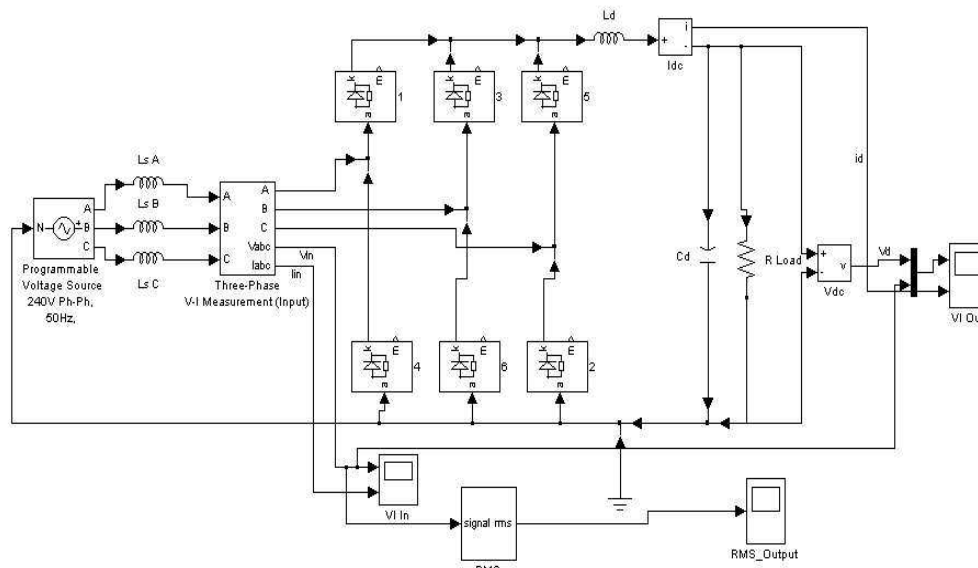


Figure 3 Diode Rectifier with Programmable Source

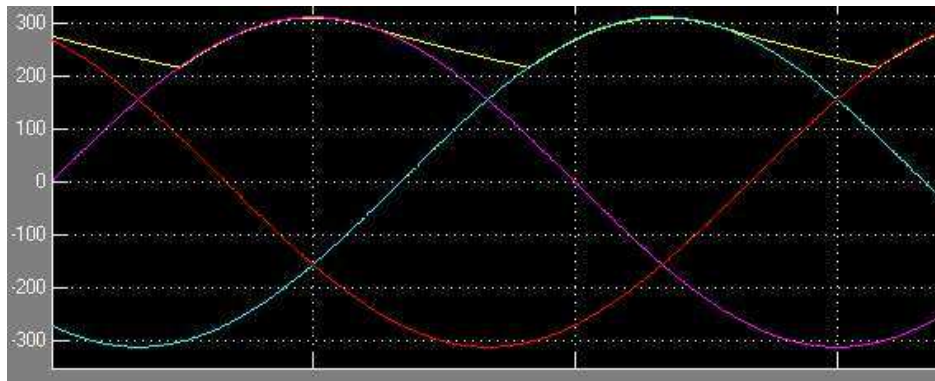


Figure 4 Input & Output Voltage Waveform

The SimPower model of the inverter as can be seen from figure 5 comprises of three separate two arm universal bridges each consisting of IGBT's semiconductors, an external PWM signal built up to provide the switching for the bridges, a fixed DC source has been used for an input for demonstration purposes instead of the DC link capacitor and boost converter, the output and input voltage waveforms of this system can be seen from figure 6. It is not expected that the final solution when built will operate identical to and produce the waveforms shown and in reality although the semiconductors are modelled with internal resistance and inductance etc. in practise substantial filtering will also be necessary.

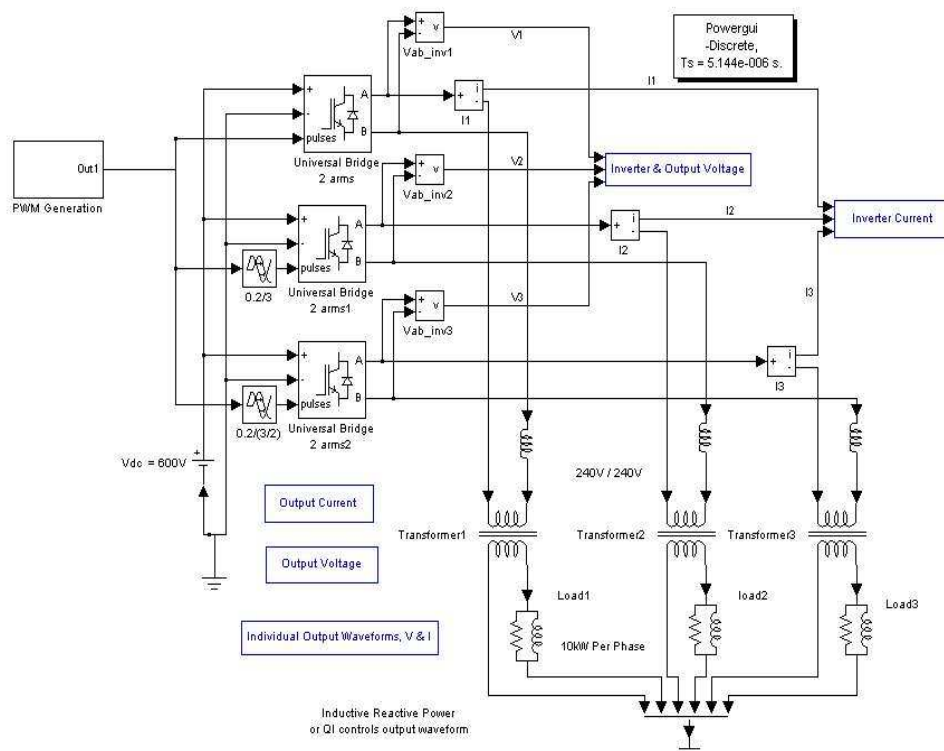


Figure 5 PWM IGBT Inverter Simulation

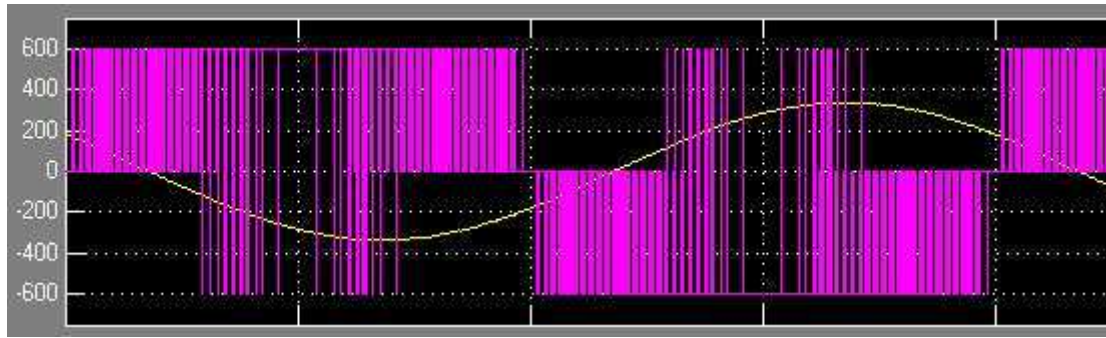


Figure 6 Inverter Input & Output Example Waveform

6. Distortion

Krien (1998) makes the point that distortion is an inevitable result of the non-linear switching behaviour of power electronics circuits, hence filtering of the output is necessary due to the switching requirement of the IGBT's inverter and boost converter.

Since power converters are based around the fundamental concept of a switching methodology, an infinite series of Fourier components called harmonics will always be present to some extent. Therefore the output of a conversion process will always contain unwanted component requiring filtering in an attempt to produce a near ideal source. Harmonics or harmonic distortion refers to this collection of unwanted components and the term ripple is often used for DC equivalent of this unwanted elements.

The term total harmonic distortion (THD) will be of most interest in evaluation the quality of the line value determined by the power, and is generally a standardised measurement of harmonics for AC applications. THD in this case is defined as the ratio of the RMS values of unwanted components in relation to the fundamental. In Europe the definition for THD is giving as:

$$THD = \sqrt{\frac{\sum_{n=2}^{\infty} c_n^2}{\sum_{n=1}^{\infty} c_n^2}} = \sqrt{\frac{f^2_{rms} - f_{1rms}^2}{f^2_{rms}}} f(t)$$

c_n = Fourier Coefficients

f = Signal $f(t)$

In order for the converter to be effective the THD absolute maximum governed by the European standard EN50160-2000 can be no more than 8%. This value is set has an absolute maximum by the European Standard, which the ESB National Grid works from, however the

grid voltage level at the point on connection is also an independent factor for the desired maximum THD.

7. Conclusion & Further Work

This paper describes the advantages of variable speed wind power generation over the more traditional fixed speed, in that 10% increased energy capture and less fluctuation from the wind turbine passing on to the grid as with standard induction generator directly coupled to the grid.

The author has presented a configuration utilizing a synchronous generator coupled to an uncontrolled diode rectifier, a boost converter and a DSP controlled inverter. Matlab's Simulink SimPower add-on models of the design are presented with their associated waveform displaying input/output voltages for the basic diode rectifier and PWM controlled inverter.

Further research is necessary in the area of grid synchronisation using the DSP control, in addition further work is crucial in order to produce a fault ride through method based on the current power converter design.

References

- [1] **Datta, Rajib and V.T. Ranganathan** (2002). "Variable-Speed Wind Power Generation Using a Double Fed Wound Rotor Induction Machine – A Comparison with Alternative Schemes", IEEE Transactions on Energy Conversion, Vol 17, No. 3.
- [2] **Svensson, J** (1996). "Possibilities by Using a self-commutated voltage source inverter connected to a weak grid in wind parks," 1996 European Union Wind Energy Conference and Exhibition, Goteborg, Sweden, 20-24 May 1996, pp. 492-495.
- [3] **Eric Hau** (2000). *Wind Turbines, Fundamentals, Technologies, Application, Economics*, Springer.
- [4] **P.W. Carlin, A.S. Laxson, E.B. Muljadi** (2001). "The history and state of the Art of Variable-Speed Wind Turbine Technology", National Renewable Energy Laboratory, NREL/TP-500-28607.
- [5] **Schreiber, D.** "Applied Designs of Variable Speed Wind Turbines and New Approaches", Application Manager, Semikron International, Nuremberg, Germany.
- [6] **A. Larsson, O. Carlson, G. Siden** (1995). "Electrical Generating Systems in Wind Turbine Applications", Stockholm Power Tech, Stockholm, Sweden, June 18-22.
- [7] **Philip T. Krien**, (1998) "Elements of Power Electronics", New York, Oxford University Press, 1998
- [8] **Jan Svensson** (1998) "Grid Connected Voltage Source Converter – Control Principles and Wind Energy Application", Technical Report No. 331, Dept of Electric Power Engineering, Chalmers University of Technology, Sweden.
- [9] **Commission of Energy Regulation (CER)** (2004), Presentation to the Joint Oireachtas Committee on Agriculture and Food, 28th January 2004.
- [10] **N. Mohan, T.M. Undeland, W.P. Robbins**, (1995) "Power electronics, converters, applications and design", J. Wiley and Sons.
- [11] **E. Muljadi, K. Pierce, P. Migliore**, (1998) Control strategy for Variable-Speed, Stall-Regulated Wind Turbines, National Renewable Energy Laboratory, NREL/CP-500-24311.
- [12] **McDade, J., Reddoch, T.W., and, Lawler, J.S** (1985) "Experimental Investigation of a Variable-Speed, Constant-Frequency Electric Generating System from a Utility Perspective." DOE/NASA/4105-1, NASA CR-174950.
- [13] **NASA**, (1982) "MOD-5B Wind Turbine System Concept and Preliminary Design Report." NASA CR-168047.

- [14] **R. Pena, J.C. Clare, G.M. Asher, (1996)** "A double fed induction generator using a back-to-back PWM converters supplying an isolated load from a variable speed wind turbine", IEEE Proceedings – Electr, Power Applications., Vol. 143, No. 5.
- [15] **Carlson O., Hylander J., (1981)** "Some methods to connect a Wind power Induction Generator to the utility Network," International colloquium on wind energy, Brighton, UK.
- [16] **Dubois, Maxime R., (2000)** "Review Of Electromechanical Conversion in Wind Turbines", Report EPP00.R03.
- [17] **Pierik J.T.G., Veltman A.T., de Haan S.W.H., Smith G.A., Infield D.G., Simmons A.D., (1994)** "A new class of converters for variable speed wind turbines", EWEC 1994
- [18] **Reeves, E.A., (1990)** Handbook Of Electrical Installation Practise, Second Edition, Blackwell Scientific Publication.
- [19] **Commission for Energy Regulation (CER),** The Moratorium on Issuing Connection Offers, December 2003

Strengthening the Practices of an Agile Methodology?

Jimmy Doody¹, Amanda O'Farrell²

¹Department of Computing, Institute of Technology Tallaght, Dublin 24

²Department of Computing, Institute of Technology Tallaght, Dublin 24

Contact email: Amanda.oFarrell@it-tallaght.ie

Abstract

An investigation into how the software development process in an agile environment (Extreme Programming) can be aided by intelligent software, leading to the development of a tool that will automate the process of standardising and clarifying source code. The tool will also speed up and aid the testing process, by producing test objects based on the source code, and by providing full test tracking, without having a negative impact on the development process. By making the coding and testing processes more automated, the research aims to evaluate the following hypotheses:

1. That the programmer's productivity can be increased by the use of such a tool

1. That the standardised and clarified source code can reinforce the practices of Collective Code Ownership, Refactoring and Pair Programming

1. That automated testing can reinforce the practices of Code Ownership, Refactoring, Small Releases and Continuous Integration, as stated by Beck (2000).

Keywords: Agile Methodologies, Extreme Programming, Programmer Productivity, and XP Practices

1 Introduction

The needs of software development projects have changed radically over the years, and as a result of this many different software development methodologies have been developed to try to address these changing needs, whilst improving the quality of the software and the predictability of the project.

The first well-known and widely used methodology, the traditional software development life cycle (SDLC), based its approach on the more traditional engineering disciplines (Burch, 1992). This approach has a strong emphasis on planning and is often referred to as being 'document driven' (Glass, 2001). Following on from this a number of methodologies have emerged which take different views on how software should be developed. Rapid application development (RAD) was one such methodology, which based its process around the production of prototypes. However it is only suited to systems that need to be developed quickly, for example, in response to a perceived opportunity to gain competitive advantage in a new market or industry (Hoffer et al., 2002). As a compromise between speed and quality the evolutionary methodologies such as incremental and spiral were developed. These methodologies combined aspects of the more traditional SDLC with aspects of RAD. Then during the 1990s came the object-oriented life cycle, this was developed to support the object-oriented programming paradigm (Martin et al., 1992). Design Patterns quickly followed; however, they try to solve a general design problem in a particular context, as opposed to defining a complete development methodology (Gamma et al., 1995).

Although these methodologies are quite different, they all (although to different levels) try to impose some disciplined process on software development in order to make it more efficient and predictable (Riehle, in Succi, et al., 2001, pp. 35). They do this by developing a detailed process with strong emphasis on planning. Fowler (2003) refers to them as engineering methodologies because they are inspired by other engineering disciplines. However, these methodologies are frequently criticised for being too bureaucratic, with so much additional work to do to follow the methodology that the development process is completely slowed down (Fowler, 2003). Indeed, Fowler (2003) refers to them as the heavy methodologies, because of the huge bureaucratic overhead associated with them.

Over the last few years, a number of new methodologies have been developed in order to address some of the problem issues that exist with the more traditional life cycles. These were originally known as lightweight methodologies, because they do not have the same administrative overhead as the more traditional methodologies (Riehle, in Succi, et al., 2001, pp. 37). However, the preferred term is agile methodologies, because although the methodology is light in weight, the Agile Alliance¹ did not want the methodologies to be referred to as 'lightweight' (Cockburn, in AgileAlliance, 2001).

2 Agile Methodologies Explained

Agile methodologies have developed as a reaction to the heavy methodologies, in an attempt to establish a compromise between too much process and no process at all (Fowler, 2003). As a result of this, agile methodologies have some differences from other methodologies. The first big difference that is usually noticed is the smaller amount of documentation. According to Fowler (2003), the agile methodologies are code-oriented, and follow a path, which says that the source code is the key part of the documentation. However, Fowler (2003), states that the lack of documentation is as a result of two much deeper differences that exist in all agile methods and these are:

¹*"Agile methods are adaptive rather than predictive"* - other methodologies try to plan out all aspects of the projects in advance, and often spend a huge amount of man hours in doing this. This works very well until things change, thus these methodologies have a tendency to resist change. Agile methodologies try to be processes that adapt, even to the point of changing themselves

¹*"Agile methods are people-oriented rather than process-oriented"* - other methodologies define a process that works no matter who is involved in using the process. The view of agile methodologies is that the skill of the development team is the most important factor and thus the role of the process should be to support the development team

¹ A consortium of experts that promote the use of agile methodologies

Taking the first point, that agile methodologies are adaptive rather than predictive, Fowler (in Succi, 2001, pp. 5-6) discussed the differences between a civil or mechanical engineering project and a software development project. According to Fowler (in Succi, 2001, pp. 5) a civil engineering project has two fundamentally different activities, performed by two very different groups of people. The first activity is design, which requires expensive and creative people. The other is construction, which requires less expensive people. Another issue is that with other forms of engineering, the models that are produced as a result of the design phase are based not only on many years of practice, but can also be subjected to mathematical analysis, and therefore schedule predictability is a given (Fowler, 2003).

So based on this, software methodologies should have a predictable schedule that can utilise people with a lower skill base, therefore we must separate design from construction – however, two problems exist with this approach. Firstly in software development, programmers must be skilled enough to question the designer's designs, especially when the designer is unfamiliar with the technologies used in the project (Fowler, in Succi, 2001, p. 5). Secondly, the design must be done in such a way that the construction is as straightforward as possible. The more traditional methodologies try to achieve this by using design notations such as UML, but Fowler (2003) questions the ability of any software designer to produce a design that can make the coding process a predictable activity. Indeed one of the main problems with any software design notation is that you will not realise that it is flawed until you have turned it into code.

He also questions the costs involved in producing extensive software design documents. In civil engineering, the cost of the design effort is about 10% of the job, with the rest of the effort going into construction. With non-agile methodologies, software design can take up to 50% of the effort, with construction taking roughly 15% of the effort. It seems that non-agile methodologies try and emulate engineering disciplines in terms of planning and predictability, and yet when it comes applying these processes to software engineering, it's apparent that the same rules are just not relevant (Fowler, 2003).

This is not a new idea. In a C++ journal published in 1992, Reeves (1992) also discussed what he calls "false parallels" with other engineering disciplines. He discusses the relationship between programming and software design and claims that the software industry as a whole has missed a subtle point about the difference between developing a software design, and what a software design really is. He claims that programming is not about building software, but about designing software, and hence source code is your design document. He also suggested that in the software development life cycle, the construction phase is actually the use of the

complier or the interpreter. Combining Reeves's thinking with his own on the subject, Fowler (2003) has concluded that:

- In software, construction can be free
- In software all the effort seems to be on design, which requires not only talented but creative people
- If design is a creative process then it cannot be easily planned, and so it cannot be easily predicted
- We should be wary of methodologies that use the traditional engineering metaphor for building software, because building software is a very different kind of activity to building a bridge

Another important issue with heavy software development methodologies is that they do not deal well with change; however, changing requirements is an unpleasant fact when it comes to developing software. If planning is based around the software requirements, and they are constantly changing, how can we get a predictable plan? According to Fowler (2003) in most cases you cannot, but that does not necessarily mean that you have no control, it just means that you should use a process that can give you control over unpredictability. This is what is meant by an adaptive process, which agile methodologies claim to be. So how exactly do agile methodologies achieve this adaptability? Fowler (2003) states that the most important factor in controlling a project in an unpredictable environment, is to know accurately where the project currently stands, and so we need some feedback mechanisms that can accurately tell us what the situation is at frequent time intervals. According to Beck (2000) the key to this feedback is iterative development, whereby you frequently produce working versions of the final system which have a subset of the required features. Lippert (2002) discusses incremental development, and states that it is important that increments are fully integrated and as carefully tested as any final delivery, because there is nothing like a series of tested, working systems to show exactly how a project is progressing. Iterative development leads to a style of planning whereby the only stable plans are for the next increment, and because of this, long-term plans are more adaptable (Fowler, 2003). However, it is worth noting that not all types of systems will be suited to iterative development, for example, an Air Traffic Control System or a Payroll system (Jeffries et al., 2001).

The second point that Fowler (2003) makes is that agile methodologies are “people-oriented rather than process-oriented”, whereas with other methodologies people are viewed as replaceable resources which play some role, and thus only the role is important, not the resource that fills it. He describes people working in this type of process as being viewed as “plug compatible programming units”, with no room for creativity or individuality, and that when you have good people, this attitude leads to lower morale – and hence lower productivity. Agile methodologies completely reject this process-oriented notion; Fowler (2003) describes

programmers as being “responsible professionals.” To back up this point he references a paper written by Alistair Cockburn, which describes people as being the most important factor in software development (Cockburn, 1999). Fowler (2003) states that if you want to retain good people you must recognise them as bright, competent, motivated professionals, and as such they themselves are best suited to decide how they conduct their own technical work. In agile methodologies, there is a clear distinction between the role of the customer and the developer, and although they are on the same team, they are considered to have different decisions to make. According to Wake (2001) for example, when practicing Extreme Programming, the developer must be able to make all the technical decisions and all time estimates, whereas the customer makes decisions on issues such as scope and priority.

3 Research Areas

3.1 Extreme Programming (XP)

To date a number of different agile methodologies have been developed, and while they all share many characteristics, there are also some significant differences, and so with this in mind the research will focus around one particular methodology. Extreme Programming (XP) has been chosen because it is strongest in its view that source code is the driving force in any development methodology. XP is the methodology which has attracted most attention; firstly because one of the main proponents of XP is Kent Beck, and secondly because some of its main practices go against what the more traditional methodologies teach. It is seen to be extreme in its views of software development and as such is considered controversial. However, it is the most widely documented of the agile methodologies and although it is extreme in some of its ideas it also claims to be open to interpretation and adaptive in nature.

Jeffries (2001) describes XP as software development based on the four values of simplicity, communication, feedback and trust. These values are quite vague however, and so they are expanded into concrete principles which are used to guide the development team as to which of the 12 XP practices² will best satisfy the four values (Beck, 2000). According to Beck (2000), these 12 practices reinforce each other – he states that none of them stand well on their own, but that each requires the others to keep them in balance. (However, this seems to mean that by ignoring one of the 12 practices, there will be a lessening in the strength of the others that it reinforces. This would indicate inflexibility when it comes to selecting which practices to use for a particular project, even though agile methodologies claim to be highly adaptive.)

² Planning Game, Metaphors, Pair Programming, Refactoring, Collective Code Ownership, On-site Customer, Small Releases, Simple Design, Testing, Continuous Integration, Sustainable Pace, Coding Standards

XP, like the other agile methodologies, believes that the key part of documentation is the source code. Indeed Beck (2000) goes as far as saying that source code is the only artefact that the development process absolutely cannot live without. As a follow-on from this statement, if we must have source code then we should use it for as many of the purposes of software engineering as possible (Beck, 2000). Thus, the first major objective for the research is to look at how the source code from a project can be used for other software engineering purposes, such as generating API documentation or creating unit tests.

3.2 Coding Standards and Simplifying Code

Lippert (2002) states that the source code should follow some coding standards as agreed upon by the programmers, so that it is impossible to tell that the code has been written by many different people. This standardisation of code is one of the main practices of extreme programming and according to Lippert (2002), can reinforce some of the other practices, such as:

- Collective code ownership - by allowing any developer to take some code and add value to it, without having to worry about different coding standards
- Refactoring - as the programmers will not have to re-format after making some changes
- Pair Programming – there will be no disagreements over different formatting and standards

We view project quality in two distinct ways: external quality, which is measured by the customer, and internal quality, which is something that is measured by the programmers (Beck, 2000). Sometimes internal quality is sacrificed in order to meet project deadlines, and ignoring coding standards is one common example of this. However, because coding standards reinforce other practices in an XP environment, ignoring them is just not an option. Following coding standards, however, is not enough when working in an extreme environment. You must agree upon and adhere to general good programming practices. According to Beck (2000), in an XP environment, good programming is programming that follows the simplest design possible. Therefore, you should only code what you need when you need it. Hence, the notion of coding for the future is considered a waste of valuable development and testing time, since the future is uncertain. His definition of simple is as follows:

- The system (code and tests) must communicate everything you want it to communicate
- The system must contain no duplicate code (Once and Once Only Rule)
- The system should have the fewest possible classes
- The system should have the fewest possible methods

Obviously taking the time to apply standards and follow the practice of keeping things simple could slow the development process. So with these issues in mind, the next objective of the

research is to look at how team-defined standardisation can be applied automatically and how the values of simplicity as defined by Beck (2000) can be upheld.

3.3 Testing

Iterative development is something that is practised by XP teams, and is enabled by having small, frequent releases of running, fully tested software. Lippert (2002) states that a simple system should be put into production quickly and the team should then release new versions as frequently as possible. The XP team should integrate and build the system many times a day, or even every time a task is completed. This in turn means that the developers should test their own code fully before trying to integrate it with the current system. According to Beck (2000) developers must continually write unit tests, which must run flawlessly to allow development to continue. Again, this could slow the development process; according to Lippert (2002) however, as with coding standards, the testing process cannot be ignored, as it reinforces other XP practices such as:

- Small releases – by continually testing, the defect rate is so small that lengthy test cycles are no longer a factor and hence frequent releases are enabled
- Refactoring – by re-running tests, programmers can instantly see if refactoring has caused any problems
- Collective code ownership – by re-running tests, programmers can check to see if a change has caused any problems
- Continuous integration – a programmer can confidently release his code for integration, when it has been fully tested

With these issues in mind, another objective of the research is to try to automate the testing process, to not only produce test classes but also to look for duplicate code, and to identify classes and methods that are no longer used.

4 Future Work

To try to answer the research questions, a prototype tool will be developed which will allow an XP team to fully define their standards and hence apply them automatically to any classes that are produced. This prototype will be written for the java language specifically and so the tool will have the option of automatically inserting javadoc comments. The tool will also automatically produce test classes for the objects defined and will look for duplicate code. It will identify methods and classes that are not used, and classes that have not been tested to date. When a class is changed in any way the tool will re-run tests on all classes, to make sure that the changes have not caused problems with any other classes.

It is hoped that the tool will be used in an industrial setting in order to test the tool and to review the impact that it can have on the coding and testing practices, and also to discover how these two practices can reinforce other XP practices.

References

- The AgileAlliance, 2001**, History: *The Agile Manifesto*, Available from <http://www.agilemanifesto.org/history.html> [Accessed 2 September, 2003]
- Beck, K. 2000**, *Extreme Programming Explained – Embrace Change*, Addison Wesley Longman, Massachusetts
- Burch, J. G. 1992**, *Systems Analysis, Design and Implementation*, Boyd & Fraser, Boston
- Cockburn, A. 1999**, *Characterizing People as Non-Linear, First-Order Components in Software Development*, Crystal Methodologies, Available from <http://crystallmethodologies/articles/panic/peopleasnonlinearcomponents.html> [Accessed 12 September, 2003]
- Fowler, M. 2003**, *The New Methodology*, Available from <http://www.martinfowler.com/articles/newmethodology.html> [Accessed 19 July, 2003]
- Gamma, E., Helm, R., Johnson, R. & Vlissides, J. 1995**, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, Massachusetts
- Glass, R. L. 2001**, *Agile versus Traditional: Make Love, Not War!*, Cutter IT Journal, 14, 12, Available from <http://www.cutter.com/itjournal/2001toc.html> [Accessed 25 September, 2003]
- Hoffer, J. A., George, J.F. & Valacich, J.S. 2002**, *Modern Systems Analysis and Design*, 3rd edition, Prentice Hall International,
- Jeffries, R. 2001**, *What is Extreme Programming*, XProgramming.com, Available from <http://www.xprogramming.com/xpmag/whatisxp.htm> [Accessed 20 September, 2003]
- Jeffries, R., Anderson, A. & Hendrickson, C. 2001**, *Extreme Programming Installed*, Addison-Wesley, United States of America
- Lippert, M. & Roock, S. & Wolf, H. 2002**, *Extreme Programming in Action, Practical Experiences from Real World Projects*, Wiley & Sons, England
- Martin, J. & Odell, J. 1992**, *Object-Oriented Analysis & Design*, Prentice Hall, New Jersey
- Reeves, J.W. 1992**, *What is Software Design*, Available from <http://www.bleading-edge.com/Publications/C++Journal/Cpjour2> [Accessed 29 August, 2003]
- Succi, G., & Marchesi, M. 2001**, *Extreme Programming Explained*, Addison-Wesley, United States of America
- Sun, 2003**, *Code Conventions for the Java Programming Language*, Sun Microsystems, Available from <http://java.sun.com/docs/codeconv/html/CodeConvTOC.doc.html> [Accessed 9 March, 2004]

A Qualitative Method for Determining the Quality of BGA Solder Joints in a Lead-Free Process

Shane O'Neill¹, John Donovan¹ & Claire Ryan²

¹ School of Engineering, Institute of Technology Sligo, Ireland

² Stokes Research Institute, University of Limerick, Ireland

Abstract

The introduction of lead-free soldering is inevitable for the electronics industry and its use poses a number of challenges. Manufacturing processes need to be re-evaluated and any reliability issue needs to be addressed. In this study the effect of lead free solder on a reflow soldering process is investigated. Experimental design techniques were used to examine a reflow soldering process using the process parameters as experimental factors. The factors included the conveyor belt speed of the reflow oven and the preheat, soak and reflow temperatures of the temperature profile. Micro Ball Grid Array (BGA) packages were used as the test components. No standard method exists to assess the quality of BGA solder joints. Solder joint quality is normally assessed using lengthy reliability tests that measure joint strength. It is highly advantageous if a qualitative assessment method was available that could determine the joint quality. This study presents a scoring method that can be used to evaluate this solder joint quality quickly and inexpensively. BGA solder joint quality was assessed using x-ray and micro section inspection techniques. This qualitative data was scored and weighted. The weighted solder joint quality scores were statistically analysed to check for effect significance. It was found that conveyor belt speed had a statistically significant effect on the weighted score. The statistical approach was verified using residual analysis. The results of the experiment demonstrate that the scoring method is a practical way of assessing BGA solder joint quality. This paper presents a unique scoring method for assessing the joint quality of BGA packages.

Introduction

The use of lead-free solders is rapidly increasing in the global electronics manufacturing industry. Until now tin-lead (SnPb) solders were the most popular alloy of choice due to their low melting point, high strength ductility, high fatigue resistance, high thermal cycling and joint integrity. SnPb solders have been used to create the electrical and mechanical connections between components and printed circuit boards (PCBs) for more than 50 years and have proved reliable. Recently, due to concerns over the environmental impact of lead, legislation has been introduced banning its use from electrical and electronic equipment. The EU directives 2002/95/EC (2003) and 2002/96/EC (2003) are two pieces of legislation that apply to EU member states and will come into effect on the 1st of July 2006. From this date electrical and electronic products sold in the EU must be lead-free. Currently there is no drop-in lead-free solder replacement for SnPb but there are many lead-free alternatives available as reported by Suraski and Seelig (2001) and Bath *et al* (2000). Since the introduction of lead-free solder is not a simple drop-in exercise, all related processes and reliability issues must be evaluated using the new solder alloys.

Ball Grid Arrays (BGAs) are electronic packages that have gained a large market-share in the electronics packaging industry. This is due to their compactness and large number of inputs and outputs that facilitates the trend toward smaller and lighter electronic products without the

loss of equipment performance. Their attractive characteristics mean they have become integral to many electronic systems from military to consumer applications.

With the advent of lead-free solder, re-evaluation of processes and the solder joint quality of BGAs has become a major research area. This current study investigates the effects of a lead-free solder in a reflow soldering process by using experimental design techniques. Visual inspection of solder joint micro-sections through x-ray is the accepted industry method of inspecting BGA solder joints. No scoring method exists to rate the quality of the solder joints from visual inspection. An inexpensive qualitative method of assessing the solder joint quality after micro-sectioning and x-ray inspection was developed. The joints were evaluated and scored against quality characteristics taken from accepted industry standards, IPC standard 610 Rev. C. Each of the quality characteristics was weighted and the weighted scores were analysed using experimental design techniques to identify any significant process factors. Residual analysis was then used to verify the statistical approach. The results of the experiment show that this scoring system is a suitable method to use when assessing BGA solder joint quality.

Preparation of Test Assemblies

A 10 x 10 array micro BGA package with lead-free solder bumps was used as the test component. The test boards were eight-layer FR-4 printed circuit boards (PCBs) with 16 BGA component positions. An Organic Solderability Preservative (OSP) board finish was used for the lead-free components. The lead-free solder paste chosen for the experiment was a 95.5Sn 3.8Ag 0.7Cu alloy, which was the same material used in the BGA solder bumps. An example of a test assembly board with BGAs attached is shown in figure 1.

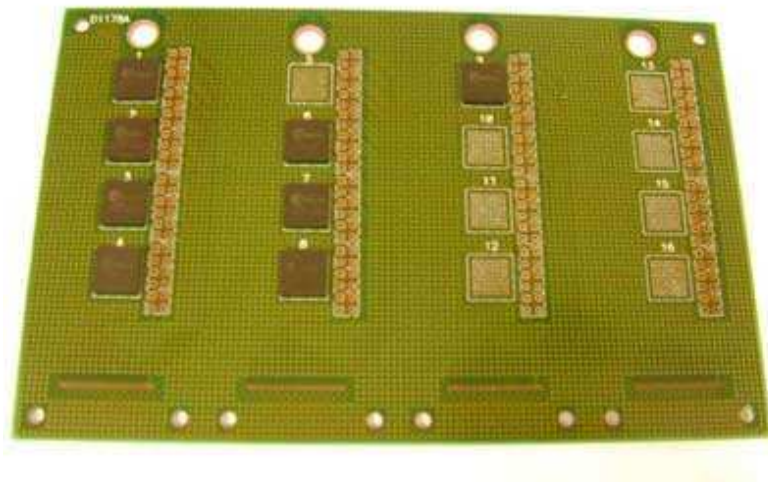


Figure 1: Test assembly board

A screen-printing process printed the solder paste on the PCBs after which eight BGA packages were placed using a BGA placement system. Each board was then reflowed at the chosen settings according to the experimental design matrix.

Experimental Design

Experimental design is a statistical technique used to help design new products and processes, optimise existing manufacturing processes and improve existing products. A 2_{IV}^{6-2} fractional factorial was used for the experimental design in the present study. This involved a total of sixteen experimental runs. One PCB containing eight BGAs was reflowed during each experimental run. The six factors in the experiment were the five temperature zones in the reflow oven and the speed of the oven conveyor belt. The factors and levels are detailed in Table 1. Factor levels were set to reflect the reflow profile recommended by the solder paste manufacturer.

	Factor	-	+
A	Conveyor Speed	12 inches/min	14 inches/min
B	Preheat temperature 1	170°C	180°C
C	Preheat temperature 2	210°C	220°C
D	Soak temperature	230°C	240°C
E	Reflow temperature 1	245°C	265°C
F	Reflow temperature 2	280°C	315°C

Table 1 Experimental Design Factors and Levels

The response of interest was the solder joint quality. There is no quantitative measure to evaluate solder joint quality so qualitative data was gathered through inspection of solder joint characteristics in accordance with IPC standard 610 Rev. C.

Method for Solder Joint Evaluation

There is no standard method used to assess the quality of BGA solder joints. Reliability tests such as accelerated temperature cycling, power cycling, or cyclic bending mechanical tests are typically used to measure solder joint strength. These are proven methods of testing but are costly and time consuming. A typical temperature cycling test carried out to assess solder joint strength could last anything up to four thousand hours as reported by Towashirapom *et al* (2002). It would be of great benefit if a fast and inexpensive method existed that could measure solder joint quality. By using visual evaluation of the solder joints and scoring of the resulting qualitative data such a method was devised in this study.

The techniques used to evaluate the solder joints were x-ray and cross section analysis. Figures 2 and 3 show examples of both.

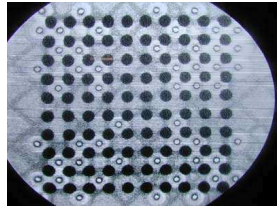


Figure 2 Solder Joint X-

Ray

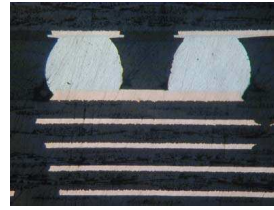


Figure 3 Solder Joint

Cross Section

X-Ray was used to examine for defects such as bridging, open joints, and solderballs. Cross sectioning was used to examine the joints in detail for solder joint formation, wetting, voids, and alignment. One BGA was cross-sectioned from each PCB. Every cross section exposed a row of ten solder joints for inspection.

Each of the ten solder joints were evaluated under the following categories and scored out of ten:

- Defects
- Solder Joint Formation
- Wetting
- Void Area
- Void Frequency
- Alignment

These categories were carefully chosen based on the guideline the IPC standard provided and knowledge of the process. A score of ten represented a bad joint and zero represented a good one. Each category was assigned a weight according to its importance as shown in Table 2. For example an open joint is categorised as a defect that would cause failure of the component immediately, accordingly Defects was assigned a high weighting. Inspection of the solder joints revealed that there were no defects on any PCB. There is an important relationship between void area and void frequency. One large void occupying 50% of the solder joint was considered more serious than several smaller voids occupying the same area. Therefore, when void frequency (Vf) was greater than zero, the experimental run weighted score denoted WS was calculated as follows:

$$WS = aD + \left(\frac{bJF + cW + \frac{xVa}{yVf} + zA}{n} \right)$$

When $Vf = 0$, Ws was calculated as:

$$Ws = aD + \left(\frac{bJF + cW + zA}{n} \right)$$

Where n = the number micro-sectioned solder balls.

Void frequency and void area were assigned a weighting of one to allow for this relationship within the equations.

Category	Category Nomenclature	Weight	Weight Nomenclature
Defects	D	0.90	a
Solder Joint Formation	JF	0.80	b
Wetting	W	0.70	c
Void Area	VS	1.0	x
Void Frequency	VF	1.0	y
Alignment	A	0.30	z

Table 2 Weighting Values

Results and Discussion

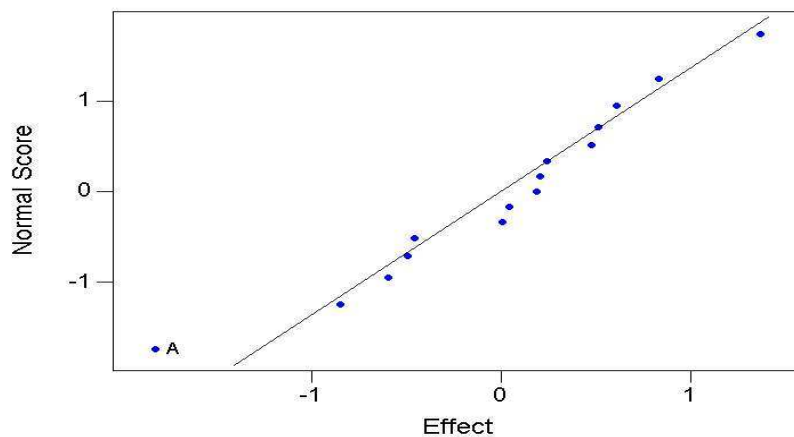
The experiment runs and weighted scores are included in the design matrix in Table 3. As this was an unreplicated experiment the analysis of the results involved using methods proposed by Daniel (1959). He suggests using probability plots to plot the effects making the assumption that the data comes from a normal distribution with a mean of zero and constant variance.

When plotted, nonsignificant effects should lie approximately on a straight line while significant ones tend to lie off the line.

<i>Run Order</i>	<i>Conveyor Speed</i>	<i>Preheat Temp. 1</i>	<i>Preheat Temp. 1</i>	<i>Soak Temp</i>	<i>Reflow Temp. 1</i>	<i>Reflow Temp. 1</i>	<i>Average Weighted Score</i>
1	12 in/sec	170°C	210°C	240°C	245°C	315°C	5.018
2	14 in/sec	170°C	210°C	240°C	265°C	315°C	4.81
3	14 in/sec	180°C	210°C	240°C	245°C	280°C	3.563
4	12 in/sec	170°C	220°C	230°C	265°C	315°C	5.235
5	14 in/sec	180°C	220°C	230°C	265°C	280°C	3.56
6	12 in/sec	180°C	220°C	230°C	245°C	280°C	4.866
7	14 in/sec	170°C	210°C	230°C	265°C	280°C	2.85
8	14 in/sec	170°C	220°C	230°C	245°C	315°C	2.04
9	12 in/sec	180°C	210°C	230°C	265°C	315°C	6.812
10	14 in/sec	180°C	210°C	230°C	245°C	315°C	3.27
11	12 in/sec	170°C	220°C	240°C	265°C	280°C	3.653
12	14 in/sec	170°C	220°C	240°C	245°C	280°C	3.105
13	12 in/sec	180°C	220°C	240°C	245°C	315°C	4.848
14	12 in/sec	170°C	210°C	230°C	245°C	280°C	8.583
15	12 in/sec	180°C	210°C	240°C	265°C	280°C	4.605
16	14 in/sec	180°C	220°C	240°C	265°C	315°C	4.788

Table 3 Experimental Design Matrix

The normal probability plot of the effects is shown in Figure 4. From the plot it may be seen that factor A (Conveyor Belt Speed) lies off the line and therefore may be considered significant. Daniels method rests on the principle of effects sparsity. This is the hypothesis that only a small proportion of the factors in an experiment have effects that are large.

*Figure 4 Normal Probability Plot of the Effects*

The main effects plot for conveyor belt speed is shown in Figure 5. It may be seen that the best response, i.e. the lowest weighted score, is achieved at the higher conveyor speed of 14 inches per second.

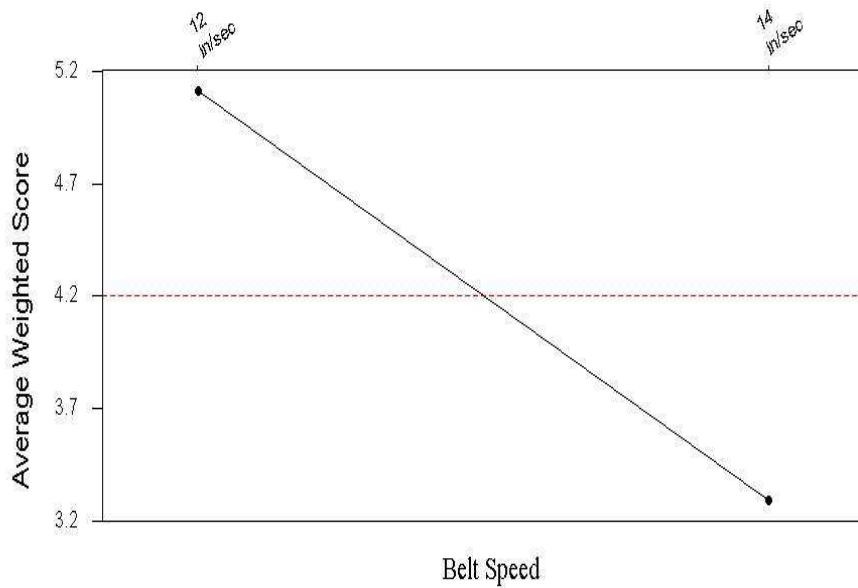


Figure 5 Main Effects Plot for Conveyor Belt Speed

ANOVA

The ANOVA table for the data is presented in Table 5. On examination of the p values factor A, conveyor belt speed is significant at the 5% level. All of the other terms were pooled to form the error.

Source	DF	SS	MS	F	P
Belt Speed	1		13.359		9.16
					0.009
Error	14	20.414		1.458	
Total	15	33.773			

Table 5 ANOVA Table

The residuals were analysed to confirm the adequacy of the model and to ensure the ANOVA assumptions were not violated. Figure 6 and Figure 7 show the normal plot of the residuals and the residuals versus the predicted values respectively. From the plots it may be seen that there is no evidence of non-normality and the equality of variance assumption was not violated.

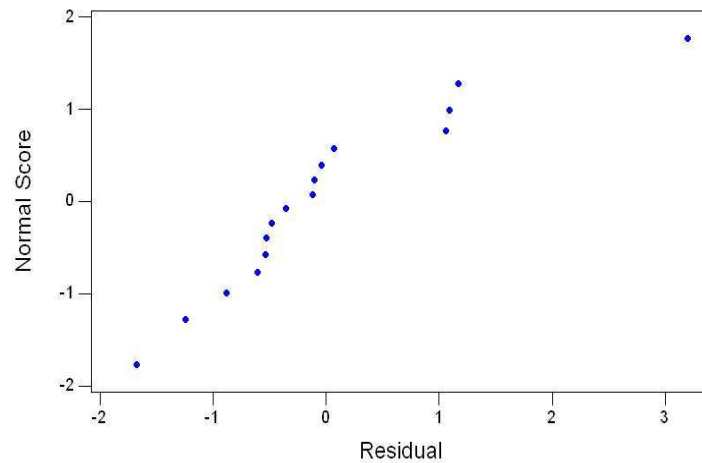


Figure 6 Normal Probability Plot of the Residuals

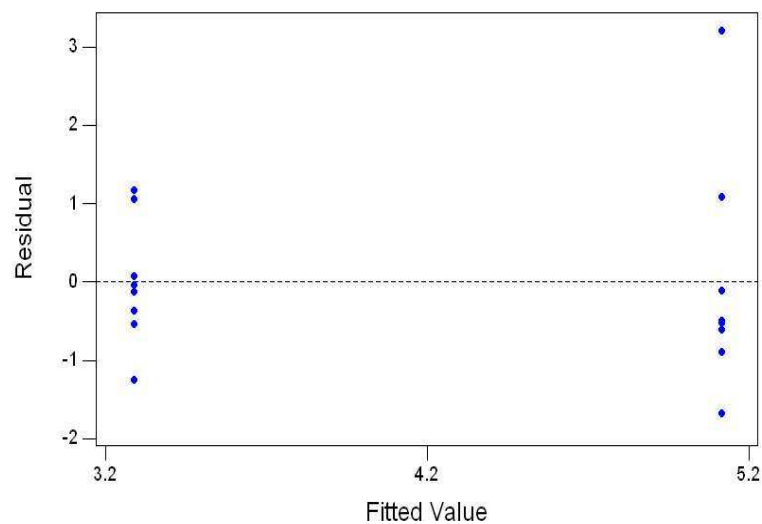


Figure 7 Predicted (Fitted) Values Versus the Residuals

Recommendation

The adequacy of the approach adopted in this study has been verified through analysis of residuals. To further validate the scoring technique presented in this paper, research into a test method designed to evaluate BGA solder joint strength must be conducted. The results presented in this paper and the results of the test to evaluate solder joint strength must then be examined for correlation.

Conclusion

The introduction of lead-free soldering in the electronics industry is coming and all related processes must be tested and evaluated. This study presents the investigation of the effects of a lead-free solder on a reflow soldering process. An experimental design was carried out on the process. Since no method of rating BGA solder joint quality exists, a scoring method was developed against accepted industry standards to assess the quality of the solder joints produced by the experiment. Statistical methods were used to evaluate the weighted score. Through experimental analysis it was shown that conveyor belt speed had an effect on the weighted score in the lead-free reflow soldering process. These results demonstrate that the scoring method is a practical way of assessing BGA solder joint quality in a confident and inexpensive manner.

Acknowledgements

The authors grateful acknowledge Enterprise Ireland for providing funding through the PEI Technologies Advanced Technology Research Programme (ATRP).

References

- Bath, J., Handwerker, C. and Bradley, E., (2000), "Research Update: Lead-Free Solder Alternatives", *Circuits Assembly*, May 2000, pp 31 – 40.
- Daniel, C., (1959), "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments", *Technometrics*, vol. 1, no. 4, pp. 311 – 340.
- Directive 2002/95/EC of the European Parliament and of the Council of 27 January 2003 on the Restriction of the Use of Certain Hazardous Substances in Electrical and Electronic Equipment (2003), Official Journal of the European Union, Vol. 37, pp. 19 - 23.
- Directive 2002/96/EC of the European Parliament and of the Council of 27 January 2003 on the Waste Electrical and Electronic Equipment (2003), Official Journal of the European Union, Vol. 37, pp. 24 – 38.
- IPC-A-610 Rev. C, pp 12-74 – 12-76, Available for purchase from IPC, orderipc@ipc.org
- Suraski, D. and Seelig, K.,(2001), "The Current Status of Lead-Free Solder Alloys", *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 24, no. 4, pp 244 – 248.
- Towashirapom, P., Subbarayan, G., McIlvanie, B., Hunter, B. C., Love, D., and Sullivan, B., (2002), "Predictive Reliability Models Through Validated Correlation Between Power Cycling and Thermal Cycling Accelerated Life Tests", *Soldering and Surface Mount Technology*, vol. 14, no. 3, pp. 51 – 60.

Application of the Hough Transform to Aid Raised Pavement Marker Detection on Marked Roadways

Colin O'Rourke¹, Catherine Deegan¹, Simon McLoughlin¹ & Charles Markham²

¹ School of Informatics and Engineering, Institute of Technology Blanchardstown, Dublin 15

² Department of Computer Science, NUI Maynooth, Maynooth, Co. Kildare

Contact email: Colin.O'Rourke@itb.ie

Abstract

A machine vision system is proposed that will identify and locate GPS co-ordinates of defective Raised Pavement Markers along road lane markings. This system will comprise of a mobile data acquisition system, and a separate offline image analysis system. In this paper we present a method for road lane marking identification, using a Hough Transform based technique. This paper describes the implementation of the Hough Transform for line detection. Using the Hough Transform algorithm road lane markings are automatically identified, given a scene acquired from a digital camera system. This knowledge is intended to be used to aid current research in the area of defective Raised Pavement Marker detection at ITB. Results of a sample dataset are presented and discussed

Keywords Machine Vision System, Hough Transform, Raised Pavement Marker, Accumulator, Global Positioning System

1. Introduction

The National Roads Authority (NRA) in Ireland is responsible for the installation and maintenance of Raised Pavement Markers (RPMs), on all roadways throughout Ireland. The main focus of this research is to develop a working prototype for the NRA that can automatically identify and locate defective RPMs. This research can be sub-divided into four parts: Image Acquisition; Image Processing/Analysis; and Fusion with GPS data, with the end goal being a fully functioning practical prototype. Currently, a stereo vision image acquisition system is under development, synchronously capturing image data via two firewire digital cameras. In this paper we focus on the image processing stage of our research, primarily concerned with a technique that can help solve the problem of defective RPM detection.

2. The Hough Transform

2.1 Introduction

In the field of image processing it is essential to be able to find and identify various objects within image sequences. Objects of interest account for various shapes (road lane markings) with straight and circular edges, that project to straight and elliptical boundaries in an image. One method of identifying features of interest within an image is the Hough Transform.

The Hough Transform (HT), developed by Paul Hough in 1962 [1] [2], has become a standard tool in the field of computer vision for the recognition of straight lines, circles and ellipses. The HT is a technique which can be used to isolate features of a particular shape within an image.

Using some curve representation, this technique transforms a set of points defined over the image space to a set of points defined over some parameter space (known as Hough space). Points in Hough space represent particular instances of a curve in the image. Therefore, the strategy used by the HT is to map sets of points from a particular instance of the considered curve, i.e. the parameterized curve, to a single point representing the curve in Hough space and, in effect, cause a peak to occur at that point [3]. The main advantages of the HT technique are that it is relatively unaffected by image noise while it is also particularly robust to missing and contaminated data, tolerant of gaps in feature boundaries.

2.2 Implementation of the Hough Transform

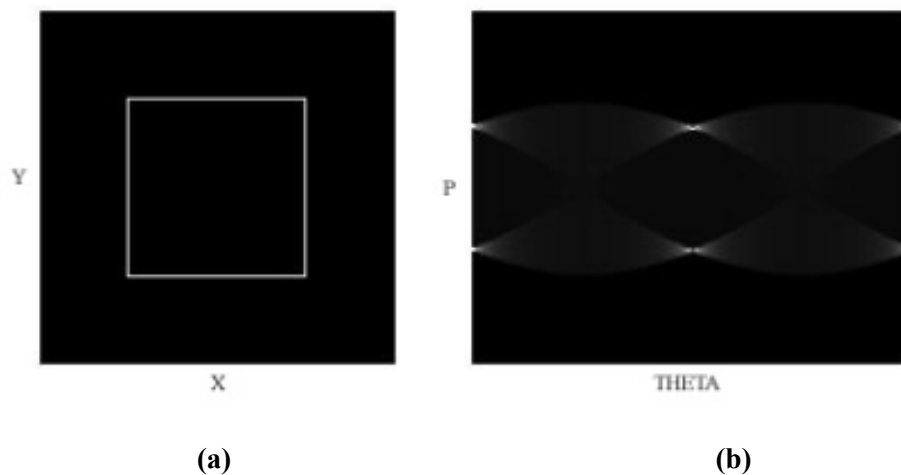


Figure 1: Hough Transform from (a) image space to (b) Hough space

The motivation for the HT technique for line detection is that each input measurement (i.e. the co-ordinate point) indicates its contribution to a globally consistent solution (i.e. the physical line which gave rise to that image point). To illustrate how the HT technique works, consider the common problem of fitting a set of line segments to a set of discrete image points (i.e. pixel locations output from an edge detector). A simple square is used as an example, illustrated in Figure 1(a), above. The corresponding Hough space is shown in Figure 1(b). In Figure 1(b) above the intensity represents peak size, ρ is represented by the vertical axis, and θ is represented by the horizontal axis.

When interpreting the information contained in Figure 1(b) it might seem that six peaks have occurred to the trained eye when in fact only four true peaks have occurred, since Hough space is periodic. This point-to-curve transformation, from image space to Hough space, represents the HT for straight lines. When viewed in Hough parameter space, points which are collinear in the Cartesian image space become readily apparent as they yield curves which intersect at a

common point. Hence each of the four peaks identified in Hough space correspond to each of the four lines that represent the square in our image space.

A straight line in an image has two ends, but the Hough transform (see Figure 1(b)) does not find end-points. Rather it finds the infinite straight lines on which the image edges lie. A part of a line lying between two end-points is called a line segment, note that all references in this paper to the term line will mean an infinite straight line. These infinite lines are worth extracting, since once found they can be followed through the image and end-points located if necessary.

2.2.1 Parametric Representation of HT

Consider a single isolated edge point (x, y) ; there could be an infinite number of lines that could pass through this point. Each of these lines can be characterized as the solution to some particular equation. The simplest form in which to express a line is the slope-intercept form:

$$y = mx + c \quad (1)$$

where m is the slope of the line and c is the y -intercept (the y value of the line when it crosses the y axis). Any line can be characterized by these two parameters m and c . Each of the possible lines that pass through point (x, y) can be characterized as having coordinates (m, c) in some slope-intercept space. In fact, for all the lines that pass through a given point, there is a different value of c for m :

$$c = y - mx \quad (2)$$

The set of (m, c) values, corresponding to the lines passing through point (x, y) , form a line in (m, c) space. Every point in image space (x, y) corresponds to a line in parameter space (m, c) and each point in (m, c) space corresponds to a line in image space (x, y) [2].

2.2.2 Accumulators

The Hough Transform works by quantizing the Hough parameter space into finite intervals (i.e. letting each feature point (x, y) vote in (m, c) space for each possible line passing through it). These votes are totaled in an *accumulator*. Suppose that a particular (m, c) has one vote, this means that there is a feature point through which this line passes. If there are two votes, this means that two feature points lie on that line. If a position (m, c) in the accumulator has n votes, this means that n feature points lie on that line. Lines for which a high number of votes are accumulated result in the occurrence of peaks in Hough space.

2.2.3 The HT Algorithm

The algorithm for the HT can be expressed as follows:

1. Find all of the desired feature points in the image
2. For each feature point:
3. For each possible line, in the accumulator, that passes through the feature point (ρ, θ)
4. Increment the (ρ, θ) position in the accumulator
5. Find local maxima in the accumulator

2.2.4 Polar Representation of HT

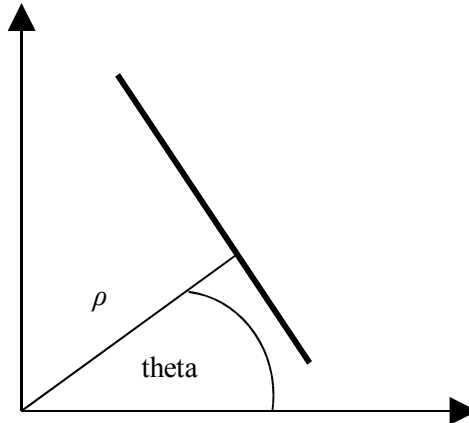


Figure 2: Parametric representation of a line

The slope-intercept form of a line has a problem with vertical lines: both m and c are infinite. Another way of expressing a line is the *normal parameterisation form*, or (ρ, θ) form:

$$x \cos \theta + y \sin \theta = \rho \quad (3)$$

One way of interpreting this is to draw a perpendicular line from the origin to the line (Figure 2). θ is the angle that the perpendicular line makes with the x -axis and ρ is the length of the perpendicular (bounded by the diagonal of the image). If θ is restricted to the interval $[0, \pi]$, then the normal parameters for a line are unique. With this restriction, every line in the x - y plane corresponds to a unique point in the θ - ρ plane [2]. Given some set $\{(x_1, y_1), \dots, (x_n, y_n)\}$ of n figure points. By transforming the points (x_i, y_i) into the sinusoidal curves in the θ - ρ plane, one can calculate the set of lines that fit the n figure points.

Curves generated by collinear points in the image space intersect and form peaks in *Hough space*. These intersection points characterize the straight line segments of the original image [2]. Instead of making lines in the accumulator, each feature point votes for a sinusoid in the accumulator. Where these sinusoids cross, there are higher accumulator values. Finding maxima in the accumulator still equates to finding the lines. There are a number of methods which one might employ to extract these bright points (peaks), or local maxima, from the accumulator array.

3. Experiments and Results

A sample dataset was collected on a national roadway. A single frame of the data used is illustrated in Figure 3(a) below. It was decided that the camera position was set as close to driver eye level as possible to allow for realistic measurements and realistic data analysis of the driving scenario. For the purposes of the HT technique all image data had a threshold function applied to convert it to a binary image, and a Sobel edge detector [6] was then applied to each frame of the image sequence. Figure 3(b) illustrates the edge map, of the original test image data Figure 3(a), returned (using a threshold level of 40).

Figure 3: (a) Test image



Figure 3: (b) Sobel edge map of test image

After the edge map was generated, the HT function was applied to the edge map rendering a Hough parameter space plot. Some of the noise was eliminated by thresholding the peaks in *Hough space*. Hence the number of legal peaks found, to be plotted was reduced to 100 in order to reduce the amount of information to be processed.

To aid our analysis of the test data Hough space was represented as a 3-D mesh of peaks, see Figure 4. It can be clearly seen from this Figure where the strong and relevant peaks lie. The strongest peak representing the centre lane of the roadway can be clearly seen as a light blue/yellow peak.

In order to allow individual identification of each road lane, the approximate angle of each line was computed in relation to the camera position, in the image space. An angular threshold was set to extract peaks corresponding to the individual lane markings.

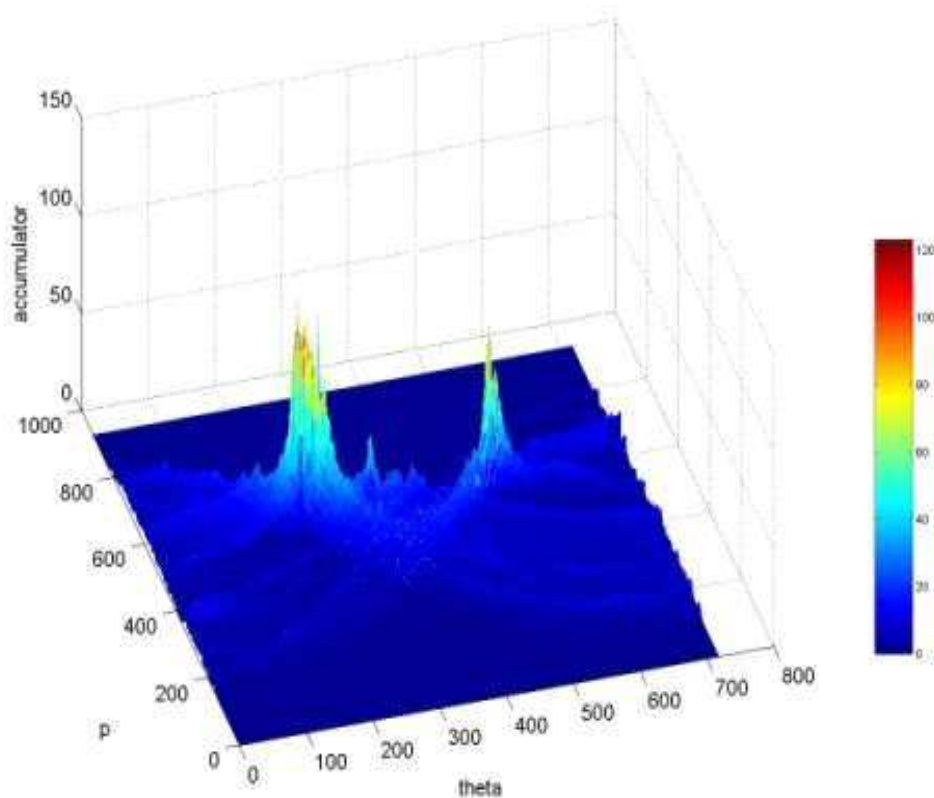


Figure 4: Hough space illustrated in a 3-D mesh

The HT technique discussed above was applied to other datasets collected from minor roads. As expected, when this technique was applied to some of the data sets acquired, the results for the right road lane were inadequate. The cause of this error was related to the camera position.

Hence these problems can be over come by introducing a second camera to concentrate on the right hand side road marking. This was due to the fact that the angle of the right hand road marking is greater (closer to horizontal) than the left and centre road markings in terms of the camera position, hence other feature points in the image (noise etc.) were causing incorrect peaks being created in Hough space.

3.1 Conclusions

This paper details a method of identifying road lanes, with the intent to reducing the image processing complexity when identifying RPMs along a roadway. The method involves applying the Hough Transform to the input data, with the result rendering the identification of the road lane(s). Knowledge of where the road lanes are will greatly reduce the amount of information needed in the data set, and allow for image analysis to be focused on a subset of the data set. It is planned to develop this prototype technique to take into account faded or defective painted road lanes, in order to identify roadways that need immediate maintenance.

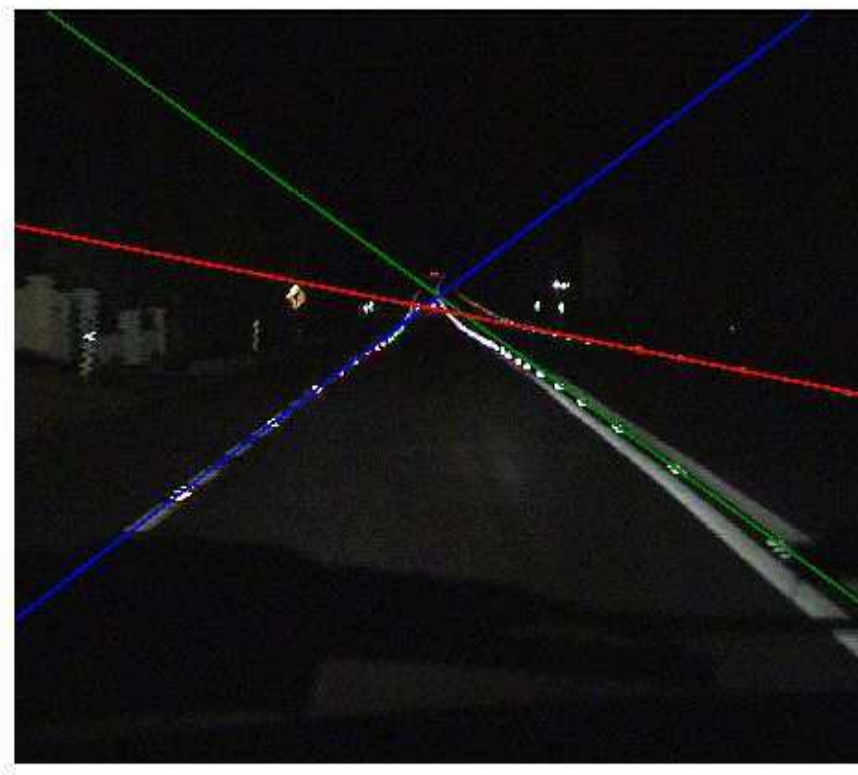


Figure 5: Inverse Hough Transform result

4. Future Work

It is planned to develop this technique further and acquire road data using a stereo vision system, which is currently near development completion. With this system it is envisaged that

the right hand road marking will produce a greater peak in Hough space allowing for more straightforward identification, and separation from unwanted noise.

The next stage of research will involve developing a robust RPM locator algorithm which can traverse along road markings and identify all defective RPMs along those lines. In order to detect RPMs along the road, a threshold must be set to allow for the contour changes in the ground plane levels of the road. Therefore a segment of the road data image will be used, where the contour of the ground plane is constant. Once the RPMs have been identified and located in relation to the vehicle (on which the stereo vision system will be mounted) this data will be fused with vehicle GPS position information to allow extraction of global RPM co-ordinate information. With this GPS co-ordinate information it is hoped that as an extension to this research initiative, a database of defective RPMs can be catalogued.

References

- [1] P. Hough, "Method and means of recognizing complex patterns" U.S. Patent 3069654, 1962
- [2] R. Duda and P. Hart, "Use of the Hough Transform to detect lines and curves in pictures", *Communications of the ACM*, vol. 15, pp. 11-15, January 1972
- [3] J. McDonald, Jochen Franz, and Robert Shorten, "Application of the Hough Transform to lane detection in Motorway Driving Scenarios", Signals & Systems Group, Department of Computer Science, NUI Maynooth, Maynooth, Ireland
- [4] S. R. Deans, "Hough Transform from the Radon Transform", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, pp. 185-188, March 1981
- [5] D. H. Ballard, "Generalizing the Hough Transform to detect arbitrary shapes", *Pattern Recognition*, vol. 13, pp. 111-122 (1981)
- [6] D. Vernon, "Machine Vision", Prentice Hall, 1991, Chap. 5

Investigation Into The Correct Statistical Distribution For Oxide Breakdown Versus The Oxide Thickness Used In Integrated Circuit Manufacture

James Prendergast¹, Eoin O'Driscoll¹, Ed Mullen².

Institute Technology Tralee¹.

Analog Devices, Raheen Industrial Estate, Limerick²

Abstract

A critical aspect of integrated circuit manufacturing is the reliability of the components, in particular the gate oxide of transistors and capacitors. Accelerated stress tests are used for reliability predictions of gate oxides. There are two statistical distributions, which can be applied to stress test failure data, namely the Lognormal or the Weibull distributions. The failure data can fit each distribution equally well. However the use of either distribution will give vastly different lifetime predictions and their correct use is crucial for accurate lifetime prediction. A statistical based test, developed with Monte Carlo data, which is designed to decide if a failure data set has an underlying Lognormal or Weibull distribution is applied to empirical Time Dependent Dielectric Breakdown (TDDB) failure tests. The TDDB tests are carried out on 7nm and 15nm thick gate oxides. The results generated show the necessity of making the correct choice between the two distributions for accurate lifetime prediction and validate the test for different oxide thickness.

I. Introduction

With increased scaling the reliability of thin gate oxides such as silicon dioxide (SiO_2) has become a very important issue in the microelectronics industry. The requirement for higher performance integrated circuits, has led to the necessity for continuous downscaling of gate oxide dimensions. Gate oxides thickness as small as 1.5 nm thick is being implemented in MOSFET's with gate lengths of only 4nm [1]. The industry often specifies that there is less than a 0.1% cumulative failure rate, for 10-years or greater, operational lifetime of semiconductors. Yet as gate oxides become thinner, the margin between the predicted lifetimes of the oxide versus this 10-year specification is decreasing rapidly [2].

II. TDDB Reliability Testing

TDDB tests are required for the determination of gate oxide reliability. These tests are typically performed on small sample sizes across multiple high temperatures and voltages. In many cases the sample size is approximately 16 units per test condition and they are generally tested to 100% failure. From these tests the failure times are generated and statistically analysed. The critical times are the T_{50} and $T_{0.1}$ percentiles. The T_{50} percentile is used to generate the thermal and field acceleration factors while the $T_{0.1}$ is used to predict the predicted lifetime based on the Arrhenius equation. The TDDB failure times have underlying statistical distributions, the two most common being the Lognormal and Weibull distributions.

The choice between the Lognormal and Weibull distribution can have major consequences on the determination of the predicted lifetime of the gate oxide. It has been reported that there is a significant difference between both distributions at lower percentiles, and that this difference increases as gate oxides get thinner [3][4][5].

Variations of the Arrhenius equation are used to extrapolate accelerated life data to operational life data. The basic form of the Arrhenius equation is given by:

$$r = A \exp\left(\frac{-E_a}{kT}\right) \dots\dots\dots(1)$$

Where; r = reaction rate, k = Boltzmann's Constant (8.63×10^{-5} eV/K), E_a = Activation Energy (eV), A = Frequency factor, T= Absolute temperature K.

Some of the variations of the Arrhenius equation are used to derive the thermal and field acceleration factors in TDDB experiments. These acceleration factors are used to extrapolate the TDDB test conditions, i.e. high temperature and voltage, to operating conditions. The acceleration factors due to voltage and temperature are expressed in equation 2 and 3 below:

$$AF_{voltage} = \exp\{\gamma (V_{stress} - V_{op})\} \dots\dots\dots(2)$$

Where: γ = field acceleration factor, V_{stress} = Voltage Stress, V_{op} = Operating voltage.

$$AF_{temp} = \exp\left[\frac{E_a}{kT_{op}} - \frac{E_a}{kT_{stress}}\right] \dots\dots\dots(3)$$

Where: E_a = Activation Energy (eV), k = Boltzman's constant (8.617×10^{-5}), T_{stress} = Stress temperature, T_{op} = Operating temperature.

Both E_a and γ are derived using T_{50} percentiles of TDDB failure data as variables.

III. Correct Distribution Test

Chi² and Kolmogorov-Smirnov tests are some general Goodness-of-fit tests. These have historically been used by Reliability Engineers to help choose the correct distribution, yet they are not specific to Lognormal and Weibull distributions.

A test was developed by Croes et al [5] to offer reliability engineers an objective tool, which would make a distinction between Lognormal and Weibull distributions. The test is based on Pearson's correlation coefficient;

$$\rho = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \dots\dots\dots(4)$$

Where; \bar{x}, \bar{y} are the respective means of x, y.

The test involves calculating the ratio between the correlation coefficient of a Weibull distribution (ρ_{wei}) for a certain failure data set, and the correlation coefficient of the Lognormal distribution (ρ_{logn}) for the same failure data set. The ratio is written as $\rho_{\text{wei}} / \rho_{\text{logn}}$. This ratio is shown to have the ability to be used as a test statistic. Using $\rho_{\text{wei}} / \rho_{\text{logn}}$ as a test statistic, Croes et al [5] devised a hypothesis test that chooses the correct statistical distribution to a certain significance level.

This test was subsequently further developed by Cain [6] for useful applications. $\rho_{\text{wei}} / \rho_{\text{logn}}$ is compared to a critical value (W_{crit}). The critical value is dependent on the size of the TDDB data set, and is tailored such that the test makes a choice given no *a priori* information on the data. If $\rho_{\text{wei}} / \rho_{\text{logn}}$ is greater than W_{crit} then the underlying distribution of the TDDB data is a Weibull distribution, and if $\rho_{\text{wei}} / \rho_{\text{logn}}$ is less than W_{crit} then the underlying distribution of the TDDB data is a Lognormal distribution. Cain [6] also calculates the probability of making a correct decision in relation to the size of the data set, as seen in *Figure 1*, assuming that all the gate oxide test structures fail during the TDDB test.

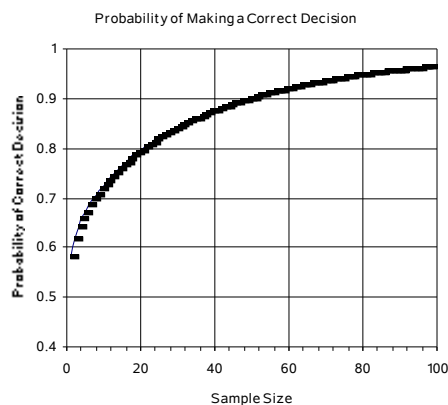


Figure 1 Probability of Correct Decision with Increasing Sample Size up to N=100.

The above tests were developed based on Monte Carlo data sets. The following sections apply empirical data to the theoretical work and points to the inconsistencies that arise if the incorrect statistical distributions are used.

IV. Experimental Details

TDDDB tests were carried out on n-type MOS capacitors with gate oxide thickness (t_{ox}) of 7nm and 15nm. The number of capacitors used and stress conditions are in Table 1 below.

Table 1.

t_{ox} (nm)	No. of Caps.	E-field Stress (MV/cm)	Temp.Stress (°C)
7	170	9	225
15	110	8	225

All capacitors were stressed till failure occurred, i.e. hard breakdown of gate oxide [7].

V. Application of Test for Correct Distribution

Figures 2 and 3 are the TDDDB failure data of the 7nm gate oxide, given a Lognormal and Weibull distribution respectively.

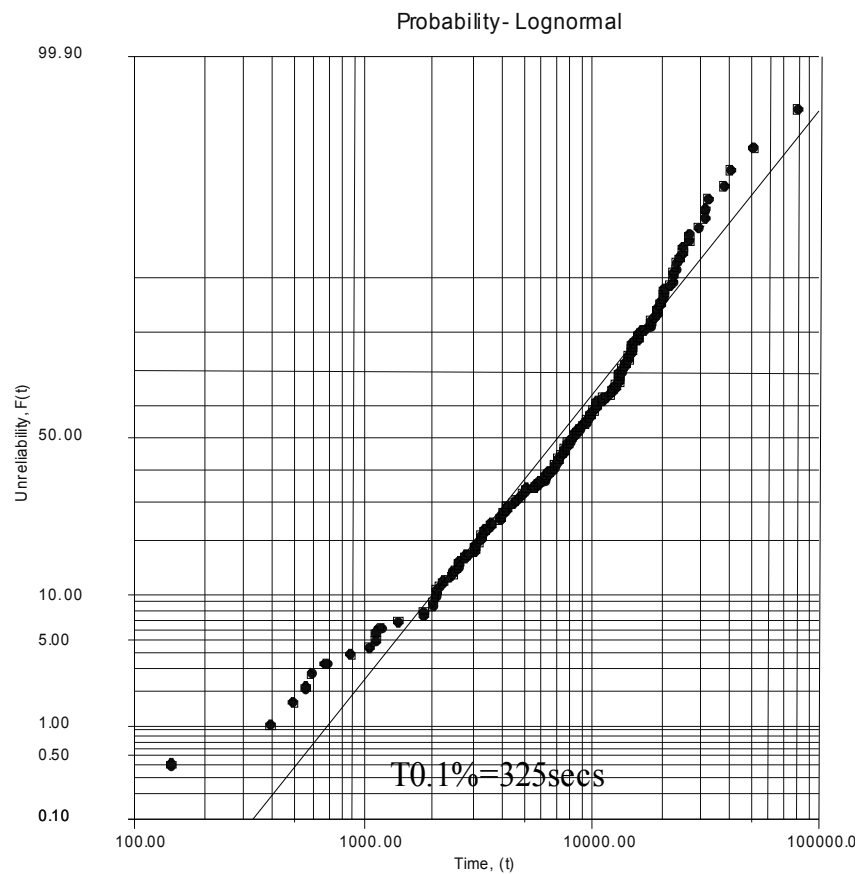


Figure 2 Lognormal distribution of 7nm TDDDB Data

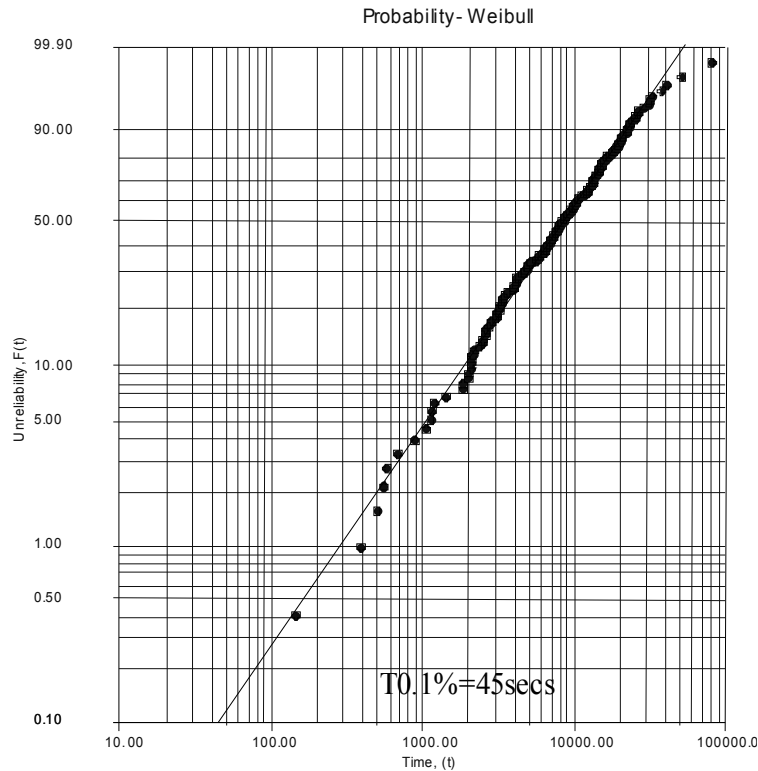


Figure 3 Weibull distribution of 7nm TDDB Data

The correlation coefficient for the Lognormal distribution ($\rho_{\log n}$) = 0.9807, and the correlation coefficient for the Weibull distribution (ρ_{wei}) = 0.9970. The ratio $\rho_{\text{wei}} / \rho_{\log n} = 1.017$. The critical value of a data set of 170 = 0.9958. $\rho_{\text{wei}} / \rho_{\log n}$ is greater than the critical value which implies that the TDDB failure times has an underlying Weibull distribution. The $t_{0.1}$ (time taken for 0.1% of population to fail) of the Lognormal is 325 seconds, and the $t_{0.1}$ of the Weibull distribution is 45 seconds. This is a difference of a factor 7.

Figures 4 and 5 are the TDDB failure data of the 15nm gate oxide, given a Lognormal and Weibull distribution respectively

Using the test as for the 7nm gate oxide, the underlying distribution for the 15nm gate oxide is found to be a Lognormal distribution. The $t_{0.1}$ of the Lognormal distribution is 850 seconds, and the $t_{0.1}$ of the Weibull distribution is 195 seconds. This is a difference of a factor 4. This difference is less than 7 nm gate oxide difference, which is in agreement with previous findings [4]. The results imply that with 93% confidence the lognormal distribution is the most suitable for the 15 nm oxide. For the 7 nm oxide there is a 95 % chance that the Weibull distribution is the correct one to use. As a result the correct statistical distribution to use depends on the oxide thickness.

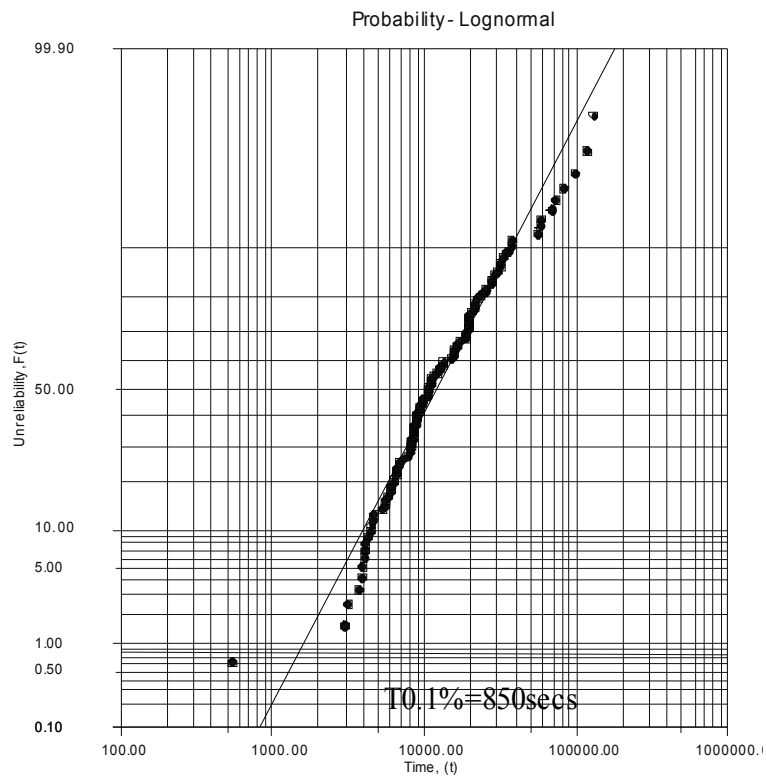


Figure 4 Lognormal distribution of 15nm TDDB Data

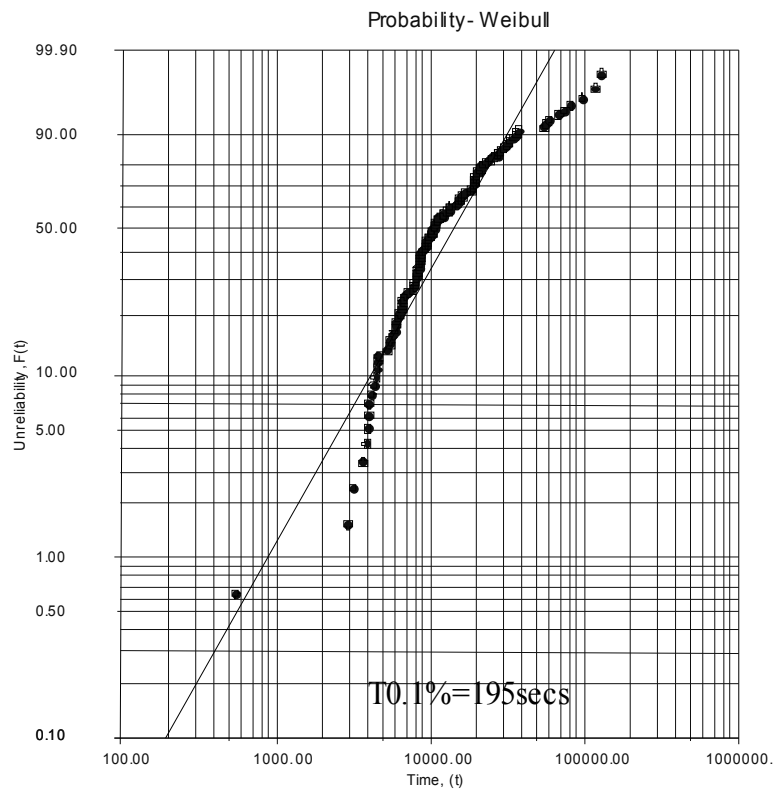


Figure 5 Weibull distribution of 15nm TDDB Data

Equation 2 and 3 can be combined and rewritten as can be rewritten such that:

$$TTF_{op} = Exp [-E_a/k (1/T_{test} - 1/T_{op})] \times Exp \gamma (V_{test} - V_{op}) \times TTF_{test} \dots \dots (5)$$

where: TTF_{op} represents the time for 0.1% fails at operating temperature and voltage conditions, e.g. 55°C and 3V, extrapolated from stress conditions. TTF_{test} is the $t_{0.1}$ percentile of the TDDB failure data. Using this equation with E_a and γ values derived from experiment, $t_{0.1}$ for operating temperature conditions are calculated for 7nm and 15nm gate oxides, using both the Lognormal and Weibull conditions, as seen in Table 2.

Table 2

t_{ox}	Distribution	T0.1% test	T0.1% use
7 nm	Weibull	45 sec	16yrs
	Lognormal	325 sec	65yrs
7 nm	Weibull	195 sec	5552yrs
	Lognormal	850 sec	9005yrs

For the 7nm gate oxide, the $t_{0.1}$ at operating temperature and field conditions is 65 years using the Lognormal distribution, and 16 years using the Weibull distribution. As the underlying distribution for the 7nm TDDB data is found to be Weibull, the use of a Lognormal distribution would give an over-optimistic evaluation of the reliability of the gate oxide.

For the 15nm gate oxide, the $t_{0.1}$ at operating temperature and field conditions is 9005 years using the Lognormal distribution, and 5552 years using the Weibull distribution. As the underlying distribution for the 15nm TDDB data is found to be Lognormal, the use of a Weibull distribution would give an over- pessimistic evaluation of the reliability of the gate oxide.

VI. Conclusion

In this paper empirical data has been presented to support the theoretical model proposed by Croes and further developed by Cain to distinguish between Lognormal and Weibull distributions. The results show how choosing the wrong distribution can lead to erroneous reliability projection. Choosing distributions arbitrarily may no longer be sufficient and care needs to be taken before reliability characterisation to determine the correct statistical

distribution. This is crucial for the microelectronics industry as gate oxides become thinner and customer's reliability expectations increase.

References

- [1] R.Degraeve, B.Kaczer, G.Groesenken, "Degradation and Breakdown in Thin Oxide Layers: Mechanisms, Models and Reliability Prediction", *Microelectronics Reliability*, Issue 39, pp. 1445-1460, 1999.
- [2] J.H. Stathis, "Physical and Predictive Models of Ultra Thin Oxide Reliability in CMOS Devices and Circuits", *IEEE/IRPS*, pp. 132-147, 2001
- [3] R.Degraeve, B. Kaczer, G. Groeseneken, "Reliability: a possible showstopper for oxide thickness scaling? ", *Semiconductor Science Technology*, Volume 15, pp. 436-444, 2000.
- [4] E.Wu, W.Abadeer, L-K. Han, S-H. Lo, G.Hueckel, "Challenges For Accurate Reliability Projections In The Ultra-Thin Oxide Regime", *IEEE/IRPS*, pp. 57-65, 1999.
- [5] K.Croes, J.Manca, W. DeCeuninck, L.Deschepper, G.Molenberghs " The time of 'guessing' your failure time distribution is over", *Microelectronics Reliability*, Issue 38, pp. 1187-1191, 1998.
- [6] S. Cain, "Distinguishing between lognormal and weibull distributions", *IEEE Transactions on Reliability*, Volume 51, pp.32-38, 2002.
- [7] J.Prendergast, N.Finucane, J.Suehle, "Investigation of the SiO₂ Area Dependence using TDDB Testing". *IEEE International Reliability Workshop*, pp.22-25, 1997.

Emotion Authentication: A Method for Voice Integrity Checking

C. Reynolds¹, L Vasiu² and M. Smith³

¹ Middlesex University, School of Computing science, Bramley Rd, Oakwood, N14 4YZ UK.

² Middlesex University, School of Computing science, Bramley Rd, Oakwood, N14 4YZ UK.

³ Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Republic of Ireland

Contact email: c.reynolds@mdx.ac.uk

Abstract

When people communicate with telephone type systems, it is often assumed that the listener would notice any modification of the speaker's voice. It is possible however to introduce small changes that would not be noticed by a listener but could modify the reading of a Voice Stress Analyser, popularly referred to as a lie detector. Existing approaches to checking the integrity of voice require significant amounts of processing or are able to detect only non-subtle modification such as change of speaker. With the advent of real time voice modification using software and hardware based signal processing, we argue that it is not possible at present to easily determine if a received unencrypted voice message has been modified in some subtle way. This is particularly important in the current climate of biometric and psychophysiological analysis. We critically investigate alternative approaches to a voice integrity check in the light of the subtle changes that might be made to a speaker's voice and propose a method that enables voice integrity checking for digital communications in a variety of scenarios.

Keywords: Voice Integrity, Emotion, Voice Stress Analysis, Voice Communication.

1.0 Introduction

Current digital communication systems are becoming integrated with local network services in a move towards greater mobility and flexibility. For example, Voice Over Internet Protocol (VOIP) is used locally within networks and assists in reducing costs. Within this new environment new technologies such as speaker identification systems are developing. These technologies are becoming more robust and the gathering and usage of biometric and psychophysiological information is increasing.

In many voice communication systems the voice information is encrypted using robust techniques and this may act to prevent third party real-time modification of the voice communication. With some new developments such as VOIP implementation, secure encryption adds to the processing overheads and the latency inherent in the communication. Therefore much of the communication traffic may run as unencrypted data to keep latency to a minimum especially where transmission delays are an issue.

It is possible to access VOIP communications using packet-sniffing approaches. This gives rise to the possibility of interception and transmission of the voice data packets, in an attack commonly called the man-the-middle-attack. Until recently the modification of a voice signal would be difficult to achieve in real time or would involve such delays or gross distortion of the original signal that it would be likely that such an attack would be detected. It is now possible to modify voice parameters subtly that give rise to possible privacy and security risks.

The low cost of secondary storage space required to store digital data is enabling long-term storage of communications sessions. Often this may be for prevention of fraud or training purposes, but in many instances, data is kept as a matter of course and may find use in consumer profiling.

With Voice Stress Analysis (VSA) becoming ubiquitous as a method of determining truthfulness and currently popular in the U.K. insurance market, it is important that any data gathered that might find later use has not been tampered with by a third party or those storing data. We propose that using the emotion cue content of a speech provides a robust method for checking message integrity and can be used to check for third party modification of the voice.

Speaker Identification and authentication would not form the basis for message integrity, as it is possible to make subtle changes to the voice that would not affect the parameters used for speaker identification. In the instance of jitter or microtremors, (Smith, 1977. Ruiz, Legros, & Guell, 1990) these are often ignored by voice authentication systems, but find extensive use in VSA devices.

1.1 Background

There has been a great deal of voice emotion research by groups such as the Geneva Emotion Research Group. Scherer pointed out the possibilities of resynthesis in better understanding emotion in the voice and a move towards a more theoretical approach to work in this area (Scherer, 1989. Scherer, Johnstone & Bänziger, 1998. Scherer, 1995. Johnstone & Scherer, 1999. Schuller Lang & Rigoll 2002).

The types of cues that might be analysed could include:

- **Pitch information**, including the mean pitch, its amplitude, its variation and rate and direction of change, in addition to the formants. These are strong harmonics that give a voice its characteristic sound.

- **Amplitude variation** during and between utterances as a whole and in different frequency bands. It is possible to consider a general envelope and also more specific peaks in different formant amplitudes.
- **Low frequency changes** otherwise known as jitter or micro tremors (Smith, 1977. Ruiz, Legros, & Guell, 1990).

VSA manufacturers suggest that the changes in these parameters during speech provides psychophysiological evidence for a speaker's emotional state and can be used to determine truthfulness of a statement by looking for stress cues. Although many researchers currently consider VSA devices be very unreliable (Janniro, M. J., & Cestaro, V. L. 1996. Meyerhoff, Saviolakis, Koenig & Yurick, 2001) such devices still find extensive use in the insurance industry in the UK. It is also possible that more accurate and robust techniques for Voice Stress Analysis (VSA) will exist in the future and this provides a motivation for developing voice integrity checks.

1.2 Watermarking and Fingerprinting Strategies

Watermarks and fingerprints are information that is hidden in a file. Watermarks are often copyright messages that are file specific and fingerprints are often unique messages, for example that might identify the file's owner. This type of message hiding is known as steganography. Providing message integrity might be achieved in a number of ways. These include

- **Fragile Watermarking.** This approach to steganography is not particularly robust, and often relies on changes in either the amplitude or frequency domain to least significant bits. This is often considered as unsuitable for voice data, as the watermarks are often unable to survive the compression process.

With traditional watermarking, the purpose is often to prove ownership and methods exist to remove the watermark data, using tools such as Stirmark (Petitcolas, Anderson & Kuhn, (1998)) often restricting the effectiveness of the watermark. In this instance the watermark is to show integrity of the file and any hacker would be required to modify the data without affecting the watermark. This is far more difficult and would require knowledge of the watermarking process.

- **Semi-Fragile Watermarking.** These schemes accept some changes to a file or to a data stream but work with respect to a threshold. This could allow basic tone control or amplitude changes, but not changing the length of the file for example. By only placing data across frequency bands that are not quantised or lost in lossy compression such as MP3 it is possible to retain the watermark even when compression takes place. It may be difficult to guarantee the integrity of data with this approach as it is possible to make changes below the threshold.
- **Robust Watermarking.** These watermarks are designed to withstand attacks and are suitable for embedding copyright information. This might be in conjunction with a web spider in order to check for marked files.
- **Content Based Watermarking.** Wu and Kuo (2001) have considered a semi-fragile watermarking scheme based on semantic content. In this instance semantic is taken to mean higher level detail such as pitch. This approach adopts a semi-fragile approach, which provides clear indication of dramatic changes in the voice as a result of modification, but would not identify changes to the high frequency energy or the addition or removal of jitter. The tests carried out involved replacing one speaker's voice with another.

2.0 Hacking

There are many opportunities for subtle modifications of emotional voice cues. This modification may be subtle and difficult to identify.

Different VSA devices use different voice emotion cues to decide on the veracity of statements made by a subject. However it is very simple to remove jitter using a very steep high pass filter or to add jitter either by modulation of the voice signal or by adding a very low frequency waveform between 8 and 15Hz. For many traditional VSA approaches this is sufficient to modify the readings of the VSA device. Devices might also use other voice stress cues such as frequency changes in the second formant or changes in high frequency energy.

To demonstrate a simple hacking approach we have developed tools such as a jitter generator (JitGen4) that can add jitter in a variety of ways (Fig 1.). Including the generation of low frequencies and the modulation of the speech signal. These have been developed using graphical modular synthesiser development tools such as SynthEdit (McClintlock, 2002) and run in AudioMulch (Bencina, Ross 2004), a VST (Steinberg) real time processing environments.

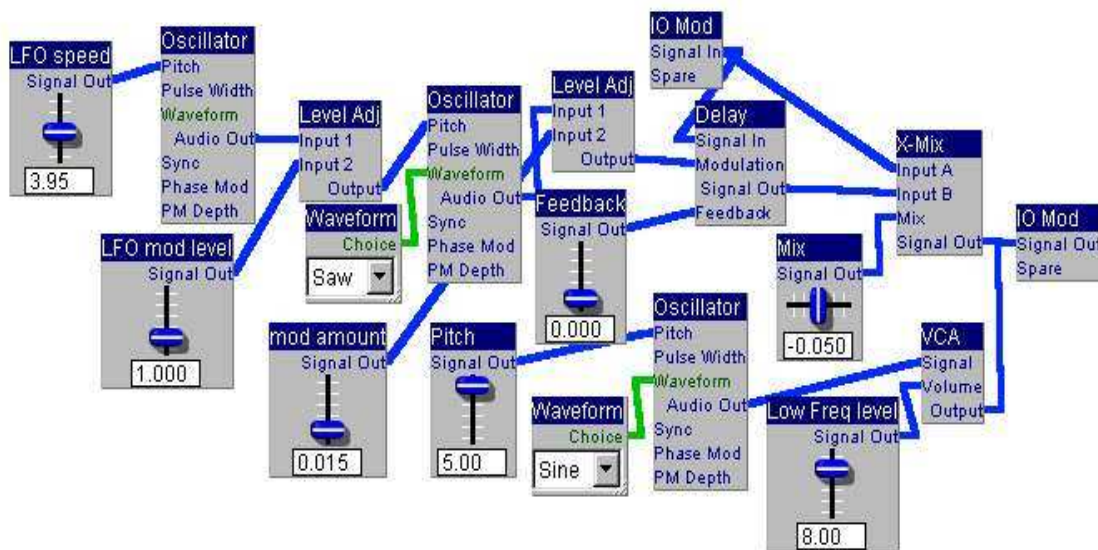


Figure 1: JitGen4 Structure.

We have developed simple filters (Fig 2.) that can also be used in the same environment together with off-the-shelf audio compressors to limit the dynamic variation for a given frequency range. If used in conjunction with expanders to increase the dynamic variation in a different frequency range, the overall envelope does not appear to change but the processing can have a significant impact on the voice emotion cues.

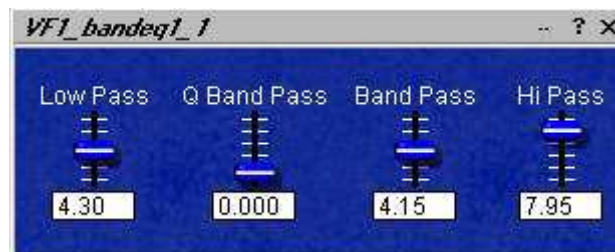


Figure 2: Filter front end showing available control.

Although these devices are crude, they demonstrate the potential of "Emotion Hacking" and associated risks to privacy. Third party use could lead to erroneous judgements made about the speaker and the semantic content of their message. This effectively leads to a breakdown in trust.

3.0 Requirements

The requirements of integrity checking are very different from the need to retain information within a file. This implies that the embedded data needs to be suitably immune to transfer

through a communications network but not robust enough to survive changes to harmonic content or jitter. General requirements are listed:

- **Blind detection:** This is where the file contains data that self validates the integrity and does not rely on an original file being available, this is of particular importance when the integrity check may need to occur in real time upon receiving a spoken message.
- **Low Distortion:** The impact of adding a fingerprint does not degrade the quality of the transmission.
- **Low Latency:** The delay caused by adding the fingerprint, encryption and validating the data on reception should not have an impact on the communication.
- **Low Data Rate:** The amount of data added by the fingerprinting process, should be small.

Low latency can be partly achieved by only having very small amounts of data that have to be encrypted. If we consider common hacking techniques such as:

- **Ciphertext only attack.** If we know the structure of a file then file predictability allows us to know when we have successfully attacked an encrypted message.
- **Known plaintext attack.** If we know part of a message, then it is much easier to attack by attempting to encrypt the known message fragment and comparing to the encrypted file.
- **Chosen plaintext attack.** This is where we encrypt a message with the public key and try to decode. This is very time consuming and is often described as a brute force attack.
- **Man-in-the-middle attack.** Intercept and pass on.
- **Timing attack.** The time taken to encrypt is used as an indication of the key size, this is a sophisticated attack and difficult to perform.

A small amount of encrypted data sent with no encrypted header is difficult to attack, in terms of breaking the encryption.

4.0 Proposed Emotion Profile Integrity Checking

To ensure low latency only a small amount of encryption is carried out. The encrypted data is a randomly generated seed that is used to generate a pseudorandom sequence. The sequence generated is used to indicate the parameters used to perform the integrity check, and at what intervals in the file. The encrypted seed is sent to the receiver who is also able to generate the same sequence and can extract the voice parameters in order to check them. The use of a seed to generate the numbers has a number of advantages. It needs no encrypted file header and is relatively small. This makes it difficult to attack using standard cryptanalysis techniques. The proposed approach is illustrated in Fig.3 at the end of the paper.

Although we are of the opinion that transmission without compression is impractical due to the high bandwidth required, we feel that as part of a communications system any insertion and extraction of a fingerprint for integrity checking should take place in the compressed file, directly prior to, and after transmission. This removes the need for the fingerprint to be robust enough to survive lossy compression.

An authentication system needs to embed a fingerprint that is both unique and that allows the integrity of the data to be checked. It needs to be able to work without user intervention and in near real time. That is, with a low enough latency to be undetectable.

We propose a voice integrity checking method based on emotional cue features. Unlike existing techniques we consider not a semantic based approach but a psychophysiological approach that could possibly be combined with a fingerprint. The approach taken is developed to indicate any modification that might have taken place. Previous work by Wu and Kuo (2001) has suggested that changes to the brightness (they might mean tone in this context) or amplification should not damage the watermark. We suggest that the listener carries out any required modification to the sound data in real time and that stored files should remain faithful to the received data. This would prevent accidental modification of emotional cues that might later be used for analysis. To ensure that this does not occur, we would suggest that fingerprinting files for integrity checking remain fragile.

If a seeded pseudorandom selection of emotion characteristics is used to generate encrypted data in either packet headers or in least significant bits, this can be extracted simply at the other end. Not every feature needs to be sent simultaneously and this makes it difficult to hack such a system, as it would identify parameters in the voice with high risk, when it comes to the use of stress analysis.

Examples of possible parameters for the integrity checking include:

- Pitch variation
- Mean energy in given frequency band
- Jitter levels
- Jitter frequency
- Distance between formants
- Ratios between formant amplitudes

5.0 Conclusion

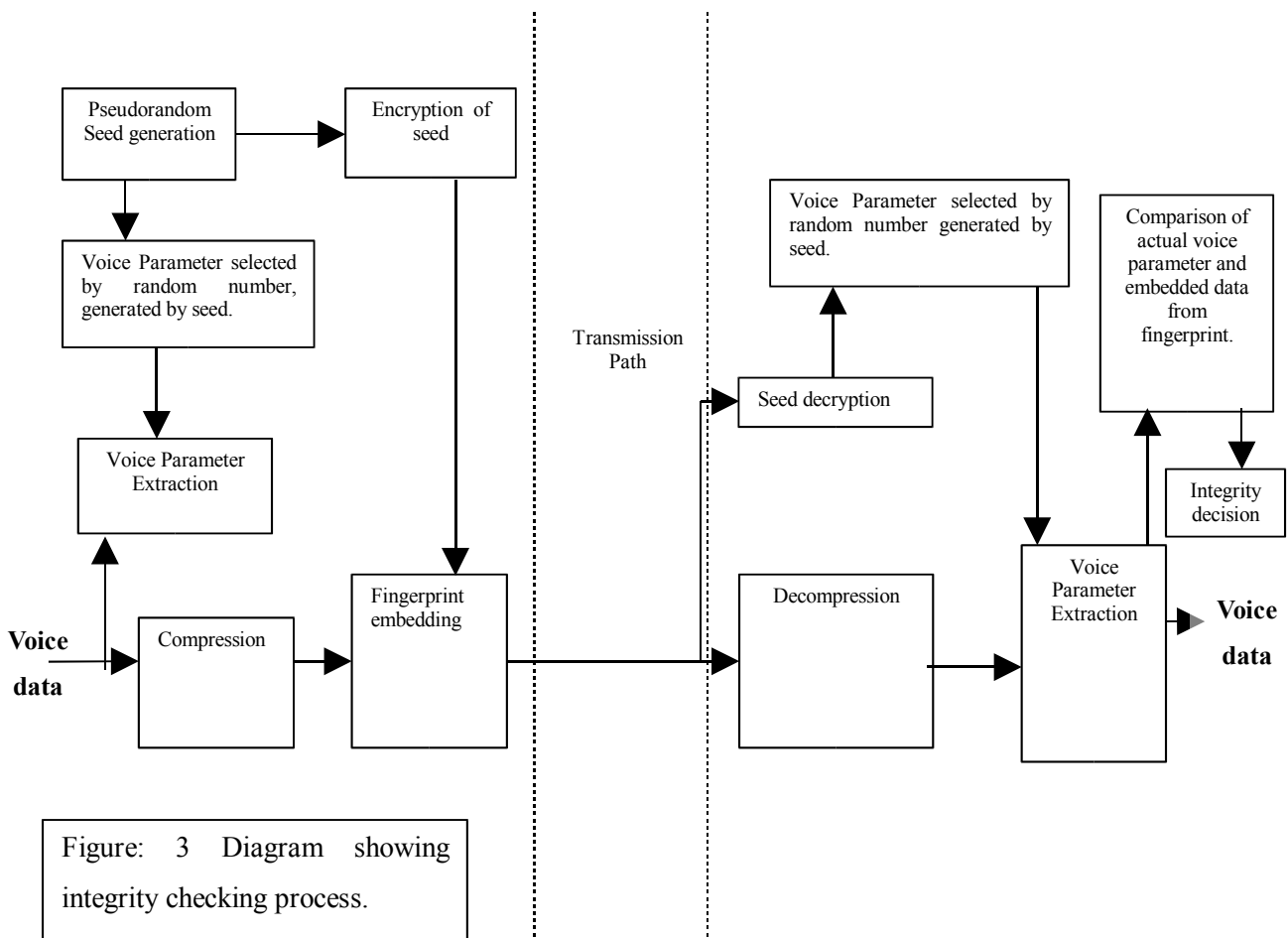
The current available techniques for voice authentication and message integrity checking are suitable for detecting many of the changes in the voice but may not be strong enough to detect the very subtle changes that can be used for modifying voice emotion cues. The proposed approach allows a combination of fragile watermarking and a secure emotion cue content-based system to provide security for the listener. Security for the speaker would require additional fingerprinting to show that the listener had not retrospectively modified the file and the fingerprint.

As VSA devices become more sophisticated and with Moore's law (Moore, 1965) still holding true and providing the potential for very low latency complex processing, the requirements of integrity and authentication checking will change. Our proposal anticipates some of these changes and provides a novel and robust approach to voice integrity checking.

The advantages of our approach include better risk analysis in deciding which aspects of the voice signal should be ignored in any stress analysis and a robust integrity check that requires less processing overheads than current strategies that use encryption.

5.1 Future Work

Work needs to be carried out to test the proposed method in real communication scenarios. We also need to ascertain the extent to which real time processing of media makes integrity checking necessary. This may be achievable by developing the tools that might be used to modify the voice and investigating their impact on VSA devices. The work in this area will hopefully lead to more rigorous methods for Voice Stress Analysis and more reliable techniques for preventing such analysis when required.



6.0 References

- Bencina, Ross** (2004), Audiomulch ver 0.9b15 2/2/2004 available from www.audiomulch.com
- Fabien A.P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn.** Attacks on Copyright Marking Systems
- David Aucsmith, Ed., *Second workshop on information hiding*, in vol. 1525 of *Lecture Notes in Computer Science*, Portland, Oregon, USA, 14{17 April, 1998, pp. 218{238. ISBN 3-540-65386-4.
- Janniro, M. J., & Cestaro, V. L. (1996).** Effectiveness of Detection of Deception Examinations Using the Computer Voice Stress Analyzer. (DoDPI95-P-0016). Fort McClellan, AL : Department of Defense Polygraph Institute. DTIC AD Number A318986.
- JitGen4** (2004) available upon request from C. Reynolds, carl9@mdx.ac.uk.
- Johnstone, T. & Scherer, K. R. (1999).** The effects of emotions on voice quality. Unpublished research report. *Geneva Studies in Emotion and Communication*, 13(3).
- McClintock, Jeff, (2002)** SynthEdit rev 0.9507 available at <http://www.synthedit.com>
- Meyerhoff, Saviolakis, Koenig & Yurick (2001)** DoDPI Research Division Staff, Physiological and biochemical measures of stress compared to voice stress analysis using the computer voice stress analyzer (CVSA). (Report No. DoDPI01-R- 0001). Fort Jackson, SC: Department of Defense Polygraph Institute, & Washington, DC: Walter Reed Army Institute of Research.
- Moore Gordon E. 1965** *Cramming more components onto integrated circuits*. Electronics, Volume 38, Number 8, April 19, 1965
- Ruiz, Legros, & Guell. (1990).** Voice analysis to predict the psychological or physical state of a speaker. *Aviation, Space, and Environmental Medicine*, (1990).
- Scherer, Johnstone & Bänziger (1998)** Scherer, K. R., Johnstone, T., & Bänziger, T. (1998, October). Automatic verification of emotionally stressed speakers: The problem of individual differences. Paper presented at *SPECOM'98, International Workshop on speech and Computers*, St. Petersburg, Russia. *Geneva Studies in Emotion and Communication*, 12(1).

Scherer (1995) Scherer, K. R., "Expression of emotion in voice and music", *J. Voice*, 9(3), 1995, 235-248.

Schuller Lang & Rigoll (2002) Automatic Emotion Recognition by the Speech Signal Björn Schuller, Manfred Lang, Gerhard Rigoll, Institute for Human-Machine-Communication, Technical University of Munich 80290 Munich, Germany, presented at *SCI 2002*. CD-ROM conference proceedings.

Smith (1977) Smith, G. A. (1977) Voice analysis for the measurement of anxiety. *British Journal of Medical Psychology*, 50, 367-373.

Steinberg. VST is a trademark of Steinberg Soft- und Hardware GmbH"

Wu. C hung-Ping and Kuo. C.-C. Jay *Speech Content Integrity Verification Integrated with ITU G.723.1 SpeechCoding*. IEEE International Conference on Information Technology: Coding and Computing (ITCC2001), pp. 680-684, (Las Vegas, Nevada), April 2001

Neural Networks for Real-time Pathfinding in Computer Games

Ross Graham, Hugh McCabe & Stephen Sheridan

School of Informatics and Engineering, Institute of Technology at Blanchardstown, Dublin 15

Contact email: Ross.Graham@itb.ie

Abstract

One of the greatest challenges in the design of realistic Artificial Intelligence (AI) in computer games is agent movement. Pathfinding strategies are usually employed as the core of any AI movement system. The two main components for basic real-time pathfinding are (i) travelling towards a specified goal and (ii) avoiding dynamic and static obstacles that may litter the path to this goal. The focus of this paper is how machine learning techniques, such as Artificial Neural Networks and Genetic Algorithms, can be used to enhance an AI agent's ability to handle pathfinding in real-time by giving them an awareness of the virtual world around them through sensors. Thus the agents should be able to react in real-time to any dynamic changes that may occur in the game.

Keywords: Neural Network, Genetic Algorithm, Pathfinding.

1. Introduction

Agent movement is one of the greatest challenges in the design of realistic Artificial Intelligence (AI) in computer games. This challenge is compounded in modern games that are becoming more dynamic in nature as a result of middleware engines such as Renderware [Renderware] and Havok [Havok]. These middleware companies allow game developers to spend more time developing interesting dynamic games because they remove the need to build custom physics engines for each game. But these new dynamic games create a strain on existing pathfinding strategies as these strategies rely on a static representation of the virtual world of the game. Therefore, since the games environment can change in real-time, the pathfinding strategy also has to occur in real-time. The two components for basic real-time pathfinding are (i) heading in the direction of a goal and (ii) avoiding any static and dynamic obstacles that may litter the path to that goal in real-time.

This paper will highlight the need for real-time pathfinding and how effectively a neural network can learn this initially at a basic level. It will then discuss the steps taken to determine the possibility of using neural networks for basic real-time pathfinding and the pros and cons of the results.

1.1 The Need for Real-Time Pathfinding

Traditionally in computer games pathfinding is done on a static scaled down representation of the virtual world that the game presents. This works fine if there is little or no change to the virtual world throughout the course of the game. This was the case in most games up until now as the sophistication of the game's real-time physics engine was limited mainly due to the time required to develop it. However games are now being built using middleware for key components of the game, including the physics engine [Havok]. Middleware is software written by an external source that has hooks that allow it to be integrated into a game developer's code.

Therefore game developers can spend much more time creating more immersible games with real-time dynamic scenes. This sounds exciting however it is being impeded by traditional pathfinding AI that operates off a *static* representation of the games virtual environment. This limits the amount of dynamic objects that can be added to games, as the pathfinding strategy will have to be fine-tuned to handle them thus adding more time to the development of the game.

To allow an AI agent to effectively navigate a dynamic world it would have to be given real-time awareness of the environment surrounding it. To achieve this with traditional methods would require running the pathfinding algorithm at every move, this would be computationally expensive, especially for the limited memory available to the games consoles of today. Therefore the AI agent will have to be given some kind of sensors that can obtain information about its surroundings. This is not difficult to implement, however a key problem arises in making the agent decipher useful information from these sensors and react accordingly in real-time without putting too much of a strain on the computers resources. Artificial neural networks are a well known AI technique that provides a potential solution to this problem.

1.2 Neural Networks for Real-Time Pathfinding

An artificial neural network is an information-processing system that has certain performance characteristics in common with biological neural networks [Fausett94]. Each input into a neuron has a weight value associated with it; these weights are the primary means of storage for neural networks. Learning takes place by changing the value of the weights. The key point is that a trained Neural Network (NN) has the ability to generalise on situations that it has never encountered [Champanand 04]. This is a particularly useful feature that should help considerably with dynamic scenes.

There are many different types of neural networks but the one particularly suited to real-time games is the Feed Forward Neural Network (FFNN) due to the speed it can process data [Fausett 94]. Therefore the extent to which a FFNN could be trained to decipher the information presented to it, from sensors attached to an AI agent in a dynamic virtual world, was investigated. These sensors will receive real-time data from the physics engine therefore giving the agent a sense of awareness about its surrounding environment.

The next stage was to come up with some system that will train the weights of the FFNN. If simple rules are to be learned then it is possible to compile a set of inputs and their expected outputs and train the FFNN through backpropagation. Otherwise reinforcement learning [Champanand 04] will be used to evolve the FFNN's weights through a genetic algorithm

(GA) [Buckland 02]. This is achieved by rewarding AI agents for following various rules that the user specifies at different time intervals. Then the AI agents are ranked according to their respective scores with the top ranking agents putting a mixture of their weights in to a lower ranking agent. This is analogous to nature's survival of the fittest model in the real world that has helped humans evolve to where we are today.

1.3 Evolving the Weights of a Neural Network

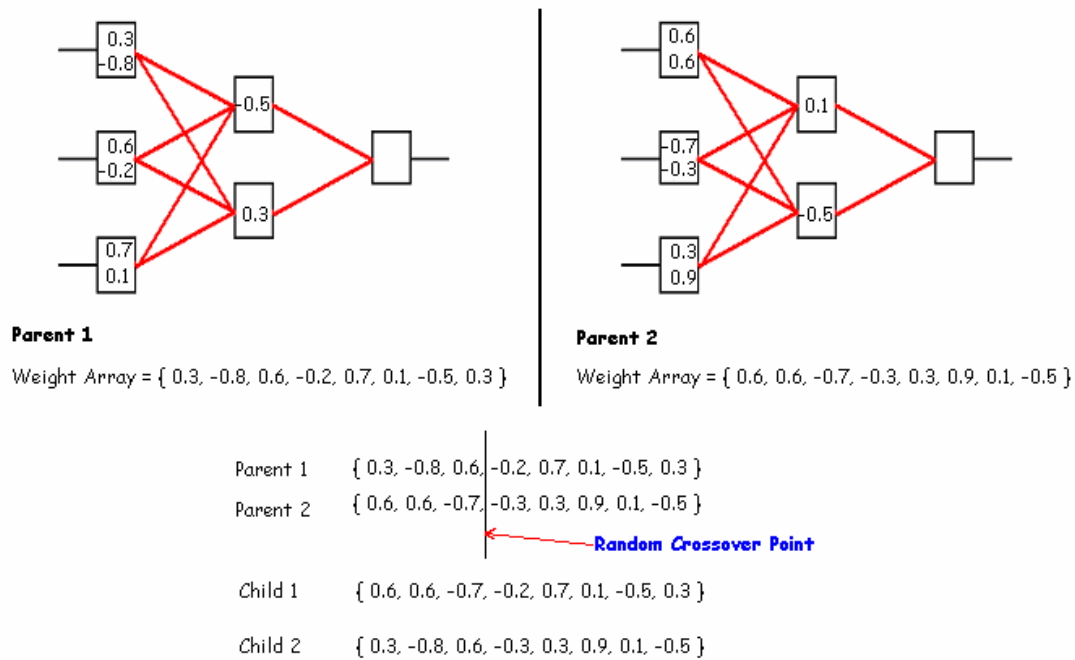


Figure 1.1

The encoding of a neural network which is to be evolved by a genetic algorithm is very straightforward. This is achieved by reading all the weights from its respective layers and storing them in an array. This weight array represents the chromosome of the organism with each individual weight representing a gene. During crossover the arrays for both parents are lined up side by side. Then depending on the crossover method, the genetic algorithm chooses the respective parents weights to be passed on to the offspring as shown in figure 1.1. Training the neural network for basic real-time pathfinding first required it to learn to (i) head in direction of goal and then to (ii) navigate around any obstacles that might litter the path. Therefore the first step will be to see if the NN can learn these two tasks separately and then finally learn both of them combined.

2 Test bed

The test bed for the experiment was a simple 2D environment (grid of square cells) in which the agents can only move either *up*, *down*, *left* or *right* one cell from their present locations. There is a boundary around the perimeter of the grid that has one of the following two properties *solid boundary* and a *wrap-around boundary*. With the *wrap-around boundary* if

the AI agent hits it they will be transported to the opposite side of the grid while with the *solid boundary* the agent is stopped from moving. Objects can also be added to the grid, which will permanently occupy their respective position in the grid thus making it off limits to the agent. The test bed also allows real-time modification to the genetic algorithms parameters so the user can observe the evolution in real-time and change these values due to observations. Thus the test bed offers the NN a simple virtual environment to learn the two different components of basic real-time pathfinding through reinforcement learning, which will be conducted through a Genetic Algorithm (GA). This was done in terms of stages of increasing difficulty that present an AI agent with different situations.

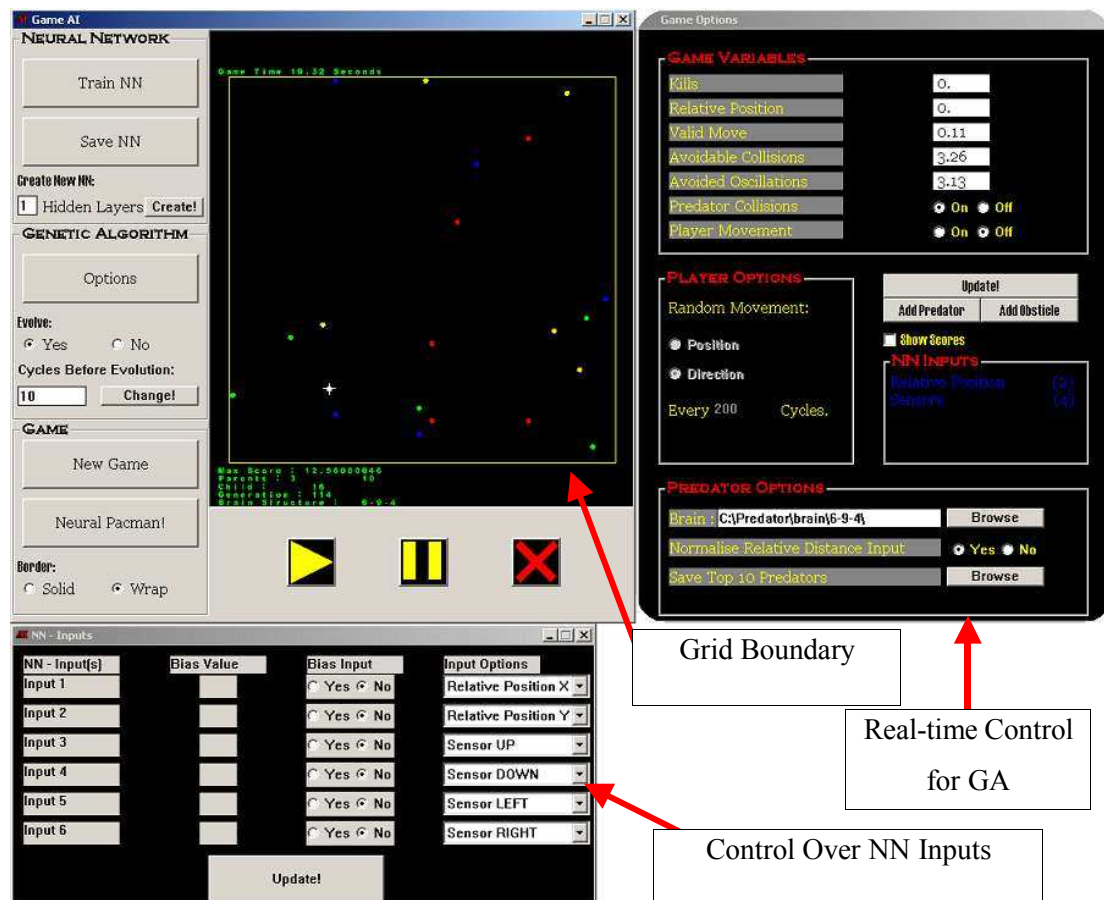


Figure 2.1

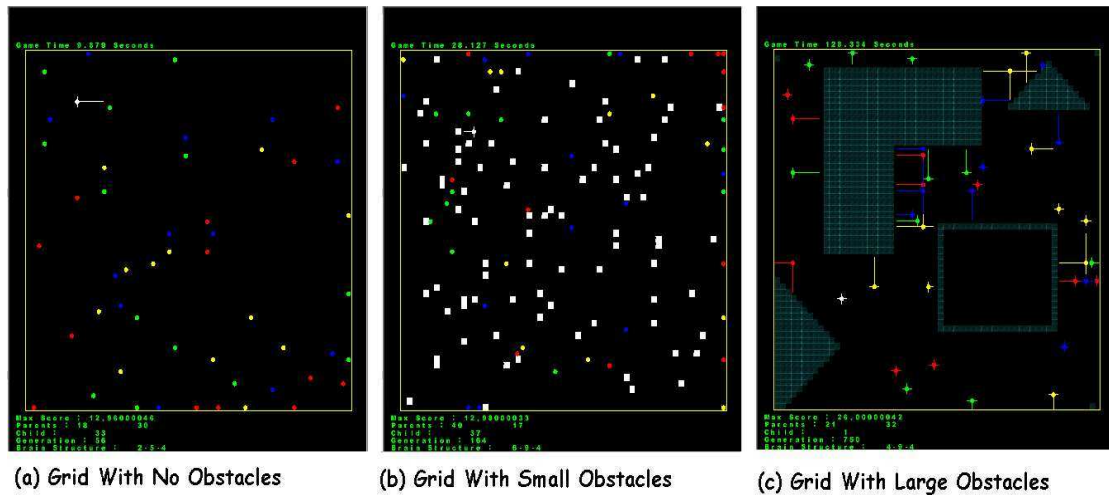


Figure 2.2

2.1 Stage One

This stage comprised of a Predator\Prey pursuit scenario with no obstacles as shown in Figure 2.2(a). The relative position of the prey was used as the input to the NN.

Neural Network Information	
Inputs (2)	Outputs (4)
<i>Relative Position of Prey on X-axis</i>	<i>UP</i>
	<i>DOWN</i>
<i>Relative Position of Prey on Y-axis</i>	<i>LEFT</i>
	<i>RIGHT</i>

The first thing that the NN was tested on was its ability to go towards a goal. The idea here is to have predator that relentlessly pursues its prey around an obstacle free space. Therefore the predator will decide which way to move via a NN that takes the relative position of the prey as its input. The NN had four outputs for each of the possible moves it can make. The output with the strongest signal was selected for the next move.

Num Inputs	Hidden Layer Neurons	Time(Sec)	Generations	Population Size
2	2	230.1	1380.6	20
2	3	141.6	849.6	20
2	4	83.5	501	20
2	5	99.5	597	20
2	6	101.9	611.4	20

Table 2.1

Table 2.1 shows the length of time and the number of generations it took for the predators to evolve to a satisfactory solution to the behaviour required with different numbers of neurons in the hidden layer. The results were compiled by getting the average time for a satisfactory result after running through the simulation ten times for each of the hidden layer configurations. As

highlighted in table 2.1 neural networks with four neurons in their hidden layer on average evolved quickest to the solution.

Since the objective of the NN was very clear and simple it was possible to train it using back propagation and GA [Fausett 94]. This task was learned very quickly through back propagation and GA but GA had the benefit of creating less predictable solutions. This simple Predator/Prey demonstrated that the NN had no trouble learning to head in the direction of a goal, which is the first requirement of real-time pathfinding.

2.2 Stage Two

This stage comprised of a Predator\Prey pursuit scenario with small obstacles as shown in Figure 2.2(b). The inputs to the NN were the Relative position and the contents of the four surrounding cells.

Neural Network Information	
Inputs (6)	Outputs (4)
<i>Relative Position of Prey on X-axis</i>	<i>UP</i>
<i>Relative Position of Prey on Y-axis</i>	<i>DOWN</i>
<i>Cell Above</i>	<i>LEFT</i>
<i>Cell Below</i>	
<i>Cell Left</i>	<i>RIGHT</i>
<i>Cell Right</i>	

This next stage aims to test if the predator can avoid obstacles that litter the path between it and the prey. To achieve this the predators NN was given more inputs so as to inform it about its immediate surrounding environment. To keep it simple the predator was given four sensors that indicated if the next cell in each of the possible directions from its present location was obstructed.

The task was learned quickly through back propagation and GA for obstacles up to two cells in size, however with obstacles larger than two cells the Predator got stuck. This indicated that the predator did not have enough time/information to avoid larger obstacles as it could only sense one cell ahead. Therefore to possibly overcome this problem the predator would need longer-range sensors i.e. sensors that scan greater than one cell in each direction.

2.3 Stage Three

This stage comprised of a Predator/Prey pursuit scenario with large obstacles as shown in Figure 2.2(c). The inputs to the NN were the relative position of the prey and with four directional sensors that can look ahead more than one cell.

Neural Network Information	
Inputs (6)	Outputs (4)
<i>Relative Position of Prey on X-axis</i>	<i>UP</i>
<i>Relative Position of Prey on Y-axis</i>	<i>DOWN</i>
<i>Sensor UP</i>	<i>LEFT</i>
<i>Sensor DOWN</i>	
<i>Sensor LEFT</i>	<i>RIGHT</i>
<i>Sensor RIGHT</i>	

The Predators could not learn the two tasks combined with any real impressive results therefore the search space needs to be reduced. One way of achieving this would be to use a hybrid neural network which is more than one neural network being used to solve the problem. i.e. a NN could work out the direction to travel towards goal and then input this into another NN that checks for obstacles.

2.4 Stage Four

This stage comprised of a Predator with four sensors that can look more than one cell ahead in each direction for obstacle avoidance as shown in Figure 2.2(c).

Neural Network Information	
Inputs (4)	Outputs (4)
<i>Sensor UP</i>	<i>UP</i>
<i>Sensor DOWN</i>	<i>DOWN</i>
<i>Sensor LEFT</i>	<i>LEFT</i>
<i>Sensor RIGHT</i>	<i>RIGHT</i>

Since there are two parts to real-time pathfinding it was decided to investigate whether the NN could learn to steer around large and small obstacles. Thus giving similar results to Renoylds steering algorithms [Reynolds99]. This time the predators were only given long-range sensors as inputs to their NN. The Predators learned to steer away from obstacles large and small. Therefore if a NN can learn the two requirements for basic real-time pathfinding separately it is a reasonable assumption that with a bit more refinement into the training procedure that it can be trained to do both of them.

3 Future Work

Future work will involve investigating the use of hybrid neural networks [Masters93] that will help break up the problem into its two components thus reducing the search space for the full problem. Our intention is then to expand the research into 3D games and benchmark the results against traditional pathfinding strategies such as A* and Dijkstra [Graham04] with regard to speed, realistic movement and ability to handle dynamic environments. The move to 3D should

not be much more of a challenge for the NN to handle as gravity constrains in most games will confine the agent to a 2D plane most of the time. Another extension to this research is to look at a path-planning component that can guide the AI agent by supplying it different goals to seek.

3.1 Conclusion

The two main problems associated with traditional pathfinding are they (i) rely on a static representation of the virtual world and (ii) rigid unrealistic movement may be observed unless fine-tuned in someway. This therefore hinders using the full power on offer from dynamic middleware components. Since a neural network (NN) can learn the two main components required for a basic level of real-time pathfinding separately with some more research into refining the training, it will be possible for a NN to learn the two together. The NN should also be able to generalise on dynamic situations that it did not encounter during training. This would offer the possibility of creating a pathfinding Application Programming Interface (API) based on a neural network that will take inputs from a games physics engine in real-time. Thus create the foundation for real-time pathfinding middleware that would allow the next generation of computer games to immerse players into much more unpredictable and challenging game environments.

References

- [Buckland02] Buckland, Mat., "AI Techniques for Game Programming", Premier Press, 2002
- [Champandard04] Chapandard, Alex J., "AI Game Development", New Riders Publishing, 2004
- [Fausett94] Fausett, Laurene, "Fundamentals of Neural Networks Architectures, Algorithms, and Applications", Prentice-Hall, Inc, 1994.
- [Graham04] Graham, Ross., "Pathfinding in Computer Games", In proceedings of ITB Journal Issue Number 8, 2004, Pages 56-80
- [Havok] www.havok.com
- [Masters93] Masters, Timothy., "Practical Neural Network Recipes in C++", Boston: Academic Press, 1993
- [Renderware] www.renderware.com
- [Reynolds99] Reynolds, C. W., "Steering Behaviors For Autonomous Characters", In proceedings of Game Developers Conference 1999 held in San Jose, California. Miller Freeman Game Group, San Francisco, California. Pages 763-782
- [RusselNorvig95] Russel, Stuart., Norvig, Peter., "Artificial Intelligence A Modern Approach", Prentice-Hall, Inc, 1995

A New Integrated Style to Teaching Engineering Mathematics at Third Level Engineering Courses

Mohamad Saleh¹ B.Sc. M.Eng., Ph.D., CEng, MIEE

Colm McGuinness² B.Sc., Ph.D., CMath, MIMA

¹School of Informatics and Engineering, Institute of Technology, Blanchardstown, Dublin 15

²School of Business and Humanities, Institute of Technology, Blanchardstown, Dublin 15

Contact email: mohamad.saleh@itb.ie

Abstract

Mathematics is the main pillar in the engineering foundation courses and the engineering profession where mathematical manipulation, modelling and simulation are used widely. However, experience in engineering courses has shown that students encounter some difficulties in mathematics, with a certain degree of disinterest and apathy. This is reflected in the mathematical continuous assessments, final exams, laboratory reports for other engineering subjects and in answering engineering numerical question-based mathematical formula.

This paper investigates a new development and the implication of two models of a CBL integrated with course lecture material. This is part of an overall integrated approach, achieved through an embedded Visual Basic mathematics programming into MS Excel. The results of this paper attempt to promote mathematics in engineering courses and provide substantial information to the educators in both mathematics and engineering at third level education.

Keywords: CBL, engineering mathematics, education.

1. Introduction

Recently there have been various attempts to improve the engineering education and teaching engineering mathematics. It has been shown that the engineering educational system at present is falling behind the manufacturing system with regard to quality within its industry (J.Perendergast, M.Saleh et.al , 2001). This suggests that urgent reform of the engineering tertiary educational system is needed, as this system is expected to provide the skilled engineering workforce for today's manufacturing technology (M,saleh, 2003). The general mathematics problem in undergraduate engineering courses was described already in (Howson. A.G, 1995 - James, D.G., 1995). At present, this general decline shows little sign of improving (Sutherland R. and Pozzi S., 1995)

P. Edwards et al.,2003, investigated a project to enhance students understanding of mathematics. This project has resulted in the translation of some *Visual Basic* programmes into Java applets delivered on the Internet Web site, *MathinSite*. A.Howrwitz et al.,2002, described a collaborative project between the Mathematics and Engineering Departments of universities. This collaboration effort involved enhancing the first year calculus course with applied engineering and science projects. The application project involved both teamwork and individual work, and required both programmable calculator and Matlab. Milton Fuller,2000, reviewed developments and trends in engineering mathematics education in Australian universities over the past number of years. He recommended that mathematics

departments in Australia should set up working groups to tackle the emerging needs of engineering education.

A. Carlos et. al.,1996, have developed a sequence of programmes to assist the engineering students to explore in depth several features of the soliton's formation, propagation, and collision. The physical origin of the solitons is the Kerr effect, which relies on a nonlinear dielectric constant that can balance the group dispersion in the optical propagation medium. These numerical routines were implemented for use with MATHEMATICATM and the results give a very clear idea of this interesting and important practical phenomenon.

M. Saleh et.al.,1992, have developed a two dimensional linear and non linear finite element computer programme as an educational aid. This was to help the engineering students to understand the practical implication of the matrix manipulation and numerical analysis in engineering design.

This paper examines the integration between mathematics, computing and applied engineering to deepening the mathematical understanding at third level engineering courses. Graphical visualization and physical application of the mathematics of real engineering problems are used to motivate the engineering student and enhance the mathematical learning process.

2. Engineering Curricula and Mathematics

Engineering is a field of science where students develop their critical, lateral thinking and creative skills to enable them to apply the knowledge they gain effectively in an innovative fashion in real life. This is based on a combination theoretical, tutorial and experimental studies of the relevant subjects in the engineering curricula. The main objective of this teaching method, in engineering, must enable the students to construct engineering concepts. Therefore, the acquisition of engineering skills, during engineering courses, depends widely on the understanding and the flexibility of the teaching methodology. This is based on extracting the rational conclusion from a combination of theoretical, tutorial and relevant practical studies. The role of mathematics in the engineering profession is mainly to idealize a real-world situation through models that aid in the analysis of the engineering problems with minimum time and cost. Models in engineering can be theoretical or empirical models. The application of mathematical techniques and numerical methods-based computing are used widely in engineering to simulate and solve engineering models. On the other hand, mathematics is

delivered in engineering courses as a series of sequential theoretical and tutorial based quiz sessions similar to the way it is delivered in any school of mathematics. Hence, mathematics is often seen among engineering students as a collection of abstract objects without direct practical reference. This discourages engineering students, as it is structured in direct contrast to other engineering subjects in the curricula, and lacks stimulation. However, the modern mathematical education of engineering students must recognise that (Norbert Grünwald and Dieter Schott, 2000):

- Understanding and handling the basics and foundations of mathematics is more important than knowing a lot the details.
- The use and interpretation of the results comes prior to being able to prove results.
- Controlling computations, calculations and estimations is more significant than being able to do computations by oneself.

3. Practical Considerations

The proposed teaching method is mainly to sustain the academic balance between engineering and mathematics. The current approach to date has been to write software that is capable of creating problems which typically involve random numbers, and detailing associated solutions. The engineering mathematic modules should be developed in collaboration with engineers and mathematicians, containing a set of mathematical problems/mini projects underlying engineering applications. Students should be introduced to the basic concepts of these problems. A mathematic laboratory should be available to students at different levels during the engineering course of study. These laboratories should be equipped with the relevant tools/materials to deliver the mathematical theoretical concepts or to support the relevant acquisition of experimental data. Students should be encouraged to solve the assigned problems/mini projects through group work. Based on the experimental and theoretical investigation of each mini project, each student has a different related problem to solve through the CBL system. This is intended to reduce the possibility of direct “cogging” by students, and enhances the sense of personal “ownership” of their particular problem. Students who get stuck can simply look at the answer provided, and unlike a book or a problem database that has limited numbers of questions, the system can then create a brand new problem and try again. They can also experiment by creating different types of problems (sometimes specifiable within the application) and looking at the solutions to see how the solution varies with the problem details.

4. Foundation Mathematics

In some respects this is a difficult subject area to tackle. Key areas of weakness for students tend to be in algebraic manipulation and transposition of formulae. To date we have not tackled this with CBL but some of our ideas are presented below. It is the inescapable symbolic nature of this area of mathematics that makes it more difficult. Many other specific areas, such as those treated later, do not need to focus on the symbolic nature of the solution techniques and so are more tractable in CBL terms using applications programming in Visual Basic or Microsoft Excel. To tackle algebraic manipulation and transposition of formulae would either involve:

- A fairly restricted problem set, or
- A fairly (very?) advanced software application, or
- Software written using an existing symbolic manipulation packages, such as MuPad, Macsyma, Maple, etc.

One idea here, whichever environment is chosen, is to present the student with varying levels of difficulty with each level introducing additional aspects of manipulation. This approach allows the student to build up confidence at each level, and in essence gives them more control to pace the learning to suit themselves.

So level 1 might contain problems that only require basic laws of algebra: Associativity, commutativity and distribution of multiplication over addition.

A sample question might be:

$$\text{Simplify } 2(x + y) - 2x$$

Then at level 2 introduce more complex manipulations and equations, for example:

$$\text{Solve } 2(x + y) = 2x - y \text{ for } x.$$

Then at level 3 introduce quotients of terms. Then at level 4 include powers, logs, e^x . Next at level 5 introduce trigonometric functions. Finally at level 6 combine them all.

One key difficulty, with a symbolic manipulation programme, is to implement the mathematical equivalence of the expected answer and the answer provided by the student. Most computer algebra systems (CAS) can test for the equivalence of two expressions. Writing a computer programme to determine whether both the student's answer is equivalent to the expected answer, and determining that they did in fact use the expected rules is probably a significant and difficult challenge. A way around this is to provide a "symbolic editor" which allows the user to manipulate the expressions using only an allowed and restricted set of operations. This

would help determine that they took the correct steps. On the downside it removes the manipulations from the “real world” somewhat, and is a complex application to create.

4. Case Studies

5.1. Harmonic and phase analysis:

This is to demonstrate the understanding of the mathematical concept of the harmonic and phase analysis in AC circuits and dynamic systems. Figure 1 shows the MS Excel application of the Sine function and Figure 2 shows the Cosine function. Simply, the value for the amplitude, frequency and phase can be altered in the target values, and accordingly the current value of the Sin/Cos is graphically changed. This current value can be visualised from the screen and compared with a standard waveform every time the target value is altered.

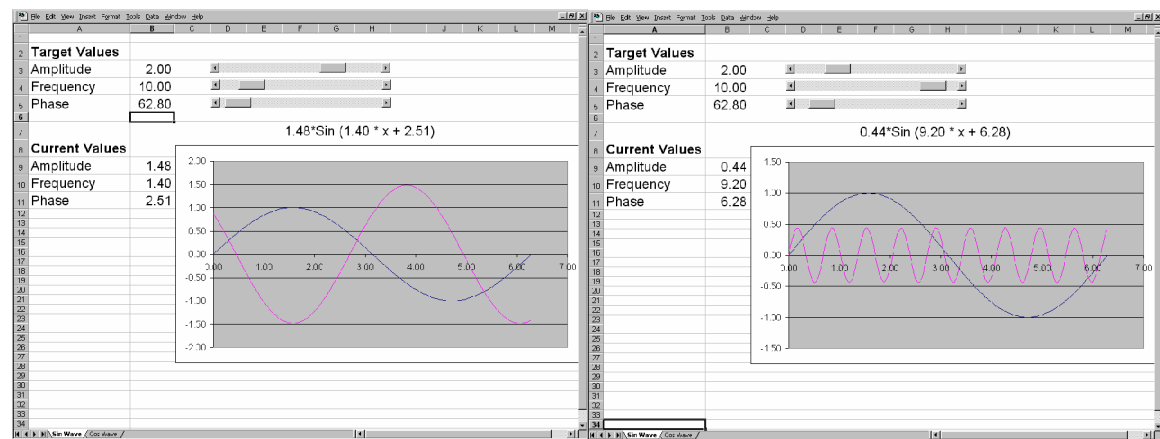


Figure 1: CBL Sine function

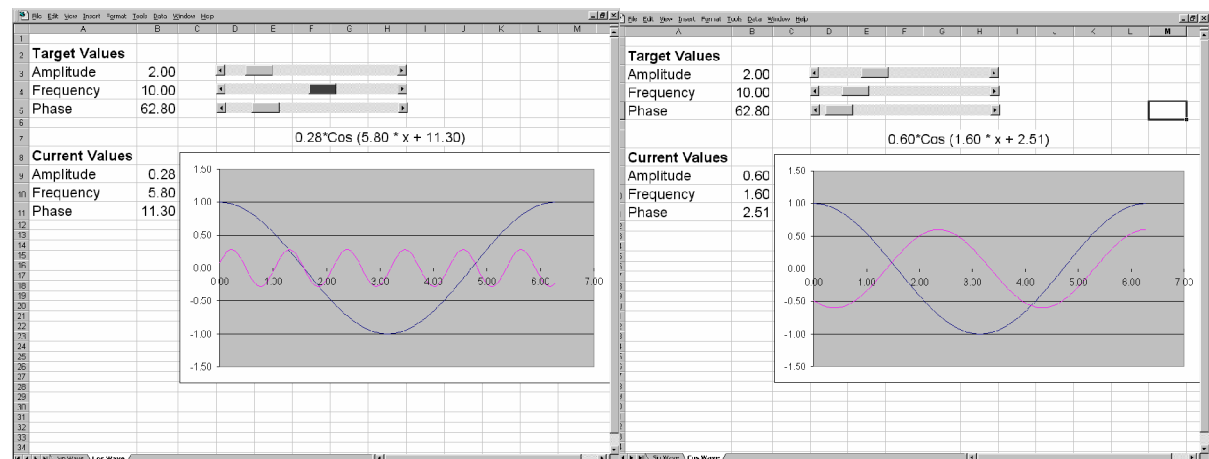


Figure 2: CBL Cosine function

5.2. Analysis of Variance (ANOVA):

This is a sample application which could be used to give students practice in single factor ANOVA. It is not intended as a standalone learning module. It is intended to be used as an

integrated course tool (i.e quality assurance, measurement system, etc.) to give students practice with real data, and a tool that they can interact with, and carry out personal investigations into ANOVA and the F distribution. Students should have prerequisite knowledge of:

- random sampling;
- frequency distributions, frequency polygons;
- probability distributions - Normal & F;
- hypothesis testing, and ANOVA.

Learning Objectives:

- To give students practical and interactive experimental experience with ANOVA from multiple samples;
- To show students how the F-distribution is generated from MSA/MSW where H_0 is true/false. The CBL of this analysis is shown in Figure 3.

5. Discussion

Remembering mathematical ideas and techniques can be a problem. Students will depend heavily upon semantic memory, for example, under examination conditions where careful and accurate recall is required. Semantic memory relates to the meanings of things, such as words or concepts. In contrast, episodic memory is the remembering of events, things that happened, places visited, based upon sequences of events. Calling on episodic memory requires event reconstruction (...and then I did this...), which, unfortunately on the one hand, can be susceptible to an individual's perceptions and interpretations but on the other hand, can complement semantic memory (Peter Edwards et.al,2003). Since the proposed integrated learning method depends heavily on the interactive visualisations. This visualisation approach can aid memory stimulation and retention. Also, the graphical visualisation of this approach, through a sequence of events, give students a deeper insight into the mathematical problem they are investigating as every time a graph is accomplished on the screen it will add some meaning to the students. This encourages students to sketch the visual graphs from the screen. The integration and the application of this approach in real engineering problems help the students to construct the engineering mathematical concept through the reflective abstraction. The teamwork in this approach allows students to abstract the engineering principles, and hence to explore the engineering profession. Consequently, studying and solving problems as a group would enable students to gain very good communicative skills.

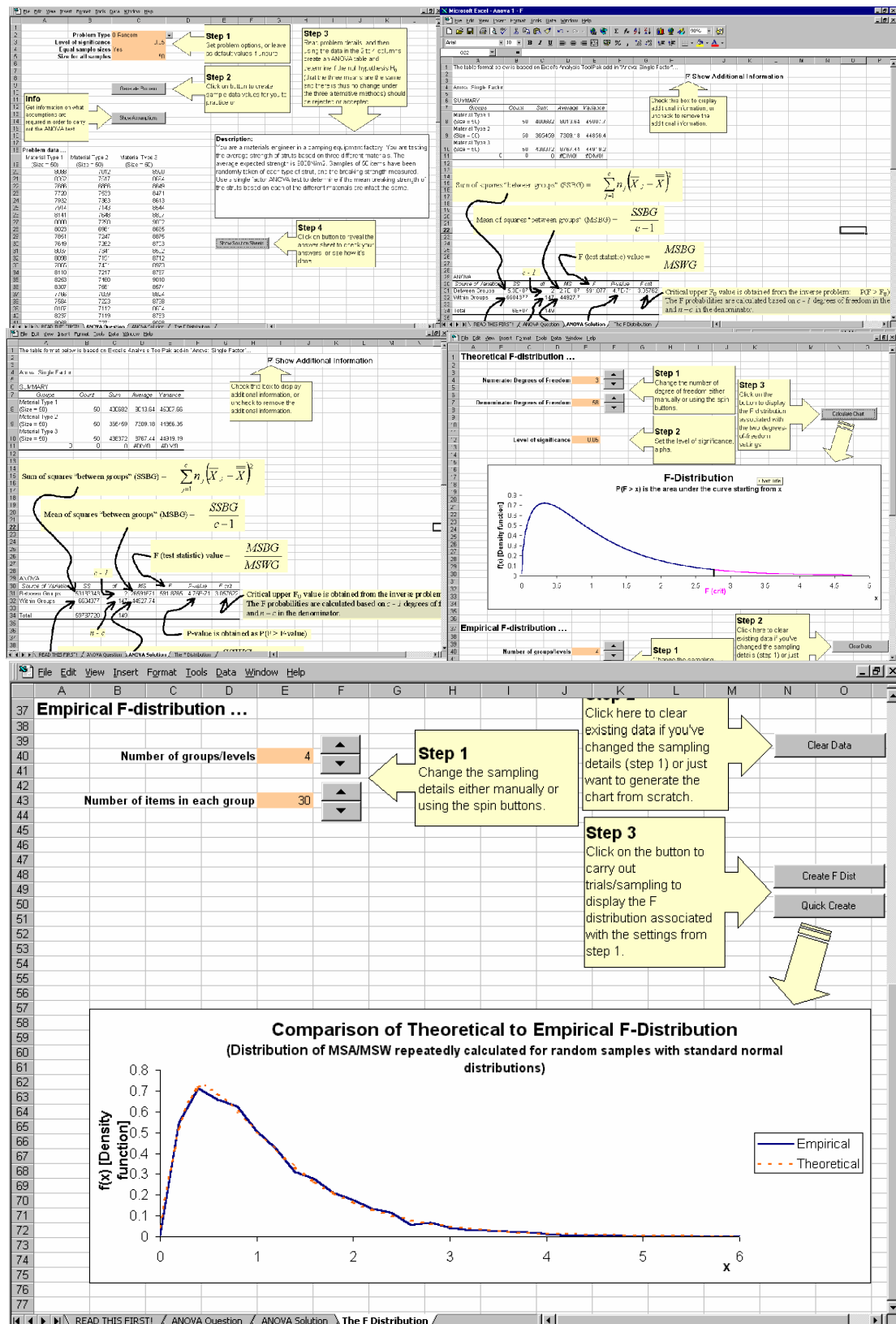


Figure 3: CBL of ANOVA

This is to ensure an effective environment for the exchange and development of ideas (Belbin, R.M., 1996). Also, this environment develops the lateral and critical thinking which allows the

students to use what is so called the How? and Why? effectively in engineering mathematical modules. The nature of the challenge in this method would stimulate the students and give them a sense of ownership and thus, motive and commit them to further study. The role of the lecturer in this method is to provide a strategic practice in which he or she should be working with the students, suggesting different methods and solutions.

6. Conclusion

Overall, the proposed integrated CBL approach has been discussed. It has been argued that this graphical visualisation approach is an effective method to learning engineering mathematics at third level engineering courses. However, it should be noted that the CBL approach does not replace the lecture entirely. It is a tool to help the student's own learning process and students need to formulate the problems in mathematical language and thinking.

7. References

- J. Prendergast, Mohamad Saleh, K. Lynch and J. Murphy (2001).** A revolutionary style at third level education towards TQM. *J. of Materials proc. Technology*, 118, pp. 362-367.
- Mohamad Saleh (2003).** Reverse Engineering: An effective approach to studying Mechatronics at undergraduate tertiary education, *IEEE ICIT*, pp 824-829.
- Howson, A.G. ,Chairman, (1995).** Tackling the Mathematics Problem. London: London Mathematical Society, the Institute of Mathematics and its Applications and the Royal Statistical Society.
- James, D.G. ,Chairman), (1995).** Mathematics Matters in Engineering. Working Group Report, the Institute of Mathematics and its Applications.
- Sutherland, R. and Pozzi, S. (1995).** The Changing Mathematical Background of Undergraduate Engineers - a Review of the Issues. London: the Engineering Council.
- Peter Edwards and Paul M. Edwards (2003)** .Tackling the Mathematics Problem with *MathinSite*", *Global J. of Engng Educ. Vol.7. No.1, pp95-102.*
- A.Howrwtz and A. Edbrahimpour, (2002).** Engineering Applications in Differential and integrated Calculus *Int. J. of Eng. Educ. Vol 18, No 1, pp 78-88*
- Milton Fuller, (April 2000).**Current Issues in Engineering Mathematics in Australia. *Proc. Of the Mathematical Educ. Of Engineers II, Australia*, pp 47-52.
- Carlos A. Caballero P. and Rui Fragassi Souza , 1996).** Observation of Solitons with MATHEMATICA. *IEEE Trans. On Educ., VOL. 39, No 1, pp 46-49*
- Mohamad Saleh et.al, (1992).** Development of a finite element computer program for two dimensional large elastic plastic deformation". *Int. Cong. Num. Methods in App. Sci., Concepcion, Chile.*
- Norbert Grünwald and Dieter Schott , (2000).** World Mathematics 2000: Challenges in Revolutionising Mathematical Teaching in Engineering #education under Complicated Societal Conditions. *Global J. of Engng Educ. Vol. 4. No.3, pp 235-243*
- Belbin, R.M.,(1996).** Management teams: why they succeed or fail. Butterworth-Heinemann; ISBN: 0750626763. 1996

Design Study of a Heavy Duty Load Cell Using Finite Element Analysis: A practical Introduction to Mechatronic Design Process

Mohamad Saleh B.Sc. M.Eng., Ph.D., CEng, MIEE

School of Informatics and Engineering, Institute of Technology, Blanchardstown, Dublin 15

Contact email: mohamad.saleh@itb.ie

Abstract

Mechatronics design process is a series of analytical brain storming operations from specification to implementation. The mechatronic products are widely available in the market for various use and applications. This is expected to develop further in the future with great competitiveness. In order to succeed in the market, mechatronic products need to satisfy the market expectations with regard to quality, fitness for purpose, customer's appeal and cost efficiency. Therefore, the design analysis techniques play a significant part in the market success of these products.

Finite Element Analysis is probably the most commonly used numerical technique for mechatronic product design. This technique has been developed over a number of years and is now available in a wide variety of packages, mostly for mainframe and workstation platforms but also for personal computer systems. Over the past few years, the range of users has broadened dramatically with a rapid need for training and education of this technique.

This paper reviews the design philosophy and modelling strategy of a mechatronic product using finite element techniques. This takes into account the design study of a 140 tonne load cell for measuring a load mechanism in metal working applications.

Keywords: FEA, transducer, Mechatronic, product design.

Introduction

Mechatronics is the synergic combination of precision mechanical engineering, electronic control and systems thinking in the design of products and processes (Reitdijk, J. A). It is therefore important that mechatronics, from the very beginning, is considered not as a separate engineering discipline but instead as an integrating systems level approach to design and operation of a wide range of complex engineering productions and processes (D.A Bradley, 1997). Thus, mechatronics, at undergraduate level, should be delivered as a compatible method to its integrated and inter-disciplinary nature. M. Saleh, 2003, discussed an effective approach in mechatronics design and its implementation at third level mechatronics courses.

Finite element method has become the most reliably effective numerical technique used to calculate the most complicated engineering problems compared to other computer aided engineering. The popularity of this technique is due to its wide applicability in both static or dynamic linear and non-linear problems. In fact, continuous fast advancement in computer hardware and software and the availability of computer power have caused a dramatic increase in the use of finite element techniques in engineering design. This has led to the belief that the development of educational and training courses in finite element techniques is crucially

important to extend the popularity of this technique for obtaining quality and reliable results of the modern product design.

The material presented in this paper is a revised and expanded version of a lead paper presented by M. Saleh, 1999, at AMPT'99 [4]. This paper discusses a practical approach for using finite element techniques for reliable design and analysis. It takes into account the systematic scenario and strategy of FEA modelling which is enhanced through a practical example. It is believed that the implementation of this approach in training and educational programmes will benefit the designers, as it improves the productivity, reduces analytical costs and hence more accountable quality design can be seen.

1. Finite element technique

Finite element is a method of mathematically modelling a components/structure for stress analysis. It requires large quantities of data which are manipulated by matrix techniques using the appropriate computer software programmes. It is not intended to give a detailed account of finite element techniques, as this is well documented in numerous text books. However, it is felt that a general overview will enhance the understanding of the subsequent work reported in this paper. Figure 1 shows the procedure of finite element technique to solve the design parameters of a cantilever loaded with a point load at its free end. This problem can be solved using two or three dimensional F.E modelling analysis. In three dimensional analysis the entire structure of this cantilever will be modelled using three dimensional elements. Often in stress analysis a full three dimensional treatment of a problem can be avoided by assuming that the structure to be adequately represented if plain stress or plane strain conditions are adapted. It is assumed that plane strain conditions are justified for the cantilever in Figure 1.a when the ratio b/h exceed 5. This means that there is no deformation in Z direction. However plane stress conditions are justified when b/h is less than 5.

Following on from this, the structure of the cantilever will be divided into a mesh of elements, as shown in Figure 1.b with the appropriate material properties, boundary conditions at the restraints and external loading types. In this mesh, each node of the individual element has two degrees of freedom (Figure. 1 .c) making a total of eight nodal displacements for the element. Similarly, each element in this mesh has eight nodal forces. The concept of element stiffness matrix $[K^e]$ is introduced by relating the nodal forces to the nodal displacements of the element Figure 1 .d. The overall stiffness matrix $[K]$ in Figure i.e can be established by assembling the element stiffness matrices of the overall structures. From the equation in Figure 1.e, the unknown displacements for each individual node in the mesh can be found. Thus, the stress and

strain for each element in the structure can be calculated using the formulas shown in Figure 1.f.

This sequence of operations can be performed using the interactive finite element commercial software programmes. These programmes, in general, feature the same philosophy with different sizes of analysis and slight variations of menu driven functions and operations. The accuracy and reliability of results depend upon the finite element modelling strategy in which choosing the appropriate element, material properties, boundary conditions and applied load must be defined precisely. In practice, the finite element solution of an engineering problem may be varied among different users using the same FE programme. Also, errors in finite element solutions can be produce by the F.E programmes. Consequently, agencies such as NAFMES (National Agency for Finite Methods and Standards) in England play a most important role with their regular survey report of tests carried out on leading finite element software programmes worldwide. These are reported in NAFEMS periodical “Benchmarks”.

The unfamiliarity with finite element methods and relevant software programmes produces erroneous answers to problems using finite element modeling. Recently, this has led to increased demand for training and education in finite element modelling techniques. This is to enable the trainees to carry out a reliable finite element analysis, under critical working conditions, of their engineering design.

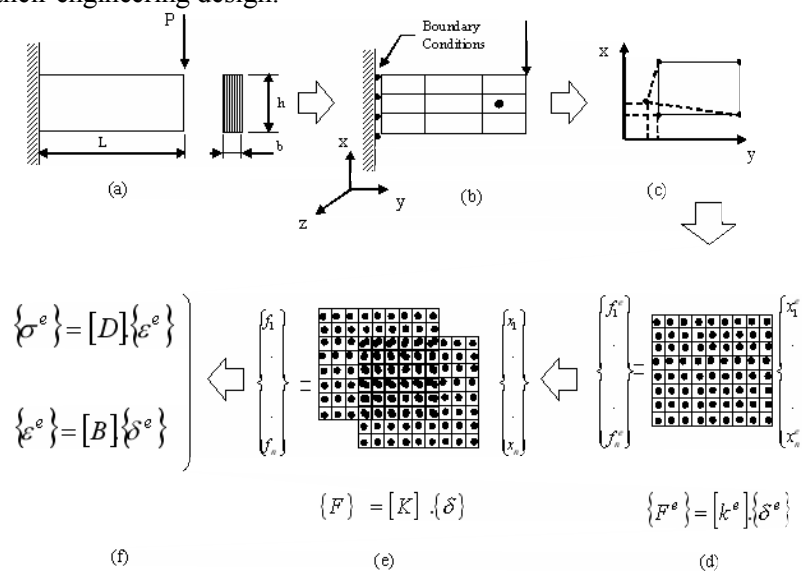


Figure 1: Procedure of finite element analysis

2. Modelling strategy

In order to enable the structure to maintain its characteristic safety in critical situations, the finite element model should be approached with a sound basic knowledge of the finite element

technique and its applications to engineering problems. Therefore, the users must have an appreciation for discretisation principles, element performance, modelling procedures and material constitutive behaviour, as the reliability of any finite element solution depends fundamentally on the consideration of these factors and the background to the field of application.

Figure. 2 shows the sequence of a systematic procedure for using finite element technique as a reliable computer-aided design and analysis in engineering. This procedure takes into account the theoretical model, experimental model and the design goal of the problem being investigated. The theoretical model considers the finite element and the conventional analytical model of the problem. The conventional analytical model is based on the traditional calculations of stress and strain in engineering design. This is to determine the design specifications, to develop the experimental model and consequently to validate the refinement of the finite element model. The finite element model is a scenario of model modifications with different types of elements, boundary conditions, material properties, mesh refinement; and theoretical design modifications. The design goal is the permissible design parameters/specifications taken from international literature or standard systems. The experimental model involves the physical development of the design in question, including the appropriate measurement system. These are to make comparative studies and to validate the finite element solutions.

Having introduced the first finite element model, comparison takes place between this model, experimental model and the conventional analytical model. This is to establish a referenced correlation between these models. The negative decision at the end of this comparison leads to more modifications/refinements of the finite element model. The positive decision means that the finite element model is valid and the theoretical modifications of the design can be made. This involves a series of theoretical changes of the design and relevant finite element modelling. The objective of this is to optimise the design functions with respect to the design goal. In this regard, the negative comparison suggests that more theoretical modifications of the design should be made. However, the positive comparison leads to applying the theoretical modifications on the actual experimental model. Following on from this, comparison will be carried out to validate the final design solution. Consequently, a new design and analysis will be introduced with practical conclusion.

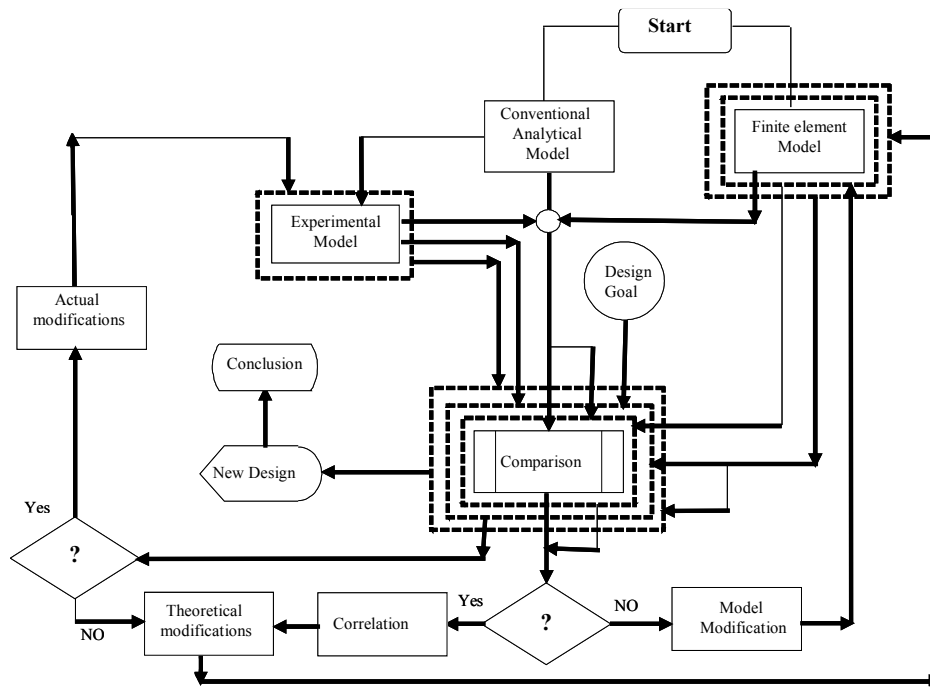


Figure 2: Modeling strategy

3. Design philosophy

In order to enhance the understanding of the modelling strategy presented in this paper, the development of the F.E model of a load cell of 140 tones, shown in Figure. 3, is described. The principle of four bridge single strain gauges is used in the design to translate the physical applied load to an electrical signal. This electrical signal can be sent to a data acquisition system based computer to further evaluation and accurate read out of the applied load. The philosophy of the design was to optimise the structural design functions of this cell so that it can perform safely under the critical loading conditions. The non-linearity portion of the output signal was considered and consequently a clamping system with the relevant accessories was implemented in the design to overcome the non-linear loading during the calibration process.

4. Finite element model

The finite element model took advantage of the axisymmetrical geometry of the load cell to reduce the time and the cost of the finite element analysis. LUSAS 10.1 FE software package was used as a computer aided design and analysis. Accordingly, an axisymmetric finite element model of the load cell was introduced as shown in Figure4.a. This employed axisymmetric elements with 2d.f7node and static faced load was applied externally. The mating elements of the load cell were modelled using the non-linear boundary conditions and the model was restrained as follows:

- (i) Nodes on centre line were restrained in X direction.
- (ii) Nodes on B1-B2 were restrained Y direction.

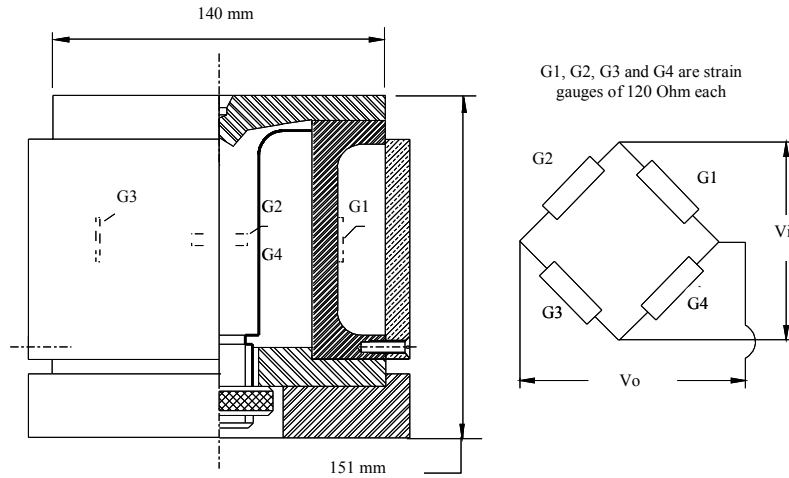


Figure 3: Load cell

Material properties were considered for tool steel (AISID2). The design goal was counted as the permissible stress for this steel. The model underwent a scenario of theoretical modifications to optimise the dimensions; to give as constant stress as possible across the thickness (t) at half of the height (L) and to minimise the concentrated stresses at radius r and r_3 . The final results of the stress distribution and deformation mode are shown in Figure.4.b and Figure4.c respectively.

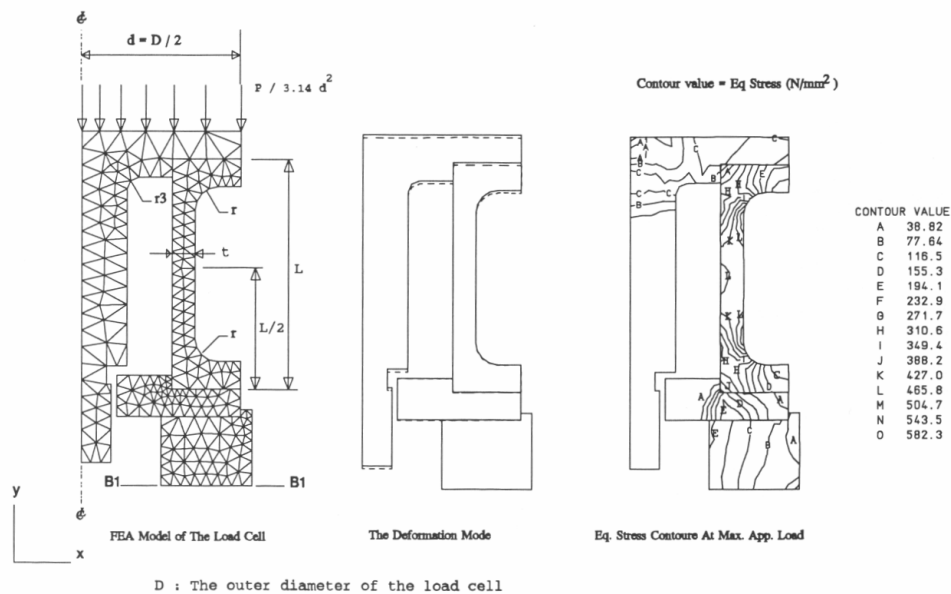


Figure 4: FEA Model and results

5. Calibration procedure

This procedure is to establish the relationship between the external applied load and the output of the load cell, and thus the sensitivity of the load cell. The calibration procedure was carried out on a 250 KN Instron machine as shown in in Figure 5. A 100 KN was considered as pre-load to overcome the no-linearity encountered during the process of calibration. Thus the sensitivity of the load cell is found as follows:

$$\text{Sensitivity} = 1.95 \mu\text{V}/\text{KN}/V_i \quad (1)$$

This mean that the sensitivity of the load cell is $1.9 \mu\text{V}$ for each KN per excitation voltages. Assuming a 6 VDC excitation voltage V_i with an amplification 200 and 100KN external applied load, then the output voltage of the load cell is:

$$V_o = 1.95 \mu\text{V} \times 100 \times 6 \times 200 \quad (2)$$

$$V_o = 0.234 \text{ Volts}$$

From equation 1 and 2, the applied load is KN is :

$$\text{Load (KN)} = V_o(\text{Volts}) / 1.95 \times 10^{-6} \times 200 \times 6 \quad (3)$$

The calibration curve of the load cell is shown is Figure 6

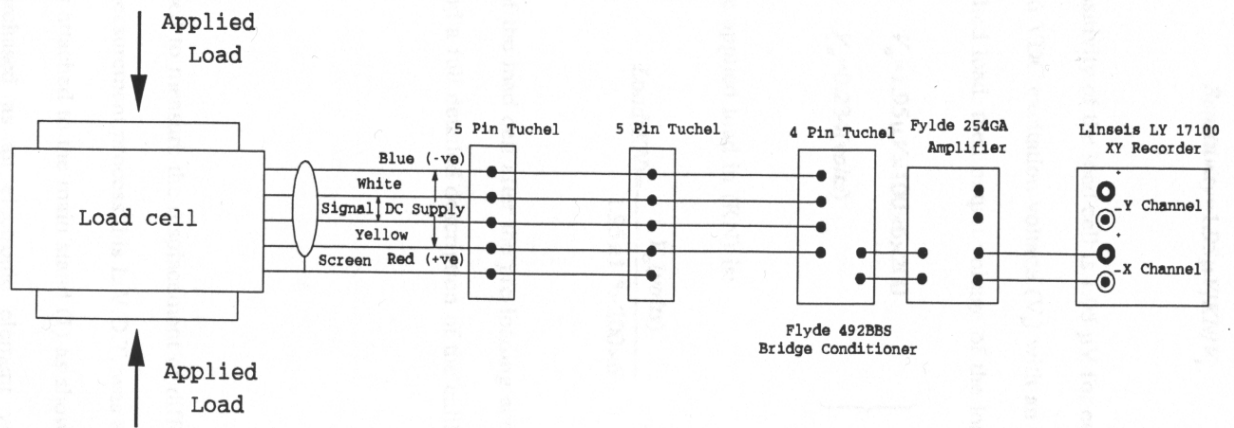


Figure 5: Calibration procedure

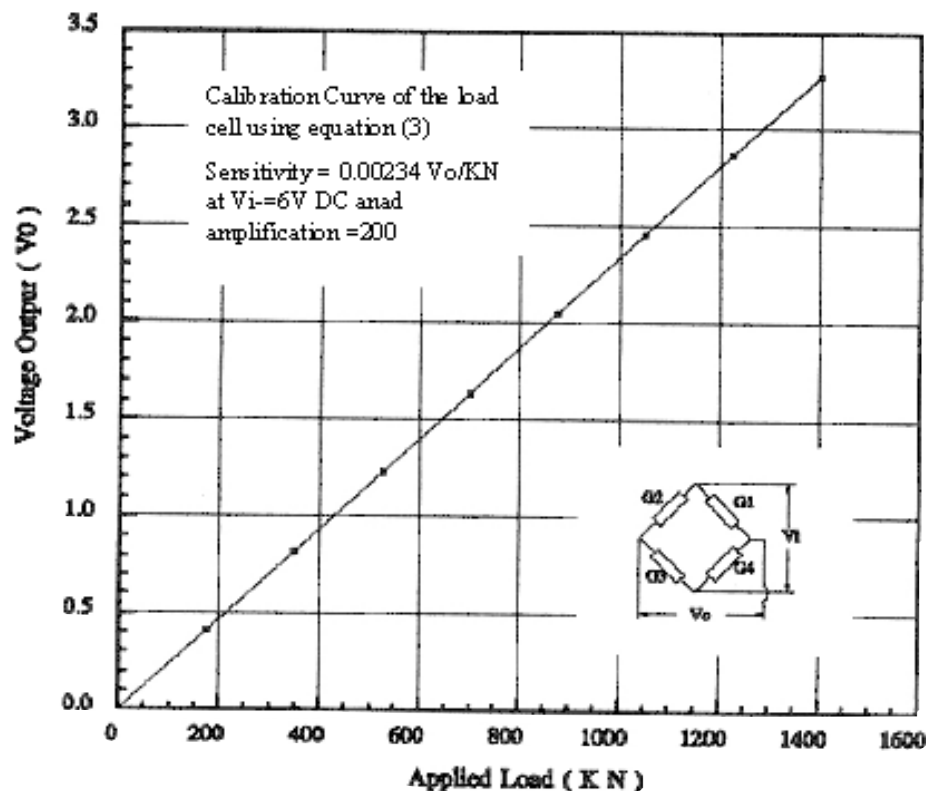


Figure 6: Calibration curve

6. Conclusion

As it can be seen a practical approach for using finite element for engineering design and analysis has been described. This was supported with a FEA modelling strategy of a functional 140 tonnes load cell as an example of mechatronic products.

It is believed that the material presented in this paper could be of great benefit for the designers who use finite element techniques as a design tool. In addition, this can be implemented and adapted in training and educational courses in third level or industrial establishments to broaden the effectiveness and the popularity of FEA among the engineering community.

7. References

- Reitdijk, J. A, (1996). Ten propositions on mechatronics. *Mechatronic syst. Eng.* 1, (1), pp. 9-10.
- D.A Bradley (1997). The what, why and how of mechatronics" *J. of Eng. Sci. Education*, 6,2, pp.81-88.
- Mohamad Saleh, (2003). Reverse Engineering: An effective approach to studying Mechatronics at undergraduate tertiary education. *IEEE ICIT 2003*, pp 824-829.
- Mohamad Saleh, (1999) A practical Use of Finite Element Techniques for Rapid design and analysis" *Proc. Int. Conf. AMPT'99*, Dublin, DCU, Ireland, pp 1607-1613

Measurement of the Frequency Response of Clinical Gas Analysers

**Kabita Shakya¹, Catherine Deegan¹,
Fran Hegarty², Charles Markham³**

¹School of Informatics and Engineering, Institute of Technology Blanchardstown.

²Department of Medical Physics and BioEngineering, St James's Hospital, Dublin 2.

³Department of Computer Science, National University of Ireland, Maynooth.

Abstract

A technique for the time and frequency response measurement of clinical CO₂ analysers has been established. The time and frequency response of several analyser systems has been determined. This paper presents the case for the routine measurement of the dynamic performance of such systems in the context of their application in high-frequency ventilation schemes. The importance of these measurements has been demonstrated in the comparison of older and newer systems in the sense that older systems demonstrate significant deterioration in performance. In the context of the use of capnographs in life-support systems, it is essential to measure and monitor the dynamic performance of such systems to ensure the appropriate monitoring of ventilation schemes. All of the units so far analysed cannot be used for high-frequency and/or paediatric ventilation schemes of over 80 breaths per minute.

Keywords: Mechanical ventilation, respirator, ET_{CO₂} capnogram, , capnography, capnometer , BPM, ventilation rate , frequency response , dynamic response, rise time, sensor response time, transit time, total system response time .

Introduction

In health care units, patients in acute illness and surgery are supported by mechanical ventilation. Mechanical ventilation is the process by which the patient's respiratory function is artificially supported using a respirator. The patient is monitored for several ventilatory parameters during this period, as the adequacy of ventilation must be ensured. End tidal CO₂ (EtCO₂) is one of them. EtCO₂ is the partial pressure or maximal concentration of carbon dioxide (CO₂) at the end of an exhaled breath, which is expressed as a percentage of CO₂ or mmHg. Measurement and monitoring of this parameter assists with appropriate ventilation of the patient as under normal respiration, a person will inhale 0% and exhale up to 5% CO₂ in any one breath.

Capnography is the method that traces out expired CO₂ versus time graphically and measures EtCO₂. The measuring device is called a capnometer (if it displays numerical value only) and capnograph (if it displays graphically as well). The waveform displayed by the capnograph is called a capnogram. A sketch of a typical capnogram is shown in Figure 6(a), with an actual capnograph trace shown in Figure 6 (b).

Determination of EtCO₂ concentration is important as this value of CO₂ (in mmHg or percentage) is accepted to be the equal to the CO₂ content of the blood and is an indicator of how efficiently CO₂ is cleared from the blood supply .

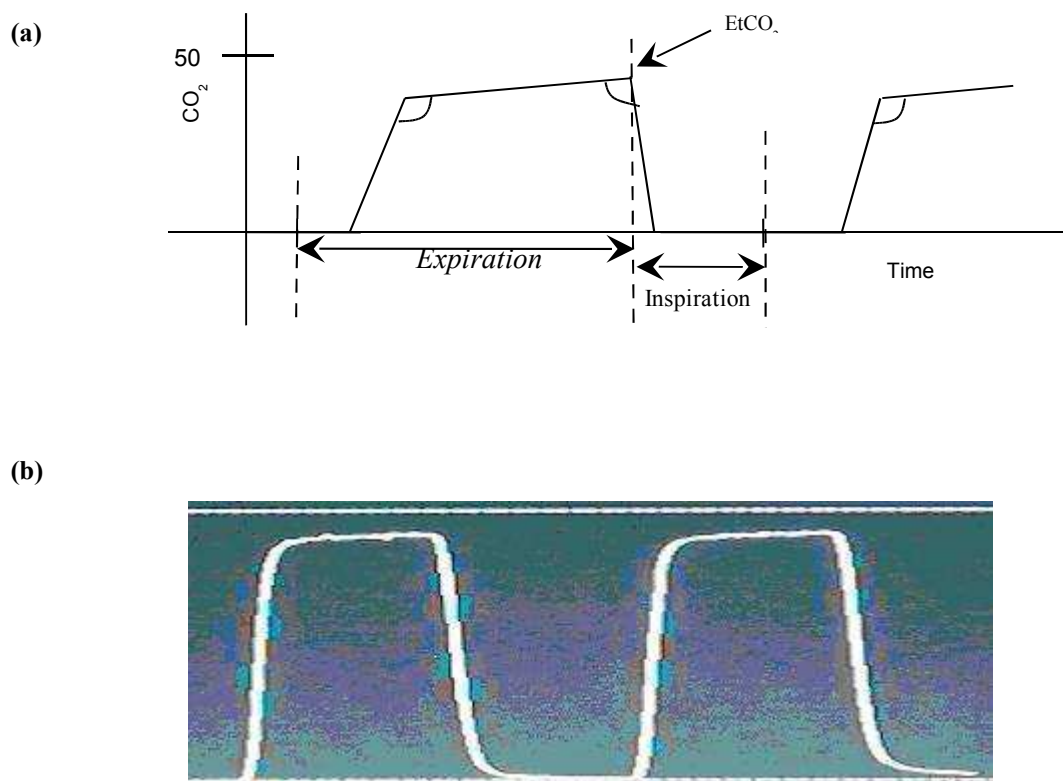


Figure 6 (a) : Normal capnogram of a healthy human. (b) Capnograph Trace

There are three basic CO_2 gas-sampling technologies for EtCO_2 analysis. They are sidestream, mainstream and microstream technologies.

Sidestream technology is the most common. In this technology, the gas sample is aspirated from the ventilator circuit and the analysis occurs away from the ventilator circuit. A pump and lengthy tubing is used to aspirate the sample from the exhaled breath.

In mainstream technology, the CO_2 sensor is positioned directly at the patient's airway. The response time is thus faster and mixing of CO_2 with fresh gas is prevented to a greater extent. However, in this case, the sampling device is bulky and difficult to keep in place so this technology is not usually the first choice for general use.

Microstream technology is comparatively a new technology. It employs a unique, laser-based technology called molecular correlation spectroscopy (MCS^{TM}) as the infrared emission source. The Microstream® emitter radiates a focused beam of infrared energy characterized by the

narrow region (0.15 μm wide) of the spectrum precisely matching the absorption spectrum of CO_2 .

1.1 The End Tidal CO_2 Analyser

The end tidal CO_2 analyser detects the CO_2 concentration in the patients' expired air. The most popular technique for this is infrared technology. Infrared technology uses absorbance spectroscopy in which the loss of electromagnetic energy is measured after the energy interacts the sample under study.

CO_2 selectively absorbs infrared light energy at a wavelength of 4.26 μm . Since the amount of light absorbed is proportional to the concentration of the absorbing molecules, the concentration of CO_2 in the exhaled breath is determined by passing that wavelength of infrared light through the sample and comparing the amount of energy absorbed with the amount absorbed by a sample that contains no CO_2 . The result is expressed either in terms of mmHg or as a percentage of CO_2 ($^3\text{PCO}_2/^4\text{P}_{\text{atm}}$). Analyser using infrared technology is called Infrared spectrograph and is more compact and less expensive than other technologies in use.

EtCO_2 analysers are used in ventilation units, some of them combined with ECG (electrocardiography) or pulse oximetry units and sometimes also as a separate unit (in handheld capnographs).

1.2 Ventilation rate

Ventilation rate is also called respiration /breathing rate. It is given as Breaths per Minute (BPM). The reciprocal of ventilation rate is Ventilation Frequency. A typical individual at rest takes about 12-18 BPM (0.2Hz-0.33Hz), but this rate can triple during hard work (0.3Hz – 0.99Hz). Generally, 15 BPM (0.25Hz.) to 30 BPM (0.5Hz) is considered to be normal in an adult. However, an infant has a higher ventilatory rate (up to 120 BPM i.e. 2Hz) and smaller tidal volume, as a low volume of CO_2 is produced. Tidal volume is the volume of air inspired and expired with each normal breath.

The ventilation frequency is said to be 'high frequency' when it is greater than 2 Hz (120 BPM) and it is said to be 'low frequency' if it is below 0.25 Hz (15 BPM).

³ Partial pressure of CO_2

⁴ Standard (Normal) atmospheric pressure (equivalent to 760mmHg)

1.3 Calibration of (CO₂) Gas Analysers

The most straightforward type of calibration for gas analysers is a 'static' calibration technique in which a known value is input to the system under calibration and the system output is recorded. The static calibration method (also called two-point method) is the current technique for calibrating EtCO₂ analysers. This method features 'zero' calibration to ambient air and 'span' calibration to a self-contained 5% (i.e. 38 mmHg) CO₂ calibration gas. Though the static method is widely accepted as a valid technique for such gas analysers, it is insufficient for describing the dynamic behavior of the analyser, which is the real time event. Dynamic calibration determines the relationship between an input of known dynamic behavior and the (time-varying) measurement system output. When a time dependent variable (in this case, CO₂ concentration) is to be measured, a dynamic calibration should be performed.

The need to measure the time response of these devices has been established by several authors -, however, to date, no techniques for or measurements of the frequency response of these analysers have been established.

Frequency and Dynamic Response Measurement of an ETCO₂ Analyser

In determining the frequency response of the EtCO₂ analyser, the maximum CO₂ concentration is considered as the amplitude value while the oscillation frequency of the CO₂ signal (the simulated ventilation frequency) is considered the frequency.

Figure 7 shows a sketch of a segment of a typical waveform obtained from the EtCO₂ analyser of a sidestream capnograph. The response time t_{response} is the total of 5t_1 and 6t_2 where t_1 (transit time) is the time taken for the sample gas to move from the point of sampling to the point of measurement while t_2 (sensor response time) is time taken for the waveform to reach 70% of its final value ($t_2 = t_{70}$) from 10% of its final value. t_2 can also be taken as the time taken for the waveform to reach 90% of the final value ($t_2 = t_{90}$) from 10% of its final value. t_{70} is generally used instead of t_{90} because the 70% point is on a steeper part of the response curve and therefore less dependent on noise. For all practical purposes t_{90} is twice the value of t_{70} . The value t_{70} can also be taken equal to the time constant (τ , 0% to 63% of the final value). In this paper t_{70} is taken as 0 to 70% of its final value.

The t_1 value generally accounts for about 89% or more of the t_{response} . Generally, t_1 and t_2 are not defined in manufacturers' specifications but these factors should be defined and specified as a

⁵ transit time also called transport delay (TD)

⁶ Sensor response time also called rise time (RT), defined as t_{70} (0 to 70%) or t_{70} (10 to 70%) or t_{90} (0 to 90%) or t_{90} (10 to 90%) - or sometimes also as time constant (τ , 0 to 63%). This paper defines RT as t_{70} (0 to 70%)

long t_1 can prolong t_2 which can result in the underestimation of EtCO_2 during high frequency ventilation. Also, a long t_1 may introduce an unacceptable delay in the total response time of the analyser.

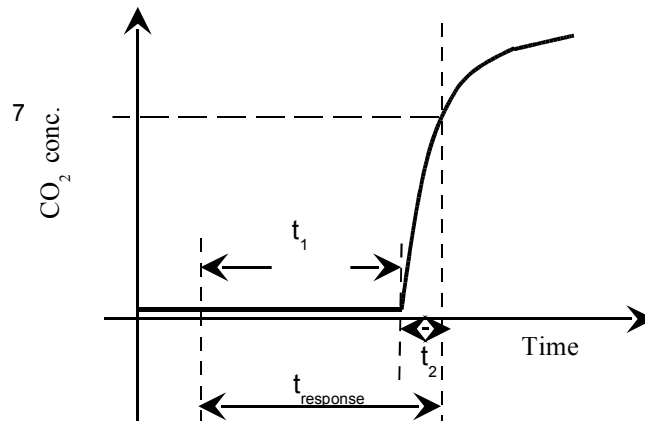


Figure 7: Sketch of a typical EtCO_2 waveform showing the transit time (t_1), sensor response time (t_2) and the total response time (t_{response}) of the sensor system.

Manufacturers generally report only the rise time (t_2) of their capnographs in the instrument manuals. In fact, the most accurate measurement of capnograph response is the total response time (sensor response time plus transit time). This response time should always be known before attempting to use the capnograph system on a non-standard (i.e. high frequency) ventilation circuit to ensure that the capnometer can accurately track the patient's respiration cycle. To date no technique for measuring this parameter has been routinely used in the clinical setting.

In this paper, the t_1 , t_2 and t_{response} values and the frequency response for a small range of capnograph systems has been presented. The techniques and analysis presented in this paper are now in use in a clinical engineering department as part of the commissioning process of new monitoring equipment.

Experimental Technique

An experimental system was developed to emulate the normal breathing patterns of a ventilated patient. The experimental setup consists basically of a CO_2 simulator and a capnograph system. The CO_2 simulator comprises of a 5% CO_2 gas cylinder, air cylinder, pressures regulators, gas valves and the connectors. The CO_2 output is finally delivered to the capnograph. The gas valves are electronically controlled via a computer program in order to simulate the respiration process. Air is used in the system to flush residual CO_2 from the valves. A block diagram of the system is shown in Figure 8. A more detailed report on this system may be found in .

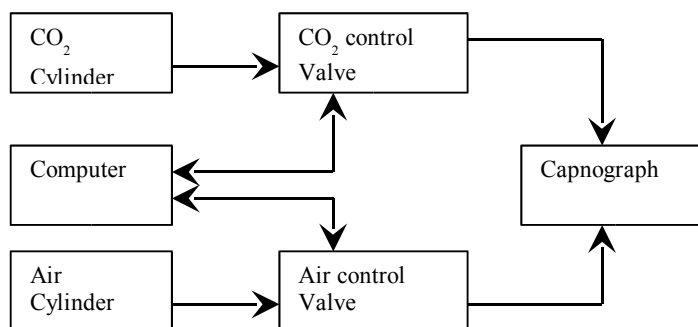


Figure 8 : Block diagram of the experimental system

Data collection from these types of monitors can be problematic as usually there is no means of accessing the raw data electronically. Data acquisition was achieved using a digital video camera to capture the waveforms as they appeared in real-time on the screen. The image data was then processed to extract the waveform data. A sample of data acquired in this manner is shown in Figure 9.

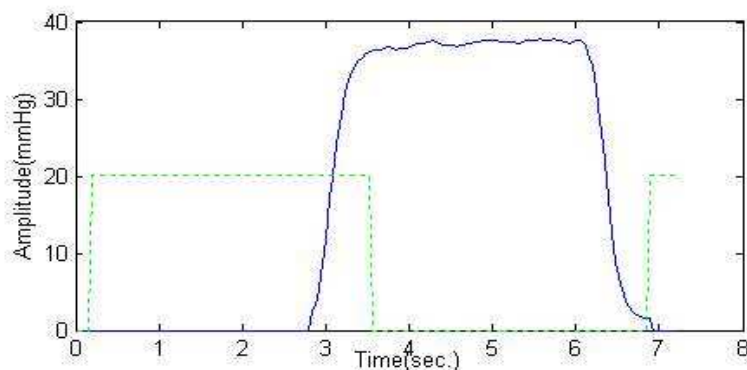


Figure 9: Single waveform (continuous line). The dashed line indicates the on/off time for the CO₂ signal.

Commissioning of Clinical Monitoring Equipment

Before new equipment is used by medical professionals, it must first be submitted to a series of tests to verify its safe and accurate operation. This commissioning procedure is one of the most important procedures carried out by the clinical engineer. The process includes visual inspection, safety tests and functional examination.

To date, the commissioning procedure specifically for CO₂ analysers has included the standard electrical safety tests and a static calibration as well as the fall time measurement of a CO₂ gas signal .

The techniques and measurements developed here will add to the functional verification procedures for the CO₂ analysers of gas monitors. This will allow the determination of the optimum working frequency range of the analyser and also enables sensor response time calculation (for checking against the manufacturers' specification), transit time calculation and hence total response time of the analyser as described in Section .

Results

Frequency Response Measurement of the Drager Capnolog™ System

The experimental system described was tested on a decommissioned capnography system [11]. The frequency response of the Drager Capnolog EtCO₂ analyser is shown in Figure 10. It is clear from this figure that the response of the analyser is compromised at respiration frequencies over 1 Hz.

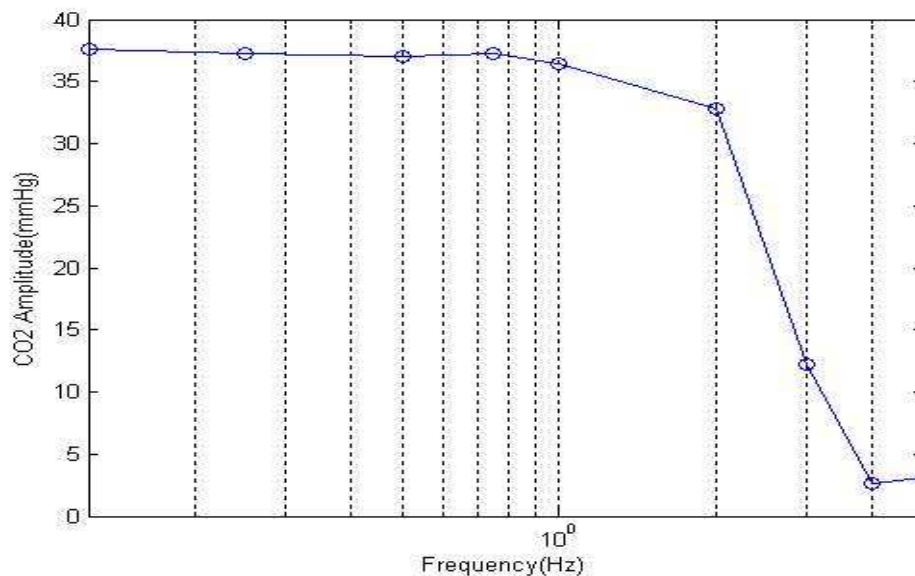


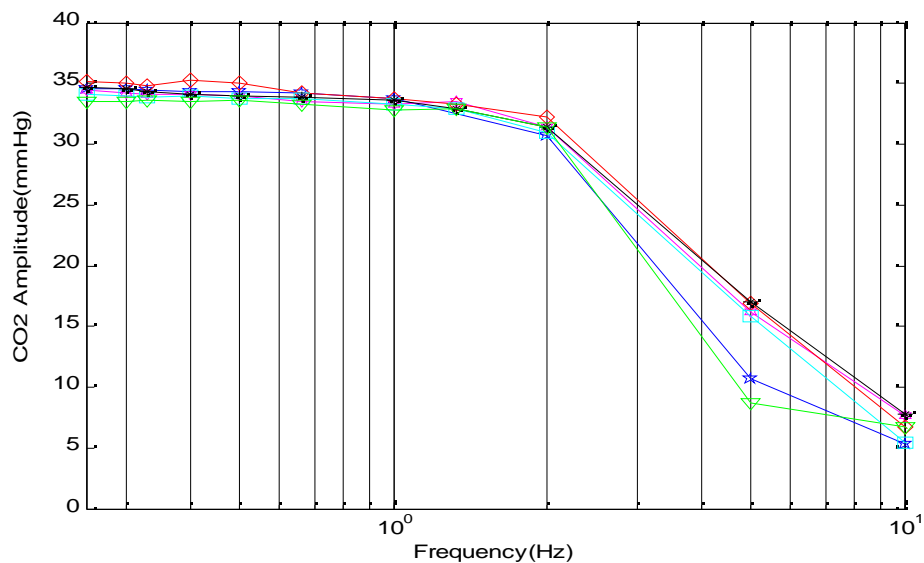
Figure 10: Amplitude versus Frequency curve for the Capnolog EtCO₂ analyser

The transit time (t_t) was found to be 2 sec and the sensor response time (t_s) was found to be 520 ms ($t_{10\%}$ to $t_{90\%}$) giving a total response time of 2528ms (including t_0 to 10%).

Frequency Response of the Datex Ohmeda Capnography System

This project has been undertaken in collaboration with the Medical Physics and BioEngineering Department at St. James's hospital. A unique opportunity to investigate and verify several identical capnographs arose when a new set were acquired and commissioned. Result from all sixteen analysers is shown in Figure 11, and in Table 1. The frequency response for all M-CAiO compact Airway modules is compared in Figure 11. The frequency responses for all Single width Airway Module, M-miniC are compared in . It is clear from Figure 11 and that,

as expected, the frequency responses are well correlated in all analysers. A point to note about this data is that the data shown in indicate a maximum CO₂ signal of 35 mmHg. Of all the analysers tested, minis could not display a full-scale signal. For this reason, it is likely that



these analysers would fail the commissioning process and be returned for manufacturer calibration or it is possible that the suction rate of these analysers is higher than that of the MCIOs and these allow the mixing of air with the 5% CO₂ sample.

Figure 11: Frequency response for Datex Ohmeda (Single width Airway module, M-miniC)

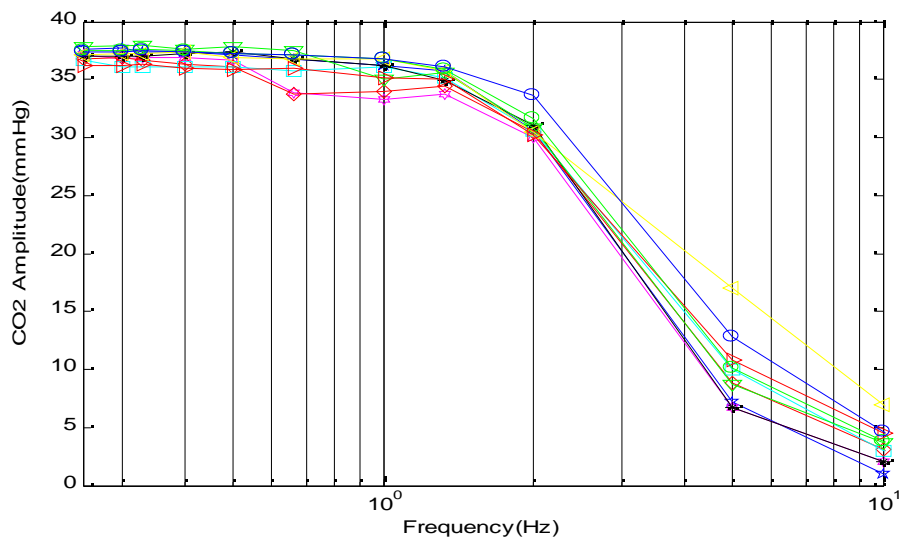


Figure 12: Frequency response of 6 capnographs (Datex Ohmeda, Compact Airway Module, M-CaiO and the single-width Airway Module, M-miniC)

The time responses for these analysers are shown in Table 1. The response time for the ‘mini-C’ module indicates a faster response time to the same signal. This is in accordance with the manufacturer’s specifications for the rise time (t_2). It is clear from this table that measurement of the

t_1 parameter is an important factor when considering a device's time response, reporting the rise time alone is clearly misleading for a sidestream type analyser as a relatively long delay is unavoidable due to the nature of the device itself.

Analyser units	Type	t_1 transit time	t_2 rise time	t_{response}
A	M-CAiO compact Airway module	2.64	0.36	3
B	M-CAiO compact Airway module	2.56	0.36	2.92
C	M-CAiO compact Airway module	2.72	0.36	3.08
D	M-CAiO compact Airway module	2.68	0.36	3.04
E	Single width Airway Module,M-miniC	2.28	0.28	2.56
F	M-CAiO compact Airway module	2.64	0.36	3
G	M-CAiO compact Airway module	2.6	0.4	3
H	M-CAiO compact Airway module	2.64	0.36	3
I	M-CAiO compact Airway module	2.48	0.36	2.84
J	M-CAiO compact Airway module	2.72	0.4	3.12
K	Single width Airway Module,M-miniC	2.28	0.32	2.6
L	Single width Airway Module,M-miniC	2.28	0.32	2.6
M	Single width Airway Module,M-miniC	2.32	0.32	2.64
N	Single width Airway Module,M-miniC	2.32	0.28	2.6
O	Single width Airway Module,M-miniC	2.28	0.32	2.6
P	M-CAiO compact Airway module	2.72	0.32	3.04

Table 1 : Time responses for six different ETCO_2 analysers.

Discussion and Conclusion

For the *Drager Capnolog* system, the working frequency range is from 0.125 Hertz to 1 Hertz i.e. it can operate for Respiratory Rates (RR) from 7.5 BPM to 60 BPM. It may not give satisfactory performance for RR greater than 60 BPM. The manufacturer's quoted respiratory rate in this case is 2-99BPM. Similarly, the sensor response time (t_r) was determined to be 520 ms ($t_{10\%}$ to $t_{90\%}$). This does not agree with the manufacturer's specified response time of 200 ms (10-90% at 200ml/min) . The transit time is not specified in the manufacturer's specification sheet though it is mentioned at one point that the gas takes approximately 1 second to reach the sensor. The measured transit time (t_t) was 2 seconds in this experiment. At this stage it is important to note that the poor performance of this analyser could be due to its age (approximately twenty years) and that it is possible that the analyser did perform as per manufacturer's specifications when first in use. The primary observation here with this system is that it is essential to monitor the performance of such analysers as drift away from the specified performance is to be expected. Frequency and time response data with the six new units were presented in Figure 11, and Table 1 for a 3.6 metre sampling line. For the M-CAiO unit, the manufacturer has quoted total system response time as 2.9 seconds with a 3 metre sampling line, including the sampling delay and⁷rise time. The CO_2 measurement rise time is quoted as <400 ms but it is not specified whether the

⁷ sensor response time is also called rise time (= t_{70} , 0% to 70% of final value)

value (of rise time) corresponds to t_{70} or t_{90} and that how the rise time is defined by them. It is assumed that the value corresponds to t_{70} (0 to 70%). Experiments are performed with 3.6 metre sampling tube so we can conclude that this type of analyser performs within the manufacturer's specification.

For the single-width Airway Module, M-miniC, the rise time is found to be 0.28 seconds, the transit time as 2.28 seconds and the total system response time as 2.56 seconds for a 3.6 metre sampling line. The manufacturer's specified total system response time is 2.4 seconds with a 3 metre sampling line. This time includes the sampling delay and rise time (specified as $<300\text{ms}$). In this case also, the rise time is not specified so it is assumed that the value corresponds to t_{70} (0 to 70%). Hence it may be concluded that this type of capnograph is also operating within the manufacturer's specified time response. However, as reported in section , the maximum signal response of this system was less than 38mmHg and so these particular units could possibly fail the complete commissioning procedure as a full scale signal is desirable for accurate ventilation monitoring. From the frequency response curves it is clear that the frequency response deteriorates over 1.33 Hz for both type of analysers. Hence it can be concluded that both of the units can function properly up to 80 breaths per minute (BPM). The manufacturer specifies 4-80 BPM for the single-width Airway module and 4-60 BPM for the M-miniC. So both of these modules are operating within specification.

In conclusion, a technique for the time and frequency response measurement of clinical CO_2 analysers has been established. The time and frequency response of old and new capnograph systems has been determined. The importance of these measurements has been demonstrated in the comparison of older and newer systems in the sense that the older system has been shown to be operating well outside specification. In the context of the use of capnographs in life-support systems, it is essential to measure and track the dynamic performance of such systems to ensure appropriate ventilation schemes are used in clinical applications.

References

- [1] Brenner JX, Westenskow DR: How the rise time of Carbon Dioxide analysers influences the accuracy of carbon dioxide measurements, *Br. J. Anaesth.* 1988; 61: 628-638
- [2] Breen Peter H, Mazumdar B, Skinner Sean C: Capnometer transport Delay- Measurement and Clinical Implications, *Anesth Analg* 1994; 78: 584 – 6
- [3] Fletcher R, Werner O, Nordstrom L, Jonson B: Sources of error and their correction in the measurement of Carbon dioxide elimination using the Siemens- Elema CO_2 analyzer, *Br. J. Anaesth.* 1983; 55:177-85
- [4] From RP, Scamman FL: Ventilatory Frequency influences accuracy of End tidal CO_2 measurements: analysis of seven capnometers. *Anesth Analg* 1988; 67: 884-6.
- [5] Pascucci RC, Achena JA, Thompson JE, Comparison of a sidestream and mainstream capnometer in infants. *Crit care Med.* 1989; 17(6): 560 – 2.

- [6] Schena J, Thompson J, Crone RK: Mechanical influences on the capnogram. *Crit Care Med* 1984; 12: 672 – 4
- [7] Anderson Cynthia T & Breen Peter H, Carbon dioxide kinetics and capnography during critical care, *Crit care* 2000; 4: 207 – 215.
- [8] Bhavani Shankar : Capnography.com (<http://www.capnography.com>)
- [9] Technology overview: Capnography (Novamatrix Products)
- [10] Frakes Michael et. al. : Measuring end- tidal carbon dioxide - clinical applications and usefulness, *Critical Care Nurse*, 2001; 25 (5)
- [11] Service Manual, Technical Manual and Operators Manual – Drager Narcomed 3
- [12] Owen Markus R, Lewis A Mark : The mechanics of lung tissue under high–frequency ventilation, *society for industrial and applied mathematics, Siam journal on applied mathematics*, 2001, 61 (5), 1731-1761
- [13] Proctor David N, Beck Kenneth C: Delay time adjustments to minimize errors in breath-by-breath measurement of Vo₂ during exercise. *Journal of Applied Physiology*, 1996, 81(6), 2495-2499.
- [14] La Valle TL , Perry AG: Capnography- Assessing end-tidal CO₂ levels, *Dimensions of Critical Care Nursing*. 1995; 14(2): 70–77.
- [15] Shakya K, Deegan C, Hegarty F : Determination of the frequency response of an End tidal CO₂ analyser , *ITB Journal ,Institute of Techonology Blanchardstown , Dublin*, 2003 ,Issue No. 8.
- [16] Garrett Lyons: Some notes on device calibration, *Trinity College, Dublin, Mechanical engg. Dept*,
- [17] Hegarty F ,Lennon B, Private communication
- [18] User's Guide for monitor setup and Reference , S/5™ Compact Anaesthesia Monitor , Datex Ohmeda

Smart Growth and the Irish Land-use Stakeholder: From Rhetoric to Reality

Dorothy Stewart,

Faculty of Built Environment, DIT, Bolton Street, Dublin 1,

Contact email: dorothy.stewart@dit.ie

Abstract

In the past decade, Ireland has undergone a period of unprecedented growth culminating in the creation of a new economic society based on service provision, a move away from a traditional agricultural base. Allied to this has been an increase in economic, social and legislative inroads into Europe and beyond. This change has brought with it challenges commonly associated with unpredicted growth such as traffic congestion, urban sprawl, access to education and a perceived lack of affordable housing. One part of the solution proposes adopting the principles that underpin the concept of "Smart Growth". This paper critically evaluates the concept of Smart Growth. In a country with increasing concerns regarding land-use and property development, this paper demonstrates possible roles for Smart Growth in the mitigation of these issues. The paper also identifies the novel dimension of the research and its contribution to the knowledge base.

Keywords: Economy, Environment, Planning and Development, Society, Smart Growth

1.0 Introduction

This paper is based on a PhD research project entitled: "*Smart Growth and the Irish Professional Land-use Stakeholder: From Rhetoric to Reality*". Although there are several published definitions of smart growth, the Urban Land Institute describes the underlying objective as follows: "*Smart growth seeks to identify a common ground where developers, environmentalists, public officials, citizens and others can all find acceptable ways to accommodate growth*" (Porter, 2002:12).

Objectives:

1. Explore how international land-use legislation and EU Directives and Regulations affects Irish land-use policy;
2. Critically evaluate the nature of the Irish political system and its influence on the formation of planning policy in Ireland in terms of long-term smart growth versus short-term political system;
3. Explore the concept of smart growth and identify if the principles of smart growth have been embraced by policies of spatial planning;
4. Identify and critically evaluate methods and processes needed to provide and implement smart growth land-use;
5. Explore the role of Futures methods and techniques such as strategic conversations and scenario planning in urban planning;
6. Create an Irish smart growth Toolkit applicable to a unique Irish context.

The fundamental aim of the research is, to ascertain the extent to which the concept of smart growth is being strategically translated into action in the Irish land-use system. The main question that has arisen thus far is: how can the long-term goals of smart growth be reconciled with the short-term

political goals of the government of the day? The research to date has identified that there are an abundance of policies and strategies that seem to support the principles of smart growth in Ireland, however, implementation in its current form does not seem to reflect the policies. Other research questions include:

1. Why are the policies and strategies that support the concept of smart growth not being strategically translated into action in the Irish land-use system?
2. What is required to move from the aspiration stage to the delivery of smarter land-use policy? and
3. How can Ireland move from rhetoric to reality in terms of smart growth?

A possible solution to effect implementation of efficient land-use policies in Ireland is to tailor strategic smart growth 'best practice' tool kits that have been developed elsewhere to the Irish situation. The end product is an 'Irish Smart Growth Toolkit'. The tools to aid policy implementation in the Irish context may be used by a range of land-use stakeholders as a means of achieving more sustainable land-use. For example, local authorities might adopt tools relating to more creative land-use zoning. Fiscal tools such as incentives for more brownfield development may be of concern to public and private developers. The tools within the kit may include visual aids, similar to the Community Imaging Survey used in the United States during public participation workshops to create a sense of visual imagery of possible land-use scenarios (Corbett, 2000). The 'Irish Smart Growth Toolkit' might also contain a tool that would enable land-use stakeholders to get a more accurate measure of the cost of urban sprawl.

To date a questionnaire has been distributed to a purposive sample of land-use stakeholders throughout Ireland including, Architects, Chartered surveyors, Planners, Planning consultants, Property developers and Property investors the aim of which was to ascertain opinions and attitudes on land-use issues in Ireland. Of the 440 questionnaires distributed, 164 were returned of which 152 were completed. The data is currently being processed with the aid of SPSS statistical package.

1.1 Background

Ireland has experienced unprecedented economic growth rates, which have averaged 8% per annum between 1994 and 2001. The emergence of the tag 'Celtic Tiger' came about in the mid-1990s and played on the term 'Tiger economies' then being used to describe the leading Asian economies (Memery, 2001). Ireland is now subject to the challenges commonly associated with unpredicted growth such as, traffic congestion, urban sprawl, access to education and a perceived lack of affordable housing. Although there has been a recent slow-down in the economy with more modest growth rates, Ireland still has higher growth rates than many European countries.

The economic boom in the 1990s in Ireland and advancements in technology has served to influence settlement patterns and how people work, rest and play. Peoples consumption patterns may be more sophisticated, nonetheless, people now want fewer hours in traffic and more opportunities to enjoy green space, and housing that is both affordable and close to jobs and activities. Furthermore, people

want healthy cities, towns and suburbs, air and water of the highest quality and a landscape that future generation can be proud to inherit. Advocates in favour of changing from the 'business as usual' (BAU) model feel that smart growth offers the best chance of attaining those goals.

From the outset it is critical to note that Irelands economy was traditionally dependent on agriculture as its main source of income and as such has never been an industrialized nation when compared to countries like Germany, United Kingdom and The Netherlands. Settlement patterns in Ireland have shifted from a past dominated by rural lifestyles to the present day situation, like many developed countries, where the majority of people reside in urban areas. *"For city planning, this transformation demands a more imaginative approach towards the way communities think, talk, plan and act creatively in tackling the urban issues they face"* (Ratcliffe, J, 2002:2). Ireland has witnessed a continual decline in agriculture and is now enmeshed in a global network of connections trading goods and services on an international scale. This requires a sophisticated and advanced array of initiatives and tools to guarantee Irelands continued competitive presence at a global level.

Despite Ireland having a buoyant economy, and relatively low levels of unemployment until recently, there is still significant poverty, as indicated by Irelands low United Nations Development Plan index. Levels of homelessness, social exclusion and inequality are also increasing, notwithstanding the high levels of economic growth, (Comhar, 2001). It is suggested that adopting the NAPS principles would go some way in addressing existing inequities in the planning and development system by ensuring policy decisions are directed towards addressing uneven development, achieving more equity and ensuring that full implementation of new policies takes place.

Furthermore it will be necessary to monitor the effectiveness of new policy initiatives using suitable benchmarks like sustainability indicators alongside the traditional gross domestic product indicators. Urban areas have a dual characteristic as victims and perpetrators of social, economic and environmental degradation. It is important to note that with 60% of the population in Ireland living in urban areas it is critical to ensure the planning and development system is sensitive to the paradox of urban living. It could be argued that in order to accommodate increased urbanisation in Ireland in a more sustainable manner, lessons could be learned from countries like Germany and the Netherlands who offer many examples of best practice. Advocates of smart growth assert that there are significant fiscal and competitive advantages to be gained from adopting smarter growth development patterns (Muro and Puentes, 2004). Allied to this are the social and environmental benefits that would accrue from smarter land-use (Pavlov, 2004).

2. Planning in Ireland

Physical planning in Ireland formally commenced with the enactment of the 1934 Town and Regional Planning Act. This act introduced a coherent system of positive and regulatory planning based on the making by the planning authority of a planning scheme, (the precursor to the development plan), which was to govern the carrying out of future development. It is

important to note that in the 1930s neither the general public nor the politicians had much interest in planning as apathy prevailed (Grist, 1999). Traditionally planning has been dominated by short-term and present-focussed decisions about space (Scannell, 1995) and (Bannon, 1989). This short-term approach is not conducive to smart growth and in fact may act as a barrier to achieving the objectives of the concept. In contrast to Ireland countries like Germany, Sweden and the Netherlands have traditionally taken a long-term approach to the planning and development process (Beatley, 2001).

Recent policy and planning initiatives include *Sustainable Development A Strategy for Ireland 1997* the Strategic Planning Guidelines for the Greater Dublin Area 1999, The Planning and Development Act, 2000, and the National Spatial Strategy 2002-2020. The Planning and Development Act, 2000, has been noted as a watershed and replaces the 1963 Act and all intervening amendment Acts to date, with the goal to counter problems in relation to imbalances in the planning and development system and challenges arising from recent social, economic and environmental factors. Most notable about this Act is the words sustainable development throughout the entire document where previously the word development appeared. Another important watershed is evident in the Local Government Act, 2001. This Act modernized and simplified the law relating to local government, in particular repealing a series of statutes dating from the early nineteenth century and providing a common legislative code applicable to all Local Authorities.

3. Smart Growth in theory

The concept of Smart Growth emerged in the United States during the 1990s from research undertaken by the Urban Land Institute (ULI). At this time the ULI was looking at ways to deal with the problems arising from urban sprawl, traffic congestion, school overcrowding and air pollution. Other issues of concern to ULI at this time were the loss of open space and skyrocketing public facilities costs. The concept is also a reaction to the apparent failure of traditional planning techniques to improve conditions. As David Crockett, the leader of sustainability efforts in Chattanooga, United States said at a speech in Cleveland: “*Every time a bulldozer cranks up it doesn’t mean we’re making money*” (Porter, 2002:41). Although the origins of smart growth are to be found in America, the worsening trends listed are not unique to the United States and are evident in many developed countries including Ireland. Smart growth is not anti-growth and instead provides solutions to address the global challenge of achieving more sustainable development defined as “*development that meets the needs of the present without compromising the ability of future generations to meet their own needs*” (Bruntland, 1987).

The concept is evolutionary and is continuously influenced by economic, environmental and social factors. In 1996 in America a broad coalition formally joined hands as the Smart Growth Network

(SGN), with members spanning real estate, advocacy and policy-making circles (Tregoning *et al*, 2002). The idea is that a community should fashion its own version of smart growth through a shared decision-making process. The concept embraces a holistic approach that accords with community interests and reasonably balances the various principles that make up smart growth in theory. The actual term ‘Smart Growth’ may be uniquely North American, but the ideas behind the concept are not. The European Union (and especially densely populated countries such as the UK and The Netherlands) has had a long history of thinking about new ways to manage growth, especially in cities. From the 1990 ‘Green Paper on the Urban Environment’ to the adoption of the ‘Strategy for Sustainable Development’ in June 2001, the European Union reaffirmed that sustainability lies among the Communities’ policy priorities (Tregoning *et al*, 2002). The concept of smart growth can be compared to other new forms of planning and development.

Some community groups view smart growth as an open invitation to public officials and developers to design too dense projects. Efforts to control development collide with strongly held values concerning property rights and home rule. This reluctance to embrace change could result in a BAU approach with little change in favour of smart growth. “The related dynamic of urban decline is more amenable to change because political forces required for change are likely as aged suburbs themselves face decline” (Bier, 2002:83). There is evidence to suggest that the present day advocates of smart growth looked to the past to get ideas for the future. In the early 20th century the ideal of a planned residential community, “The Garden City”, was devised and promoted by the English town planner Ebenezer Howard in *Tomorrow: A Peaceful Path to Social Reform* (1898). The book was a response to the need for improvement in the quality of urban life. Howard felt that overcrowding and congestion due to uncontrolled growth since the Industrial Revolution had marred urban life. Howard had the gift of persuading practical businessmen that his idea was financially sound and socially desirable. Howard’s “Garden City” concept could be compared to smart growth, where there is mixed-use development, a town centre and open space conservation whilst adopting a more holistic approach to the planning and development process. These features are evident in the Garden Cities of Letchworth and Welwyn Garden City established in England at the beginning of the 20th century and prove that good design can produce enduring values and lasting streams of profit in social, economic and environmental terms.

Traditional planning of urban form was based on short-term economic gain, and the need to find quick solutions to deal with the ills of overcrowding in the inner city slums. The dominant planning ideology of this period was based on low-density, mono-use, and lack of diversity and flexibility. The legacy of this planning era is present day auto-dependent commuter lifestyles. Tregoning *et al* (2002) discusses how the term ‘sprawl’ has become a popular pejorative, shorthand for poorly planned growth that consumes precious open space and mars the landscape with ugly development. It is blamed for constant traffic jams, crowded schools and a host of other ills that afflict fast-growing communities. But while people from all walks of life agree on the consequences of this growth pattern that originated in the 20th century, they rarely see themselves as part of the problem – or the solution.

Many gravitate to the outer edges of suburbia without fully accounting for its trade-offs and contradictions (Tregoning *et al*, 2002).

For planners and environmentalists who hope to counteract the societal forces behind sprawl, it helps to keep that perspective in mind. For most people land-use issues reside in the here and now, in their own backyard and in a time frame that can be pencilled in on the calendar, not in some distant place or future. They think about different facets of planning and development in isolation, not as interrelated pieces of a big picture. In North America, the smart growth movement has emerged as the most promising attempt yet to make these connections. Advocates share many of the same goals as of earlier anti-sprawl efforts fought under the banner of sustainability- with a key difference. Their language and methods are more pragmatic and inclusive. Instead of appealing almost entirely to environmental sensibilities, as much North America sustainability discourse does, they wrap the discussion around basic quality of life issues (Ibid). Recent urban design and planning theory attach considerable importance to the concept of mixed-use in achieving sustainability, lower reliance on private vehicular use, and achieving more vibrant urban areas for the long-term. There is however, scepticism that whilst mixed-use developments are desirable, they are, nevertheless, difficult to achieve. In an article in the journal, *Urban Design International*, Hall argues that local development plans commonly work from a paradigm based upon two dimensional uniform land use allocations, (Hall, 2000). This approach has difficulty in coping with mixed-uses, urban design principles, urban history and the more general pursuit of more compact and sustainable settlements. Furthermore this approach does not provide an adequate basis for public participation. Alexander and Tomalty (2002) argue that in practice, local residents may oppose mixed-use projects because they will generate noise, parking difficulties or other nuisances. Municipalities are increasingly interested in performance-based zoning as a way to address this issue. Performance-based zoning regulates land-use based not on proposed use, location and dimensions of the development, but on the basis of the actual impacts it will have on the neighbouring residents and businesses. It allows any land use to locate adjacent to any other use, provided it satisfies predetermined performance standards (noise, dust, hours of operation, views, etc) (ibid).

3.1 Key principles of smart growth

In 1996 the Smart Growth Network defined the principles of Smart Growth as follows:

1. Mix land uses,
2. Take advantage of compact building design,
3. Create a range of housing opportunities and choices,
4. Create walkable communities,
5. Foster distinctive, attractive communities with a strong sense of place,
6. Preserve open space, farmland, natural beauty and critical environmental areas,
7. Provide a variety of transport choices,
8. Make development decisions predictable, fair and cost effective,
9. Encourage community and stakeholder collaboration in development decisions, and

10. Infill redevelopment, and adaptive use in built-up areas.

4.0 Proposed plan for the duration of the research

Respondents who participated in the recent survey were invited to take part in further data collection for further research. A total of 60 respondents indicated a willingness to participate further in the research. Preliminary examination of the survey results presented a broader and deeper remit for the research. It is contended that the quantitative dimension of the survey data could be further enhanced, expanded and elaborated upon by further qualitative data. The valued opinions of land-use stakeholders may be of great value in order to identify tools or processes required to successfully implement smarter land-use policy in Ireland. The research is currently at the theory building stage and it is envisaged to collect further data by means of futures methods of enquiry such as focus groups, strategic conversations and scenario planning in the next six months. According to Ratcliffe (2003) such scenario-based plans will progressively become integrated forums where the objectives of many sectors are synergised and synchronised. Future studies can simply mean any exploration of what might happen and what we might want to become. It contributes to an overall understanding of and approach to the future and its methods. Future studies is subject or questions oriented, for example what are the critical technologies that will have the greatest influence over the next 25 years? Futures research means the use of techniques to identify systematically the consequences of policy options and to identify alternative futures with policy implications for decision makers.

The Smart Growth Network that was established by the Urban Land Institute in 1996 has to date published two reports entitled “Getting to Smart Growth” (2000) and “Getting to Smart Growth Two” (2003). Both reports contain a hundred policies for implementation of the smart growth principles. These implementation policies may form the basis for the 'Irish Smart Growth Toolkit'. Another potential resource for the research is the upcoming report by Sir John Egan in the United Kingdom, using 7 criteria for sustainable communities, due to be published in mid April 2004. The 'Irish Smart Growth Toolkit' will be tested for robustness and suitability using a pilot process with land-use stakeholders. This wind tunnel testing of the 'Irish Smart Growth Toolkit' will identify who would use the toolkit, its benefits and be a means to effect more efficient land-use than was hitherto the case. Allied to this may be the use of suitable benchmarks like sustainability indicators to evaluate the effectiveness of the 'Irish Smart Growth Toolkit'.

5.0 Contribution to knowledge base

The concept of smart growth emerged in the 1990s in the United States from research undertaken by the Urban Land Institute. The founding principles are based on ways to deal with the ills associated with urban sprawl. The principles are closely aligned to the principles of sustainable development in terms of economy, environment and society. The world is becoming more urbanised. Ireland is by no means an exception to this trend and this is further evident in the research conducted by Hughes (2003) on Dublin as a city-state in the 21st century. It is generally agreed that there is now and will be a continual need for research into land-use issues based on their impacts on the economic,

environmental and social aspects of societies now and in the future. The World Watch Institute (2004) identifies a 'knowledge gap'. Despite the abundance of policy initiatives that seem to support the concept of smarter land-use, there seems to be a distinct lack of knowledge of what smarter land-use is and how to achieve it. The proposed 'Irish Smart Growth Toolkit' it is suggested would offer a user-friendly vehicle as a means for Ireland to achieve more sustainable land-use.

Although a globally accepted blueprint for smart growth does not exist, means to facilitate implementation of the concept in individual societies may vary from country to country. The success of smart growth ultimately will depend on its adaptation to each country's unique political, cultural and market dynamics and development trends. To date, the smart growth movement has focussed on state, regional and local reforms. There may be a need for a National Smart Growth Agenda according to Katz (2002) who stated that to attain measurable success smart growth will need to address at least five distinct challenges in the coming years:

1. the spatial distribution of affordable housing;
2. expanding housing opportunities for middle class families in the city, suburbs and more affordable housing near job centres, (for example through zoning policies);
3. significant policy reforms at all levels of government (for example building code changes);
4. construction of new affordable housing in fast-growing areas where jobs are increasingly concentrated and requiring a change in rules;
5. regional diversity; since smart growth first and foremost focussed on changing the basic laws and practices that govern both patterns in 50 states and in thousands of local jurisdictions.

6.0 Novel dimension to the research

Over the last couple of years Ireland has updated a wide range of land-use policy documents and strategies. The National Spatial Strategy 2002-2020 represents a watershed in spatial analysis being the first strategy that looks at Ireland as a whole. The Local Government Act, 2001, brought about the modernisation of local government in an attempt to be more transparent and effective and to decentralise power to a more local level. A National Biodiversity Plan was also launched for the first time in Ireland in 2002. Furthermore, the Planning and Development Act, 2000, has replaced the word development with the words sustainable development throughout the entire document. It would appear that Ireland has indeed made a huge commitment in terms of achieving a more holistic and integrated approach to planning and development. The research being conducted is novel in that it is concerned with the most up to date strategies and policies in land-use in Ireland to date. During the continual review of literature it has become apparent that there is a deficit in the availability of literature on the concept of smart growth as applied to the Irish situation. This represents a limitation but also presents an novel opportunity to generate greater awareness of the concept by conducting the research.

The research coincides with huge local, national and global changes in a world where greater attention is being given to environmental and social factors and more recognition is being given to the

inextricable link between economy, environment and society. Another possible benefit of the research may be an enhanced awareness and shared understanding of smart growth issues across a wide range of stakeholder groups.

The research proposes the use of traditional research methods alongside the use of ‘futures methods’ such as prospective through scenario planning to facilitate the adoption of the principles of smart growth. The study of the future is a multi-disciplinary examination of change in all major areas of life to find interacting dynamics that are creating the next age (Glenn, 1994). It is only in recent years that the benefits of applying futures methods to the discipline of planning have been recognised.

The research coincides with Ireland’s EU presidency from the first of January until 30th June 2004. Ireland is in a strategic position to advance the concept of smart growth further in Ireland and throughout the rest of Europe, especially at this critical time when ten new Member States will join in May 2004 (adding an extra 100 million people to the European Union). What will the agenda be during the period of this presidency? According to the Taoiseach Bertie Ahern on 14th January 2004, the theme of the Irish presidency will be “*Europeans Working Together*”, a theme which captures a vision of the people of the European Union working as a partnership, striving together to achieve our common goals and objectives. The Irish presidency has placed sustainable growth and social cohesion at the very centre of its work programme (EU Presidency, 2004). To advance the concept of smart growth in Ireland and the European Union the following ingredients are essential:

1. vision;
2. entrepreneurship;
3. specialisation;
4. social cohesion; and
5. governance. (All the ingredients to be found in successful competitive cities) (Ratcliffe, 2003).

Objective 2 of the thesis suggests the need to decouple politics from the planning process in Ireland. According to Memery (2000:80), in 1987 policy was put in place by the Fianna Fáil Government to develop a competitive economy, “but failed to take cognisance of the housing (and transport infrastructure) requirements for such growth”. The legacy of this oversight is evident in the current socio-economic and environmental challenges now facing Ireland, and some would argue that it is too late to change. More joined up, integrated and holistic thinking by relevant stakeholders may help reverse the current trends.

As stated earlier in this paper, there is a need to adopt a more long-term approach to planning and development and this is reflected in recent policies and strategies such as the NSS. Allied to this is the insignificant role of regional planning in Ireland until recently and how this contrasts to our European counterparts whose success it is argued, has been founded on adopting a more regional approach to planning and development. Furthermore, the establishment of City/County Development boards and Strategic Policy Committees brought about under the Local Government Act, 2001, has facilitated

more participatory democracy in contrast to the traditional representative model (Grist, 2003). Ravetz (2000) states that local authorities occupy a strategic position as *'catalysts of change'* in terms of planning and development.

The United Nations Conference on Environment and Development (UNCED) in 1992 was the launching pad for Agenda 21, which is a non-legally binding authoritative statement of the principles for a global consensus for the 21st Century (Grubb, 1993). The document advocates achieving objectives and goals through a planned, democratic and co-operative process. It singles out local government as having a special role in educating, mobilising and responding to the public to promote sustainable development (DoELG, 2001). Local Agenda 21 (LA21) is a process that facilitates sustainable development within a community. *"Because so many of the problems and solutions being addressed by Agenda 21 have their roots in local activities, the participation and co-operation of local authorities will be a determining factor in fulfilling its objectives. Local authorities construct, operate and maintain economic, social and environmental infrastructure, over-see planning processes; establish local environmental policies and regulations, and assist in implementing national and sub-national policies"* (UNEP, 1992).

Greater public participation is greatly facilitated by the process of Local Agenda 21 and also affords people the opportunity to participate in the decision making process about issues that affect peoples lives. As stated from the outset, the research crosses the disciplines of economics, environment and society. Collaboration is at the heart of the concept of smart growth with all stakeholders taking an active role in the planning and development process. Modern day planning and development embraces more participatory democracy. Both government and the private sector now see Public/private partnerships as a viable means to develop required infrastructure in Ireland. Interestingly, a recent relaxation of EU rules on State borrowing has opened the way for a number of infrastructure projects like upgrading the M50, the Dublin Metro and the Cork School of music. The State may allow the private sector finance the projects and spread the cost over periods of up to 20 years, considerably reducing the impact on the Government finances. More relaxed rules apply where the project is financed by private investors who assume a significant element of the risk associated with the project (McManus, 2004).

7.0 Conclusion

This paper presented evidence that would suggest a need to adopt a more integrated, holistic and long-term approach to planning and development if the goal of more efficient land-use is to be realised. The paper suggests that one part of the solution proposes the creation of an 'Irish Smart Growth Toolkit' as a means to accommodate inevitable growth that is economically viable, friendly to the environment and enhances quality of life. It could be argued, however, that mobilising support for smart growth in Ireland and achieving it will not be easy and is not inevitable. Changes will be difficult and controversial and will require leadership, a willingness to innovate and collaboration among all stakeholders will be required. Individual special interests must be put aside in the joint

pursuit of sustainable urban forms. The paper also suggested the need to learn from the good examples of the past with continual evaluation and monitoring to establish if the objectives of smart growth are being met. The paper suggested the need for increased awareness about the concept of smart growth. One way to achieve this is through further education, discussion and greater participation. The proposed plan for further research, the contribution to the knowledge and the novel dimension to the research was also examined. Smart growth advocates have a large toolkit of time-tested programmes and regulatory tools they can use to try and overcome obstacles. Experimentation and courage to break from the short-term BAU model will be necessary. Should the needs of present generations be met by adopting the principles that underlie the concept of smart growth it follows that future generations will inherit sustainable societies. Ultimately, the smart option is dependent on meeting the needs of the present in a sustainable manner.

References

- Bannon, M. J. (1989) *Planning The Irish Experience 1920-1988*, Wolfound Press, Dublin.
- Beatley, T. (2001) *Green Urbanism: Learning From European Cities*, Island Press, Washington.
- Bruntland Commission (1987) *Our Common Future*, World Commission on Environment and Development, University Press, Oxford.
- Corbett, J. and Corbett, M. (2000) *Designing Sustainable Communities: Learning from Village Homes, Davies, California*, Island, London.
- Glenn, J. C. (1994) Introduction to Research Methodology Series, AC/UNU Millennium Project, Venezuela.
- Grist, B. (1999) *An Introduction To Irish Planning Law*, IPA, Dublin.
- Grist, B. (1983) *Twenty Years Of Planning, A Review Of The System Since 1963*, An Foras Forbatha, Dublin.
- Grist, B. and Macken, J. (2003) *Irish Planning Law Factbook*, Thomson Round Hall, Dublin.
- Grubb, M. (1993) *The 'Earth Summit' Agreements: a guide and assessment: an analysis of the Rio '92 United Nations Conference on Environment and Development*, Earthscan and the Royal Institute of International Affairs, London.
- Hall, P. and Pfeiffer, U. (2000) *Urban Future 21*, Spon, London.
- Howard, E. (1989) *To-morrow: A Peaceful Path To Real Reform*, Swan Sonnenschein, London.
- Hughes, B. (2003) *Regional Planning Guidelines*, Dublin Institute of Technology, Hughes, Brian.
- Irish DoE (2001) *Towards Sustainable Local Communities: Guidelines on Local Agenda 21*, Department of the Environment and Local Government, Stationery Office, Dublin, Ireland.
- Katz, J. B. and Lang, E. R. (2002) *Redefining Urban And Suburban America: Evidence From Census 2000*, Brookings Institution Press/ Brookings Metro Series, Washington, DC.
- McManus, J. (2004) *Boost For Metro Plan As EU Eases Rules On Borrowing*, Irish Times Newspaper, Dublin.
- Memery, C. (2001) The Housing System and the Celtic Tiger: The State Response to a Housing Crisis of Affordability and Access, *European Journal of Housing Policy*, 1(1), pp.79-104.
- Mosser, C. (1971) *Survey Methods In Social Investigation*, Heinemann, London.
- Mumma, A. (1995) ***Environmental Law: Meeting UK and EC requirements, Mc Graw Hill, Berkshire.***
- Muro, M. and Puentes, R. (2004) ***Investing in a Better Future: A Review of the Fiscal and Competitive Advantages of Smarter Growth Development Patterns, A Discussion Paper Prepared by The Brookings Institution Center on Urban And Metropolitan Policy, Washington.***
- OECD (2003) *Economic Survey In Ireland*, OECD, Paris.
- Pavlov, A. (2004) Land Values and Sustainable Development, paper presented to RICS Foundation, Simon Fraser University, Canada, RICS, U.K.
- Porter, D. (2002) *Making Smart Growth Work*, Urban Land Institute, New York.
- Ratcliffe, J. (2002) *Imagineering Cities: Creating Future 'Prospectives' for Present Planning*, CIB Tokyo Conference 2002, Dublin.
- Ratcliffe, J. (2003) *"Competitive Cities: Five Keys To Success"* Greater Dublin 'Prospective' Society, A Futures Academy Background Paper, Dublin.

- Ravetz, J. (2000) *City region 2020: Integrated Planning For A Sustainable Environment*, Earthscan, London.
- Scannell, Y. (1995) *Environmental And Planning Law In Ireland*, Blackrock Round Hall Press, Co Dublin.
- Tregoning, H. Agyeman, J. and Shenot, C. (2002) Sprawl, Smart Growth and Sustainability, *Local Environment*, 7 (4), pp. 341-347.
- UNEP (1992) *Agenda 21*, United Nations Environment Programme, UN, Stockholm.

Web sites

<http://www.esri.ie/content.cfm?t=Irish%20Economy&mid=4>
http://www.smartgrowth.net/Home/sg_Home_fst.html
<http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,contentMDK:20150219~menuPK:34458~pagePK:64003015~piPK:64003012~theSitePK:4607,00.html>

Others

- Environment 2010: Our Future, Our Choice/ 6th Environment Action Programme, 2001-2010 European Commission, Luxembourg: Office for official publications of the European Communities, 2001.
- The 1973 Council on Environmental Quality Guidelines for the content of Environmental Impact Assessments states in paragraph 1500.8, part (a) section 4; Planning and Development Act 2000, Dublin Stationery Office.

Soft, Vertical Handover of Streamed Multimedia in a 4G Network

Ger Cunningham, Philip Perry and Liam Murphy

Dept. of Computer Science, University College Dublin, Belfield, Dublin, Ireland

{[ger.cunningham](mailto:ger.cunningham@ucd.ie), [liam.murphy](mailto:liam.murphy@ucd.ie)}@ucd.ie , perry@eeng.dcu.ie

Abstract

In this paper the soft, vertical handover of streamed multimedia in a 4G network is considered. We propose a soft handover solution in which the mobile client controls the handover. This solution requires no modifications to existing wireless networks. The second stream required for the soft handover is duplicated just above the transport layer, rather than requiring the server to play out a second stream that needs to be synchronised with the existing stream. Such a scheme is outlined, and the results are presented that show how the scheme functioned in an emulated environment.

I. INTRODUCTION

Fourth-generation (4G) wireless communication systems [1] will be made up of different radio-networks providing access to an IPv6 [2] based network layer. In densely populated areas, for example, 3G will augment ubiquitous 2.5G networks by providing higher bit-rate access. In hotspots and in corporate campuses, Wireless LANs will complement these systems further by providing even higher bit-rates. Other wireless access networks envisaged for 4G include satellite networks, fixed wireless access (e.g. IEEE 802.16) and PANs (Bluetooth, 802.15, UWB).

Multimedia is expected to be a main application of 4G networks. However, multimedia streams can be sensitive to packet loss, which in turn can result in video artefacts. Such packet loss can often occur when there is an interruption to a connection when a user is moving between networks that are autonomous.

In cellular networks such as GSM, a call is seamlessly handed over from one cell to another using hard handover without loss of voice data. This is managed by network based handover control mechanisms that detect when a user is in a handover zone between cells and redirect the voice data at the appropriate moment to the *mobile node* (MN) via the cell that the MN has just entered.

In 4G networks a handover between different networks is required. A handover between different networks is usually referred to as a *vertical handover*. As 4G networks are comprised of independent networks, there may be no network based handover control mechanism. Therefore, a hard handover may not be possible for 4G as multimedia data may be lost. Instead, soft handover can be used. This ensures that data is not lost by allowing the MN to attach to two networks simultaneously during the handover period, and the same data is sent

to the MN via the two networks during this period. The price paid for soft handover is the use of resources in two networks, rather than one, but only during the handover phase.

This paper outlines a scalable soft handover system that enables a MN to control the handover of a multimedia stream in a 4G network. The rest of the paper is laid out as follows. The next section examines related work, while the following section describes the proposed scheme in detail. Section IV presents results from an emulated environment. The final section presents the conclusion.

II. RELATED WORK

Mobile IP [3] has been proposed as a solution for mobility support in IP networks. Mobile IP uses hard handover. Handovers are slow and packets can be lost during the handover procedure [4]. It is therefore unsuitable for the handover of streamed multimedia.

The mobility support provided in the Session Initiation Protocol (SIP) [5] has also been proposed for real-time communication [6,7]. SIP is an application layer protocol for establishing and tearing down multimedia sessions. It can help provide personal mobility, terminal mobility, and session mobility. SIP uses hard handover and “in-flight” packets can be lost during the handover period. In [6] it was estimated that packet loss can happen for up to 50ms during the handover period. So SIP does not provide seamless handover of streamed video in networks.

The scheme proposed below uses soft handover and so is appropriate to the needs of streamed multimedia.

III. PROPOSED SYSTEM

The architecture of the proposed system is shown in Figure 1. The system consists of a multimedia server that is connected to the Internet and a MN that is streaming a multimedia session from the multimedia server while roaming from wireless access network to wireless access network. The server has a handover agent that handles the soft handover. The MN also has a handover agent. The server’s handover agent is located between the transport layer and the normal play out function of the server, while the MN’s handover agent is located between the transport layer and the normal decoding function. The MN has 2 radio interfaces, each with its own IP address.

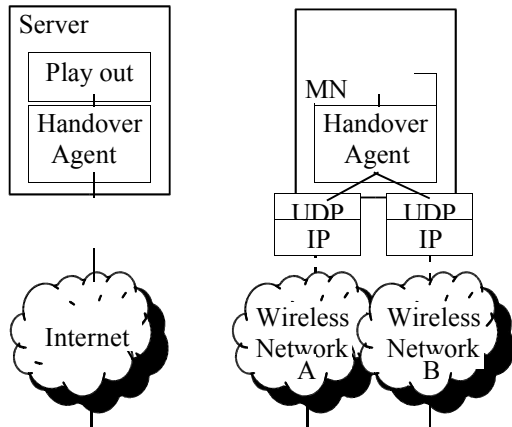


Figure 1. Proposed handover scheme.

Initially the MN is receiving the multimedia content from the server via one wireless network, and the other radio interface is looking for another wireless network. When it discovers another wireless network, the MN sends a `START_HANDOVER` command to the server's handover agent. On receiving this command, the server's handover agent duplicates each multimedia packet it receives for this multimedia session from the play out part of the server, until it receives the `END_HANDOVER` command from the MN (Figure 2). The `START_HANDOVER` command supplies a number of parameters to the server's handover agent: a session id, an IP address and port number. The session id pertains to the current multimedia streaming session in the MN, and is used by the server's handover agent to decide which packets to duplicate. The IP address and port number refer to the IP address and port number that will be used by the MN's radio interface in the wireless network just discovered. The server's handover agent uses these values with the duplicated packets when it inserts them into the UDP transport layer. Therefore the MN receives two streams from the server during the handover period, one through each wireless network, enabling the MN's handover agent perform a soft handover. The MN's handover agent decides which packets to pass on to the decoder, the original packets or the duplicate packets.



Figure 2. Soft Handover Protocol.

The mobile client controls the handover and no modifications are required to the existing wireless networks that might make up a 4G network. The duplication of the streams in the server's handover agent relieves the server from having to play out a second multimedia session that must be in sync with the current multimedia session being played out. The duplication of the packets adds overhead in the server. However, if it is acceptable to make changes to the network architecture, the server's handover agent can be placed on a proxy, freeing the server fully of the overhead in performing a soft handover.

IV. EMULATION RESULTS

To validate the concept of the proposed system, a prototype system was implemented and tested in the test bed shown in Figure 3.

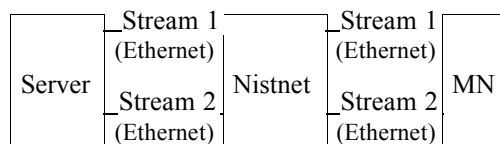


Figure 3. Emulation Test Set-up.

The test bed was based on Nistnet [8], which is a network emulation tool that is frequently used for emulating network conditions at the IP layer. It can emulate network conditions (e.g. packet delay) on a per path basis. The server and MN were connected to the computer containing the Nistnet package using Ethernet links. Two separate paths were used to emulate two wireless access networks, and the MN had two IP addresses.

The server transmitted a stream (Stream 1) of Real Time Protocol (RTP) packets [9] at 100kbps using the User Datagram Protocol (UDP) through the upper of the two paths shown in Figure 3. Nistnet was used to apply delays to the packets it received. These delays were consistent with the delays typically experienced by data packets in a handover zone between cells. After a period of time, the MN sent the START_HANDOVER command to the server (Figure 4). The server's handover agent then turned on Stream 2 through the lower path shown in Figure 3, by duplicating the RTP packets and using the IP address and port number that were received in the START_HANDOVER command. After another period of time, the MN sent the END_HANDOVER command to the server. On receiving this command the server's handover agent then turned off Stream 2 (the duplicate stream) and simultaneously switched Stream 1 to the lower path by applying the IP address and port number that were previously used by Stream 2 to Stream 1's packets.

Figure 4 shows the streams received at the MN, and the points when the MN sent the START_HANDOVER and END_HANDOVER commands. It shows that the MN's handover agent received Stream 1 and Stream 2 during the handover period, enabling it to perform a soft handover of the RTP stream.

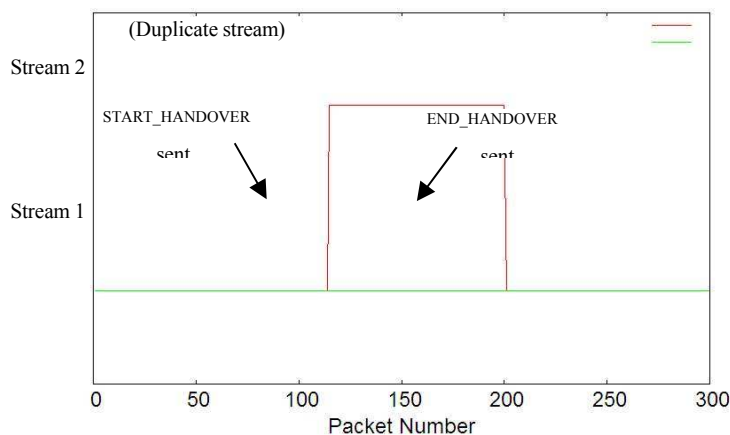


Figure 4. Emulation results.

V. CONCLUSIONS AND FUTURE WORK

In this paper we explained why soft handover is appropriate for the handover of streamed multimedia in 4G networks. We described a scheme to perform the soft handover that enabled the MN to control the handover and required no modifications to the existing stack architectures of wireless networks. Finally, we showed how we demonstrated the scheme using an emulated environment.

Our future work will focus on how the scheme functions in real wireless access networks.

ACKNOWLEDGEMENTS

The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged. Thanks to Dr. Sean Murphy for reviewing the paper.

REFERENCES

- [1] Hui & Yeung, "Challenges in the Migration to 4G Mobile Systems", IEEE Comm. Mag., Dec 2003
- [2] R. Hinden. "Internet Protocol, Version 6 (IPv6) Specification", Internet-Draft, Oct 1994
- [3] C. Perkins, "IP mobility support", RFC (Proposed Standard) 2002, IETF, Oct 1996
- [4] Stephane & Aghvami, "Fast Handover Schemes for Future Wireless IP Networks: a Proposal and Analysis", VTC, May 2001.
- [5] M. Handley et al, "SIP: session initiation protocol", RFC (Proposed Standard) 2543, IETF, Mar 1999
- [6] Wedlund & Schulzrinne, "Mobility Support using SIP", Proc. ACM WoWMoM'99, USA, Aug 1999
- [7] Wedlund & Schulzrinne, "Application Layer Mobility Using SIP", Mobile Computing and Communications Review, Volume 1, Number 2, 2001
- [8] Nistnet emulation tool, <http://snad.ncsl.nist.gov/itg/nistnet/>
- [9] Schulzrinne et al, "RTP: A Transport Protocol for Real-Time Applications", IETF RFC 1889, Jan 1996

Questions of Ethical Responsibility in the Research of Unaccompanied Minors

Oonagh Charleton & Dr. Celesta McCann James

School of Business and Humanities, Institute of Technology Blanchardstown

Oonagh.charleton@itb.ie

Abstract

This paper presents a general discussion on ethical considerations in qualitative research in the applied social sciences. It reflects on the ethical dilemmas posed to this researcher prior to initiating a field-work process and during the methodological structuring process. The intention is to promote discussion on issues of ethics when researching new types of service user groups with attention to the value of ethics committees.

1 Introduction.

Both qualitative and quantitative paradigms in social science research lay weight to ethical considerations in the research process. As researchers we evaluate and consider power relationships. We do this by creating the means to question our social realities and social relationships where we then analyze, measure and account for societal relationships and phenomena.

Statements like these find uncomfortable seats next to those concerning the manufacture of power in the social sciences. The accepted constitutions of ethical responsibility that exist within each social scientific community, the laws by which scientists abide by to protect their research, the respondents and the social milieu in which they are situated, are central in assessing how searching for truths in contemporary science can affect individuals and communities for better or worse.

This paper aims to contribute to discussion concerning issues of research ethics in the applied social sciences, notably with reference to a new social care service user, the unaccompanied minor. It seeks to consider that with an increasing number of ethnic minority communities establishing themselves in Ireland, ethics as an element of research practice needs to be further investigated. Its purpose is not to review the extensive literature available on research ethics but to consider ethical issues that may evolve out of a research process involving an unaccompanied minor service user. This paper cannot aim to evaluate what ethical conflicts have emerged whilst in the field as this phase of data gathering has not yet commenced. However, it seeks to examine the conflict that occurs when researchers have to presume certain

ethical guidelines are appropriate, prior to engaging with the service user, while all the time uncertain as to how their role in the process may affect the service user.

II Ethics in the Social Sciences

Social science research by its very nature is a tool or collection of tools designed to gather information and generate knowledge about people and their social lives. In addition to this it sets out to gather information and generate knowledge about society itself and the structure of social relationships. Sarantakos' (1993:17) defines three basic motives in the gathering of data. The first motive, 'educational', sets social research on a path to *educate* the public so that it can form a 'qualified opinion on the issue studied.' The second, 'magical', involves offering *credibility* and *respect* to certain views 'through statistics or even the mere presence of findings.' Finally, 'tactical', aimed at 'delaying decision or action' for as long as the research is underway.

The aims or driving forces of most types of social research vary with each researcher and their research question, however Sarantakos (1993:16) identifies nine main aims that researchers usually refer to:

- To *explore* social reality for its own sake or in order to make further research possible.
- To *explain* social life by providing reliable, valid, and well documented information.
- To *evaluate* the status of social issues and their effects on society
- To *make* predictions.
- To *understand* human behaviour and action,
- To *emancipate* people.
- To *suggest* possible solutions to social problems.
- To *empower* and *liberate* people.
- To *develop* and/or *test* theories.

What these aims may suggest is that social scientific research is a dynamic process where the epistemological contributions to society outweigh problems or hardship caused by the research process. There are cautionary tales however, within the social sciences as to how research practice can adversely affect the participants or the community in general.

Holloway and Walker (2000:71) examine the importance of ethical issues in the social care research process and provide a case study revealing damage to research participants. They explain:

“A study by Langer and Rodin (1976) set out to examine the effects of giving more control over volunteer visitors to elderly people in residential care settings. By the end of the study, the group that had control over the visits were less depressed and more active than the control group. However, on returning a year later, the research team were shocked to find that the death rate in the experimental group during the intervening period was double that of the control group and those still alive were more depressed.”

This example is used by Holloway and Walker to highlight some unforeseen effects of ‘interfering’ in peoples lives for the purposes of ‘well intentioned’ research. Ryan (1997:3) cites Kellehear’s (1989) experiences of ethical concerns that emerged as a result of interviewing the terminally ill, where she points to the need for researchers to recognize that the nature of some research questions may provoke responses that are ‘traumatic’ for participants.

There are generally accepted ethical standards in professional research practice that span the majority of research communities, from Europe and the United States to Australia and New Zealand. Many of these standards are borrowed or evolve from public documents like the Charter of Fundamental Rights of the European Union, Ratified at Nice in 2001, the UN Convention of Human Rights, The Helsinki Declaration, the Sociological Association of Ireland’s Ethical Guidelines and General Principles, to name but a few. If National Standards for individual countries fall short of detailing ethical considerations for specific groups, or researchers are not required to submit research questions and methodologies to university or other ethics committees, self-regulation is accepted as the norm. However many individual third level institutions and research centres produce their own codes, normally based on those accepted by the wider research community.

In Ireland, no centralized Government forum or council exists within the social sciences that might consider ethical issues in the research of humans and their social worlds. The ‘Sociological Association of Ireland’ however, use and base their codes of research practice on those produced by the British Sociological Association, the American Sociological Association, and the Australian Sociological Association. General principles cited by the SAI (2001:1) include:

- Professional Competence
- Integrity
- Respect for Human Rights, Diversity and Equality

- Social Responsibility

Dublin Institute of Technology has gathered some of their ethical guidelines for students from the Helsinki Declaration 1964 (amended 2002) and the International Code of Marketing and Social Research Practice (1995). D.I.T's Ethics Research Committee has strict procedures in place where all researchers must submit an ethics declaration along with their proposal.

All research involving pharmaceutical preparations, pregnant women, women in labour, persons under the age of 18, persons with physical or mental disabilities, or other vulnerable categories or members of ethnic or other minority groups must present an ethics declaration. The sitting committee assesses and evaluates each proposal before accepting or rejecting a student for registration.

Some institutions have quite structured and phased evaluation systems where research proposals are scrutinized by what that SAI might define as Professionally Competent committees.

Harvard University has in place a standing committee that serves as the institutional review board for the faculty of arts and sciences. With a flowchart detailing ethical considerations integrated into its system, each researcher must consult and determine if their research question and proposed methodology require panel review.

As it stands, if my research in the area of Unaccompanied Minor Service users was initiated in Harvard University, I would be obliged to submit to the ethics committee for review at phase 1 of the chart. 'Minimal risk' as stated above also includes the possibility of causing 'mental harm or discomfort' which as discussed earlier is difficult to establish prior to engaging with this unique service user in the field.

While standards like these are there and available to peruse as a researcher, ethical considerations that emerge before or throughout a research process are often only regulated by the researcher alone, or in consultation with their supervisor. This usually occurs in institutions that do not have resident committees to sanction research.

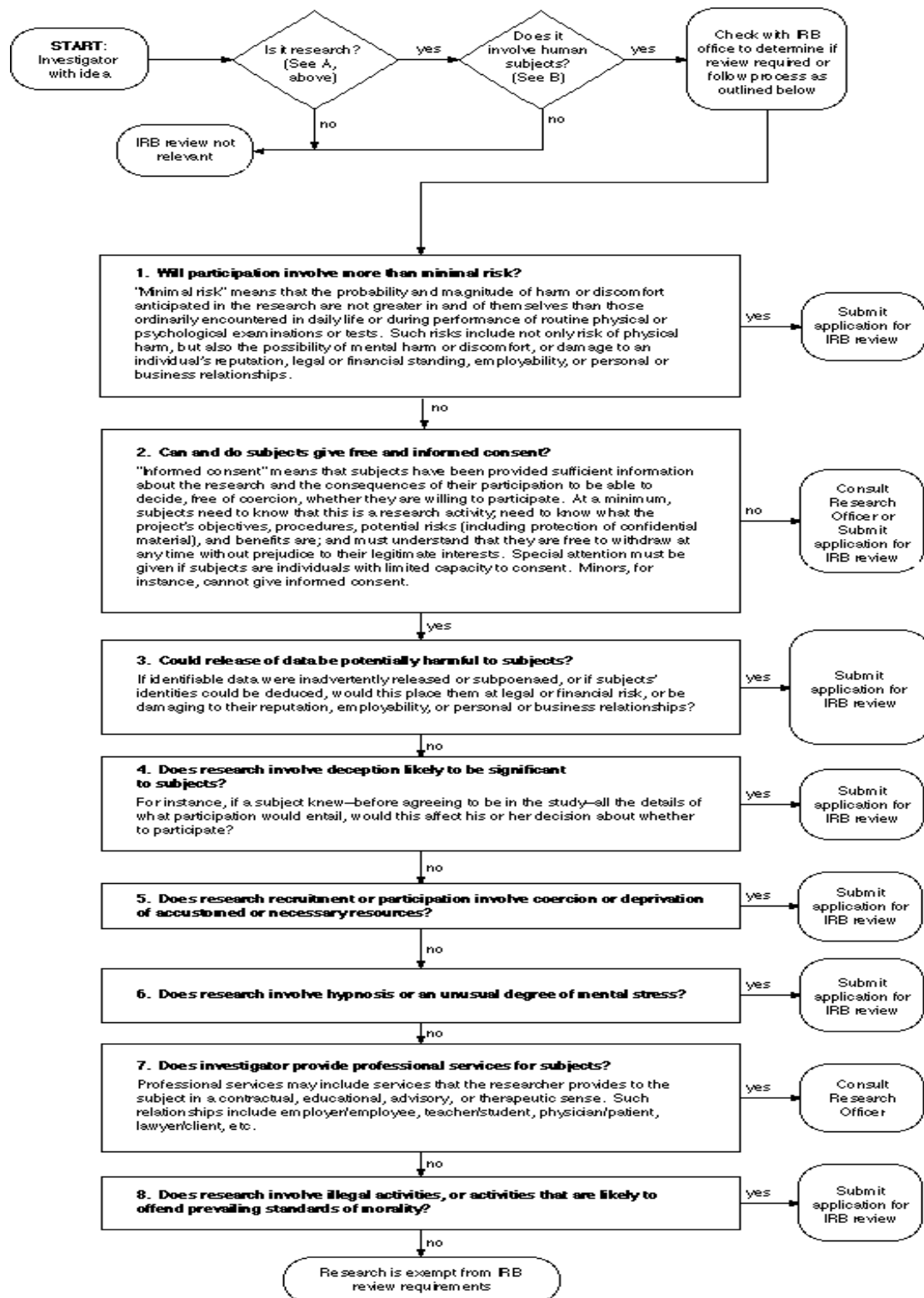


Figure 1: Harvard Human Subjects Committee Flowchart

(Source <http://www.fas.harvard.edu/~research/flowchart.html>)

III Unaccompanied Minors

The Irish social science research industry is busy and minority ethnic communities as Ryan (1997:2) comments are ‘increasingly the subject of research undertaken by students at universities, professional researchers and researchers commissioned by government agencies.’ So what of the unaccompanied minor as a ‘research subject’? What ethical questions emerge when looking at the status and social life of this social care work client group, from a research perspective?

Unaccompanied minors have been defined by ‘The Separated Children in Europe Program’ (SCEP) (2000:3) as:

“Children under eighteen years of age who have been separated from both parents or their previous legal or customary caregiver. Separated children (unaccompanied minors) may be seeking asylum because of fear of persecution or the lack of protection due to human rights violations, armed conflict or disturbances in their own country. They may be victims of trafficking for sexual and other exploitation, or they may have traveled to Europe to escape conditions of serious deprivation.”

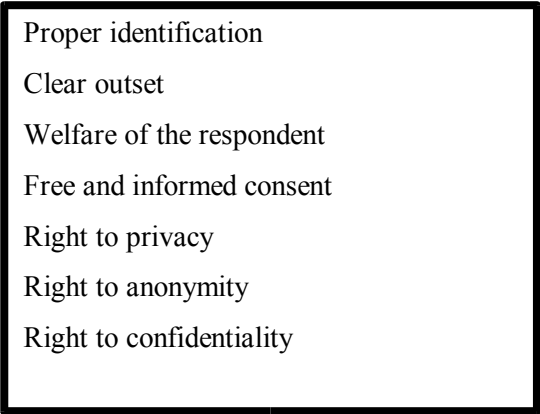
While this definition is generally accepted at a European level each member state may have its own definition. In Ireland, the East Coast Area Health Board (ECAHB) (2002: 66) defines this service user group as ‘children under the age of 18 who are identified as being unaccompanied by their parents/guardians or customary caregivers when they arrive in Ireland to seek asylum.’ The ECAHB is primarily responsible for the total care needs of these service users including the provision of ‘appropriate, immediate and ongoing care placements and social, medical and psychological services,’ as well as family tracing services. Veale et al (2003:16) identify these children as having a consistent profile of ‘emerg[ing] from countries experiencing armed conflict, political repression or the breakdown of civil society.’ If one can assume from Veale et al (2003:16) the UNHCR (2001:2) and the SCEP’s (2000) Statement of Good Practice that this client group is in fact vulnerable, ethical considerations take on added dimension.

Ryan (1997:1) clarifies the position of the role of the researcher when engaging in the evaluation of research ethics or ethical responsibility. She contends that, ‘essentially research ethics refer to the responsibility of the researcher to ensure that the participants in the research are not harmed by the research.’ She continues, ‘hence it is the responsibility of the researcher

to ensure that he or she has fully explicated the possible intended and unintended consequences of the research being carried out.’

What one can ultimately discern from this statement is that each social researcher has relative autonomy over the methodological strategies that they wish to employ. Sarantakos (1993:21) confirms this, defining the process as one that is based on ‘mutual trust and cooperation as well as accepted conventions and expectations.’ He pushes the issue further here stating that, ‘on the basis of this, researchers enter the field with relatively few limits and many options for action.’ While this freedom provides ‘great opportunity,’ he argues, ‘it can also have an adverse effect on participants.’

Understanding the adverse effects of research on participants, namely the unaccompanied minor service user in this case, requires analysis of principles governing the researcher-respondent relationship. Sarantakos (1993:24) delineates these carefully, considering:



- Proper identification
- Clear outset
- Welfare of the respondent
- Free and informed consent
- Right to privacy
- Right to anonymity
- Right to confidentiality

While all of these are pertinent to research that involves the study of human beings in both the natural and social sciences, issues can arise when research subjects do not possess a traditional voice or a vocabulary conducive to effectively questioning the incentives of the research community. Ryan (1997:2-3) identifies this within the traveling community in Ireland where she claims that as a consequence of little formal education, members of this community possess minimal technical vocabulary, and as a result have little power with respect to how research is carried out and how the data is disseminated. She argues here that research communities should have regard for the impact of research and concern themselves with the ways in which research can ‘affect the researched communities.’

So what of my role as a researcher when qualitatively engaging with a new type of service user in the applied social sciences? Where do I fit into the process? What part will I play in constructing the knowledge? Will my very presence negatively affect the service user? Will my

presence, for example, raise questions in their minds as to their environment, identity, status, or security? Will my presence cause alarm, anxiety or even stress? Could this consequently affect the validity of the data? How will dissemination or publication of the findings affect the service user or indeed the service providers? As a researcher, how do you rationalize shifting balances, affecting policies, affecting change in standards of living, standards of care, moral codes and inbuilt practices? And what of the service user group? As Ryan (2003:4) argues, ‘what happens when elements of research could be used to discontinue funding?’

With these types of questions in mind, there are no accepted ethical principles or standards exclusive to this service user type in Ireland, simply because this field is as yet relatively unexplored in the Applied Social Sciences. There are no guidelines as to what effect the researcher may or may not have on issues like identity, sense of security or levels of stress in the service user. What has emerged as part of the preliminary analysis is a sort of ‘chicken and egg scenario’. This service user group has not yet been qualitatively researched in an Irish Social Care setting so there are no unique ethical guidelines that have emerged as a result of peer reviewed findings. So it may be that only after a data gathering process involving critical self-reflection is complete, that these ethical considerations emerge and guidelines for this service user become available. As it stands, this paper cannot aim to evaluate what hasn’t yet been done.

IV Consent

Issues of consent are pertinent considerations when engaging in social science research. The Australian ‘National Health and Medical Research Council’ (NHMRC) (2002:1) look at this quite well, extensively evaluating ethical concerns relating to children as significant participants. While they are in general agreement with the premise that research is necessary in order to advance knowledge, they argue that research should only be conducted with children when certain criteria are met.

With respect to Sarantakos’ fourth principle of *free and informed consent*, the NHMRC (2002:2) concerns are interesting considerations. Outlined in the Human Research Ethic Handbook, the National Statement on Children and Research requires that ‘the child’s consent be sought if he or she is able to make a decision.’ This raises a number of concerns with regard to statutory law in Australia where ultimately, the National Statement admits to contradictions between research laws and statutory requirements. In addition to this, questions concerning ‘consent to research’ if the child is not in a position to make a decision appears to be left exclusively with the parent(s) or guardian(s) of the child. Ultimately however, it is taken for granted that the parents or guardians of the child should be approached before the child is

approached. In the case of children accommodated by statutory organizations, consent is required from those who have a legal duty of care to the child, in addition to the child themselves. McCauley and Brattman (2002:27) look at this in an Irish residential care context and question the appropriateness of seeking consent from care professionals who work with these children. Requesting clarification, they argue that it arises “*in particular with children and young people experiencing poverty and/or other forms of social exclusion.*”

Of those unaccompanied minors in Ireland who cannot consent to participating in a research process due to age, communications difficulties etc. are dependent on their care managers to provide consent on their behalf. With this in mind, managers are then the ‘last bastion’ of access. These personnel can in themselves act as research ethics committees where the researcher must provide declarations of ethical propriety. In the case of researching minors who’s sole guardians are the state, it is to be expected that individual care managers will require evidence of good ethical consideration prior to access being granted.

V Conclusion

Unaccompanied minors have been traveling to Ireland in greater numbers from 1999 to the present day. As a result Irish based research projects that began around this time may only start to emerge over the next few months and years. Data and methodologies employed may reveal ethical concerns and considerations that are unique to this client group and their social worlds. If and until this evolves, research can only follow the ethical guidelines established by the greater research community or develop their own set of ethical guidelines with respect to the client.

When questioning ones ethical responsibility with regard to researching this client group, there is a need to strike a balance between the ethical dilemmas and the search for knowledge that may lead to improvements in the provision of care. Remembering my ethical dilemmas delineated in section II, the questions remain. Will my presence cause alarm, anxiety, and stress? What part will I play in constructing knowledge? Will my presence raise questions in their minds as to their value, identity or security? How do I marry the human, emotive *me* with the academic, scientific ‘*me*’. How do I reconcile these ethical concerns with my humane research objectives? The objectives here, being a desire to seek improvements in conditions and policy, and to provide generalisable knowledge to educators and students of social care in the applied social sciences.

Whether research simply contributes to implementing strategies of good practice, or whether it fundamentally alters relationships, locating the epistemological gaps with regard to ethics should be central to the agenda of the social research scientist.

References

- Holloway, I & Walker, J (2000) *Getting a PhD in Health and Social Care*, Oxford: Blackwell Science
- Le Riche, Pat & Tanner, K (eds) (1998) *Observation and its Application to Social Work*, London: Jessica Kingsley Publishers
- May, T (2001) *Social Research, Issues, Methods and Process*, Buckingham: Open University Press
- Sarantakos, S (1993) *Social Research*, London: MacMillan Press Ltd.
- Veale, A., Palaudaries, L., Gibbons, C. (2003) *Separated Children Seeking Asylum in Ireland*, Dublin: Irish Refugee Council
- East Coast Area Health Board (2002) Annual Report
<http://www.erha.ie/news/1534.php?nCatId=276>
- International Code of Marketing and Social Research Practice (1995)
http://www.iccwbo.org/home/news_archives/1997/advertcodes.asp
- McAuley, K., & Brattman, M (2002) *Hearing Young Voices: Consulting Children and Young People Including Those Experiencing Poverty or Other Forms of Social Exclusion, in Relation to Public Policy Development in Ireland. Key Issues for Consideration*, Dublin: Open Your Eyes to Poverty Initiative
<http://www.youth.ie/research/c6.html>
- Ryan, Lorna (1997) *Researching Minority Ethnic Communities. A Note on Ethics*.
<http://www.ucc/units/equality/pubs/minority/ryan.htm>
- Sociology Association of Ireland
<http://www.ucd.ie/sai/saiethic.html>
- Separated Children in Europe Program (SCEP) (2000)
<http://www.separated-children-europeprogramme.org/global/framed.asp?source=english/goodpractice/booklet/statementgoodpractice.pdf>
- The Standing Committee on the Use of Human Subjects in Research (no date of publication available) *Human Subjects Committee Review Requirements*, Harvard University.
<http://www.fas.harvard.edu/~research/flowchart.html>

Web Enabled Embedded Devices

Brian Myler and Dr. Anthony Keane

School of Informatics and Engineering, Institute of Technology Blanchardstown, Dublin 15

Anthony.Keane@itb.ie

Abstract

The trend in manufacturing of computerised control systems has been to miniaturise the components while increasing the functionality of the systems. This has led to the development of small inexpensive hand-held computer devices coupled with the availability of a user friendly application development language, Java and public cost-effective communication networks has given the developer a programmable web-enabled embedded device. This paper investigates the steps involved in programming the Tiny InterNet Interface platform and analyses the limitations imposed by miniaturisation on this device.

Introduction

Historically we have seen the early stand-alone computational machines quickly gave way to large and expensive central processing computers that allowed many users to run programs and communicate using remotely connected dumb terminals. In time, the microcomputer answered the demand for generalised computing with cheap processing power under the control of the individual. Networking these microcomputers allowed the individuals to communicate via email and to share files, like on the expensive mainframes. The programs and technology protocols that were developed to allow networking of computers has evolved into a global system, called the Internet, where any web-enabled device can participate, Manders et al. (2002). Embedded devices are small computing control systems that have many parts in common with computers like the CPU, memory, circuitry, power source and interfaces, among others. The main differences lie in the limitations of the components and the purpose of the device. Many computers are designed as general purpose machines offering multipurpose usage whereas embedded devices are often designed as stand-alone control systems with a particular simple role, like an alarm system, control of a washing machine, fridge, etc. Advanced embedded control systems are used by car manufacturers, military and manufacturing industry, especially where automation is required.

Today, over 90% of all microprocessors are used for real-time and embedded applications. Manufactures have long recognised the convergence of technologies and communications but have been prevented for exploiting it due to the unavailability of low-cost high bandwidth networks. Also there was no standardisation across the industry of hardware or software development tools. And embedded systems require real-time programming skills which tend to be found on specialist courses for systems engineers. Recently, some manufacturers are starting to provide a simple, flexible and cost effective means to design a wide variety of hardware devices able to connect directly to corporate and home networks by using a combination of a small but powerful chipset with a Java programmable runtime environment.

These platforms of embedded devices can now easily be networked using Ethernet interfaces or using a built-in serial port that can drive an external modem. By allowing them to attach to the Web, client browser screens can be indirectly interfaced with sensors and other embedded systems for management and control. One such communication service is the GSM network where embedded systems can send notification messages via short message service (SMS) or receive data the same way. SMS is well suited for interactions of mobile and ubiquitous computing devices when the input from users can be restricted to simple decisions, confirmations, or input that is based on a set of predefined answers. Sun Microsystems originally developed Java (Arnold et al. 2000) to address the need of a high-level programming language with build-in tools (Application Programming Interfaces) for web-enabled devices. Combining the flexibility of Java with the Web technologies gives the industry and developers a standard set of tools to easily and cheaply create web-enabled embedded monitoring and control systems.

Description of Tini-InterNet Interface (TINI)

Dallas Semiconductor created a microcontroller chipset to provide system designers and software developers with a software development platform to interface with wide variety of hardware devices and to be able to connect directly to networks. TINI is based on the [DS80C390](#) microcontroller which integrates support for several distinct forms of I/O including serial, 1-Wire and Controller Area Network (CAN) bus. The chipset contains flash ROM for the runtime environment and static RAM for system data file storage. The hardware is accessed by the software developer using Java's application programming interfaces while the chipset provide for processing control, communication and networking capabilities, see figure 1.

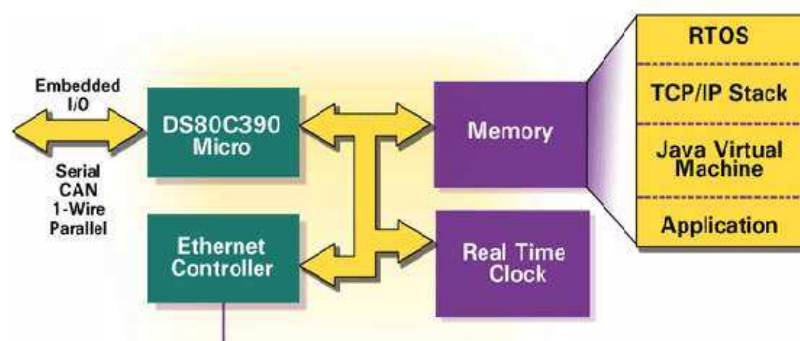


Figure 1: Components of the embedded system

The TINI platform allows everything from small sensors and actuators to factory automation equipment and legacy hardware access to the network. The combination of broad-based I/O capability, a TCP/IP network protocol stack and object-oriented programming environment

enables programmers to easily create applications that provide local and remote control of TINI-based devices through standard network applications such as Web browsers, see figure 2.

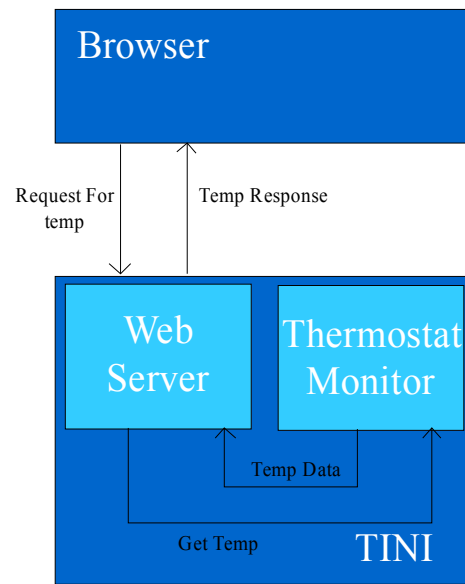


Figure 2: Software components of system

Developing on TINI

Remote Temperature Sensor Monitoring and Reporting System

To investigate the TINI platform, it was decided to build a simple temperature monitoring system that would allow remote monitoring to take place via any browser on the Internet and also for the system to inform the user of exception events, like a temperature above a threshold, using the SMS service. This project involved interfacing several hardware devices to the TINI platform, programming the board to accept the signals, analyse the signals and create an output signal informing of the results.

Configuration of TINI

The TINI platform consists of a €50 chipset consisting of a 33MHz processor, 1MB RAM, and various interfaces including RS232 DTE and DCE, Ethernet 10Mbps, 1-wire and iButton, see figure 3. TINI requires a 6V DC power supply which allows it to be battery operated. Initially, the firmware and operating system are downloaded to the TINI board from a PC using the RS-232 interface. Once this is completed you can give an ip address to TINI and access the board via TELNET and the Ethernet interface.

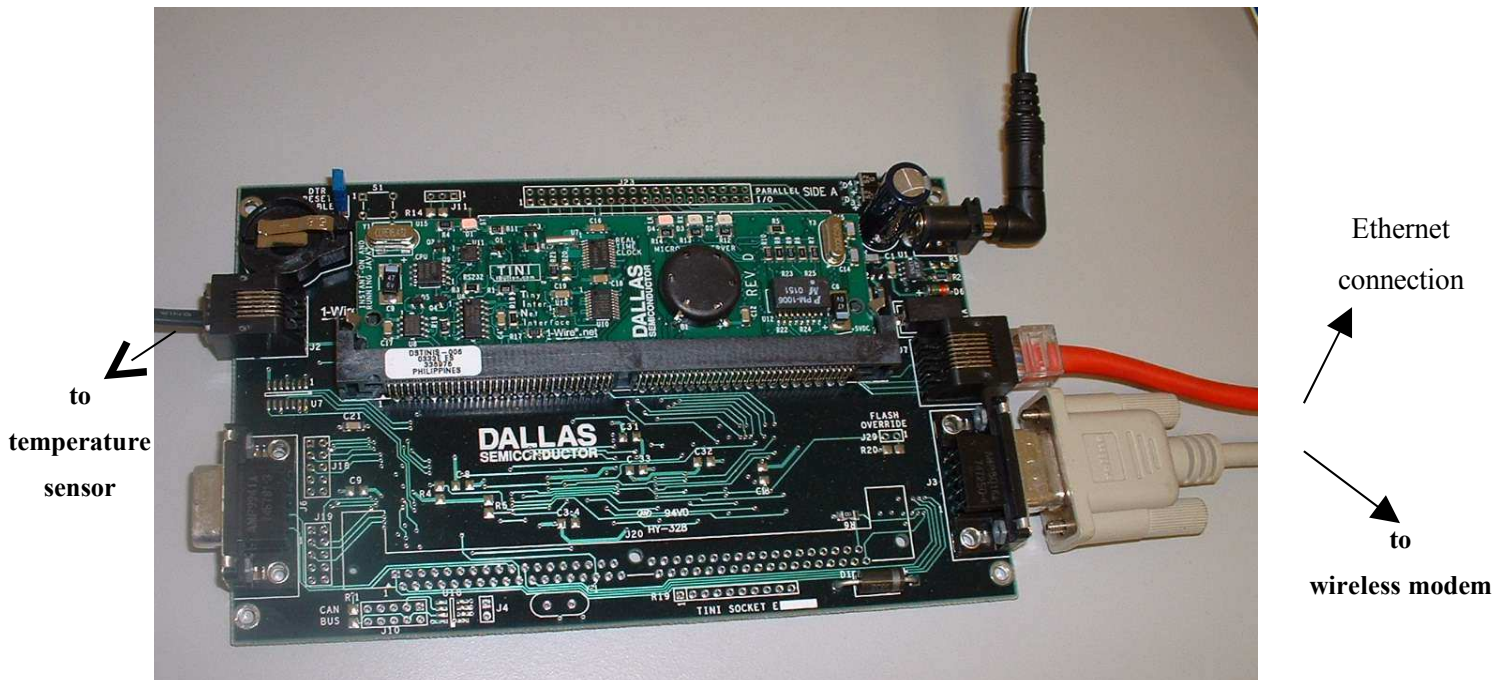


Figure 3: TINI board with attachments

The TINI OS incorporates a FTP server which allows applications to be developed on a host machine and the class files copied to the TINI board for execution. Although TINI supports java, it will not run class files, instead it uses another format called tini files. The difference is a class file will hold the compiled bytecode of a single Java class or interface whereas the tini file can contain several Java classes, the TINI native libraries, and associated resource files.

The tini file is created by converting the Java classes from class and jar files, and copying in other resources. Converting Java classes mainly involves modifying the constant pool. Many UTF8 strings are removed, including the class names that would otherwise support Java's dynamic classloading and reflection features. The memory space constraint imposes the need for compact code thus requiring the support for the bytecode verifier and the debugging information to be removed.

Two utilities provided by Dallas Semiconductors for creating tini files are the TINIConvertor and the BuildDependency. The TINIConvertor is used to build the TINI binary and can be used directly if the program does not depend on any classes that are not included in the standard TINI API, otherwise BuildDependency must be used. This process builds in extra dependencies before calling TINIConvertor to build the binary.

Another way of automating the build process of TINI applications is to use TiniAnt, which is an extension to Apache Ant the cross platform build tool. TiniAnt adds a new task type to ant for building large TINI projects with complex dependencies.

Choosing a Web Server

Servertec Internet Server TINI Edition is a full-featured Application/Web Server written entirely in Java designed to run on Dallas Semiconductor TINI boards. It has the following features that make it a good choice; it is platform independence and uses open standards, gives high performance using multi-threading and has a full-featured Servlet engine. Fault tolerance is provided as crash protecting and recovery technology automatically traps, recovers from and logs exceptions. Additionally Servertec Internet Server TINI Edition protects the integrity of the server environment by preventing exceptions occurring in one request handler from affecting other requests that the server is processing. The other advantages are the size, less than 50KB, and is easy to install.

Figure 4 is a screen capture image showing the client browser screen that is remotely connected with the TINI web server and temperature data from the sensor is relayed. This is an example of remote monitoring. The application running on the TINI board can analyse the temperature to see if it is within a range and activate an exception message if the measured data exceeds a threshold value. The exception report can be sent either by landline connection or wireless via SMS should the fixed line not be available. This is an example of affordable redundancy in communication links.

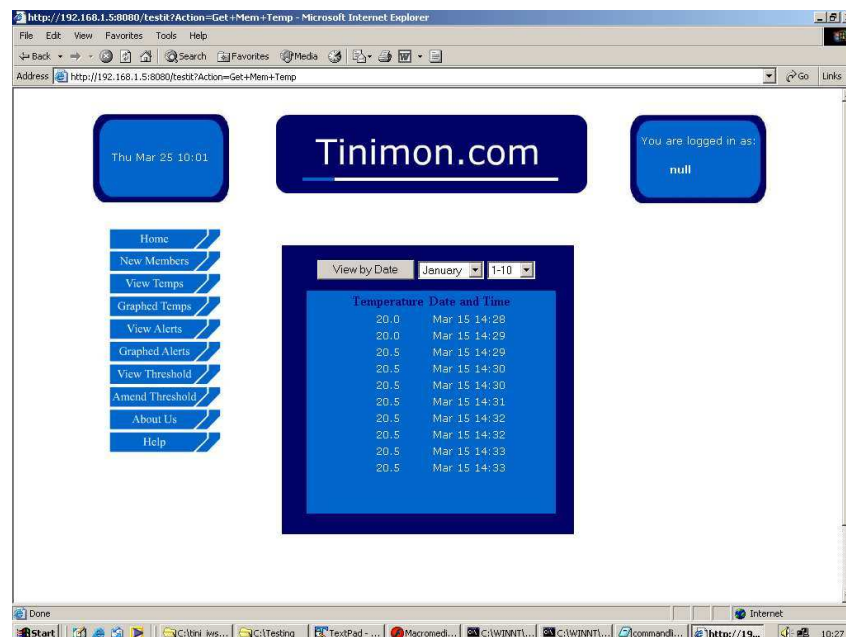


Figure 4: Web browser page

Limitations of TINi

The development limitations of the TINi platform were mainly due to the hardware configuration where the small memory and slow processor speed were easily saturated if multiple signals from different sensors and multiple applications are competing for the resources. Application developers today are used to the generous resources available on PCs and can afford to ignore the bounds of memory and processing power since the typical application's requirements falls short of the limits. When confronted with an embedded system with limited physical resources and only a subset of the JDK API's available, the developer is forced to tighten up their coding design and become more efficient in implementing their application. This can be achieved by building and debugging the application in stages rather than trying the whole thing at once. TINi has proved to be an excellent inexpensive networked development platform where proof-of-concept ideas could be tried out, inexpensively and quickly. Also, the modular nature of the TINi platform could be used to increase available resources by daisy chaining several TINi boards together, with each board being responsible for a different sensor.

Conclusions

The TINi platform's integrated I/O demonstrates the flexibility that web-enabled embedded chipsets with ease in programming using the Java technology can dramatically increase the development cycle. Many manufacturers have followed this approach by providing development platforms of their own, similar to TINi but increasing the capacity of the resources provided on the chipset. The common thread with all the platforms is Java and Web availability. Recent developments in the mobile communication networks have opened a new affordable public gateway to the services that can be provided by and from embedded systems. We have demonstrated the usefulness of having a programmable inexpensive web enabled embedded device that can easily be configured and interfaced to a sensor. Coupling this with the available access using multiple public communication links allows for an effective monitoring, information and control system.

Acknowledgements

The authors wish to thank Declan Barber for the loan of the wireless modem in the testing of this project and Conor Gildea for the useful discussions and helpful advice regarding the configuration of the TINi board system.

References

K. Arnold, J.Gosling and D.Homes

The Java Language, Boston: Addison-Wesley, 2000

M.F.A.Manders, P.J.F.Peters, J.J.Lakkien and L.M.G.Feijls

Taxonomy, Architecture and Protocols for Web-enabled Embedded System

Proceedings of PROGRESS Embedded Systems Symposium 2002

<http://www.ibutton.com/TINI/>

<http://www.maxim-ic.com/TINIplatform.cfm>

Developing Real-Time Multimedia Conferencing Services Using Java and SIP

Gavin Byrne and Declan Barber

Institute of Technology Blanchardstown, Ireland

gavin.byrne@itb.ie declan.barber@itb.ie

Abstract

This paper examines Java's suitability in creating real-time multimedia communications-based applications in Next Generation Networks (NGNs). We investigate some of the current enabling technologies provided by the Java platform which are concerned with the rapid development of real-time communications-based products and services. In particular, we look at creating a multiparty conferencing and collaboration service using the Session Initiation Protocol (SIP) and the JAIN Framework and present an approach which models multiparty conferencing applications by separating signaling and media transfer functionality. We map our model through the design stage to an implementation in Java. This paper is based on real experiences derived from work on an applied research project which is concerned with the development of a collaborative system which allows multiple distributed scientists to graphically analyse a common set of data obtained from mobile sensors in a virtual conference environment. Potential applications areas include education, emergency response services, gaming and any general collaborative application

Introduction

The Internet is steadily becoming more capable of providing real-time media distribution (voice or video) between participants. This functionality represents a significant enhancement to more traditional asynchronous conferencing functionality typified in message boards and chat rooms. One reason for this improvement is the increased bandwidth availability arising from broadband access and the promise of 3G mobile systems such as the Universal Mobile Telecommunications Systems (UMTS). This increase in bandwidth will continue to be a key factor in the increased use of media rich real-time conferencing. Another reason is the increasing support for real-time media transfer and signaling made possible by the trend towards convergence in previously diverse networks. Another third reason is the increasing support for real-time media transfer and signaling provided by open and standard Internet Protocols. Although Java has already been widely adopted for developing high-level business applications which leverage widespread Internet protocols such as HTTP, the complex and proprietary nature of underlying network technologies has meant that the creation of more flexible and granular communications services is still largely the domain of service provider or network equipment vendor personnel using highly specialised languages. This is changing with the emergence of new internet protocols and the Java Intelligent Networks Framework (JAIN), which increasingly allows third party developers to develop and deploy lower level communications services using high-level programming techniques.

This article suggests that Java is an excellent choice for developing end-to-end applications and services for the NGN (Next Generation Network) and will enable third party developers to offer new and innovative services independently of service providers. We describe our

approach to the creation of one such service using Java and the JAIN Framework and make observations on our experiences that may be more generally applicable to a range of other created services. This paper makes certain assumption:

- Bandwidth availability will steadily increase from end-to-end and decreasingly represent a technical constraint
- Convergence in the NGN will be based firmly on open standards and the TCP/IP protocol stack in particular
- Applications leveraging existing and emerging Internet Protocols will dominate

The JAIN Framework

The objective of the Java Intelligent Networks Framework (JAIN) [3, 5] is to provide service portability, convergence and secure access to integrated networks. JAIN builds on Java portability by standardizing the signaling layer of the communications networks into Java language and defining a framework in which services can be created, tested, and deployed. The JAIN initiative brings new opportunities for both developers and service providers, enabling them to create services (without rewriting) for the different network implementations and interfaces in a multi-vendor environment. The JAIN initiative proposes to do this by specifying Application Programming interfaces (APIs) which provide access to functional objects such as Call Control or User Location Objects as well as underlying signaling objects (e.g. SIP, H.323 and MGCP). This essentially allows application developers easy access to functional and signaling interfaces. In the context of our multiparty conferencing service, the direct access to signaling using SIP was a critical feature in service development.

Fundamental Design Issues

Our analysis of design principles and functional requirements is summarized in the following description of our conceptual model. Our conferencing system had the following general design aims:

- **Scalability** in the number of (distributed) users
- **Efficiency** in its use of network and node resources
- **Simplicity** in the entire service creation process and especially implementation
- **Extensibility** for the easy and rapid deployment of modifications or new services
- **Interoperability** of applications and underlying services based on NGN models.

Specific conferencing functional requirements include:

- Peer-to-peer and multiparty conferencing in a distributed NGN environment.
- Dynamic and flexible conference signaling and media transfer

- Real-time Voice/Video Support: it must allow effective real-time conferencing.
- Application platform independence

The adoption of open and standard internet protocols for real-time media transfer and signaling (specifically the Real-Time Protocol, RTP, the Real-Time Control Protocol, RTCP, and the Session Initiation Protocol, SIP) combined with the convergence of public and private networking technologies towards the internet protocol stack in general ensures that most of our design principles are achieved. From an application developer's perspective, these protocols are highly accessible through Java and its associated Application Programming interfaces (APIs). The two main functions within a conferencing system are Signaling (for call establishment, the addition/deletion of call parties and call termination) and Media Transfer. For simplicity and efficiency, our design further separates the signaling functionality into User Registration and Call functions. Consequently, our preferred approach uses a hybrid signaling architecture (Figure 1): a centralized signaling architecture supports user registration and directory support while a distributed signaling architecture allows peer-to-peer and peer-to-multi-peer calls. A distributed architecture supports peer-to-multiper media delivery. This is scalable for large numbers of participants and makes efficient use of bandwidth and processing.

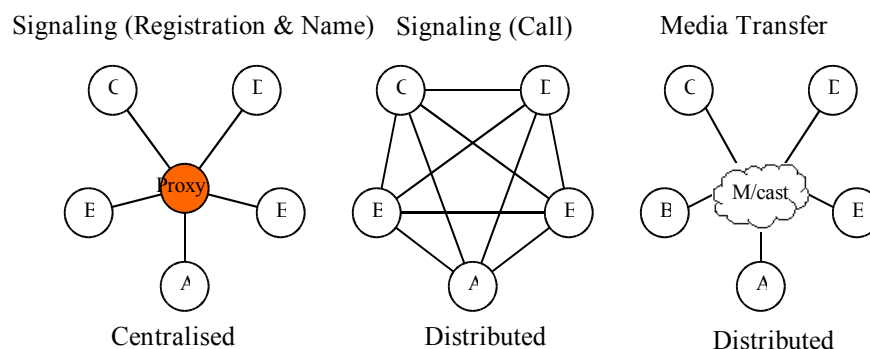


Figure 1: Conferencing Model based on Unicast Signaling and Multicast Media

Evaluating Java Technologies for Multiparty Conferencing

Java strongly supports the development of new applications and services which leverage the functionality of the new powerful Internet signaling and media transmission protocols. Java's platform independence, downloadability, mobility and object oriented structure has already led to its adoption for use in telecommunication applications and is destined to play a critical role in the development of internet-based electronic commerce systems. Java provides support for a wide variety of internet protocols such as HTTP (within applet/servlet packages), SIP (within JAIN Framework), RTP (within JMF package), IP (within net package), which allow development of inter-networked applications. The Java API's of most importance in creating multimedia conferencing applications are:

The Java Network package (java.net.*): Through the java.net package, Java provides the ability to create both unicast and multicast sockets for the transmission and receipt of data. The ability to create multicast sockets will be an advantage in our conferencing and collaboration application where identical data is being sent to multiple recipients as multicast is far more bandwidth and processor efficient (and therefore scalable) than having to open up multiple unicast sockets for the same data.

The Java Media Framework (including javax.media.*, javax.sound.*): The Java Media Framework (JMF) is a set of Java APIs for developing multimedia applications. In this project JMF provides the necessary methods for the transmission and receipt of real-time media streams using Real Time Protocol (RTP) and the Real Time Control Protocol (RTCP).

The Java Intelligent Network Framework (including javax.sip.*, javax.sdp.*): The Java Intelligent Network Framework (JAIN) includes a set of Java technology based APIs which enable the rapid development of Next Generation communications-based applications and services on the Java platform. By providing a new level of abstraction and associated Java interfaces for service creation across point-to-point, circuit-switched (PSTN, ISDN), and packet/cell-switched (X.25, Frame Relay, ATM) networks. Importantly for us, JAIN provided the only practical means of accessing the signaling using a high-level language and of separating the registration aspect of the signaling from the call-establishment.

Session Initiation Protocol (SIP)

SIP [9] is an application layer signalling protocol which provides call set-up, modification, and termination, as well as other services. Importantly, participants can communicate using multicast, unicast or a combination of both. As an application layer signaling protocol used in a distributed architecture, SIP is best suited to meet our scalability, real-time, simplicity and extensibility design requirements. The Session Description Protocol (SDP) is used in conjunction with SIP for exchanging session capabilities (ability to send/receive audio or video, supported codecs, etc.).

We believe that SIP will be the protocol of choice for Next generation Networks and we have chosen SIP to develop our multimedia conferencing application because of its easy integration with existing IETF protocols, simplicity, mobility, scalability, ease of development, extensibility and deployment in the core and at the edge of the enterprise and support for multicast, unicast or a combination of both (all documented and discussed in [4, 8, 9, 10]).

From Analysis to Design

In order to map our Unicast/Multicast model to a design and implementation we have used Sties and Keller's Independent Service Description Model [11] which consists of abstract descriptions for endpoints, communication channels, and communication relations. This is an abstract service model that allows the standardized description of the specific characteristics of a service while abstracting from any network and implementation dependant details.

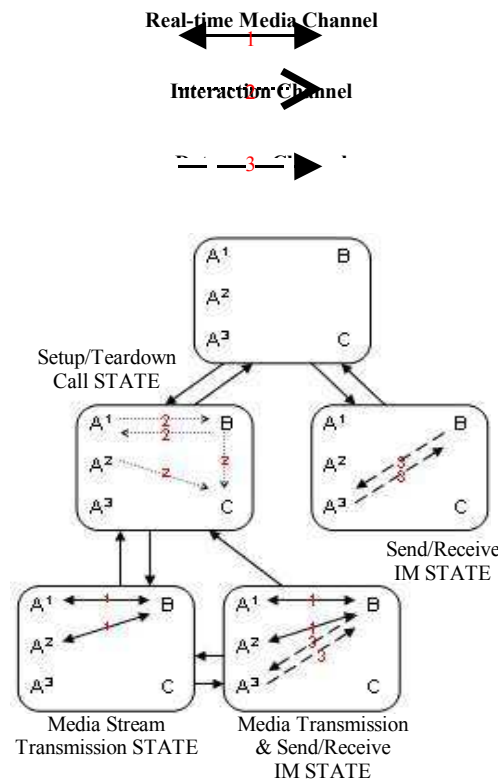


Figure 2: Finite State Machine

This approach begins by determining the functions of the service from the user's point of view: peer-to-peer call, conference call or instant messaging (labeled as A^1 , A^2 and A^3 respectively in Figure 2), and then determining the endpoints needed within the service: a conference server (labeled C) and another human participant (labeled B). The next step is the identification of communications channel types. There are three identified in our service, as shown below. The final step is to identify all Single Communication Relations, i.e. Endpoint - Communication Channel relations that might occur in our service (identified as the states in Figure 2). We then summarise the Single Communication Relation States found into an FSM (Finite State Machine that summarises our system).

We now employ the State pattern described in [13] which can be used to translate an FSM into a set of high level abstract classes and relationships. We begin with a class named UserAgent,

which creates and manages ‘manager’ classes for each of the areas of functionality which we can now identify using our FSM (call setup/teardown, media transfer, and Instant Message exchange). The UserAgent class passes events and method calls to the appropriate manager to deal with, then simply updates a global state variable and awaits further events. Figure 3 illustrates these classes in a UML diagram.

Implementation

From our analysis and design, we have identified what states our SIP user agent can be in at any one time, as well as what classes and objects need to be implemented. Our SIP user agent is implemented as a pure Java applet, which is digitally signed to enable the applet some permissions which are usually outside the Java security sandbox (such as access to hardware like microphones and web cams), allowing it to be used in thin web clients.

The key features implemented in our user agent thus far are the ability to **A)** Register with a SIP proxy/registrar server, **B)** to make voice and video calls, **C)** to send and receive Instant messages (Interoperable with MSN Messenger, etc.), **D)** to add contacts to a buddy list with presence capabilities (which informs the user when contacts are on or offline), **E)** to make conference calls using unicast call signaling to the conference server and multicast media transmission, and **F)** to collaboratively analyse remote sensor data stored in a web server database through graphs and charts with other users in either conference calls or peer-to-peer calls allowing users to highlight (by drawing on the graph) interesting findings which is replicated to all interested parties.

Through the JAIN SIP 1.0 and JAIN SDP APIs we have been able to harness the simplicity of SIP for call signaling, as well as buddy lists [6, 7] and Instant Messaging [1] (using the SIP MESSAGE and SUBSCRIBE extensions) in our conferencing system. The user agent and conference server are developed on top of the publicly available NIST (National Institute of Standards and Technology) pure java sip stack implementation [12].

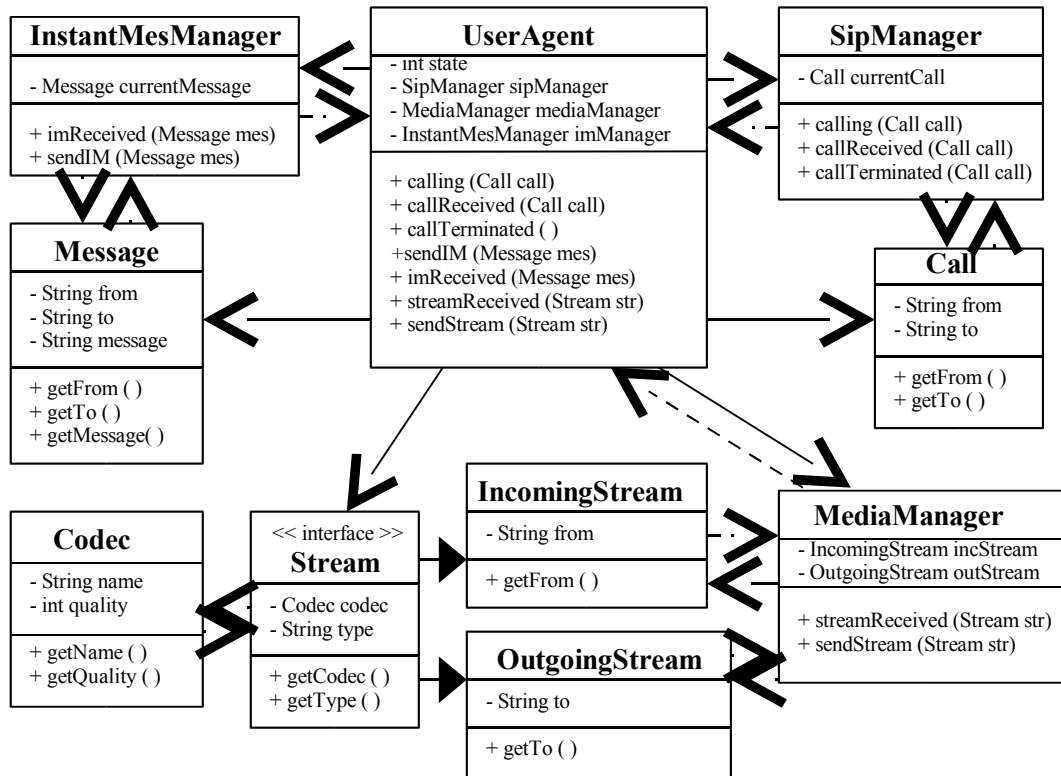


Figure 3: UML Class Diagram

Instead of developing our own SIP proxy server at this stage, we are using a proxy server which was freely available, again from NIST.



Figure 4: Implementation Screenshots

Users wishing to join the conference can simply send a standard SIP INVITE message to the conference server which in turn can choose to authenticate the user or simply send an immediate ACK reply to set-up the call. Users currently participating in a conference who would like to invite other users into the conference can send a SIP REFER message with the URI of the conference server, inviting them into the call (this REFER message could alternatively be sent to the conference server with the intended recipient's URI). Each

participant makes a normal peer-to-peer SIP call to the conference server using unicast signaling to register initially or to resolve a SIP name to an IP address. A unicast call is then made to the peer. Once the call is established, media is sent and received on a multicast connection. Other registered peers can be invited or proactively join the established multicast conference. The role of the conference server is to act as the centralized manager of the conference, and to maintain a signaling dialog with each participant in the conference

Summary of Observations on Using Java & JAIN for Conferencing services

Strengths and limitations of using Java technologies for conferencing service creation include:

- o Java's ability to send and listen for data on multicast sockets was a major benefit.
- o The JAIN framework provides a solid set of APIs for the implementation of new communications-based services and will increasingly enable developers to rapidly create and deploy new services without specialist knowledge.
- o The real-time streaming abilities provided by the JMF are slow to initialize and begin streaming, but once started provide a perfectly adequate solution. It compares poorly to similar Java telephony applications which make use of native code for access to hardware, for example the DLL (Dynamic Link Library) for the windows platform. These manage almost instantaneous responses. The use of native code in our clients would contradict our platform independent model, but in order to provide realistic response times, we may have no choice.

Conclusions

- o Java's platform independence, downloadability, mobility and object oriented structure has already led to its adoption for use in telecommunication applications and is destined to play a critical role in the development of internet-based electronic commerce systems. But it is Java's ability to leverage existing and emerging internet-based open protocols, and the fact that it is now enabling third party developers with little or no knowledge of underlying network infrastructures to develop and offer new services independently of service providers, that will no doubt make it the language of choice for developing applications and services for the NGN.
- o The ability to replicate multimedia conferencing capabilities currently available in the circuit-switched environment will not be a sufficient driver for the enterprise to adopt IP based multimedia conferencing. The ability to rapidly develop and deploy enhanced, and previously unavailable, services which leverage IP and related intelligence will be the driving force behind the evolution of NGNs. Java and the JAIN framework with its power to implement emerging internet protocols such as SIP is a key catalyst in this evolution.

References

- [1] B. Campbell et al., "SIP Extensions for Instant Messaging", IETF DRAFT, work in progress, 2002.
- [2] Sun Microsystems, "JAIN Service Creation Environment (SCE) API Specification".
- [3] Sun Microsystems, "Java Advanced Intelligent Network, The JAIN API's".
- [4] I. Miladinovic, J. Stadler, "Multiparty Signalling using the Session Initiation Protocol"
- [5] P. O'Doherty, M. Ranganathan. "JAIN SIP tutorial".
- [6] J. Rosenberg, "The Future of SIP and Presence", 2003.
- [7] J. Rosenberg et al., "SIP Extensions for Presence", IETF DRAFT, work in progress, 2001.
- [8] J. Rosenberg, H. Schulzrinne, "Modles for Multi Party Conferencing in SIP", IETF DRAFT, work in progress, 2001.
- [9] H. Schulzrinne et al., "SIP: Session Initiation Protocol", IETF DRAFT, November 2000.
- [10] H. Schulzrinne et al., "Centralized Conferencing using SIP", IETF DRAFT, November 2000.
- [11] P. Sites, W. Keller, "A Generic and Implementation Independent Service Description Model".
- [12] National Institute for Standards and Technology, "SIP Reference Implementation", <http://snad.ncsl.nist.gov/proj/iptel/>
- [13] R. Martin, UML Tutorial: Finite State Machines, Engineering Notebook Column, 1998.

Convergence Technologies for Sensor Systems in the Next Generation Networks

Conor Gildea and Declan Barber

Institute of Technology Blanchardstown, Ireland

conor.gildea.byrne@itb.ie declan.barber@itb.ie

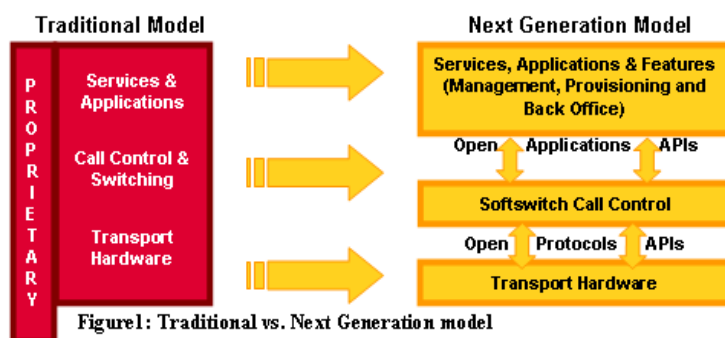
ABSTRACT

This paper describes an approach to the internetworking of sensory nodes in a converged network environment. This preliminary investigation of sensory network creation is driven by a joint applied research project which seeks to establish the feasibility of the real-time remote monitoring of animal welfare while in transit between Ireland, Europe and the Middle East. This paper examines the use of Java to create sensor services in converging architectures which leverage the Internetworking protocols and describes our implementation of such a system.

Keywords: Java, NGN, Converged Services, Sensor Networks, SIP, SLEE

1. INTRODUCTION

Traditional centralized static models have been applied to intercommunication between relatively unintelligent sensor nodes and intelligent management stations. Recent trends are making it increasingly feasible to move away from this centralized model to a more distributed one, even to a point where a mobile sensor network could be considered as an ad hoc network of autonomous nodes.



The impact of Moore's Law has led to the concentration of greater processing power, memory and storage (and consequently increased levels of intelligence) on small devices. The Internet, static switched and mobile networks are converging around the TCP/IP model and Internet protocols are providing a framework for the deployment of new applications and services across this converged space. Internet Protocol (IP) enables the internetworking of disparate network nodes by providing a standardized addressing scheme and path determination techniques. Higher layer mechanisms can provide reliability, signaling and quality of service support. One potential outcome of these trends is the repartitioning of capabilities and responsibilities within a sensory network to a more distributed model as intelligence spreads outwards from the centre to the edge of the network. Sensory nodes can now be independent computing platforms capable of peer-to-peer communication and of interacting with interim network nodes in order to provide previously unavailable services. These could operate across

a global inter-network and even between non-heterogeneous sensing nodes. Java provides a platform independent application development environment and network operating environment for code mobility and increasingly provides APIs and frameworks for internet protocols implementation and telecommunications and internetworking support. Although limitations still exist in the intelligent services supported across the Internet, new protocols and services are emerging which address these shortcomings. This paper seeks to discuss the potential impact of these developments on sensor networks.

2. Modern Embedded Systems

An embedded system (*Figure 2*) is a device that contains programmed logic on a chipset that is used to control one or more functions of the device. It usually has more limited computational power, storage and interface functionality than a desktop platform.

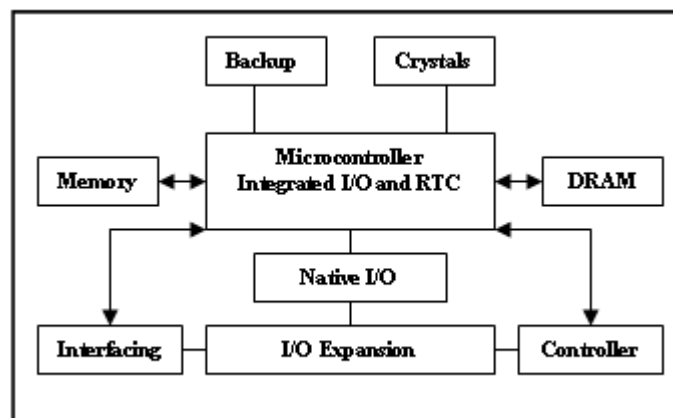


Figure 2: General architecture for embedded system

A real-time embedded system is often required to provide deterministic performance, often in a mission critical environment. This real-time behavior of embedded systems service logic is normally *event-driven* rather than conventional enterprise behavior. In comparison to enterprise computing, embedded systems use relatively thin components and perform lightweight asynchronous transactions at reasonably high frequencies. Unlike typical enterprise system business application logic, the most computationally intensive activities in embedded systems are generally more related to input/output (I/O) operations than database access related operations. The trend towards convergence in the networking world indicates that embedded systems will increasingly operate in a pervasive computing environment using wired and wireless transport technologies with IP connectivity for communication. General-purpose embedded systems with TCP/IP stack support have already emerged. Even with modest processing power, these can provide powerful event-driven distributed architectures with better functionality, scalability, availability, performance, manageability and security characteristics than ever before.

2.1 Suitability of Java

Java is a relatively new language that has much to recommend it over other high-level programming languages, including: ***Simplicity***: Some of the more difficult aspects of programming in higher-level languages, such as the need to use memory pointers and manage garbage collection, have been removed. ***Platform Independence***: In principle, Java supports 'write-once-run-anywhere' development using the idea of a Java Virtual Machine (JVM). ***Object Oriented***: Java is designed to be object-oriented throughout and has an extensive class library available in the core language packages. ***Multi-Threading***: Lightweight processes, called threads, can easily be spun off to perform multiprocessing and can take advantage of multiprocessors where available. ***Robustness & Dynamic Binding***: Exception handling is built-in and strict data type checking is enforced. Local variables must be initialized. In Java, the linking of data and methods to where they are located is done at runtime. ***Security***: Java is more secure as no memory pointers are used; a program runs inside the virtual machine sandbox. The security manager determines what resources a class can access such as reading and writing to the local disk.

2.2 Java Support for Internetworking

Many standard extensions and class libraries are provided in Java which supports the development of socket-based, client-server based and other distributed applications. Features like ports and servlets permit the rapid development of network applications based on Internet protocols while Remote Method Invocation (RMI) and the Java Messaging Service (JMS) support more complex distributed systems and can be used to effectively leverage distributed computational power to complete more complex tasks.

3. The Java Intelligent Networks Framework

A key enabler in rapid application and service development and deployment is the availability of open and standard APIs that span NGN technologies but that abstract from the specifics of underlying data transport technologies. The JAIN framework is an extension of Java, and specifies a number of open and extendable Java technology APIs that support the rapid development of Next Generation communication-based products and services on the Java platform.

Although JAIN is primarily a specifications framework, it has also provided a number of reference implementations that allow developers to access communications functions such as Call Control, Mobility Management and User Interaction in a signaling protocol-neutral way

while also providing more granular access to the underlying signaling protocols if needed all through high-level programming techniques. JAIN also defines a service creation environment and explicitly defines a Service Logic Execution Environment (SLEE).

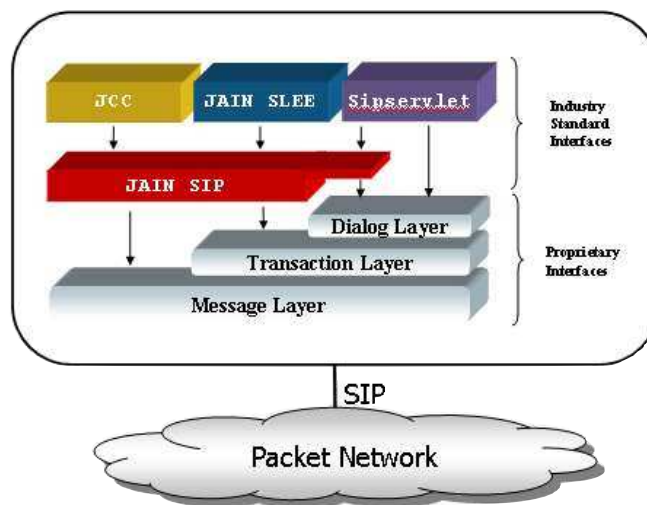


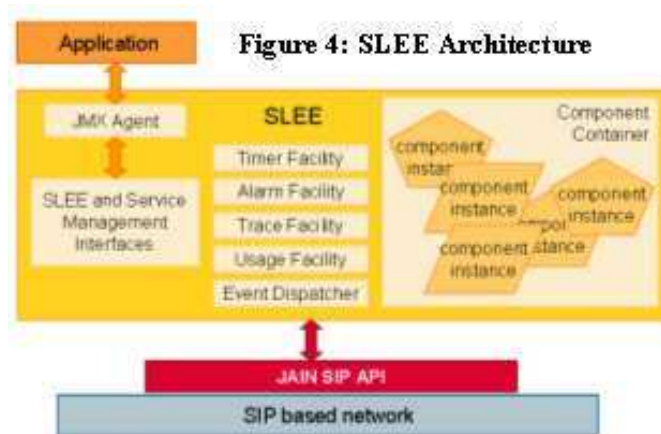
Figure 3: JAIN and SIP

SLEE is tightly mapped to event-driven services such as a Call Control, Alarming or other automated service and is eminently suitable for the execution of service logic and signaling in specialized event-driven engines. While SLEE is still at the specification stage, JAIN reference implementations for protocols such as SIP signaling moves Java into the carrier grade telecommunications domain today and promises much more for the future. From the perspective of embedded systems, the flexibility of the JAIN SIP API specification and the small size of the SIP toolkit identify this as a potentially rich application layer signaling solution for mobility management and other services. JAIN remains very much a work in progress, although we have been able to achieve a considerable amount using the SIP and Other APIs, nevertheless aspects like performance, stability or failure situations handling have to be addressed and investigated properly before attempting to port the solution to embedded platforms.

3.1 Rationale for SIP

SIP is a signaling protocol, however, and is not designed to be used to transfer data/media. It normally relies on RTP/RTCP to transfer the media. This separation of signaling from the data exchange is an important characteristic that will make it possible to use different paradigms and modes of operation for signaling, control messaging and data transfer as appropriate. In typical SIP operation, SIP signaling is used to establish a transfer session, the session characteristics are negotiated between the end devices using SDP and the media is transferred using RTP/RTCP.

Simplicity: An important attribute for any protocol is that it must be simple to provide value added services. SIP enable service providers to rapidly deploy new services without lost in complex implementations. **Scalability:** SIP is a very scalable protocol. It works from end-to-end across the LAN and the WAN. It does not rely solely on multicast/broadcast technologies to reach a destination endpoint. It has a strong concept of routing (leveraging HTTP routing) which enables a packet to traverse from source to destination using intermediate existing routes; hopping from one node to another till it reaches its final destination. Further SIP can operate on both UDP and TCP which allows SIP based servers to scale well. **Flexibility/Extensibility:** In SIP it is very simple to add extensions to support new features. The protocol is defined in a way that any provider can define extensions easily to the existing grammar set to add features which may not exist in the core SIP specification. **Registration/Location:** In SIP it is not necessary that a calling device needs to know exactly where to locate the called device. A device registers its current location with a management node. **Security:** SIP provides both authentication and encryption to provide end-end security.



Event Notification: SIP has been extended to introduce SUBSCRIBE and NOTIFY messages which enable elements to “subscribe” to certain events and can be notified when they occur.

Unicast/Multicast Support: when used in conjunction with the Session Description Protocol (SDP), a separate protocol designed to negotiate session parameters, SIP can establish connections which will use unicast or multicast delivery.

3.2 JAIN SLEE

JAIN SLEE is high performance event processing platform suitable for event driven applications. It supports both simple and complex telecommunications applications. The SLEE framework is independent of underlying networks and it portable, robust and allows for reusable applications. (see *Figure 4*).

3.3 Embedded System Support

The original Oak language from which Java derived was intended for embedded applications. The combination of platform independence and the adaptability of Java that allows it to work on micro-sized platforms by shedding non-essential code make it suitable for developing embedded system logic for a wide range of devices.

Three main approaches currently exist for developing embedded systems applications: J2ME, Embedded Java and Personal Java. The J2ME is a Java platform targeted at consumer electronics and embedded devices. It consists of a Java Virtual Machine (JVM) and a set of APIs for providing a complete runtime environment for the target device. The J2ME technology has two primary kinds of components: configurations and profiles. A configuration is composed of a low-level API and an optimized JVM. Two configurations are available: **Connection Limited Device Configuration** (CLCD), which is targeted at environments where 128-512Kb of memory is available for the Java environment and applications and **Connected Device Configuration** (CDC), which is targeted at environments, where more than 512Kb, usually about 2Mb, of memory is available for the Java environment and applications. EmbeddedJava includes tools that allow developers to configure and compile runtime environments that contain only those fields and methods necessary for a particular application's needs. Developers can use EmbeddedJava for a variety of products, including process controllers, sensory networks, instrumentation, office printers and peripherals, and networking routers and switches. PersonalJava is an upward-compatible subset of Java dedicated to consumer and embedded devices, and specifically designed for building network-connectable consumer devices for home, office, and PDA use.

3.4 Java Real-time Development

The Java Technology Model with Real-Time Extensions further leverages the capabilities of a real-time operating system (RTOS) to achieve the promise of hard real-time computing. By coupling the Real-Time Java Virtual Machine with an RTOS and giving the developer new mechanisms to separate hard real-time and soft real-time threads, objects and memory, Java will be capable of addressing the requirements faced by real-time and non-real-time applications. The extensions do not introduce new keywords or make syntactic extensions to the Java programming language, allowing the developer to utilize current tools. Real-time systems are found in embedded applications as well as other applications that require a deterministic time behavior. RTSJ was developed by specified for development by the Java Community Process (JCP). This extension package targeted a number of areas that needed to be addressed for Real-Time applications. These areas are; ***real-time threads, asynchronous***

events, interruptible non-blocking I/O, access to physical memory, scheduling, garbage collection handling and timers. Java real time aspect is still very much a work in progress. Most embedded VM and hardware vendors seem to focus their efforts on J2ME and have no plan to implement RTSJ. RTSJ reference implementation is only a partial implementation and is not suitable for serious real-time applications.

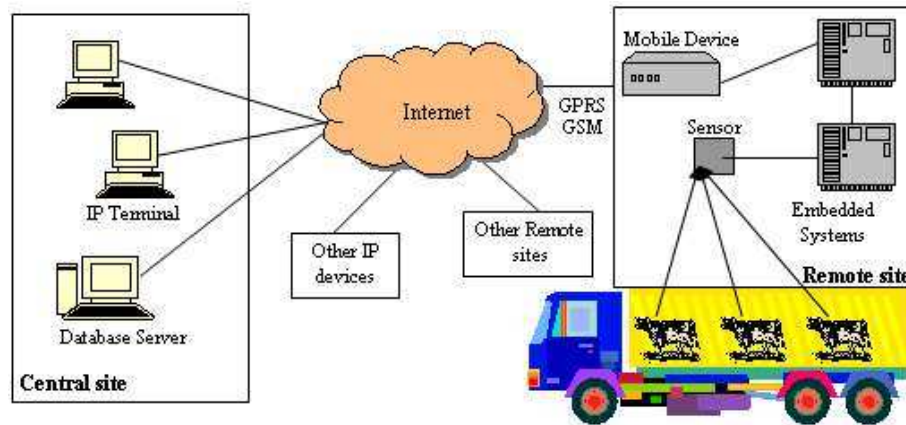


Figure 5: Applications for Embedded Systems

4. ANALYSIS

Our applied task is to create of remote sensor monitoring system that allows multiple independent sensor platforms to be monitored and be interrogated from any Internet station. The purpose of the system is to monitor the welfare of livestock in transport by monitoring the animal's heart-rate, temperature and environment factors. This system is composed of a number of embedded devices each connected to a specific sensor or a mobile modem for connection to the outside world. The sensors are responsible for monitoring ambient and internal values. The remote monitoring system needs an embedded function to report back to a central site for permanent storage of data and scientific analysis. This reporting is going to be *event-triggered*. An example of such an event would be if the sensor reading is outside a defined threshold or if the data file size has exceeded a predefined size. The collected data is sent back to the central site for permanent storage and analysis. Once the data is stored at the central site, the scientists are free to examine the data further and are able to correlate locational information with stress levels.

5. IMPLEMENTATION

The current system topology is composed of at least two sites. The remote site is made up of a number of a number of IP nodes interconnected using Ethernet for LAN based connectivity and GPRS for WAN based connections. The remote site system is made up of a number of Java enabled embedded systems. The next section outlines a breakdown of the main core components.

5.1 Communication among embedded systems:

The communication between the embedded systems is carried out using SIP J2ME. SIP is used to maintain data convergence among each of the connected devices, through the use of Instant Messages which are exchanged on timed or triggered bases. These Instant Messages are also routable outside the LAN through the use of public proxies. The goals of SIP J2ME are; it enables terminals supporting CLDC to run SIP enabled applications, it is firmly build and the CLDC Generic Connection framework. Another important factor is the API size small and to keep the number of created objects is at a minimum. This is very important in an embedded environment when memory and processor power is at a premium.

5.1.1 Instant Message Design Example

Instant messaging is defined as the exchange of content between a set of participants in real time. We will consider short simple textual messages only. Although forms of Instant

Message F1	Message F4
<pre>MESSAGE im:user2@domain.com SIP/2.0 Via: SIP/2.0/UDP user1pc.domain.com From: im:user1@domain.com To: im:user2@domain.com Call-ID: asd88asd77a@1.2.3.4 CSeq: 1 MESSAGE Content-Type: text/plain Content-Length: 30 [1,105041071103,16.62,0,11,23]</pre>	<pre>SIP/2.0 200 OK Via: SIP/2.0/UDP user1pc.domain.com From: im:user1@domain.com To:im:user2@domain.com;tag=ab8asdasd9 Call-ID: asd88asd77a@1.2.3.4 CSeq: 1 MESSAGE Content-Length: 0 Note that most of the header fields are simply reflected in the response. The proxy receives the response, strips off the top Via, and forwards to the address in the next Via, user1pc.domain.com and the result message is F4</pre>

Table1: Instant Message exchange

Messaging have been in existence within intranets and IP networks for quite some time, most implementations are proprietary and there is no Internet Application protocol specifically designed to support this function. Messaging between the nodes of a real-time mobile sensor network could be considered as an Instant Messaging application. Such an application could be implemented by using SIP but without requiring the establishment of a call. There is currently a proposal to extend the SIP specification by adding a new MESSAGE method. This method supports both the addressing and the transfer of any MIME type content between nodes but does not require prior call establishment. A MESSAGE request may traverse a set of SIP proxies using a variety of transport mechanism (UDP, TCP) before reaching its destination. The destination for each hop is located using the address resolution rules detailed in the SIP specifications. During traversal, each proxy may rewrite the request address based on available routing information. This method leverages Routing like functionality (the pre-pending of proxy information in this case) to provide a reply path. Provisional and final responses to the request will be returned to the sender as with any other SIP request.

5.1.2 Data Retrieval from Sensors:

The system is made up of a number of sensors which are primarily connected using open standard interfaces, which lead to the creation of a general package (API) for sensor reading and manipulation. The package is used for sensor polling, data storage and compression. Each of these attributes can be configured through the use of configuration files.

<pre># serial port - that is connected to the sensor serial=serial0 # read for new data(specified in minutes) read=1 # mechanism for exchange exchange=tftp,ftp,tcp # filename containing compressed data tftp_file=test.zzz</pre>	<pre># log file that will contains the readings log=duck_info.dat # mechanism used for compression compression=lzw,readings,ascii_table # address of tftp server tftp_server=10.1.1.100 # send data to server after x concurrent reads send=7 # sensor1 threshold threshold 37.0</pre>
--	--

Table2: Configuration example from sensor package

Once the data has been collected from the sensors it is converted into a generic PDU (protocol description unit). This generic PDU is very important because it means that all the data is in a common form and data processing is greatly reduced making it easier when the data is being permanently stored in the database at the central site. The PDU string contains not only the sensory data, but also a lot of meta-information about the sending device.

The PDU is in the form of hexa-decimal octets or decimal semi-octets. The PDU is encapsulated inside a SIP instant message. The PDU enables data homogeny and the SIP instant messaging provides application layer signaling.

The combination of these two components enables the rapid developed of sensor based networks, which is stable, secure and extensible once the device has conformed to the PDU format.

Octet(s)	Description
07 34 56	CRC (cyclic redundancy check)
99 30 92 51 61 95 80	Time stamp (semi-octets)
00	Sensor type: alarm/response/update
9B	Sensor ID/Sensory interval
FD	Data coding scheme/Data compression algorithm (00 - default coding/no compression)
0A	Length of payload.
E8329BFD46979EC37	Sensor data: 8-bit octets representing 7-bit data

Table 3: Sensor PDU breakdown

Code-Listing1: Sensor API code snippet

```

Sensor.S1.activate();
Sensor.S1.addSensorListener (new SensorListener() {
public void stateChanged (Sensor src, int value) {
LCD.display(value);
int data_changed = src.readSensorValue();
}    public boolean passivate()
    {
        return Sensor.S1.passivate();
    }
});

```

The interval delay for polling for new data is set in the configuration descriptor as well as the defined threshold for that sensor and when the *readSensorValue()* is called the threshold levels are checked and if the threshold is breached then an *alarm type* Sensor Type is generated. If its reading is within the defined level then a *response type* is generated and the reading is logged. The final case is if the next reading has the same sensor reading then an update type is generated and the reading is logged.

6. CONCLUSIONS

We are entering a new phase in the development of distributed sensory networks. Embedded processing is becoming powerful enough to tackle an ever-widening range of applications. Wireless and wired networking is becoming ubiquitous, cheap, and low-power so that we can envision interconnecting all our embedded processors. Modern embedded systems can support either partial or entire TCP/IP stacks and will be inter-networked over the NGN using Internet protocols. This means that more distributed architectures will be possible in the area of mobile sensor networks, by leveraging emerging low-level access internet protocols such as SIP. The sheer breadth of scope of the Java initiative seems set to encompass the previously disparate areas of open-standard internetworking, embedded systems, real-time logic and high-level service development making it a key ingredient in a converging technologies world. Another advantage of Java is that it is becoming increasingly possible to develop new sensory based services quickly which were previously in the realm of the network operators. ***An embedded device is becoming just another IP node on a mobile network.*** IP Communication and open protocols are the crux of NGN's, investment and converged applications. Java provides convergence at the application level and it is increasingly possible for non-specialist 3rd parties to create value added sensor system services without detailed knowledge of the underlying network complexities.

7. LOOKING FORWARD

We are currently working on data retrieval using a SIP based call from the sensor network to the central site and to further leverage the capabilities of Instant Messaging enabling them to be routed through public proxies to be received on any IP enabled device running the SIP stack. We are also investigating the feasibility of the transfer of Real-time video across the GPRS backbone encapsulated in a RTP (real-time protocol) socket from the sensor network to the central site.

8. REFERENCES

- [1] Arjun Roychowdhury & Stan Moyer, Instant Messaging and Presence for SIP Enabled Networked Appliances, 2002.
- [2] Wolfgang Kellerer, Intelligence on Top of the Network: SIP based Service Control Layer Signaling, 2002.
- [3] O'Doherty, Java Technology for Internet Communications, 2003.
- [4] M.Satyanarayanan, Pervasive Computing: Vision and Challenges, 2001.
- [5] M.Satyanarayanan, Pervasive Computing: Vision and Challenges, 2001.
- [6] J.P Martin-Flatin, Push vs. Pull in Web-based Network Management, 1999.
- [7] Johannes Stadler, A Service Framework for Carrier Grade Multimedia Services using PARLAY API's over a SIP system, 1999.

Implementing Test Patterns to Dynamically Assess Internet Response for Potential VoIP Sessions between SIP Peers

Declan Barber, Gavin Byrne & Conor Gildea

Institute of Technology Blanchardstown,

Declan.barber@itb.ie

Abstract

The capability of VoIP to provide internet telephony is limited by the lack of homogeneous quality of service (QoS) mechanisms in the Internet. Whereas approaches which reserve QoS resources will work well in an end-to-end managed environment, they are not automatically suited to the heterogeneous nature of the Internet. It may be possible to adopt the 'chirp-sounder' approach uses in establishing the optimal frequency channel for a high frequency (HF) radio transmission which dynamically samples a range of possible transmission channels and uses the echoing of a an established test pattern to ascertain the quality of the potential links. The optimal 'channel' can then be selected for transmission. By repeating the process at intervals during the call, transparent handover can be achieved if the current channel deteriorates. This article asks if such an approach can be adapted to suit voice over IP telephony across the internet, specifically in relation to the Session Internet Protocol (SIP). SIP is an Internet-based protocol for establishing real-time end-to-end conference calls between peers. It already includes a mechanism, through the Session Description Protocol (SDP), of establishing the lowest common media capability available on both peers, but currently has no mechanism for establishing if the proposed media connection has adequate latency or packet loss performance to support real-time voice packets. This article asks if SIP should be extended to include such functionality and proposes the adoption of a client/server based measurement-based approach to control call admission.

Introduction

The Internet Engineering Task Force (IETF) Session Initiation Protocol (SIP) and the associated Session Description Protocol (SDP) are emerging as simple but effective protocols for establishing real-time single and multiparty voice or multimedia calls in IP networks. In terms of IP-based real-time multimedia transfer across the Internet, voice traffic is more sensitive to *loss* and *delay* than video, even though it requires far less bandwidth. If voice over IP (VoIP) is to become a realistic replacement for standard circuit-switched telephony services, users must experience the same consistently high-quality service they have become used to with traditional circuit-switched telephony. A lot of work is being done to develop quality of service (QoS) mechanisms for the Internet which will provide bandwidth guarantees and servicing priority for voice. A number of mechanisms have emerged for providing QoS for VoIP traffic. These mechanisms ultimately rely on marking voice packets so that bandwidth and improved packet servicing can be assigned to them as they cross an increasingly intelligent Internet, rather than merely providing 'best-effort' IP delivery. A related but significantly different approach, which draws on traditional circuit-switched telephony, is to try and establish if an adequate call-path is available and only to allow the call to occur if it is. This article asks if such an approach is suitable for SIP based VoIP calls, and if so, how the SIP/SDP protocol standard might be extended to include such a call admission mechanism.

SIP Overview

SIP is a relatively-new (1999) simple ASCII-based signalling protocol that uses application-layer requests and responses to establish one-to-one or conference multimedia communication between peer end-nodes on an IP network. It is an Internet Engineering Task Force (IETF) standard. It is compatible with existing Internet protocols and extensions such as HTTP, SMTP, LDAP and MIME and ultimately supports two important aspects of multimedia calls: **call signalling** and **session management**. Once a session has been established, SIP the IETF standard Real Time Protocol (RTP) is used to transfer media. The operation of a SIP-based call can be summarised as follows:

- A calling node (A), determines the location of the desired peer node (B) using standard Internet address resolution, name mapping and redirection
- SDP is used to determine the lowest common media capability that both A and B have – this is primarily a matter of agreeing a common codec from a set of available codecs
- A uses SIP to identify if B is available or not
- If the call is possible, a two-way RTP session is established between A and B
- The voice or other media is transferred using RTP/RTCP unicast or multicast traffic.
- SIP supports the transfer, call-holding, addition of other peers or other conference type functions during the call
- SIP terminates the call when media transfer is complete

SIP peers are called User Agents (UAs) and can operate as a client (call requesting party) or as a server (call responding party). A SIP peer would normally only operate as one or the other during any given session. Users are identified by unique SIP addresses based on the existing IETF formats but with a SIP flag e.g. *sip:joe.bloggs@itb.ie*. Users register their address (which is bound to the corresponding IP address of whichever terminal is used to register) with a local SIP Registration server which makes the information to a location server as required. Other types of SIP servers include proxy servers, for call-forwarding and SIP redirect servers, which call location servers to determine the path to the called peer and informs the calling peer when of the path when it learns it. Typical SIP peers include SIP-based applications or SIP phones. SIP gateways provide call control, translating calls between SIP endpoints and non-SIP endpoints (e.g. PSTN or GSM phones).

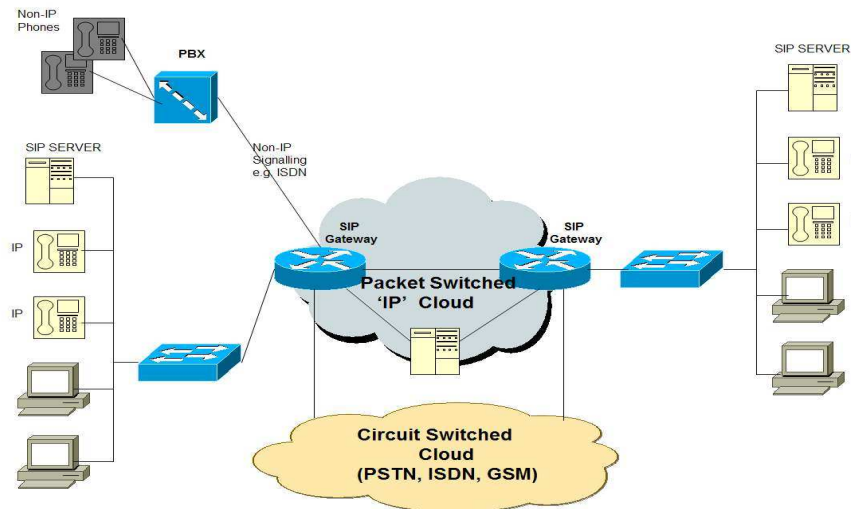


Figure 1: VoIP (SIP) Infrastructure with Gateway Support to Circuit Switched Networks

A transitional and convergent SIP-based VoIP Internetwork, incorporating support for existing local Private Branch Exchange (PBX) telephony and gives access both to packet and circuit switched wide area networks is shown below in Figure 1. This is the reference architecture we have adopted for this research.

QoS issues for VoIP

Voice over IP (VoIP) is extremely sensitive to loss and delay, even more so than video. In order for VoIP transmissions to be intelligible at the receiver, the voice packets should not be subject to excessive loss, (which is essentially a bandwidth issue) or variable delay (which is essentially a congestion issue). If VoIP is to become a realistic replacement for standard circuit-switched telephony services, users must experience the same consistently high-quality service they have come to expect from existing telephony. Some guidance in relation to VoIP includes:

- Voice packets are usually transmitted every 20ms
- Sufficient bandwidth provisioning should include consideration of both voice payload (e.g. 64 kbps for standard PCM voice) and associated IP header size (e.g. a further 16kbps).
- Voice packet round-trip time should not exceed 250 milliseconds
- Packet loss should be less than 1%
- Jitter (i.e. variable delay) must be minimised – delay variation should not exceed 100 ms, and ideally should be less than 10 ms

In general, VoIP can guarantee high-quality transmission of voice only if the bandwidth exists and voice packets are given priority over other less-sensitive traffic. The first step in providing QoS for voice packets is to mark and classify the packets so that network devices can discriminate between the time-sensitive voice packets and non-sensitive packets as they traverse the network and create different service levels for each class. Marking can be achieved in a number of ways. A common static method involves setting the IP *Precedence* bits in the IP header (the first three bits of the Types of Service (ToS) field in the IP header). This technique has now been extended to all marking using the first six-bits of the ToS field. These six bits represent the Differentiated Services Code Point (DSCP) and can be used to provide improved differentiated services to packet flows. The first three bits in DSCP are called the *class selector* bits and are compatible with precedence – the higher the decimal value, the higher the priority assigned to the packet. The next two bits are used to define the drop precedence (in the case that the network is congested and the final ‘set’ bit is used to indicate that the packet has been classified. If a device near the edge of a network has already identified a packet as being a VoIP packet (typically based on the protocol and port number in use) and marked it with an appropriate precedence or DSCP as it enters the internetwork, subsequent network devices can then classify the traffic by matching these bits. The appropriate QoS can then be applied.

Once traffic has been placed into QoS classes based on their QoS requirements, you can then assure priority treatment through an intelligent and configurable queuing mechanism. For VoIP traffic, the key issue is to ensure minimal latency, packet-loss and variation in delay. A range of queuing techniques exist but the preferred option for VoIP is a *priority* queuing scheme where packet classes can be prioritised to be sent before other less sensitive traffic and yet the remaining bandwidth can still be managed to meet lower priority traffic requirements. Typically between three and four queues are configured for High, Medium, Default and low priority classes of traffic. The network device process scheduler services the priority queue first (this can be limited to a configured maximum packet-rate) and the remaining queues are then serviced, using the remaining bandwidth (to a configurable proportional use-level). When the remaining bandwidth is apportioned to the non high-priority queues using an algorithm such as the Weighted-Fair-Algorithm, and packet-class is used to decide which queue the packet is placed on, this queuing technique is known as Low Latency Queuing (LLQ).

Even when voice packets have been classified and queued using low latency techniques, problems may still arise by the voice packets get trapped behind larger data packets in an environment in which bandwidth constraints enforce even minimal service-levels to non-priority queued traffic. A key contributor to variable delay is the transmission of large data packets

between small voice packets on a network. If the configured data packet size is such that it may hold a voice-packet in a transmission queue and force it to exceed the acceptable delay for voice packet intervals at the receiver, then the data packets need to be fragmented. On low speed links, the data fragment should typically take less than 10ms to transmit but should never be lower than the VoIP packet size. The VoIP packets can then be interleaved between them, minimising variation in delay at the receiver.

A final technique worth referring to is the compression of the IP RTP header in point-to-point VoIP calls (e.g. between gateways). IP RTP can reduce the 40 byte header of a voice packet to just two bytes, thereby significantly reducing the bandwidth required to transfer voice. Although this comes at the price of increased processing, it can be of particular value on point-to-point low speed links.

In VoIP the amount of bandwidth required for a call will depend primarily on the codec selected and typically ranges from 80 kbps (for 64kbps encoded speech using G711 a-law or u-law encoding) to 26 kbps (for 8 kbps encoded speech using G.729 encoding). This can be reduced somewhat if header compression is used for the RTP packet to the range 67 to 11 kbps.

Will the Internet Support his Call?

In traditional and mobile telephony, a placed call may be rejected by the local exchange if the circuit-switched connection cannot be made for resource shortage reasons. When a call is placed you either get a guaranteed dedicated connection or you get no connection at all. It is possible to transfer this rationale to packet-switched voice calls. In VoIP, it may be better to deny a VoIP call than to allow it to proceed in a network where the requisite bandwidth and QoS resources are not available at the time the call is placed. If the call went ahead, it would experience unacceptable and intermittent service, resulting in packet loss and excess latency. It could also affect other existing voice calls detrimentally. This is different from the QoS techniques discussed above insofar as it takes place *before* voice packets belonging to a requested call are allowed to be transmitted – it is basically a process to make an informed decision on whether to allow a call to proceed or not, or even to discern from a number of available routes the most feasible route for VoIP traffic. It is typically made based on one or a combination of local parameters, estimates of network congestion and the known shortage of requested QoS resources. As such, it could prevent excess voice traffic from getting onto the network and thereby also protect existing voice calls from being adversely affected by new calls.

Such decisions can be made based on local mechanisms such as the state of the local node and its interfaces, call volume restriction or some other locally known or configured parameters. Although a valuable decision-making component, to allow a call to proceed based solely on local mechanisms to and without any knowledge of network congestion is incomplete. A more evolved approach is to calculate the resources needed before a call is made and to request that the required resources are reserved for the call. This means that each network device along the call-path sets aside a subset of its resources to support the call and if any device cannot, the decision to abort the call may be made. Based on the response to the request for resources, a more informed decision can be made on whether to allow the call proceed or not. This approach is appropriate in an environment which is managed by a single administration from end-to-end but is not appropriate from a heterogeneous environment like the Internet. A compromise between using mechanisms using solely local information and resource reservation schemes could include a range of measurement-based techniques which gauge the network congestion in advance of making a call and make a decision based on the current *network* state. Unlike resource-based mechanisms, these techniques do not guarantee service resources and the measurement provides a basis for *estimating* if the state of the network will support the call. Although the local mechanisms are generally always pertinent to CAC decision-making, resource-based CAC is only possible if the calling/called parties are fully aware of the network topology and have some access to the intermediate backbone devices. This makes it impractical for the Internet, at least at the present time. Test pattern packets could be sent across the network to the destination node, which then return the pattern to the source node. By measuring the round-trip response, the source can estimate the loss and delay characteristics on the network path at the current time. Such techniques can essentially be independent of the network topology and will work transparently across the backbone without requiring any service management cooperation. For this reason, we are suggesting that any SIP-based VoIP deployment that uses the Internet as its backbone should adopt a combination of local and measurement-based techniques for decision-making. Based on the measured values and the estimation of network loss and delay characteristics, the call can either be allowed to proceed, refused or rerouted.

The Basic Research Idea

Although traditional ‘ping’ packets will give a rough evaluation of the network resources, a more appropriate approach, from a voice perspective, is to use a test pattern based on a series of realistic voice RTP packets which have been derived from the lowest common codec identified by the Session Discovery Protocol between the voice peers. Once the two RTP

channels have been established by SIP the test pattern would be sent from the source to the destination and 'bounced' back to the source. The returned test pattern could then be measured to establish the level of network congestion and the network's ability to support the call. The primary parameters measured to establishing the go/no-go decision or even route selection would include packet loss and delay characteristics. Measuring these characteristics would give a strong indication of bandwidth availability and congestion.

In VoIP, the amount of bandwidth required for a call will depend primarily on the codec selected and typically ranges from 80 kbps (for 64kbps encoded speech using G711 a-law or u-law encoding) to 26 kbps (for 8 kbps encoded speech using G729 encoding) for packet header and payload. This can be reduced somewhat if header compression is used for the RTP packet to a range of 67 to 11 kbps. Different approaches could be taken to assembling the test pattern packets: packets for a specified test pattern file could be dynamically generated by the selected codec scheme once SDP has identified the common codec or a library of prepared test packets could be available. Alternatively, the appropriate test pattern could be selected from a library of codec specific test patterns already existent on the peer. The client server character of a SIP user agent provides an appropriate underlying request/response paradigm between the peers. A call originating client transmit could send a series of RTP packets, with the appropriate precedence value 'typically 5 or '101' for voice) or DSCP value, requesting the server to echo the test pattern. The test pattern is delivered across the network to the destination SIP 'server' which responds by retransmitting what it has received to the source 'client'. The port numbers used for the connection correspond to the UDP port numbers already decided for the potential RTP voice session or a specific port could be used. Figure 2 below describes the basic process.

The bandwidth perceived by the client can be considered as:

$$\text{Network Bandwidth Perceived by SIP Client} = \frac{\text{Byte count transferred}}{T_{RT}}$$

Where T_{RT} = round-trip time

T_{RT} can be measured with a time that starts when the last byte of the test pattern has been transmitted and which stops when the last byte has been received back. This simple equation could be improved by subtracting the overhead taken by the destination to process the test pattern. This processing time will be dependent on the processing speed of the terminals in use e.g. it may be a workstation with a soft client or a SIP-enabled IP phone. For a given test pattern, it will be possible to derive good approximations based on the number of packets needed to create the test pattern for a specific codec, the number of bytes per packet, the

number of bytes required for data Link Layer de-encapsulation/re-encapsulation and whether compression is used or not.

$$\text{Network Bandwidth Perceived by SIP Client} = \frac{\text{Byte count transferred}}{T_{RT} - T_P}$$

Although not an accurate measurement of real network bandwidth, this perceived bandwidth measurement provides a reasonable basis for estimating of the network current availability. The second measurement to be taken requires the source client to compare the echoed test pattern sequence with the original test pattern in order to establish the packet loss and data corruption. Packet loss above 1% is considered to be unacceptable and would serve as a basis for attempting to reroute or cancel the call.

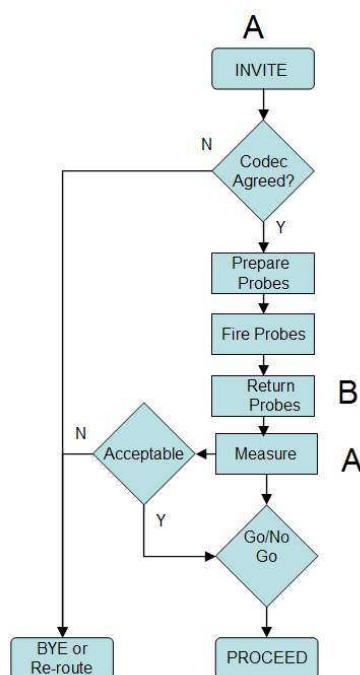


Figure 2: Process flow for establishing network state prior to permitting a voice call

Preliminary Conclusions: Potential Benefits & Limitations of this approach

The main advantage of this type of approach is that it is suitable for use across a backbone network such as the Internet, which some non-homogeneous QoS resources in place, but which you are not in a position to reserve or control without service provider intervention. It is an end-to-end solution, which has no reliance on QoS support in the intermediate service provider networks. These measurements can provide an initial basis for assessing whether a call should proceed. Implementing the process will only require a small amount of code to be added to the SIP application and would sit well with the client-server characteristics of the SIP user-agent. This is consistent with transitional convergent models where IP connectivity forms the lowest common service denominator. It is possible to conceive of this approach interacting with

routing intelligence in order to test multiple available routes (not just the best route indicated by the routing protocol which is usually bandwidth driven) on an application specific basis. In this way, a call with lower bandwidth but better delay characteristics might be chosen in preference to a route solely selected on bandwidth availability. A disadvantage is that it would only work between SIP peers and would not support calls between SIP and non-VoIP terminals. Furthermore, taking these measurements at the beginning of a call and establishing that conditions are satisfactory for the call to proceed at that point is no guarantee that the network will support the call for the entire duration of the call. It only serves as an *assessment* of the capability of the network to support the call at that time. One possibility would be to update the measurements at intervals throughout the call and if the measurements seem to be deteriorating towards a critical threshold, to seek an alternative connection to which the call can then be dynamically transferred in a manner that is transparent to the user. This is analogous to a mobile phone call channel hand-over when moving between cells or when experiencing difficulty. SIP currently doesn't support this functionality but as a protocol, which supports multi-party conferencing, it would not be difficult to extend it to do so. Averages could be maintained on a peer-to-peer basis to identify repetitive success and failure patterns to which different codec strategies could then be applied. For organisations with repetitive patterns of call traffic, the approach could be made proactive to determine network performance systematically between common call points. There is an obvious overhead in terms of delaying the actual call, while the assessment is made, and if subsequent test patterns are sent during a call.

References

- Sugih Jamin , Peter B. Danzig , Scott J. Shenker , Lixia Zhang, A measurement-based admission control algorithm for integrated service packet networks, IEEE/ACM Transactions on Networking (TON), v.5 n.1, p.56-70, Feb. 1997
- Matthias Grossglauser, David N. C. Tse, A framework for robust measurement-based admission control, IEEE/ACM Transactions on Networking (TON), v.7 n.3, p.293-309, June 1999
- H. Schulzrinne et al., "SIP: Session Initiation Protocol", IETF DRAFT, November 2000.
- H. Schulzrinne et al., "Centralized Conferencing using SIP", IETF DRAFT, November 2000.
- Ulysses Black, Advanced Internet Technologies (Prentice Hall 2001)

Auto Generation of XLIFF Translation Documents from Proprietary File Formats

Kieran O'Connor & Geraldine Gray

kieran.o'connor@itb.ie

Institute of Technology Blanchardstown, Dublin 15, Ireland

Abstract

The handling of proprietary documents by localisation vendors is time consuming and error prone, and represents a significant challenge to localisation projects. Vendors with many customers, each with its own set of document formats, must potentially support a document format set numbering in the hundreds or thousands. This paper describes an approach to automating the extraction of translatable text from a variety of file formats. The solution is based on XLIFF, language parsers, and XML transformations.

Keywords: Localisation, grammar, regular expression, JavaCC, JTree, XLIFF, XSLT, XML, XML Schema, transformation, translation, language, parse.

1 Introduction

Localisation is the process of adapting text to specific target audiences in specific geographic locations [WorldLingo, 2004, section: Glossary of Terms: Localization]. It is a fundamental process for many industries expanding or product selling into foreign regions. Hence there is a large market for vendors who provide translation services. Making the translation process as efficient and streamlined as possible provides benefits both to the document owner and to the vendor. The owner realises a quicker translation turnaround, and the vendor leverages its efficient process in acquiring and retaining business. [Rubric 2000].

Current challenges in the localisation process include insufficient operability between tools; lack of support for an overall localisation workflow; the number of file formats localisation developers have to deal with; and the large number of proprietary intermediate formats. [8th Annual International Localisation Conference 2003]

To address some of these challenges, and OASIS technical committee, whose objective was to address the lack of standards in the localisation industry, has developed an XML based standard called XLIFF -XML Localisation Interchange File Format. The specification provides the ability to mark up and capture localizable data and interoperate with different processes or phases without loss of information. [XLIFF 2004] Version 1.1 of the standard was submitted to the Oasis standards review process in November 2003. The significance of XLIFF is that translation tools can now be developed which accept input in a common format, i.e. XLIFF, enhancing operability between tools.

The research documented in this paper was done in collaboration with a local IT company, who are developing a product that implements a localisation workflow based on the XLIFF standard. However as stated above, one of the major challenges of localisation is the large number of file formats presented by

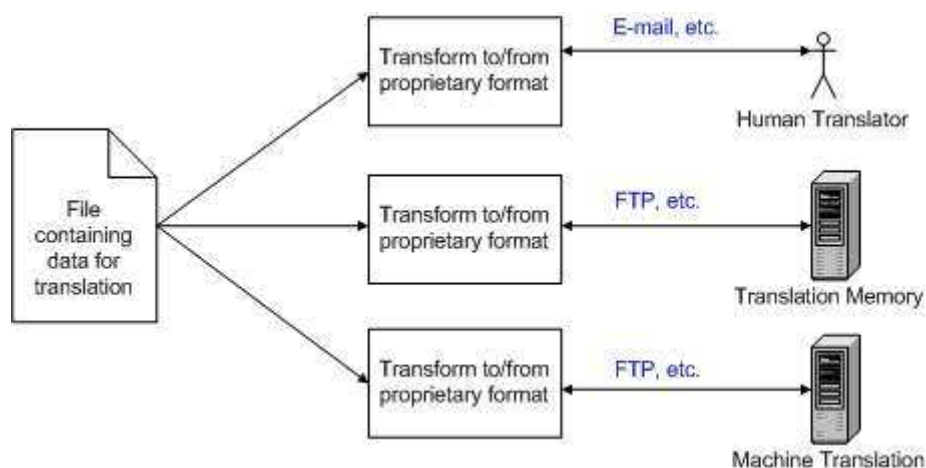
customers for translation. Therefore an XLIFF based localisation workflow requires pre-process of a proprietary file format to extract translatable text, and insert it into an XLIFF document. Once the workflow is complete, and the translated text has been added to the XLIFF document, post processing is required to insert this translated text back into the propriety file.

The focus of the work documented in this paper was to investigate if it is possible to automate these pre-processing and post-processing steps, given the variety of file formats presented for translation. At the time of writing, there was no documented work in this area. It is out belief that ITB, in conjunction with a local technology company, are the first to publish a technique for automating these pre-processing and post-processing steps.

2 State of the Art

Currently the typical translation process consists of a proprietary document submission to a localisation vendor for processing. This could be done, for example, online or through the use of physical storage devices. The document is manually examined to determine its type and to identify the text contained within which is translatable. Therefore there is a requirement for the examiner to understand the document contents. The translatable text is then stored in a file, such as a name/value pair file. The file is then passed through various tools, such as machine translation and translation memory, until as close to 100% translation as possible is achieved. Specialised human translators may also be used when necessary. The tools used for the machine translation and translation memory tasks each require the input data to be of a particular format, depending on the implementation. Thus file transformations may be required at each step. When the translation process is complete the original document must be re-constituted complete with the new translations.

The diagram below illustrates the basic translation process where proprietary file and exchange formats are used. The file passed to each translation component must be in a format understandable by receiver. Thus it must be transformed.



▪ Figure 1 – Proprietary document localisation

The transformation of a file from the proprietary formats supplied by the customer, to the proprietary format required by the translation tool, can be a costly and complicated task. When new file formats are introduced so also must a new transformation rule set be developed.

3 Methodology

This research has identified two categories of input file formats that cover the majority of input files presented for translation. The first category is documents that can be described using a language grammar. The second category is documents in XML format and hence can be described by an XML Schema.

3.1 Grammar

3.1.1 Backus-Naur Form (BNF) & Extended Backus-Naur Form (EBNF)

Language rules, manifested by grammars, are a method of describing the expected lexical and syntactic structure of a document. Backus-Naur Form (BNF) and Extended Backus-Naur Form (EBNF) are standard ways of describing complex languages using grammars. Particularly, they are a standard means of describing languages that comprise a large number of possible valid expressions. The following is an example of a simple EBNF grammar that describes how integers may be combined to form a valid expression e.g., 5+8.

```
simpleLanguage ::= integerLiteral ("+" | "-" | "/" | "*") integerLiteral
integerLiteral ::= [0-9]+
```

3.1.2 JavaCC Grammar

This research utilises the JavaCC grammar format, as specified for the JavaCC parser generation tool (detailed in section 4). This format is specific to JavaCC but is closely related to EBNF. There is a grammar repository of existing JavaCC grammars available at the following address: <http://www.cobase.cs.ucla.edu/pub/javacc/>.

3.2 XML Schema

XML documents are described using an XML Schema. The XML schema defines the contents of an XML document as a valid tree structure of elements and their children. The following shows the definition of a complex element that contains two child elements.

```
<xsd:complexType name="PolicySyncRq">
  <xsd:sequence>
    <xsd:element ref="RqUID" minOccurs="1" maxOccurs="1"/>
    <xsd:element ref="PolicyNumber" minOccurs="1" maxOccurs="1"/>
  </xsd:sequence>
</xsd:complexType>
```

4 Technologies

Various technologies are utilised in this research. Some are directly related to localisation such as XLIFF, others are more general such as JavaCC and XSLT. This section gives a brief overview of each technology used.

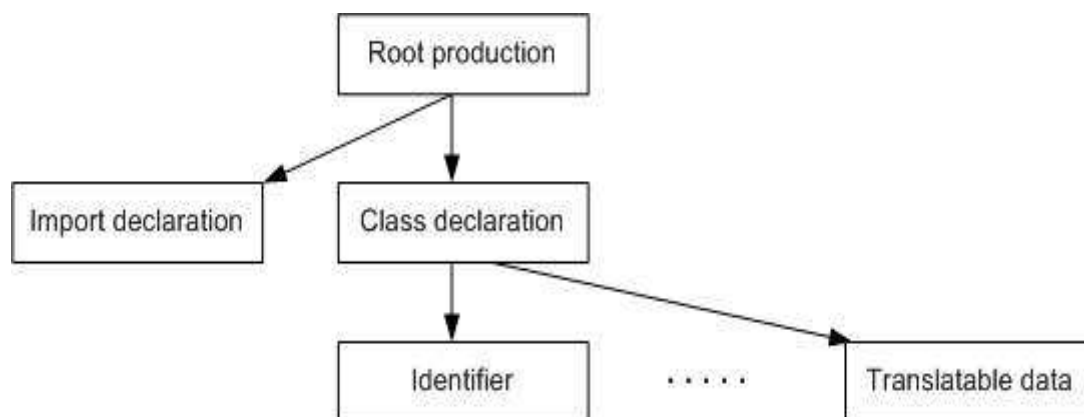
4.1 XML Localisation Interchange File Format (XLIFF)

XLIFF is an emerging XML specification aimed at the localisation industry. It has been adopted, and is currently under review by the standards organisation OASIS. The purpose of XLIFF is to standardise the format in which translation information is held in transmission between applications and/or organisations. It allows the mark-up of localisable data into a common, globally understandable format that enhances the larger translation process by providing data format transparency between interoperating parties. The XLIFF specification is available online at: <http://www.oasis-open.org/committees/xliff/documents/cs-xliff-core-1.1-20031031.htm>. This research will use XLIFF as the standard for representing translation information.

4.2 Java Compiler Compiler (JavaCC) & JJTree

JavaCC is a tool that takes as input a grammar and creates a generic parser that, when given a source document, will evaluate the syntactic correctness of the same. JJTree is a pre-processor tool accompaniment to JavaCC. JJTree creates a tree representation of a source document parse when used in conjunction with JavaCC. The tree can then be traversed and evaluated node by node at runtime. As nodes are encountered that correspond to translatable information they can be handled. The tree can also be unparsed, i.e., written out to a file node by node, hence recreating the original document. This allows for intelligent document alteration at runtime.

Below is a diagram that outlines how a Java programming language document might be represented as a parse tree (a full parse would probably contain hundreds of nodes). The root production is present in all JavaCC grammars. Child elements of this production essentially represent the contents of the document.



▪ Figure 2 – Parse tree

4.3 XSL Transformations (XSLT)

XSLT is one part of the Extensible Stylesheet Language (XSL) family of specifications for XML document transformation and presentation. The other members are: XML Path Language (XPath) and XSL Formatting Objects (XSL-FO). XSLT is a language for transforming one XML document into another XML document [XSL Transformations, 1999, section: Abstract], or into another format completely, such as HTML, PDF, or SQL statements. XSLT files contain templates which are matched to elements using patterns (expressed in XPath). These templates define change rules to apply to the source XML tree (matched using XPath) such as renaming and reordering XML elements. A transformed XML file can be completely unrecognisable from the original depending on the transformations used. Below is a simple XML file pre and post transformation with the XSLT file that was used to facilitate the change. This is a simple example but XSLT contains many directives and functions to create more complex transformations.

```

Original:  <?xml version="1.0" encoding="UTF-8"?>
           <prescription>
             <medicalPrescription>
               <id>123456789</id>
               <name>John Doe</name>
               <address>57, Nowhere Avenue</address>
             </medicalPrescription>
           </prescription>

Transformed: <?xml version="1.0" encoding="UTF-8"?>
            <prescription>
              <medicalPrescription>
                <id>123456789</id>
                <name>John Doe</name>
                <address>HELLO</address>
              </medicalPrescription>
            </prescription>

Transform:  <?xml version="1.0" encoding="iso-8859-1"?>
            <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
            version="1.0">
              <xsl:output method="xml" indent="yes"></xsl:output>
  
```

```

        <xsl:template match="/"><xsl:apply-
templates/></xsl:template>
        <xsl:template match="//*">
            <xsl:copy>
                <xsl:apply-templates/>
            </xsl:copy>
        </xsl:template>
        <xsl:template match="//address">
            <xsl:element name="address">
                <xsl:value-of select="HELLO"/>
            </xsl:element>
        </xsl:template>
    </xsl:stylesheet>

```

Instruction for the XSLT processor to find all elements named 'address' (denoted by the XPATH statement //address).

Instructs the processor to set the value of the contents of the elements found (<address> elements) to 'HELLO'.

5 Case 1 – Grammar Based Parser Generator

5.1 Objective

The objective is to automatically generate a parser from a given grammar. The purpose of using parser technology is to allow the document type to be described as a grammar, create the parser once, and use the same parser to process all documents of that type. The parser created must be capable of the following:

- Examine a source document to identify the data contained within which is translatable.
- Create an XLIFF document with the translatable text contained.
- Create a skeleton document. Post translation, the translated text must be merged with this skeleton document to recreate the format of the original. The skeleton file is similar to the original except that, in place of the translatable text extracted to the XLIFF document, there are markers inserted that are used to merge with the translated text.

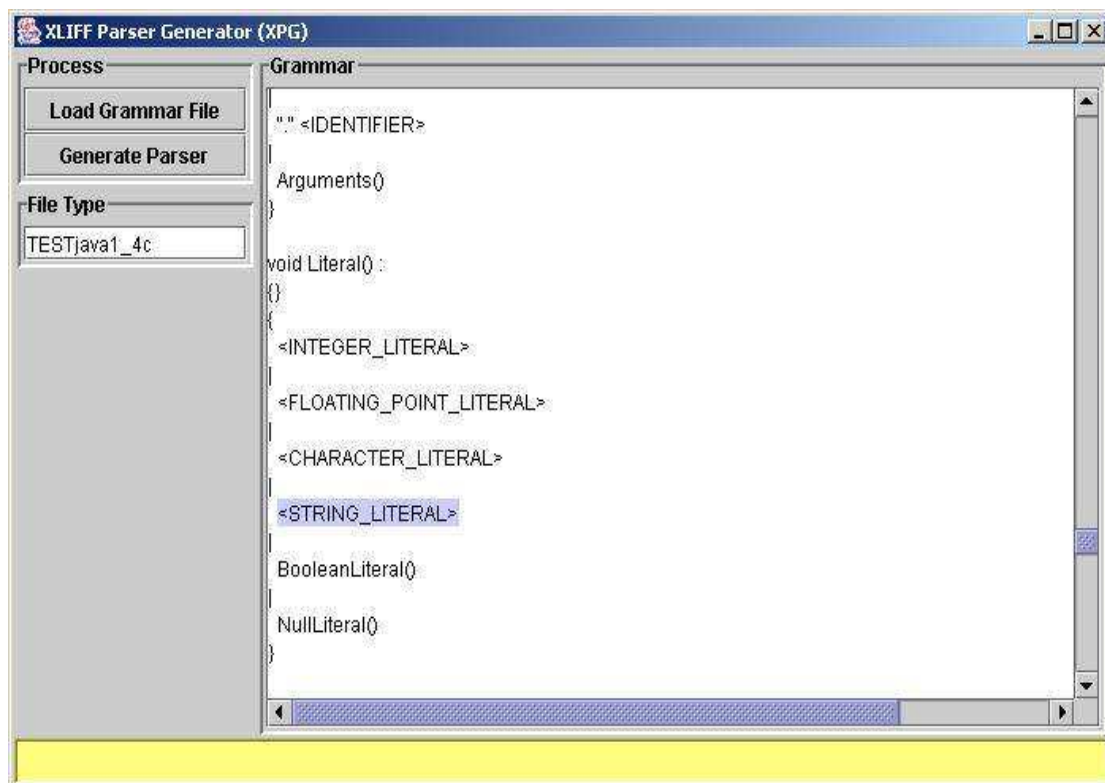
5.2 Parser Creation

Creating a parser for a specific grammar involves the steps outlined below.

1. Identify, or create a grammar representation of the document type in question.
2. Identify, within the grammar, the token that defines the data that will need to be extracted for translation. For example, the '`<STRING_LITERAL>`' declaration within the '`literal()`' production represents a translatable string for a normal programming language. JJTree directives and processing instructions are inserted into the grammar based on the selection. This creates a JJTree grammar.
3. The pre-processor JJTree is run against the new grammar to create tree handling classes. These generated classes are used for such purposes as representing tokens encountered in the tree traversal as nodes of the parse tree. Also outputted by the JJTree tool is a regular JavaCC grammar file.
4. JavaCC is run against the new JavaCC grammar file from the previous step to produce the remaining files needed for the parser. The combination of classes now available constitutes the parser.

5.2.1 JJTree Directives and Processing Instructions

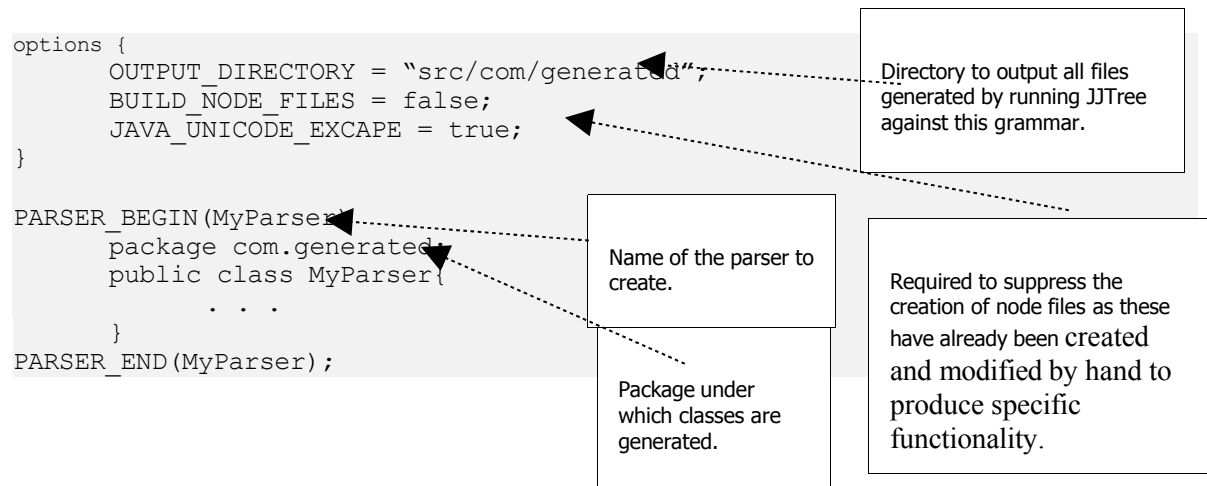
In step 2 above, JJTree directives are inserted into a grammar based on user input. These form the processing instructions that the generated parser uses to create the parse tree, identify translatable text, and consequently output the XLIFF and skeleton files. Using a graphical user interface the user loads a grammar and selects from it the token that represents translatable text. Figure 3 shows the user interface used, with a grammar loaded and the '<STRING_LITERAL>' token highlighted.



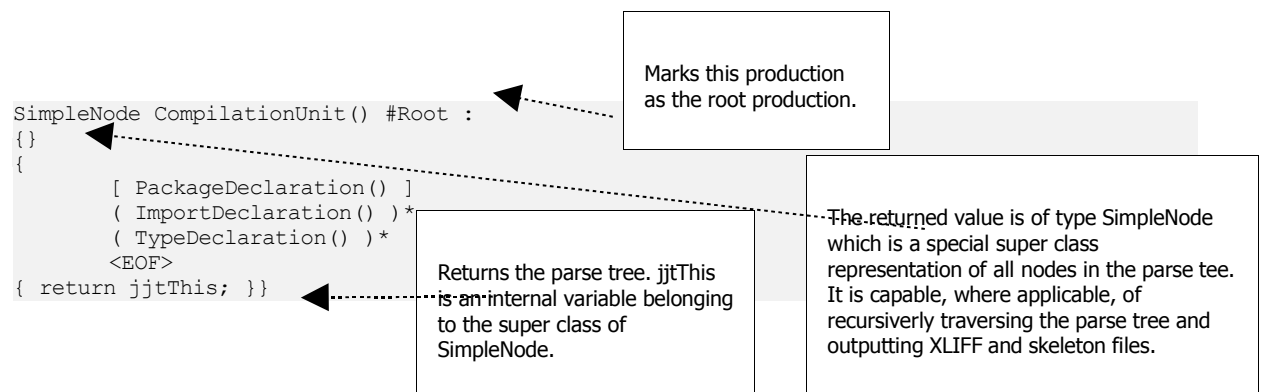
▪ Figure 3 – Parser Generator User Interface

With this selection made by the user the parser can be generated by hitting the 'Generate Parser' button. This action inserts the required JJTree directives into the original grammar and processes this new grammar as described in the following paragraphs.

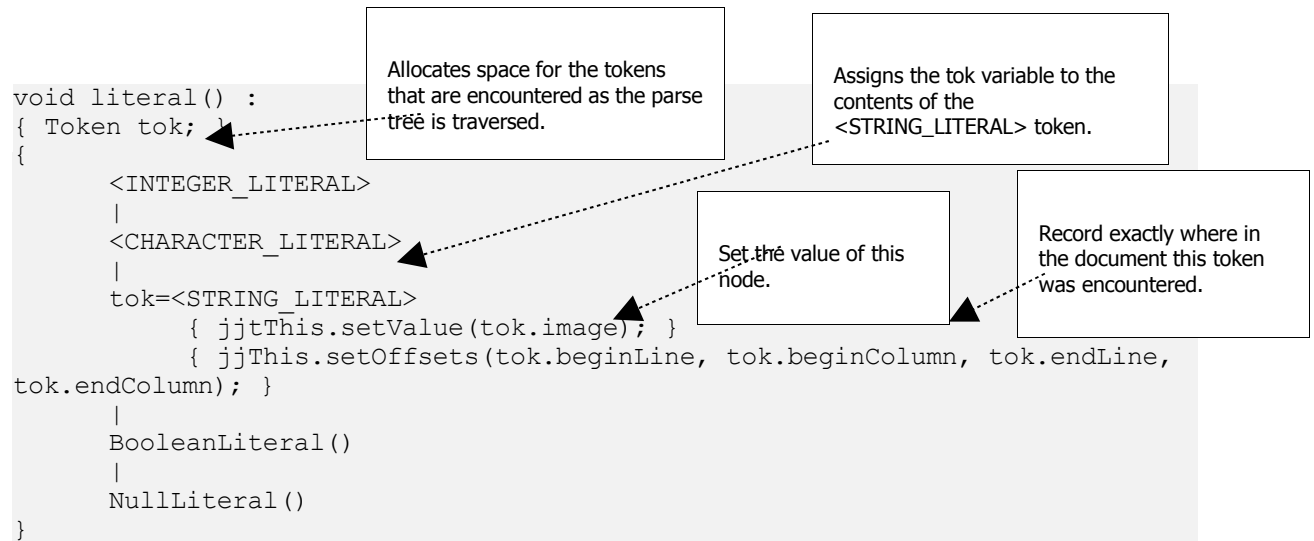
The options section, at the beginning of the grammar, is updated with processing instructions. The parser code section is updated with the name of the parser to create and the package to use for the generated classes.



The root production in a grammar is the production from which all other language productions hang (as illustrated by figure 2). JJTree directives must be inserted at the root production of the grammar to allow for the creation of the parse tree. The following code snippet shows a root production called 'CompilationUnit'.

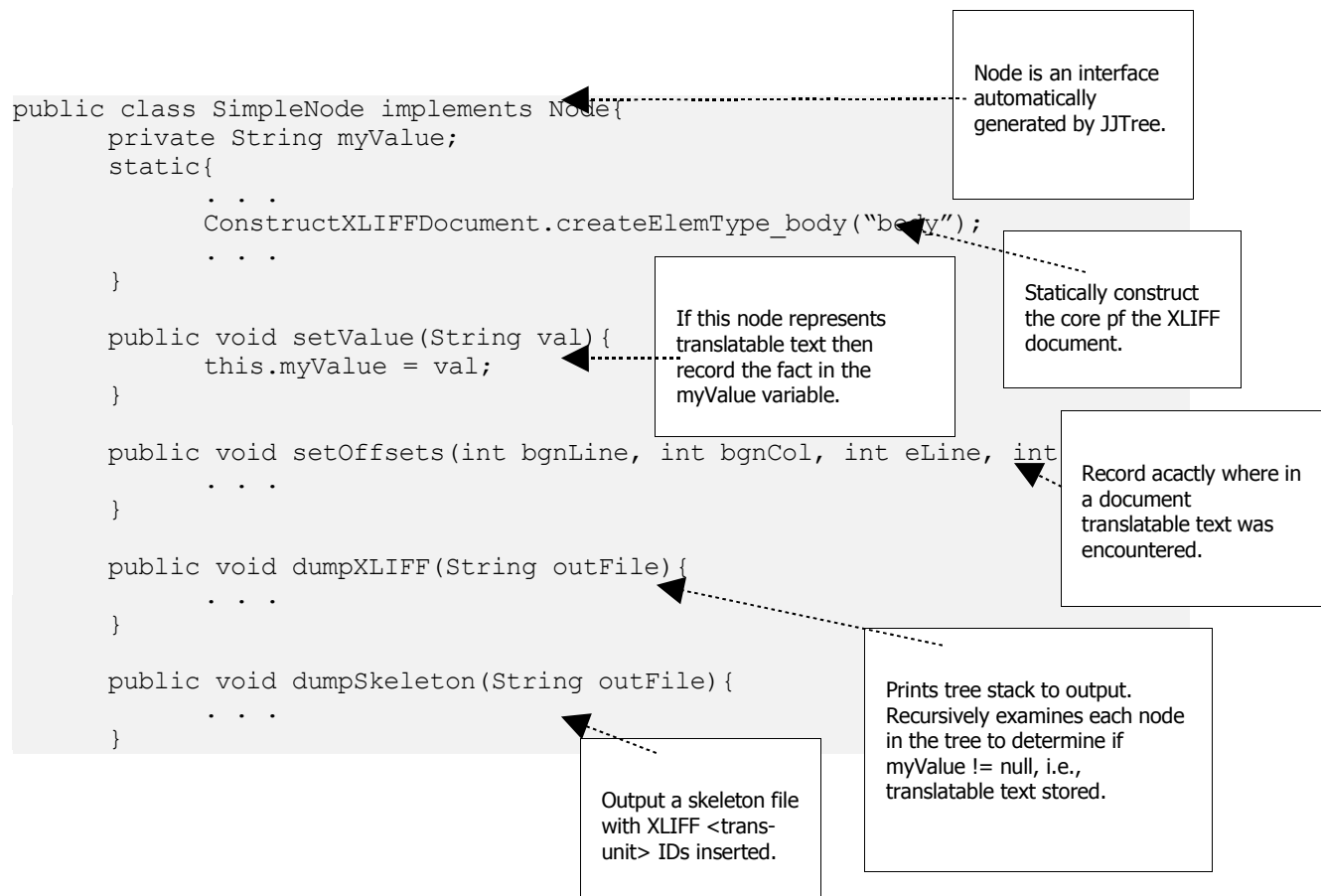


The SimpleNode class is a node super class that contains utility methods used when processing the parse tree node by node. Say for example that the user selected the '<STRING_LITERAL>' token as representing translatable text as illustrated in the user interface diagram above. The following code is inserted within the production that the selection belongs to.



5.2.2 The SimpleNode Class

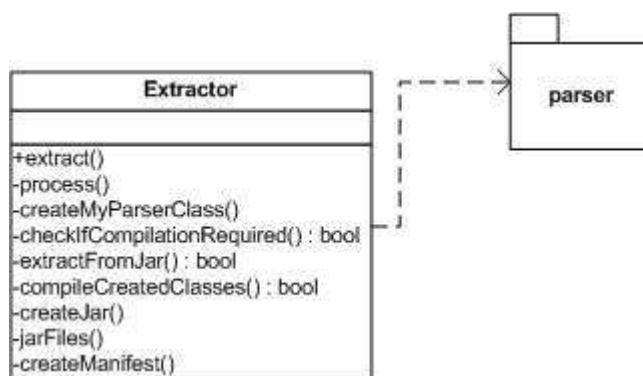
The SimpleNode class itself, which maintains node information (and hence translatable text information), and outputs the XLIFF and skeleton files, is described in the following code snippet.



5.3 Using the Parser

The parser is used by an extractor class whose purpose is to provide functionality for creating XLIFF and skeleton documents. The extractor class uses the parser to recursively traverse the parse tree created for the inputted document and outputs the relevant XLIFF & skeleton files.

The extractor class is described by the following diagram. It contains one publicly available method, `extract`. This method takes as parameters the URL of the source document, and the name of the parser to use. It then uses the private methods at its disposal to process the document.



▪ Figure 4 – Extractor class

The result of running the extractor class against a file using a previously create parser is a valid XLIFF document with translatable data suitably inserted, and a skeleton file.

```

XLIFF:    <trans-unit id="123456789">
           <source>my translatable text</source>
           <trans-unit id="98734">
           <source>yet more translatable text</source>
           </trans-unit>

Skeleton: public static void main(String args[]){
           String s1 = "<trans-unit id=123456789>";
           System.out.println("<trans-unit id=98734>");
           }
  
```

5.4 Post translation

A typical XLIFF document will go through phases of translation until the data contained is fully translated. At the end of this process there will be an XLIFF document that contains original data and the data's resultant translation within the specified XML element '`<trans-unit>`'. The following is an example of a translated XLIFF document.

```

<trans-unit id="123456789">
  <source>hello</source>
  <target>bonjour</source>
</trans-unit>
<trans-unit id="436554">
  <source>how many</source>
  <target>combien</target>
</trans-unit>

```

A merge class takes the XLIFF document, and specifically the translations, and merges them with the skeleton file created during the extraction process, to produce a final translated document. This process is accomplished using XML processing technologies to process the XLIFF document and regular expression matching to process the skeleton document

6 Case 2 – XML Schema Based Transformer Generator

6.1 Objective

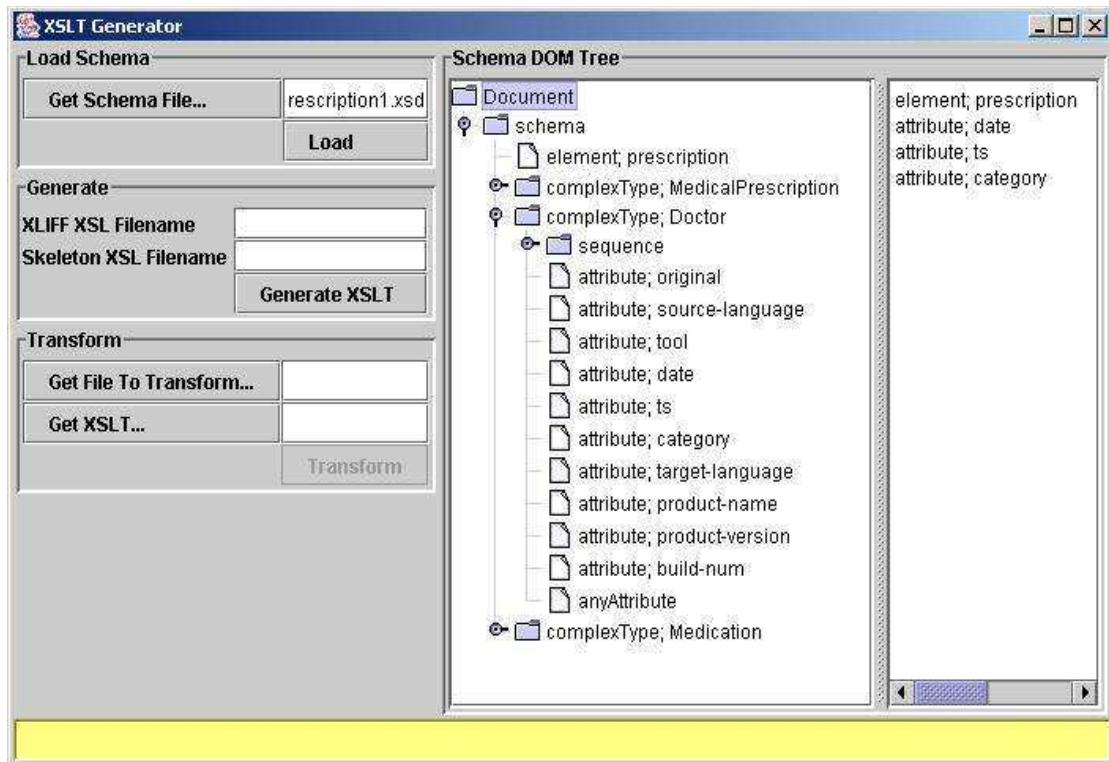
In the XML Schema based approach the objective is to identify XML elements and attributes, from an XML Schema, that are considered to contain translatable text. From this information two XSLT files are created. The first converts an XML document into an XLIFF document complete with translatable text suitably inserted. The second converts the original XML document into a skeleton used for merging.

6.2 Transformer Creation

Creating the two files used to process an XML document for localisation follows the following steps:

1. Identify within an XML Schema the elements and attributes that can contain translatable text.
2. Generate the necessary XSLT files based on these selections from the step above.

The task of identifying the elements and attributes within an XML Schema which can contain translatable text is simplified by using a graphical user interface. The user will load an XML Schema into the interface. This creates a tree structure of the information. The user can then navigate to and select the required elements and attributes. The following diagram shows the user interface.



▪ Figure 5 – Transformer Generator User Interface

6.2.1 XSLT Directives and Processing Instructions

Based on the selections made in the user interface, two transformation files are created; one for creating the XLIFF, the other for creating the skeleton file. The following code snippet shows the XSLT generated for converting an XML file conforming to the format described by a given XML Schema into XLIFF based on the user selections. The XSLT generated for creating the skeleton file will be similar.

```
<xsl:stylesheet xmlns:xsl=http://www.w3.org/1999/XSL/Transform version="1.0">
  <xsl:output method="xml" indent="yes"></xsl:output>
  <xsl:template match="/">
    <xlf:xliff version="1.1" . . . >
      <xlf:file datatype="rtf">
        <xlf:header/>
        <xlf:body>
          <xsl:for-each select="//Doctor/@category">
            <xlf:trans-unit id="{generate-id()}">
              <xlf:source>
                <xsl:value-of select="."/>
              </xlf:source>
            </xsl:for-each>
          </xlf:body>
        </xlf:file>
      </xlf:xliff>
    </xsl:template>
  </xsl:stylesheet>
```

Create this XLIFF construct.

Examine each element called 'Doctor' with attribute 'category'.

Insert the value into this XLIFF construct.

6.3 Using the Transformations

1. Run the first XSLT file against the original XML document to create an XLIFF file.
2. Run the second XSLT file against the original XML document to create a skeleton file.

Having created the initial stylesheets, any XML document conforming to the XML Schema in question can be transformed to create the necessary XLIFF and skeleton files.

6.4 Post Translation

See section 5.3. The merger application for this is the same as that for the grammar based approach.

7 Conclusion

The prototype developed by this research has successfully translated both programs of a recognisable grammar, and XML files from a previously processed XML schema. Using language parsers and XSLT transformations to enable efficient and automatic handling of multiple document types by localisation vendors improves the translation process efficiency. No longer must vendors spend time examining and extracting data from every document that requires translation. This research shows that documents can be processed by type rather than individually. Also, as a larger set of parsers and/or transformation files is built up over time, the less likely it is that the vendor will encounter an unsupported format. This is particularly true for proprietary file formats that are unlikely to change over time.

Reference

WorldLingo (2004)

Glossary of Terms, Retrieved March 2004, from WorldLingo website
<http://www.worldlingo.com/resources/glossary.html>

XSL Transformations (1999)

XSL Transformations (XSLT), Version 1.0. Retrieved March 2004 from W3C website
<http://www.w3.org/TR/xslt>

Rubric (2000)

Regina Born, Computer Associates Tom Shapiro, Rubric, Streamline The Localization Process, Software Business, March 2000

8th Annual International Localisation Conference (2003)

Tony Jewtushenko, Principle Product Manager, Oracle and chair OASIS TC XLIFF, "XLIFF – the XML based standard for Localisation File Format", 8th Annual International Localisation Conference, Localisation Research Centre, University of Limerick, 2003

XLIFF (2004)

<http://www.xliff.org>



<http://www.itb.ie>