Electronic Engineering and Computing Technology

Lecture Notes in Electrical Engineering

Volume 60

For further volume:
http://www.springer.com/series/7818

Sio-Iong Ao • Len Gelman
Editors

Electronic Engineering and Computing Technology

Springer

Editors

Dr. Sio-Iong Ao
International Association of Engineers
Unit 1, 1/F
37-39 Hung To Road, Hong Kong
publication@iaeng.org

Professor Len Gelman
Cranfield University
School of Engineering
Dept. Process & Systems Engineering
Cranfield, Beds
United Kingdom MK43 0AL

ISSN 1876-1100
ISBN 978-90-481-8775-1 e-ISBN 978-90-481-8776-8
DOI 10.1007/978-90-481-8776-8
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010925339

Printed on acid-free paper

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

Preface

A large international conference in Electronic Engineering and Computing Technology was held in London, UK, July 1–3, 2009, under the World Congress on Engineering (WCE 2009). The WCE 2009 was organized by the International Association of Engineers (IAENG); the Congress details are available at: http://www.iaeng.org/WCE2009. IAENG is a non-profit international association for engineers and computer scientists, which was founded originally in 1968. The World Congress on Engineering serves as good platforms for the engineering community to meet with each other and exchange ideas. The conferences have also struck a balance between theoretical and application development. The conference committees have been formed with over 200 members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries. The conferences are truly international meetings with a high level of participation from many countries. The response to the Congress has been excellent. There have been more than 800 manuscript submissions for the WCE 2009. All submitted papers have gone through the peer review process, and the overall acceptance rate is 57%.

This volume contains 61 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include Control Engineering, Network Management, Wireless Networks, Biotechnology, Signal Processing, Computational Intelligence, Computational Statistics, Internet Computing, High Performance Computing, and industrial applications. The book offers the state of the art of tremendous advances in electronic engineering and computing technology and also serves as an excellent reference work for researchers and graduate students working on electronic engineering and computing technology.

International Association of Engineers

Dr. Sio-Iong Ao
Professor Len Gelman

Contents

Chapter 1
On the Experimental Control of a Separately Excited DC Motor

Rubén Salas-Cabrera, Jonathan C. Mayo-Maldonado,
Erika Y. Rendón-Fraga, E. Nacú Salas-Cabrera,
and Aaron González-Rodríguez

Abstract This paper presents the experimental implementation of a controller for the rotor position of a separately-excited DC Motor. The control law was designed by using a dynamic model for the motor and a dynamic model for the load torque. The torque model was proposed in order to improve the performance of the closed loop system. Discrete state feedback, discrete integrator and a discrete state observer are implemented and calculated by employing a real time control tool that works under a Linux operating system. Custom-made digital, analog and power electronics designs are the fundamental components of the closed loop implementation.

Keywords Real time · data acquisition · DC motor · electronics

1.1 Introduction

This work deals with the experimental implementation of a state space controller for a DC motor that has separate winding excitation. In other words, the field and armature windings of the electric machine are fed from independent sources, Krause et al. [7]. A brief review of the literature follows. An interesting contribution is presented in [12]. It is related to the speed control of a series DC motor. Zhao and Zhang [12] use a nonlinear dynamic model for deriving a backstepping-based control law. The emphasis is on the analysis of the controller. Jugo [5] uses RTAI-Lab and Scilab for implementing a transfer function-based controller of a DC motor. The main contribution in [5] is related to testing several open source real time tools (RTAI, RTAI-Lab, Comedi, Scilab, xrtailab). Experimental results regarding high gain observers are shown in [1]. RTAI-Lab and Scicos/Scilab are used for the real

R. Salas-Cabrera (✉), J.C. Mayo-Maldonado, E.Y. Rendón-Fraga,
E.N. Salas-Cabrera, and A. González-Rodríguez
1501-1ro. de Mayo Avenue Pte., Cd. Madero, Tams., Mexico
e-mail: salascabrera@aol.com; jcarlos_mayo@hotmail.com;
erika.yuridia.rendon@hotmail.com; nacu_salas@hotmail.com; aaronglzrod@yahoo.com.mx

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_1,

time implementation. The system to be tested in [1] was a series DC motor. The contribution of this work is associated with the real time experimental implementation of the closed loop system. In other words, several hardware and software components were designed, i.e. some analog, digital and power electronics circuits were implemented. Important components of the experimental system are: a personal computer, a PCI-6024E data acquisition card [9], comedi driver library for Linux [11], a free open source real time platform [2], a custom-made power electronics converter, an incremental encoder and signal conditioning circuits for measuring the rotor position.

1.2 Modeling

The equations that establish the DC motor behavior are obtained by using fundamental electrical and mechanical laws [7], this is

$$V_f = r_f i_f + L_{ff} \tfrac{d}{dt} i_f \tag{1.1}$$

$$V_a = r_a i_a + L_{aa} \tfrac{d}{dt} i_a + L_{af} i_f \omega_r \tag{1.2}$$

$$J \tfrac{d}{dt} \omega_r + T_L = L_{af} i_f i_a \tag{1.3}$$

$$\tfrac{d}{dt} \theta_r = \omega_r \tag{1.4}$$

Notation for parameters and variables is given in Table 1.1. Parameters involved in (1.1)–(1.4) were calculated once we analyzed the experimental results of several transient and steady state tests.

Table 1.1 Notation for variables and parameters

θ_r	Rotor position
ω_r	Rotor speed
i_a	Armature current
i_f (0.46 Amps)	Field current
V_f	Field voltage
V_a	Armature voltage
r_f (309.52 Ohms)	Field resistance
r_a (6.615 Ohms)	Armature resistance
L_{aa} (0.0645 H)	Armature inductance
L_{ff} (14.215 H)	Field inductance
L_{af} (1.7686 H)	Mutual inductance
J (0.0038 kg/m^2)	Inertia
T_L	Load torque
K_0 (0.20907), K_1(−9.8297)	Coefficients of load torque

1.2.1 Load Torque Model

For this implementation, an experimental-based load torque dynamic model is proposed. The idea of proposing this model is to improve the performance of the closed loop system. Let us propose a state equation for the load torque, this is

$$\frac{d}{dt}T_L = K_0\omega_r + K_1T_L \tag{1.5}$$

In order to obtain parameters in (1.5), we rewrite Eqs. (1.3) and (1.5) as

$$\frac{d}{dt}\begin{bmatrix}\omega_r \\ T_L\end{bmatrix} = \begin{bmatrix}0 & -\frac{1}{J} \\ K_0 & K_1\end{bmatrix}\begin{bmatrix}\omega_r \\ T_L\end{bmatrix} + \begin{bmatrix}\frac{L_{af}i_f}{J} \\ 0\end{bmatrix} i_a \tag{1.6}$$

where the armature current i_a and rotor speed w_r are interpreted as the input and output of this subsystem, respectively.

To identify parameters in (1.5) an experimental test was carried out. During this test, the field winding was fed with a constant current at $i_f = 0.46$ amps. It is clear that under these conditions state equation in (1.6) becomes a linear time-invariant description. Basically, the test consists of applying a random sequence to the armature terminals. In order to obtain the experimental transient variables, a tool called FIFO that is included in the real time platform was used [2, 3]. The experimental setup used for identifying parameters in state equation (1.6) consists of an ATmega8 microcontroller for defining the random sequence, a Nana Electronics SHR-100 hall effect sensor for measuring the armature current, a Pepperl+Fuchs incremental encoder for measuring the rotor speed, a National Instruments data acquisition card, RTAI-Lab real time platform and a custom-made PWM H-bridge converter. Diagrams of the experimental setup for parametric identification can be found in [10]. Additionally, reference [10] contains the source program of the microcontroller for defining the random sequence. Once we obtained the transient experimental data, we utilized Matlab for the off-line processing of the data. In particular, prediction-error approach is used to calculate the numerical version of the state equation in (1.6). By comparing (1.6) and its corresponding numerical version, load torque parameters K_0 and K_1 are defined.

1.2.2 State Space Representation

Mostly of the physical systems presents non-linear dynamic behavior. However, under some conditions they may be considered to have a linear time-invariant representation. This is the case of a DC motor with a field winding that is fed from a constant source. Substituting the numerical parameters specified in Table 1.1 into

(1.2)–(1.5) and employing 5 kHz as a sample rate, it is possible to obtain the nominal linear time-invariant discrete time dynamic model. Thus we have

$$\begin{bmatrix}\theta_r(k+1)\\ \omega_r(k+1)\\ i_a(k+1)\\ T_L(k+1)\end{bmatrix} = \begin{bmatrix}1 & 0.000200 & 0.000004 & -0.000005\\ 0 & 0.000045 & 0.042383 & -0.052578\\ 0 & -0.002497 & 0.979644 & 0.000065\\ 0 & 0.000041 & 0.9\times10^{-6} & 0.998035\end{bmatrix} \begin{bmatrix}\theta_r(k)\\ \omega_r(k)\\ i_a(k)\\ T_L(k)\end{bmatrix} + \begin{bmatrix}4.4\times10^{-9}\\ 0.0000659\\ 0.0030691\\ 9.201\times10^{-10}\end{bmatrix} V_a(k) \tag{1.7}$$

$$\theta_r(k) = \begin{bmatrix}1 \; 0 \; 0 \; 0\end{bmatrix}\begin{bmatrix}\theta_r(k) \;\; \omega_r(k) \;\; i_a(k) \;\; T_L(k)\end{bmatrix}^T \tag{1.8}$$

where (1.7) is the discrete time state equation and (1.8) is the discrete time output equation.

It is clear the following definition for the state $x = \begin{bmatrix}\theta_r \; \omega_r \; i_a \; T_L\end{bmatrix}^T$, the output to be controlled $y = \theta_r$ and the input of the system $u = V_a$. State equation (1.7) assumes that parameters are linear time-invariant. Similarly, the field winding is supposed to be fed by a voltage source that keeps constant the field current at 0.46 amps. This is the reason for not including the state equation of the field winding. Experimental results in Section 4 will show that the nominal model in (1.7)–(1.8) contains the fundamental dynamic characteristics of the system necessary to design a successful experimental control law. Parametric uncertainty will be addressed by using an integrator, i.e.

$$x_I(k+1) = x_I(k) + e(k) = x_I(k) + y(k) - r(k) \tag{1.9}$$

where x_I is the integrator state variable, r is the setpoint and $y = \theta_r$ is the output to be controlled.

1.2.3 State Observer

On the other hand, since the experimental setup contains a PWM H-bridge converter some switching noise occurs. In addition, some problems regarding constructive imperfections of the incremental encoders make the measurement of the rotor speed unreliable [8]. In order to avoid these issues, we implement an experimental discrete-time state observer to estimate the rotor speed, armature current and even the load torque [4]. The state space equation of a standard linear time-invariant discrete time observer can be written as

$$\tilde{x}(k+1) = G\tilde{x}(k) + Hu(k) + K_e[y(k) - C\tilde{x}(k)] \tag{1.10}$$

The observer gain K_e is a 4x1 constant vector associated with the output error $y(k) - C\tilde{x}(k) = \theta_r(k) - \tilde{\theta}_r(k)$. As it is explained in the next subsection, desired poles are basically computed by specifying the desired time constant. Once the set of desired poles are obtained, a standard procedure for calculating the gain K_e is applied. This gain can also be computed by using the Scilab/Scicos *ppol* command. In this particular case, gains in (1.10) were chosen such that the desired observer poles become

$$z_{1,2,3,4} = \; 0.994017964, 0.994017963, 0.994017962, 0.994017961$$

this is

$$K_e = \begin{bmatrix} 0.0015523 \\ 0.1544085 \\ -0.0392419 \\ -0.0014389 \end{bmatrix} \tag{1.11}$$

It is important to emphasize that for the purpose of implementing the real time closed loop system, the rotor position is the only state variable to be measured. The state feedback uses the state variables (including the load torque) provided by the experimental discrete-time state observer. Armature current and rotor speed were measured just for the off-line parametric identification of the load torque model.

1.2.4 Controller

Pole placement technique is used to calculate the gains associated with the state feedback. Armature voltage $u(k) = V_a(k)$ is now defined by the following standard state feedback

$$u(k) = -\begin{bmatrix} K_I & K \end{bmatrix} \begin{bmatrix} x_I(k) \\ x(k) \end{bmatrix} \tag{1.12}$$

where K_I is the gain corresponding to the integral state variable, K is the gain vector associated with the state variable of the original system x. These gains are calculated following a standard procedure [4]. Gains can also be computed by using the Matlab command *place*. On the other hand, it is well known that the relationship between a s-plane pole having a particular time constant and the corresponding z-plane pole is given by $z = e^{-\frac{1}{\tau}T}$, where τ is the time constant and T is the sampling period. If $\tau = 0.1$ and $T = 0.0002$ s, thus the corresponding pole in the z-plane becomes $z_1 = e^{-\frac{1}{0.1}(0.0002)} = 0.998001998$. Similarly, the rest of desired controller poles can be obtained

$$\begin{aligned} z_{1,2,3,4,5} = {} & 0.998001998, 0.998001997, 0.998001996, 0.998001995, \\ & 0.998001994 \end{aligned}$$

thus the corresponding controller gains become

$$\begin{aligned} [K_I, K] = {} & [0.0006168, 1.2288494, -0.6467532, -4.021708, \\ & -2.4009488] \end{aligned} \tag{1.13}$$

1.3 Experimental Setup

1.3.1 Real-Time Platform

Linux and RTAI-Lab are the open source tools for real time tasks that we employed to solve the control algorithm [2, 6]. The resulting executable program is able to provide correctness of the calculated values and to accomplish strict time requirements. In this case, the control algorithm was computed every 0.0002 s (5 kHz). The real time program consists of a discrete-time state feedback including an integrator, a full order discrete-time state observer and a rotor position sensing algorithm. This program is able to access the analog/digital input/output terminals of the data acquisition card. In this particular work, a National Instruments PCI-6024E data acquisition card is used. The source program is compiled by using the RTAIcodegen tool which is included in the RTAI-Lab package [2]. The Knoppix 5.0 Linux distribution contains RTAI-Lab. Scilab/Scicos is a graphical environment that includes a library of blocks that can be employed to simulate the closed loop system [3]. Scilab/Scicos are used to simulate the closed-loop system as a first step of the implementation. Then, the Scilab/Scicos simulation blocks are switched to the corresponding RTAI-Lib real-time blocks. The main source program is shown in Fig. 1.1. In order to be able to show it in just one figure we used several subroutines that are called super blocks. They are a tool for organizing the structure of the program. A super block is similar to any other block, however its operation is defined by a custom-made Scicos code. The main program can be divided in three parts. The first part is the super block that includes the position measurement algorithm. The second part is the super block that computes the state observer. The last part calculates the state feedback and solves the integral state space equation. Closed

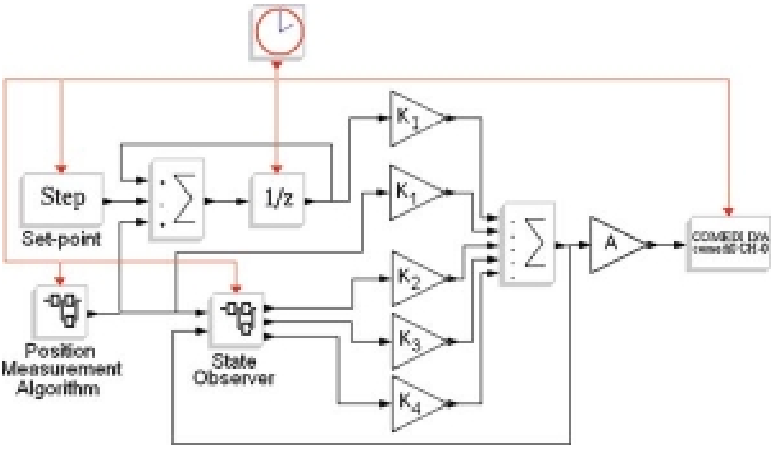


Fig. 1.1 Main Scicos program including several super blocks (subroutines)

loop gains K_I, K_1, K_2, K_3 and K_4 are defined in (1.13). In addition, the program includes the set-point block, the Comedi block (DAC) and a gain denoted by A. This gain scales the numerical value of the calculated armature voltage.

1.3.2 Rotor Position Measurement Design

A design for measuring the rotor position is implemented. There are several components of this hardware/software design. One of them is the Pepperl+Fuchs encoder that is physical attached to the motor shaft. The second one is an ATmega8535 microcontroller-based signal conditioning circuit and the third one is the real time software that calculates the rotor position from the digital signals provided by the microcontroller. The ATmega8535 microcontroller is programmed for being used as a 16-bit binary UP/DOWN counter. A signal conditioning circuit is necessary to determine the shaft direction, which is implemented by using a flip-flop integrated circuit. The flip-flop uses the pulses of channels A and B of the encoder and provides a binary 1 or 0 depending on the shaft direction. Channel A (or B) of the encoder is then connected to a micro-controller terminal. In this case, the encoder generates 1,024 pulses per revolution. The micro-controller count goes up or down depending on the flip-flop binary signal. The complete electronics diagram of the rotor position measurement setup is shown in Fig. 1.2. The micro-controller provides a 16-bit binary count that represents the rotor position. The algorithm to interpret the 16-bit binary count is coded in the Scilab/Scicos source program. A part of this algorithm is shown in Fig. 1.3. The algorithm consists of multiplying each bit (1 or 0) by its corresponding value in the decimal system. Once the results are added a decimal measurement of the rotor position is obtained.

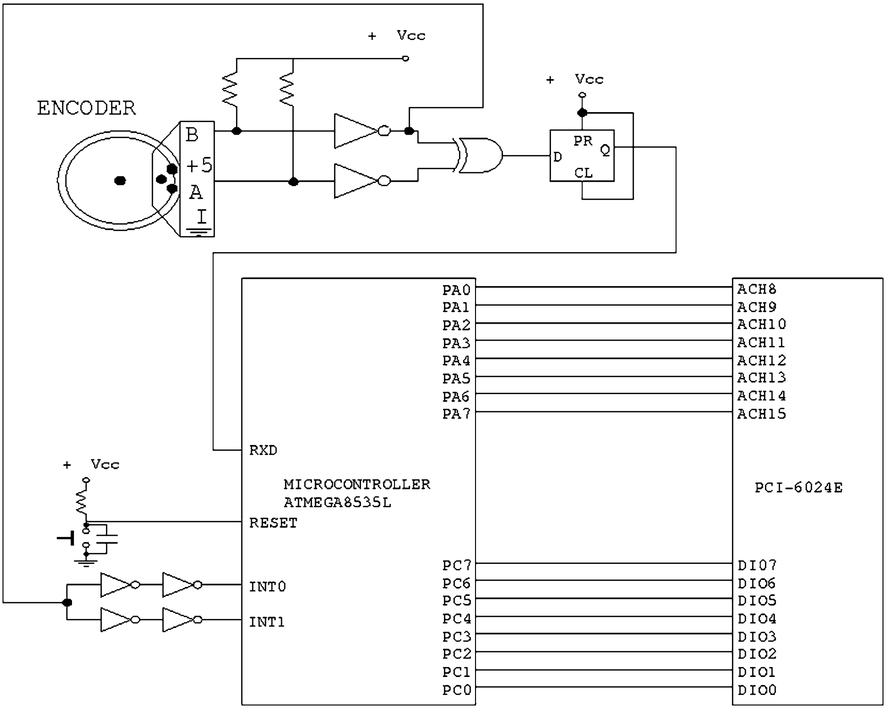


Fig. 1.2 Rotor position measurement setup

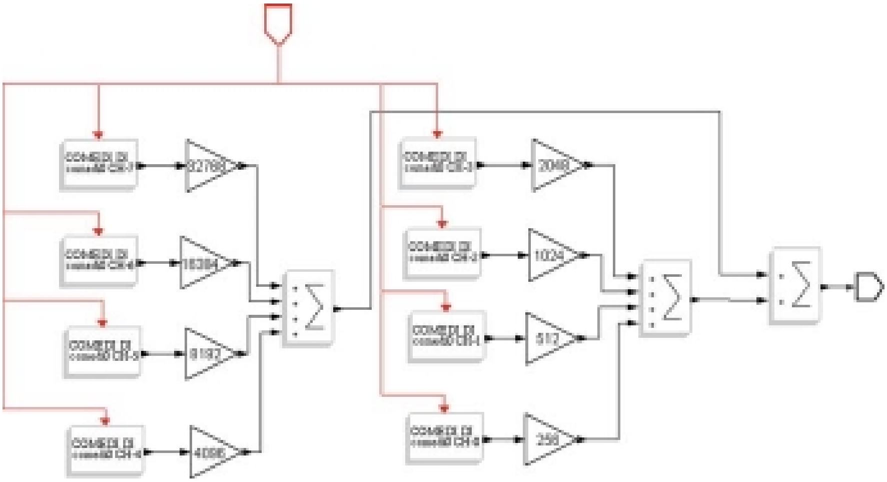


Fig. 1.3 Subroutine associated with the digital inputs of the position measurement algorithm

A final super block called *Position Measurement Algorithm* (see Fig. 1.1) was created to include the described code. It is important to mention that the rotor position setup is able to measure positive and negative values. In order to accomplish this feature, the microcontroller was programmed to have an initial count that is located at the middle of the 16-bit count range.

1.3.3 Power Converter

A Pulse Width Modulation-based MOSFET H-Bridge converter was designed and implemented. This type of power electronics device is commonly used to drive DC motors when bidirectional speed/position control is needed. The numerical value of the armature voltage is defined by the state feedback and calculated by the computer. This value is written by the real time program to one of the analog output channels of the data acquisition card. The power electronics converter and motor are shown in Fig. 1.4.

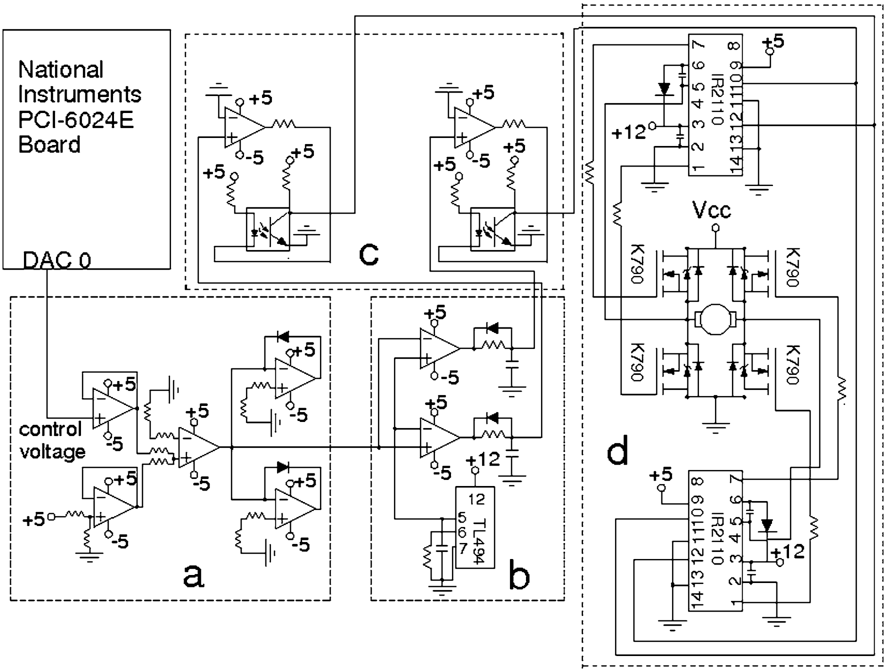


Fig. 1.4 Power converter and DC motor. (**a**) Signal conditioning. (**b**) PWM and switching delay. (**c**) Isolation. (**d**) H bridge and DC motor

1.4 Experimental Results

This section presents the experimental transient characteristics of the closed loop system. The armature voltage is computed as a function of estimated states excepting the rotor position, which is the only state variable that is measured. Estimated states, including the load torque, are obtained by solving the full order discrete-time state observer in (1.10). Initial conditions of the integrator and observer state variables were set numerically equal to zero. The experimental trace shown in Fig. 1.5 illustrates the dynamic characteristic of the rotor position following a 25.1328 *radian* (4 *revolution*) reference command. The simulated model-based trace of the rotor position is also shown in Fig. 1.5. It is clear that these signals (simulated and measured) are similar to each other. Initially, the rotor position was at zero radians. The position begins to increase immediately until the position error is close to zero, which occurs approximately at 1.2 s. In order to save these transient data, a real-time block called FIFO was used [2,3]. Experimental observer-based variables were also saved by using FIFO. One of these variables is shown in Fig. 1.6. The armature voltage computed by the real time program is depicted in Fig. 1.7.

For the purpose of testing the experimental closed loop system in a more demanding operating condition, we installed an inertial disk. The original value of the inertia was 0.0038 kg/m^2, see Table 1.1. The new inertia is 0.0398 kg/m^2. In other words, the new parameter (inertia) is about 10.5 times the original value. It means a significant increase of a particular parameter of the mechanical subsystem. It is clearly connected to the output to be controlled (rotor position). In addition, the inertial disk was intentionally slightly loose to the shaft, therefore the parameter varied during the test around the non-original value. The original gains that were calculated by using the nominal parameters in Table 1.1 were not changed during

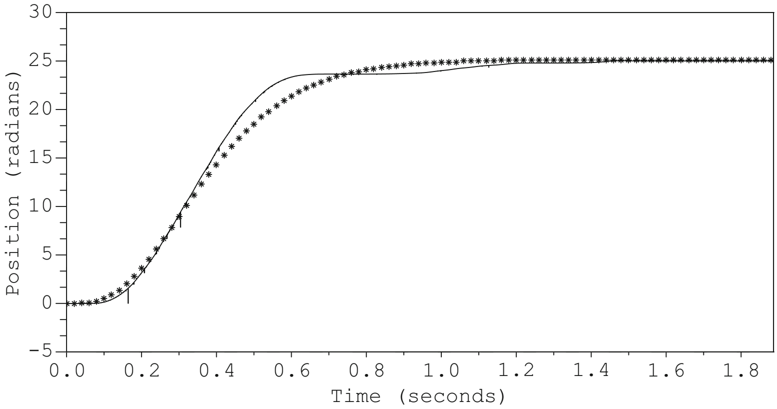


Fig. 1.5 Rotor position: – Experimental measured, ** Simulated model-based

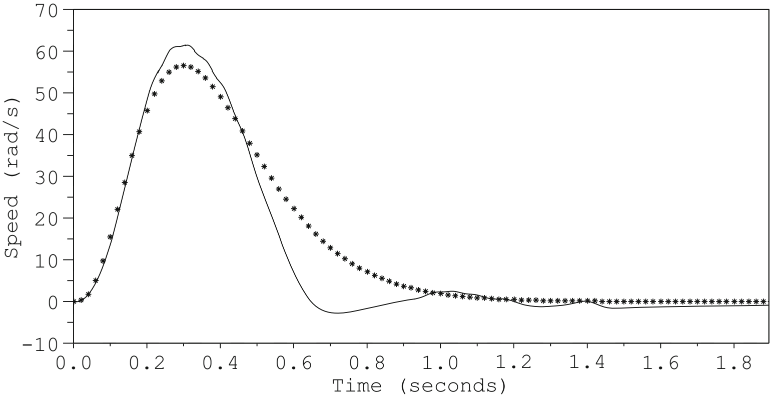


Fig. 1.6 Rotor speed: – Experimental observer-based, ** Simulated model-based

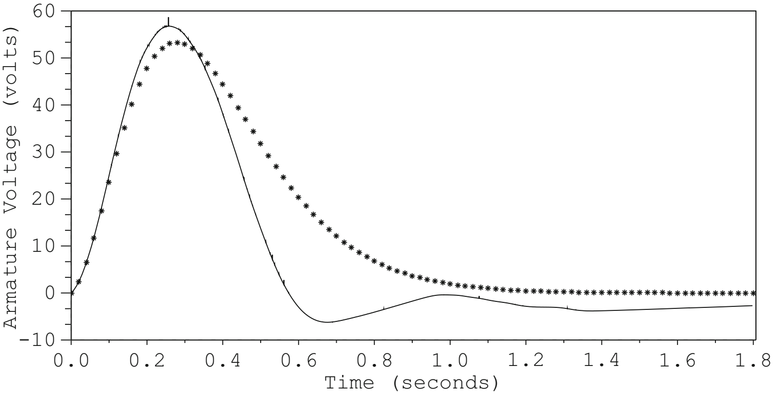


Fig. 1.7 Armature voltage: – Experimental, ** Simulated model-based

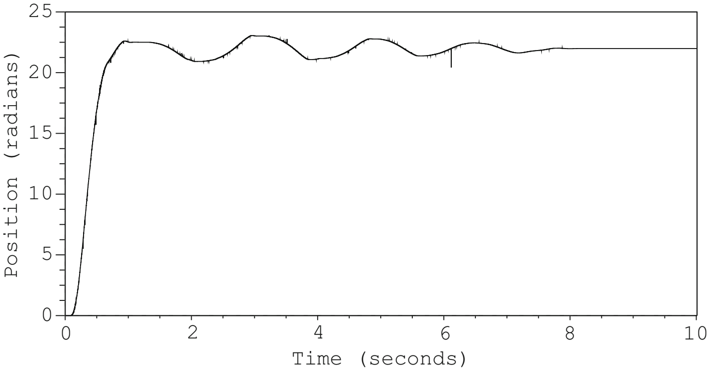


Fig. 1.8 Experimental measured rotor position when a significant change of a parameter occurs (10.5 times the original value approx.)

this test. Even under that demanding condition the experimental closed loop system was able to follow a 21.99 *radian* (3.5 *revolution*) reference command, see Fig. 1.8.

References

1. Boizot, N., Busvelle, E., Sachau, J.: High-gain observers and Kalman filtering in hard real-time. RTL 9th Workshop, 2007
2. Bucher, R., Mannori, S., Netter, T.: RTAI-Lab tutorial: Scilab, comedi and real-time control. http://www.rtai.org. 2008
3. Campbell, S.L., Chancellier, J.P., Nikoukhah, R.: Modeling and simulation in Scilab/Scicos. Springer, New York (2006)
4. Franklin, G.F., Powell, J.D., Workman, M.: Digital control of dynamic systems. Addison Wesley, Menlo Park, CA (1998)
5. Jugo, J.: Real-time control of a DC motor using Scilab and RTAI. Scilab INRIA Rocquencourt, 2004
6. Knoppix: Open source Linux distribution. Knoppix 5.0, 2004
7. Krause, P.C., Wasynczuk, O., Sudhoff, S.D.: Analysis of electrical machinery and drive systems. Wiley, IEEE Press Power Engineering (2004)
8. Merry, R., van de Molengraft, R., Steinbuch, M.: Error modeling and improved position estimation for optical incremental encoders by means of time stamping. American Control Conference, 2007
9. National-Instruments: PCI-6024E National Instruments Manual (2006)
10. Rendon-Fraga, E.Y.: Rotor position control of a DC motor using a real time platform. M.Sc. thesis (in spanish). Instituto Tecnologico de Cd. Madero. Cd. Madero, Mexico (2009)
11. Schleef, D., Hess, F., Bruyninckx, H.: The control and measurement device interface handbook. http://www.comedi.org, 2003
12. Zhao, D., Zhang, N.: An improved nonlinear speed controller for series DC motors. IFAC 17th World Congress, 2008

Chapter 2
MASH Digital Delta–Sigma Modulator with Multi-Moduli

Tao Xu and Marissa Condon

Abstract This chapter proposes a novel design methodology for a Multi-stAge noise SHaping (MASH) digital delta–sigma modulator (DDSM) which employs multi-moduli. The structure is termed the MM-MASH. The adequacy and benefit of using such a structure is demonstrated. In particular, the sequence length is lengthened if the quantizer modulus of the first-order error feedback modulator (EFM1) of each stage are co-prime numbers. An expression for the sequence length of an MM-MASH is derived.

Keywords Delta–sigma modulation · sequence length · multi-moduli · co-prime numbers

2.1 Introduction

The digital delta–sigma modulator (DDSM) sometimes acts as the controller of the multi-modulus frequency divider in the feedback loop of the Fractional-N Frequency Synthesizers [4]. Since the DDSM is a finite state machine, when the input is a DC rational constant, the output is always a repeating pattern (limit cycle) [2, 8]. The period of the cycle is termed the sequence length. For this type of input, the quantization noise is periodic. When a sequence length is short, the power is distributed among spurious spurs that appear in the DDSM output spectrum. Hence, there is a desire to break short sequences. Dithering [10, 11] is one of the most commonly employed methods to break the short sequence length. However, it requires extra hardware and inherently introduces additional inband noise. Recently,

T. Xu (✉) and M. Condon
RF Modelling and Simulation Group, Research Institute for Networks and Communications Engineering (RINCE), School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland
e-mail: tao.xu3@mail.dcu.ie; marissa.condon@dcu.ie

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*, Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_2,

some design methodologies have been proposed to maximise the sequence length. Borkowski [2] summarises the maximum period obtained by setting the initial condition of the registers in an EFM. Hosseini [7] introduced a digital delta–sigma modulator structure termed the HK-MASH with a very long sequence. The period of the HK-MASH is proven by mathematical analysis [6, 7]. This paper proposes a novel architecture to further increase the modulator sequence length.

This chapter proposes that the modulus of each quantizer in a DDSM is set as a different value from each other. Note that each quantizer has only one modulus. Furthermore, all of the moduli are co-prime numbers. Co-prime numbers [3] are two real numbers whose greatest common divisor is 1. They do NOT have to be prime numbers. A novel design methodology based on this concept is proposed. It will be shown to increase the modulator period and consequently reduce the effect of quantization noise on the useful output frequency spectrum of the DDSM.

In Section 2.2, the architectures of the classic MASH DDSM and the HK-MASH are reviewed. In Section 2.3, a novel structure is proposed that results in the maximum sequence length. The expression for the sequence length is derived as well. The simulation results are shown in Section 2.5.

2.2 Previous MASH Architectures

The architecture of an lth order MASH digital delta–sigma modulator (DDSM) is illustrated in Fig. 2.1. It contains l first-order error-feedback modulators (EFM1). $x[n]$ and $y[n]$ are an n_0-bit input digital word and an m-bit output, respectively. The relationship between them is

$$mean(y) = \frac{X}{M} \tag{2.1}$$

where X is the decimal number corresponding to the digital sequence $x[n]$ [1], i.e., $x[n] = X \in \{1, 2, \ldots, M\}$, and M is the quantizer modulus which is set as 2^{n_0} in the conventional DDSM.

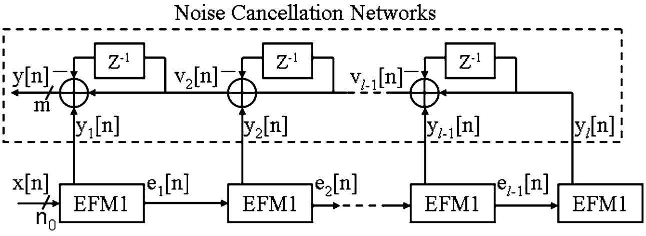


Fig. 2.1 MASH DDSM architecture

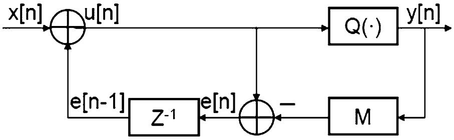


Fig. 2.2 EFM1: first-order error-feedback modulator

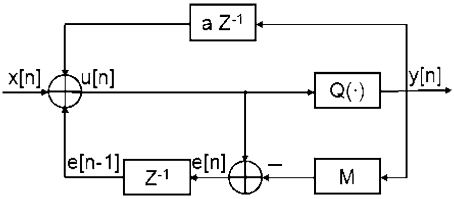


Fig. 2.3 The modified EFM1 used in HK-MASH

The model of the EFM1 is shown in Fig. 2.2. This is a core component in the make-up of the MASH digital delta–sigma modulator (DDSM). The rectangle Z^{-1} represents the register which stores the error $e[n]$ and delays it for one time sample. $Q(\cdot)$ is the quantization function:

$$y[n] = Q(u[n]) = \begin{cases} 1, & u[n] \geq M \\ 0, & u[n] < M \end{cases} \tag{2.2}$$

where

$$u[n] = x[n] + e[n-1]. \tag{2.3}$$

The maximum sequence lengths for this structure are 2^{n_0+1} or 2^{n_0+2} when the modulator order is below 5, which is found from simulations [2]. To achieve both of these sequence lengths, the first stage EFM1 must have an odd initial condition. This is implemented by setting the register.

The architecture of the modified EFM1 used in the HK-MASH is illustrated in Fig. 2.3. The only difference between it and the conventional EFM1 in Fig. 2.2 is the presence of the feedback block aZ^{-1}. a is a specifically-chosen small integer to make $(M - a)$ the maximum prime number below 2^{n_0} [7]. The sequence length of it is $(2^{n_0} - a)^l \approx (2^{n_0})^l$. This value will be compared with that of the proposed novel MM-MASH in Section 2.4.

2.3 The Adequacy and Effect of the Multi-Modulus MASH-DDSM

We assume that the quantizer modulus of the ith stage EFM1 in an lth order MASH-DDSM is represented by M_i, where $i \in \{1, 2, \ldots, l\}$. It shall first be shown that the MM-MASH is an accurate modulator. Then the effect of the multi-moduli on the sequence length shall be investigated mathematically.

2.3.1 The Suitability of the Multi-Modulus MASH-DDSM

In a fractional-N frequency synthesizer, the static frequency divider is controlled by the average value of the delta–sigma modulator output, $mean(y)$. The goal is to show that in an MM-MASH, $mean(y)$ is affected only by the quantizer modulus of the first stage EFM1, M_1, and is independent of the moduli in other stages. With this being true, having a multi-modulus architecture does not affect the accuracy of the digital delta–sigma modulator. Hence, it is a suitable digital delta–sigma modulator. Required to prove:

$$mean(y) = \frac{X}{M_1}. \tag{2.4}$$

Proof. As seen in Fig. 2.1, at the output of the last adder,

$$v_{l-1}[1] = y_{l-1}[1] + y_l[1] - y_l[0] \tag{2.5}$$

$$v_{l-1}[2] = y_{l-1}[2] + y_l[2] - y_l[1] \tag{2.6}$$

$$\vdots$$

$$v_{l-1}[N] = y_{l-1}[N] + y_l[N] - y_l[N-1] \tag{2.7}$$

where N is assumed as the sequence length of the MASH delta–sigma modulator. Adding all of the above equations yields:

$$\sum_{k=1}^{N} v_{l-1}[k] = \sum_{k=1}^{N} y_{l-1}[k] + \sum_{k=1}^{N} y_l[k] - \sum_{k=0}^{N-1} y_l[k] \tag{2.8}$$

where the period of y_l is assumed as N_l. As seen in Fig. 2.1, the output of the MASH modulator is obtained by simply summing and/or subtracting the output of each EFM1. Hence, the period of the MASH DDSM is the least common multiple of the sequence length of each stage. In other words, N is a multiple of N_i, where N_i is the period of the ith stage EFM1 and $i \in \{1, 2, \ldots, l\}$. It follows that

$$\sum_{k=1}^{N} y_l[k] = \sum_{k=0}^{N-1} y_l[k]. \tag{2.9}$$

Thus (2.8) becomes:

$$\sum_{k=1}^{N} v_{l-1}[k] = \sum_{k=1}^{N} y_{l-1}[k]. \tag{2.10}$$

Similarly, each of the other adders' output is obtained as

$$\sum_{k=1}^{N} v_{l-2}[k] = \sum_{k=1}^{N} y_{l-2}[k] \tag{2.11}$$

$$\vdots$$

$$\sum_{k=1}^{N} v_{2}[k] = \sum_{k=1}^{N} y_{2}[k] \tag{2.12}$$

$$\sum_{k=1}^{N} y[k] = \sum_{k=1}^{N} y_{1}[k]. \tag{2.13}$$

Each side of (2.13) may be expressed as

$$\sum_{k=1}^{N} y[k] = N \cdot mean(y) \tag{2.14}$$

$$\sum_{k=1}^{N} y_1[k] = K \sum_{k=1}^{N_1} y_1[k] \tag{2.15}$$

where N_1 is the sequence length of y_1, K is an integer and $N = K \cdot N_1$. Since y_1 is the output of a first-order delta–sigma modulator EFM1,

$$\sum_{k=1}^{N_1} y_1[k] = N_1 \cdot mean(y_1) = N_1 \cdot \frac{X}{M_1}. \tag{2.16}$$

On substitution of (2.16) into (2.15), the right-hand side of (2.13) becomes

$$\sum_{k=1}^{N} y_1[k] = K \cdot N_1 \cdot \frac{X}{M_1} = N \cdot \frac{X}{M_1}. \tag{2.17}$$

By substituting (2.14) and (2.17) into (2.13), the average value of the MASH DDSM output y is determined as

$$mean(y) = \frac{X}{M_1}. \tag{2.18}$$

■

2.3.2 The Effect of the Multi-Moduli on the Modulator Sequence Length

It is required to prove that the sequence length of the MASH modulator depends on the product of all the quantizer moduli. The expression for the lth order MASH DDSM sequence length is

$$N = \frac{M_1 \cdot M_2 \cdot \ldots \cdot M_l}{\lambda} \tag{2.19}$$

where λ is a parameter to make N the least common multiple of the sequence length of each stage N_i.

In addition, if the following two conditions, C1 and C2, are satisfied:

1. X and M_1 are co-prime numbers
2. $\{M_1, M_2, \ldots, M_l\}$ are co-prime numbers

then the sequence length of the MASH DDSM attains the maximum value:

$$N_{max} = M_1 \cdot M_2 \cdot \ldots \cdot M_l. \tag{2.20}$$

Proof: In the first-stage EFM1 shown in Fig. 2.2,

$$\begin{aligned} e_1[1] &= u[1] - y_1[1]M_1 \\ &= X + e_1[0] - y_1[1]M_1 \end{aligned} \tag{2.21}$$

$$e_1[2] = X + e_1[1] - y_1[2]M_1 \tag{2.22}$$

$$\vdots$$

$$e_1[N_1] = X + e_1[N_1 - 1] - y_1[N_1]M_1 \tag{2.23}$$

where $e_1[0]$ is the initial condition of the register. The sum of all of the above equations is

$$\sum_{k=1}^{N_1} e_1[k] = N_1 X + \sum_{k=0}^{N_1-1} e_1[k] - \sum_{k=1}^{N_1} y_1[k]M_1. \tag{2.24}$$

Since in the steady state, the first EFM1 is periodic with a period N_1 [5],

$$\sum_{k=1}^{N_1} e_1[k] = \sum_{k=0}^{N_1-1} e_1[k]. \tag{2.25}$$

Hence, (2.24) may be modified to

$$\sum_{k=1}^{N_1} y_1[k] = \frac{N_1 X}{M_1}. \tag{2.26}$$

In practice, the input DC X is set as $0 < X < M_1$. So in order to make the right side of (2.26) an integer, the minimum nonzero solution of N_1 has to be

$$N_1 = \frac{M_1}{\lambda_1} \tag{2.27}$$

where λ_1 is the greatest common divisor of M_1 and X. If M_1 and X are co-prime numbers, λ_1 equals to 1.

If the process of (2.21)–(2.26) is repeated with the second EFM1, the sum of its output which has a period N_2 is obtained as

$$\sum_{k=1}^{N_2} y_2[k] = \frac{\sum_{k=1}^{N_2} e_1[k]}{M_2}. \tag{2.28}$$

If the relationship between the sequence lengths of the first and second stages is

$$N_2 = K_1 N_1 \tag{2.29}$$

(2.28) becomes

$$\sum_{k=1}^{N_2} y_2[k] = \frac{\sum_{k=1}^{K_1 N_1} e_1[k]}{M_2}. \tag{2.30}$$

Since e_1 is periodic with the sequence length N_1 [7],

$$\sum_{k=1}^{N_2} y_2[k] = \frac{K_1 \sum_{k=1}^{N_1} e_1[k]}{M_2} \tag{2.31}$$

where

$$\sum_{k=1}^{N_1} e_1[k] = N_1 \cdot mean(e_1). \tag{2.32}$$

Recalling (2.27),

$$\sum_{k=1}^{N_1} e_1[k] = \frac{M_1 \cdot mean(e_1)}{\lambda_1}. \tag{2.33}$$

On substitution of (2.33) into (2.31), the following expression is obtained

$$\sum_{k=1}^{N_2} y_2[k] = \frac{K_1 \cdot M_1 \cdot mean(e_1)}{\lambda_1 \cdot M_2} \tag{2.34}$$

Normally $mean(e_1)$ is a decimal fraction. However, if both sides of (2.33) are multiplied by λ_1, the result is

$$\lambda_1 \sum_{k=1}^{N_1} e_1[k] = M_1 \cdot mean(e_1). \tag{2.35}$$

Thus, $M_1 \cdot mean(e_1)$ is always an integer.

Then the minimum solution of K_1 so that the right-hand-side of (2.34) is an integer is obtained as

$$K_1 = \frac{\lambda_1 M_2}{\lambda_2} \tag{2.36}$$

where λ_2 is the greatest common divisor of $\lambda_1 M_2$ and $M_1 mean(e_1)$. Substituting (2.27) and (2.36) into (2.29), the sequence length of the second stage is

$$N_2 = \frac{M_1 M_2}{\lambda_2}. \tag{2.37}$$

If M_1 and M_2 are co-prime numbers, the greatest common divisor of $\lambda_1 M_2$ and $M_1 mean(e_1)$ is λ_1, i.e., $\lambda_2 = \lambda_1$. Hence,

$$N_2 = \frac{M_1 M_2}{\lambda_1}. \tag{2.38}$$

When X and M_1 are also co-prime numbers, λ_1 equals to 1. Thus the maximum sequence length for y_2 is obtained as:

$$N_{2\_max} = M_1 M_2. \tag{2.39}$$

Continuing in this manner, the sequence length of the ith effective stage EFM1 in an lth order MASH modulator is

$$N_i = \frac{M_1 M_2, \ldots, M_i}{\lambda_i} \tag{2.40}$$

where $i \in \{1, 2, 3, \ldots, l\}$ and λ_i is the maximum common divisor of $\lambda_{i-1} M_i$ and $M_1 M_2, \ldots, M_{i-1} mean(e_{i-1})$. Note that when $i = 1, mean(e_0) = X$ and $\lambda_0 = M_0 = 1$.

If $\{M_1, M_2, \ldots, M_i\}$ are co-prime numbers, M_i and $M_1 M_2, \ldots, M_{i-1}$ have to be co-prime numbers as well. Thus $\lambda_i = \lambda_{i-1}$. Since M_{i-1} and $M_1 M_2, \ldots, M_{i-2}$ are also co-prime numbers, $\lambda_{i-1} = \lambda_{i-2}$. Repeating this manner, it is follows that

$$\lambda_i = \lambda_{i-1} = \ldots = \lambda_1. \tag{2.41}$$

Then the sequence length of the ith EFM1 becomes

$$N_i = \frac{M_1 M_2, \ldots, M_i}{\lambda_1} \tag{2.42}$$

where λ_1 is the greatest common divisor of X and M_1. In practice, if the input X and M_1 are set as co-prime numbers, the maximum sequence length of the ith stage EFM1 is

$$N_{i\_max} = M_1 M_2, \ldots, M_i. \tag{2.43}$$

Since N is the least common multiple of $\{N_1, N_2, \ldots, N_l\}$, as is proven in Section 2.3.1, the sequence length of the MASH DDSM is obtained as

$$N = \frac{M_1 \cdot M_2 \cdot \ldots \cdot M_l}{\lambda} \tag{2.44}$$

where λ is the least common multiple of $\{\lambda_1, \lambda_2, \ldots, \lambda_l\}$.

When $\{M_1, M_2, \ldots, M_l\}$ are co-prime numbers, (2.41) is true. Then

$$N = \frac{M_1 \cdot M_2 \cdot \ldots \cdot M_l}{\lambda_1}. \tag{2.45}$$

In addition, if X and M_1 are co-prime numbers as well, λ_1 becomes 1. Thus the maximum sequence length is

$$N_{max} = M_1 \cdot M_2 \cdot \ldots \cdot M_l. \tag{2.46}$$

■

2.4 The Proposed Structure and Simulation Results

A novel structure for the MASH digital delta–sigma modulator (DDSM) employing multi-moduli is proposed in this section. It is termed the Multi-Modulus MASH – MM-MASH. As illustrated in Fig. 2.4, M_i represents the quantizer modulus in ith stage of MM-EFM1.

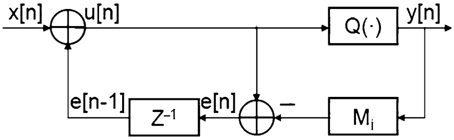


Fig. 2.4 MM-EFM1: The modified first-order error-feedback modulator used in MM-MASH

Table 2.1 Some sample moduli of the third order MM-MASH

Word length	M_1	M_2	M_3
5 bit	31	32	29
8 bit	251	256	255
9 bit	509	512	511
10 bit	1,021	1,024	1,023
11 bit	2,039	2,048	2,047

M_1 is set as a prime number below 2^{n_0}. This is to make X and M_1 always co-prime numbers and therefore satisfy the first condition, C1, stated in the previous section. This condition must be satisfied to maximise the sequence length of the MASH DDSM and to make the sequence length independent of the value of input. In order to maintain the modulator output accuracy, the value of the input DC X is adjusted to

$$X = M_1 \cdot mean(y) \tag{2.47}$$

where $mean(y)$ is the required output to control the static frequency divider in a fractional-N frequency synthesizer.

In an lth order MM-MASH, there are l co-prime numbers around 2^{n_0} that need to be found in order to satisfy the second condition, C2. The higher the modulator order, the greater difficulty in finding suitable values for these moduli. Fortunately, the most popular MASH DDSM in modern communication systems is the third order one [2]. Note that all of the quantizer moduli should be chosen no bigger than 2^{n_0} to avoid necessitating extra hardware. Some quantizer moduli chosen by the authors for a third order MM-MASH are given in Table 2.1.

2.5 The Simulation Results

All of the models of the EFM1 and MASH DDSM are built and simulated in **Simulink**. The simulation confirms that the average value of the output $mean(y)$ equals to $\frac{X}{M_1}$. The sequence length of the MASH DDSM is determined using the autocorrelation function [2]. Figure 2.5 shows that the sequence length of a third-order 5-bit MM-MASH is 28,768 and this equals $M_1 \cdot M_2 \cdot M_3$ as given in Table 2.1. The sequence length is only 64 from a third-order 5-bit conventional MASH. The sequence lengths of the HK-MASH and the MM-MASH are compared in Table 2.2. The MM-MASH achieves a longer sequence when the word length is 8, 9, 10 and 11, and hence is deemed superior.

The power spectral density [9] of the third-order 9-bit MM-MASH and a dithered [10] third-order 9-bit conventional MASH DDSM is compared in Fig. 2.6. It is evident from the figure that the MM-MASH is significantly more effective than the conventional MASH DDSM at the important useful lower frequencies.

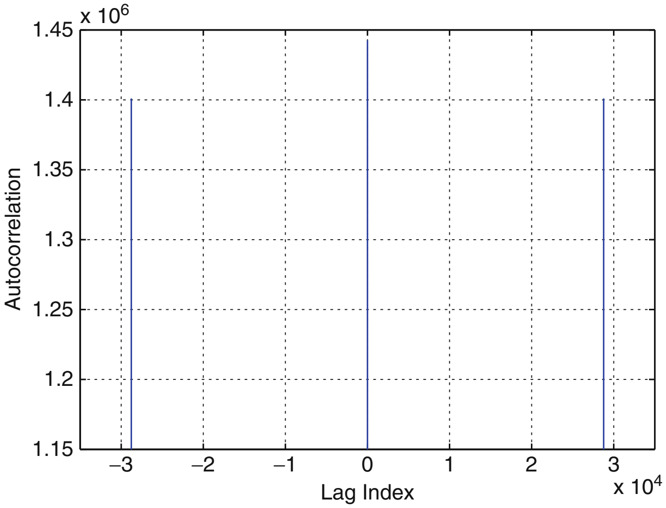


Fig. 2.5 The autocorrelation result for the third-order 5-bit MM-MASH

Table 2.2 A comparison of the sequence lengths for the HK-MASH and MM-MASH

Word length	HK-MASH	MM-MASH	Difference
8 bit	15.81×10^6	16.39×10^6	$+0.58 \times 10^6$
9 bit	131.87×10^6	133.17×10^6	$+1.3 \times 10^6$
10 bit	1.06×10^9	1.07×10^9	$+10 \times 10^6$
11 bit	8.48×10^9	8.55×10^9	$+70 \times 10^6$

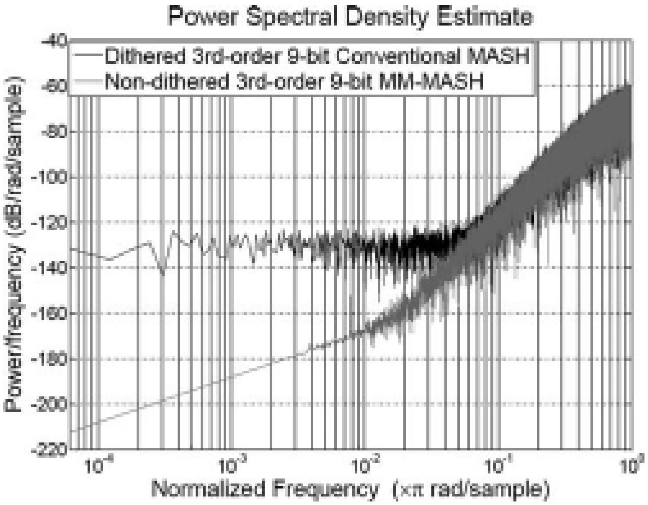


Fig. 2.6 The power spectral density of the dithered conventional MASH DDSM and non-dithered MM-MASH

2.6 Conclusions

A novel structure for the MASH digital delta–sigma modulator employing the multi-moduli (MM-MASH). This method employs different moduli in each stage of the EFM1. It is proven that the multi-modulus architecture is suitable because the output of the MASH modulator is only dependent on the quantizer modulus of the first stage EFM1 and independent of the others. The expressions for the sequence length of the EFM1 of each stage and for the complete MASH DDSM are derived. There are two conditions given that must be satisfied to yield the maximum modulator period. The outcome is a structure with an increased sequence length and hence an improved noise performance.

References

1. Borkowski, M.J., Kostamovaara, J.: Spurious tone free digital delta-sigma modulator design for dc inputs. In: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 5601–5604. Kobe, Japan (2005)
2. Borkowski, M.J., Riley, T.A.D., Häkkinen, J., Kostamovaara, J.: A practical Δ-Σ modulator design method based on periodical behavior analysis. IEEE Trans. Circ. Syst.-II: Express Briefs **52**(10), 626–630 (2005)
3. Crandall, R.E., Pomerance, C.: Prime Numbers: A Computational Perspective. Springer, Germany (2001)
4. Gu, R., Ramaswamy, S.: Fractional-*N* phase locked loop design and applications. In: Proceedings of the 7th International Conference on ASIC, pp. 327–332. Guilin, China (2007)
5. Hosseini, K., Kennedy, M.P.: Mathematical analysis of digital mash delta-sigma modulator for fractional-*N* frequency synthesizers. In: Proceedings of the IEEE Ph.D. Research in Microelectronics and Electronics, pp. 309–312. Otranto (Lecce), Italy (2006)
6. Hosseini, K., Kennedy, M.P.: Mathematical analysis of a prime modulus quantizer mash digital delta-sigma modulator. IEEE Trans. Circ. Syst.-II: Express Brief **54**(12), 1105–1109 (2007)
7. Hosseini, K., Kennedy, M.P.: Maximum sequence length mash digital delta-sigma modulators. IEEE Trans. Circ. Syst.-I: Regular Papers **54**(12), 2628–2638 (2007)
8. Kozak, M., Kale, İ.: Rigorous analysis of delta-sigma modulators for fractional-*n* pll frequency synthesis. IEEE Trans. Circ. Syst.-I: Regular Papers **51**(6), 1148–1162 (2004)
9. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing. Prentice-Hall, Upper Saddle River, NJ (1999)
10. Pamarti, S., Galton, I.: Lsb dithering in mash delta-sigma d/a converters. IEEE Trans. Circ. Syst.-I: Regular Papers **54**(4), 779–790 (2007)
11. Pamarti, S., Welz, J., Galton, I.: Statistics of the quantization noise in 1-bit dithered single-quantizer digital delta-sigma modulators. IEEE Trans. Circ. Syst.-I: Regular Papers **54**(3), 492–503 (2007)

Chapter 3
Sensorless PM-Drive Aspects

**Christian Grabner, Johannes Gragger, Hansjoerg Kapeller,
Anton Haumer, and Christian Kral**

Abstract The development procedure of permanent magnet drives for sensor less operation beginning from standstill under overload conditions has to consider different design aspects coevally. First, the robust rotor position sensing by test signal enforces a design with a strongly different behavior of the spatial dq-oriented differential inductance values. Therefore, the interior rotor magnet array arrangement is from principle predestinated for the controlled sensor less mode including standstill. Fortunately, in order to reduce costs, the distinct reluctance torque capability of such interior magnet arrangement is additionally used for a significantly increased torque by applying a pre-oriented stator current space vectors within the quasi-steady control.

Keywords Sensorless vector control · machine design · inductance modeling · nonlinear saturation effects

3.1 Introduction

The overall design process of modern high performance and cost effective permanent magnet drives is still a complex multi physic task, as shown in Fig. 3.1 [1]. In order to achieve both a high efficiency and a simple controllability of the motor, many proposed designs have to be investigated systematically by the ‘virtual design’ approach based on manifold numerical commercial available simulation tools [2]. During the electromagnetic machine, the electrical power converter and the software based control design process a lot of high sophisticated simulations take place in order to verify the desired performance behavior in advance [3, 4].

C. Grabner (✉), J. Gragger, H. Kapeller, A. Haumer, and C. Kral
Electric Drive Technologies, Austrian Institute of Technology,
Giefinggasse 2, 1210 Vienna, Austria
e-mail: christian.grabner@ait.ac.at; johannes.gragger@ait.ac.at; hansjoerg.kapeller@ait.ac.at; anton.haumer@ait.ac.at; christian.kral@ait.ac.at

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_3,

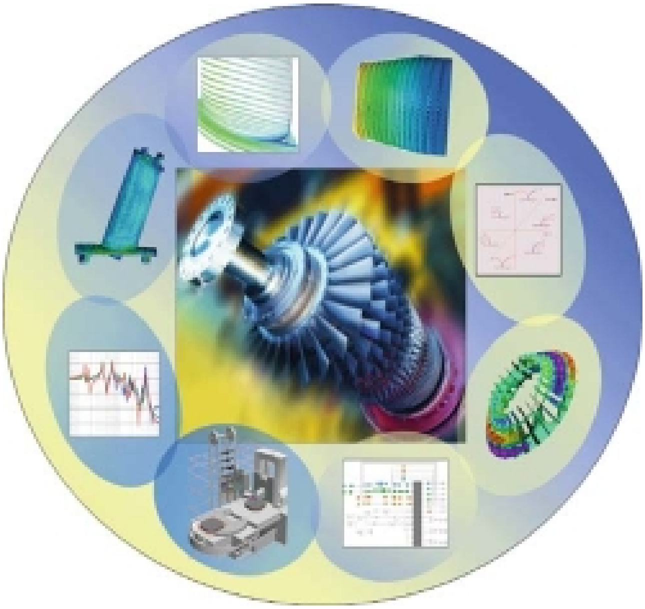


Fig. 3.1 Multi-physical design process of drives

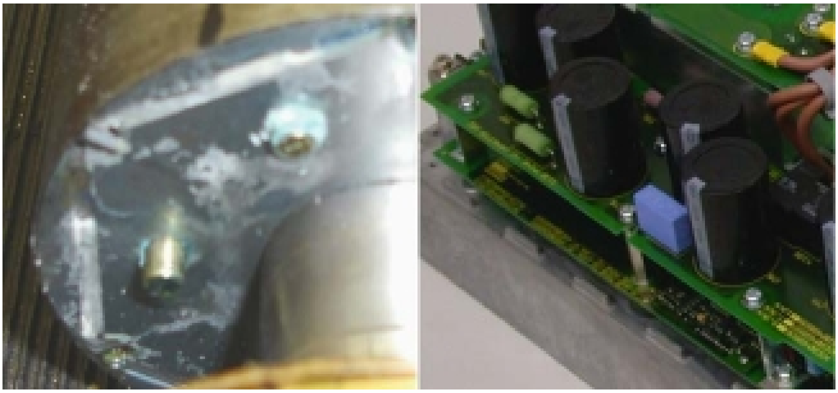


Fig. 3.2 Typical interior permanent magnet machine construction (*left*) and power electronic with DC-link (*right*)

It takes implicit account of important geometric and crucial nonlinear magnetic properties of the machine, as well as the converter topology, which mainly govern the relevant drive performance aspects [5, 6]. Moreover, the inclusion of basic control features within the circuit approach allows the simulation of the complete drive system in advance. A deeper insight into the 4.5 kW permanent magnet machine and the according power converter with the implemented control algorithm is given in Fig. 3.2, where the typical interior permanent magnet structure as well as the

Table 3.1 Machine properties

Rated power	Copper losses	Iron losses	Friction, fan losses	Balance losses
4.5 kW	246 W	103 W	55 W	38 W

Table 3.2 Converter properties

Rated power	Pulses	Output current	Frequency	Efficiency
5 kVA	2 kHz	14 A	0–200 Hz	95%

DC-link capacitors of the power module are obvious. The rated values of machine and converter are partially listed in Tables 3.1 and 3.2.

3.2 Sensorless Control Mode

The vector control mode operates the permanent magnet motor like a current source inverter driven machine applying a continuous current modulation [7]. Therefore, a very precise knowledge of the rotor position is necessary. This could be achieved by processing the response of the machine to arbitrarily injected high frequency test signals, whereby special anisotropic properties of the machines rotor are required. The vector control software itself is adapted to the hardware system depicted in Fig. 3.2.

3.2.1 Nonlinear Motor Model

The practical realization of the control structure is fortunately done within the rotor fixed (d, q) reference system, because the electrical stator quantities can there be seen to be constant within the steady operational state [8]. The stator voltage and flux linkage space vectors are therefore usually formulated in the (d, q) rotor reference frame as

$$\underline{u}_{Sdq}(\tau) = r_S \underline{i}_{Sdq}(\tau) + \frac{d\underline{\psi}_{Sdq}(\tau)}{d\tau} + \mathrm{j}\omega\underline{\psi}_{Sdq}(\tau), \tag{3.1}$$

$$\underline{\psi}_{Sdq}(\tau) = \underline{\underline{l}}_S(\underline{i}_{Sdq})\underline{i}_{Sdq}(\tau) + \underline{\psi}_{Mdq}, \tag{3.2}$$

whereby $\mathrm{r_s}$ denotes the normalized stator resistance. The inductance matrice is written as

$$\underline{\underline{l}}_S = \begin{pmatrix} l_{Sd}(i_{Sd}) & 0 \\ 0 & l_{Sq}(i_{Sq}) \end{pmatrix}. \tag{3.3}$$

It considers, e.g., only main inductance changes in the spatial d-direction due to the d-current component of the stator current space vector, whereas influences due to the q-current component are neglected in (3.3). With respect to the q-direction, the same aspects are valid in (3.3). Consequently, the relation (3.3) has a symmetrical diagonal shape, whereas cross-coupling effects due to saturation effects are omitted. The inclusion of some basic identities for the magnet flux space vector in the (d, q) system, namely

$$\frac{d\underline{\psi}_{Mdq}}{d\tau} = 0, \quad \underline{\psi}_{Mdq} = \psi_M + \mathrm{j}0 \tag{3.4}$$

reduces the system (3.1–3.3) to the simpler set of equations

$$u_{Sd}(\tau) = r_S i_{Sd}(\tau) + l_{Sd,\mathrm{diff}} \tfrac{d}{d\tau} i_{Sd}(\tau) - \omega l_{Sq} i_{Sq}(\tau), \tag{3.5}$$

$$u_{Sdq}(\tau) = r_S i_{Sq}(\tau) + l_{Sq,\mathrm{diff}} \tfrac{d}{d\tau} i_{Sq}(\tau) + \omega l_{Sd} i_{Sd}(\tau) + \omega \psi_M. \tag{3.6}$$

Thereby, (3.5 and 3.6) distinguishes between main inductance values l_{Sd}, l_{Sq}, governing the rated operational state, and differential inductances

$$l_{Sd,\mathrm{diff}} = l_{Sd}(i_{sd}) + i_{Sd} \frac{d}{di_{Sd}} l_{Sd}(i_{Sd}), \tag{3.7}$$

$$l_{Sq,\mathrm{diff}} = l_{Sq}(i_{sq}) + i_{Sq} \frac{d}{di_{Sq}} l_{Sq}(i_{Sq}), \tag{3.8}$$

which are mainly influencing the voltage drops due to the (d, q) related time-dependent current changes in (3.5 and 3.6). It is thereby obviously from e.g. (3.7) that there is a direct vice-versa relation between the differential inductance $\mathrm{l}_{\mathrm{Sd,diff}}$ and the main inductance l_{Sd}, if the saturation dependency from the d-current component is well known. Unfortunately, both Eqs. (3.5 and 3.6) are always directly coupled without the exception of standstill. That fact is very unsuitable for the design of the current controller. Thus, with regard to the used vector control topology, a more favorable rewritten form of (3.5 and 3.6) as

$$\begin{aligned} u_d(\tau) &= u_{Sd}(\tau) + \omega l_{Sq} i_{Sq}(\tau) \\ &= r_S i_{Sd}(\tau) + l_{Sd,\mathrm{diff}} \frac{d}{d\tau} i_{Sd}(\tau), \end{aligned} \tag{3.9}$$

$$\begin{aligned} u_q(\tau) &= u_{Sq}(\tau) - \omega l_{Sd} i_{Sd}(\tau) - \omega \psi_M \\ &= r_S i_{Sq}(\tau) + l_{Sq,\mathrm{diff}} \frac{d}{d\tau} i_{Sq}(\tau). \end{aligned} \tag{3.10}$$

with (3.7 and 3.8) is commonly introduced.

3.2.2 Simplified Block Diagram of Closed-Loop Control

The control schema in accordance to (3.9 and 3.10) is realized by using the decoupling network of Fig. 3.3 in Fig. 3.4 with respect to the (d, q) axis notation as a two-step overlaid cascade structure [7, 8].

The decoupling phenomena (3.9 and 3.10) is build up with both fictive voltages u_d, u_q according to Fig. 3.3. So, both (d, q) axes related controller in (3.9 and 3.10) can be adjusted independently from each other. The outer speed cascade allows adjusting a pre-set speed value n_{def}, after getting smoothed by a PT1-element.

The output of the PI speed controller is also smoothed by a PT1-element and restricted by the thermal I^2t-protection in order to avoid thermal damages. The PI speed controller has a moderate sampling rate and determines the demanded q-current component. In a simplified application, the drive is operated with a default zero d-current component in order to achieve maximal torque output. The actual measured electrical phase currents are transformed to the rotor fixed (d, q) reference system and continuously compared to the demanded d- and q-current components at the innermost current cascade structure. With regard to the d- and q-axis separation, the generated PI-current controller output voltages u_d, u_q are almost seen as fictive quantities in Fig. 3.5, from which the de-coupling-circuit given in Fig. 3.6 calculates the real demanded stator voltage components u_{sd}, u_{sq} for the normal operation afterwards.

3.2.3 Injected High Frequency Test Signals

Test signals of very high frequency are superimposed to the demanded stator voltage components u_{sd}, u_{sq} and are further treated within the power stack ASIC in order to generate the according space vectors in Fig. 3.4. Thereby, the desired current test signals i_{test} are almost related in the d-direction of the estimated (d′, q′) reference

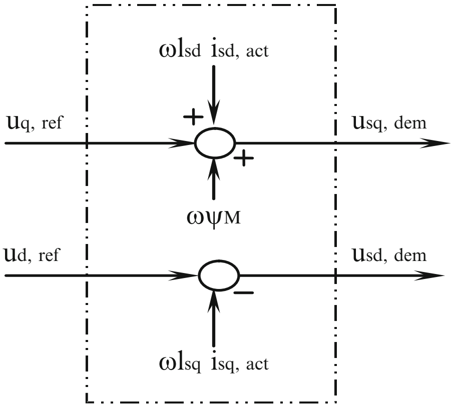


Fig. 3.3 Block diagram of the decoupling-circuit

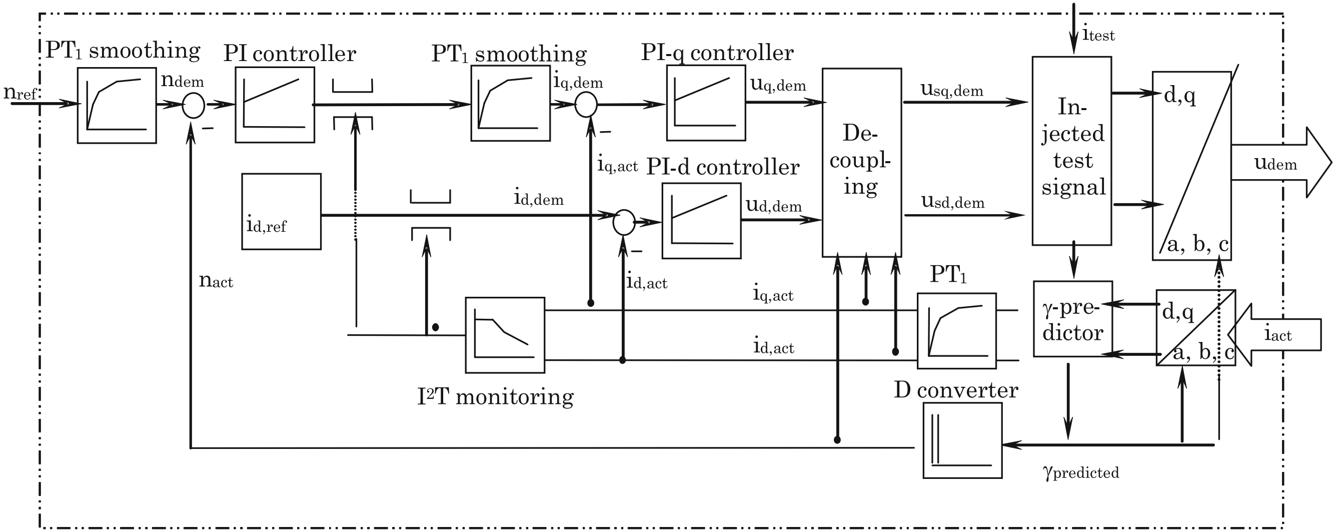


Fig. 3.4 Outline of important devices of the vector controlled permanent magnet motor by neglecting any feed-forward loops

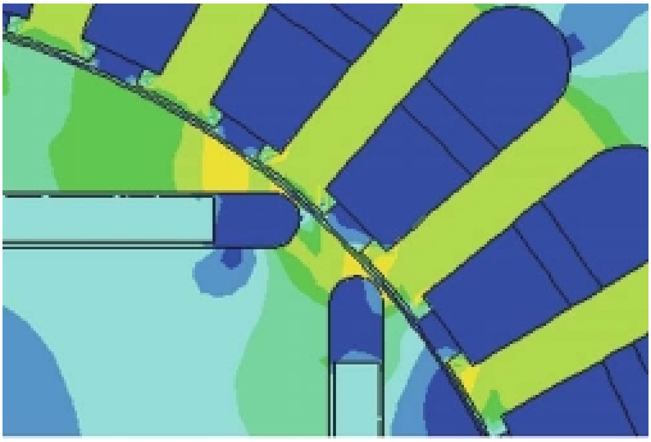
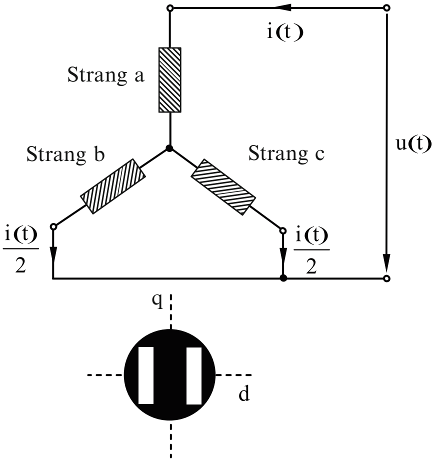


Fig. 3.5 Standstill set-up for the determination of saturable main and differential d-inductances taking account also account of the joke saturation due to the embedded magnets of a electrical two-pole machine

Fig. 3.6 Standstill set-up for the determination of saturable main and differential q-inductances of a electrical two-pole machine



system of the given rotor in Fig. 3.4. Thereby the estimated (d′, q′) and the real (d, q) system could differ and hence, the difference is thereby treated by a PLL.

If the measured high frequency content of the q-current component of the PLL is zero, both systems (d′, q′) and (d, q) match. So, the position would be correctly estimated. But if there occurs also a d-current component due to the applied test vector in a wrong direction, a difference between (d′, q′) and (d, q) would occur. For the correctness of those algorithm it is essential, that the influence of the saturation dependent parameters (3.7 and 3.8) of the machine are laying between safety limiting values.

3.3 Space Dependent Inductances

Inductances are governed by various aspects, whereby the geometric properties and the occurring magnetic saturation of the lamination have a significant impact [9]. Moreover, the magnetic circuit is almost pre-saturated due to the inserted embedded permanent magnets. So, the main task is to set-up a straightforward numerical procedure for the evaluation of those equivalent circuit characteristics with the transient Finite Element approach. Thus, the inductances (3.7 and 3.8) are derived indirectly from (d, q) space dependent flux-current characteristic.

The proposed origin design of flux barriers between the magnets shown in Fig. 3.7 influences the previous parameters (3.7 and 3.8) as well as the torque output and therefore also the drive performance. The desired ratio between both inductances (3.7 and 3.8) can be obtained by the choice of the air gap length, the degree of coverage of the permanent magnets based on the pole pitch of the machine, and the flux barriers between the poles [10, 11]. With respect to mechanical aspects circular ends are always favorable in Fig. 3.7. An optimum distance between the rotor surface and the hole of the magnets has to stand with the mechanical strength and the magnetic saturation. The sheet thickness of the laminated is the smallest possible distance. A smaller distance results in problems with the punching process of the lamina [12].

3.3.1 Set-Up for the Saturation Dependent Differential Inductances Evaluation

The saturation dependent values of (3.7) have to evaluated in advance with respect to the spatial directions in order to verify the robustness of the drive. In particular

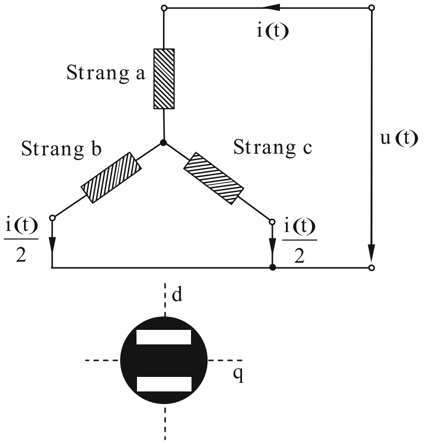


Fig. 3.7 Detail of the rotor design concerning stray fields

the sensor less mode at standstill even at very high currents, which are e.g. necessary for a high pull-out torque have thereby to be covered. If the known q-axis of the rotor is aligned with the q-axis of the stator current space vector, no effect due to the permanent magnet flux (3.4) in the q-axis has to be considered under those assumptions (3.3). The relations (3.7) are therefore valid. The pre-saturation of the lamination due to the magnets is indirectly considered. Neglecting slotting effects, the necessary flux-current characteristic in the q-axis in (3.2), the proposed machine design is treated within the numerical simulation by Finite Elements as a single phase test set-up as shown in Fig. 3.6. If the system is simply fed by a sinusoidal time-dependent voltage, the obtained time-dependent current consumption is mainly non-sinusoidal due to the iron saturation. But the courses are symmetrical to the abscissa.

Contrarily, if the d-axis of rotor coincides with the d-axis of the stator current space vector in Fig. 3.5, the numerical calculated time dependency of the d-current inside the machine represents a non-sinusoidal course with a constant offset, caused by the magnetized magnets within the lamination.

3.3.2 Saturation Dependent Differential and Main Synchronous Inductances

With the aid of commercial numerical field calculation tools, the obtained flux-current relation (3.3) and consequently the differential and the main inductance (3.7) courses are straightforward derived with respect to the (d, q) notation as shown in Figs. 3.8 and 3.9. Thereby, saturation within higher current magnitudes enforces that both d- and q-characteristics become very similar, which means, that the test signals response is equal in each direction and no distinction could be achievable. The dependency of the main inductance shown in Fig. 3.9 shows along the complete current range a distinct difference. Thereby, both relations (3.7 and 3.8) could

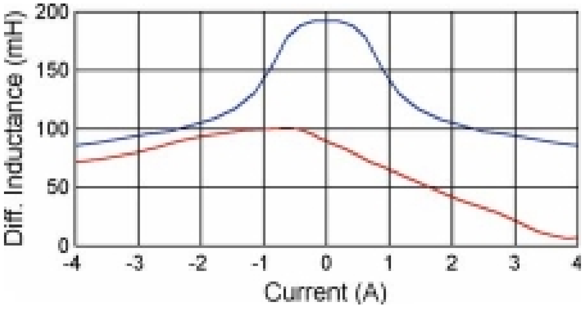


Fig. 3.8 Dependency of d-axis differential inductance from d-current (*red*) and q-axis differential inductance from q-current (*blue*)

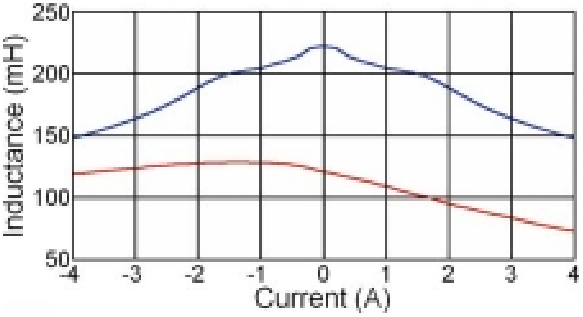


Fig. 3.9 Dependency of d-axis main inductance from d-current (*red*) and q-axis differential inductance from q-current (*blue*)

be clearly verified. The previous desired reluctance effects due to different d- and q-inductance values over the complete current magnitude in Fig. 3.9 are fortunately usable even for full- or overload capabilities.

3.4 Conclusion

The design of permanent magnet drives for sensor less operation has to consider different aspects. First, the robust sensing capability of test signals enforces different differential inductances with respect to the (d, q) notation. The influence of occurring saturation to that behavior has to be restricted. In order to improve the torque output due to the usage of the reluctance torque, the anisotropic inductance properties should be significantly different. Thus, a negative d-axis current increases thereby the effective torque of the motor up to e.g. 15%. Moreover, several secondary effect, such as cogging torque, torque ripple, non-sinusoidal shape of no-load voltage and losses, which are worsening the performance of such innovative drives, must be restricted to limiting values. Thus, a coupled transient electromagnetic-mechanical finite element calculation method with directly coupled external circuits is useful, in the case of the voltage waveform, torque ripple and additional losses. The influence of higher harmonics on the design steps becomes visible by this method.

References

1. Nowotny, W., Lipo, T.A.: Vector Control and Dynamics of AC Drives. Clarendon, Oxford (2000)
2. Hendershot, J.R., Miller, T.J.E.: Design of Brushless Permanent Magnet Motors. Oxford University Press, Oxford (1994)

3. Bush, K.G.: Regelbare Elektroantriebe: Antriebsmethoden, Betriebssicherheit, Instandhaltung. Verlag Pflaum, München (1998)
4. Salon, J.S.: Finite Element Analysis of Electrical Machines. Cambridge University Press, Cambridge (1996)
5. Davat, B., Ren, Z., Lajoic-Mazenc, M.: The movement in field modeling. IEEE Trans. Magn. 21(6) (1985)
6. Kohnke, P.: Theory Reference Released Version 6.1. Ansys Inc., Canonsburg, PA (April 2002)
7. Vas, P.: Electrical Machines and Drives: A Space-Vector Theory Approach. Clarendon, Oxford (1996)
8. Vas, P.: Vector Control of AC Machines. Oxford University Press, Oxford (1990)
9. Vas, P.: Parameter Estimation, Condition Monitoring, and Diagnosis of Electrical Machines. Clarendon, Oxford (1993)
10. Domack, S.: Auslegung und Optimierung von permanent- erregten Synchronmaschinen mittels Steuerverfahren und der Methode der finiten Elemente. Aachen, Verlag Shaker (1994)
11. Kiyoumarsi, A., Moallem, M.: Optimal shape design of interior permanent-magnet synchronous motor. IEEE Conference on electric machines and drives, pp. 642–648 (2005)
12. Bödefeld, T.H., Sequenz, H.: Elektrische Maschinen, Eine Einführung in die Grundlagen. Wien, Springer-Verlag, New York, pp. 230–236 (1971)

Chapter 4
Design of Silicon Resonant Micro Accelerometer Based on Electrostatic Rigidity

Zhang Feng-Tian, He Xiao-Ping, Shi Zhi-Gui, and Zhou Wu

Abstract The structure characteristics and working principle of silicon resonant micro-accelerometer based on electrostatic rigidity are presented. Dynamic characteristics of double-ended tuning fork (DETF) in the sensor are analyzed. Force equilibrium equations of mass and DETF are built respectively for with or without acceleration, through which the relationship between DETF resonant frequency and acceleration is obtained. The influences of folded supporting beams linked with proof mass and gap between capacitive parallel plates on the sensor sensitivity are analyzed, and finally a resonant micro accelerometer with sensitivity of 60 Hz/g is designed and fabricated with bulk-silicon dissolved processes.

Keywords Electrostatic rigidity · resonance · accelerometer · bulk-silicon dissolved processes

4.1 Introduction

Resonant principle has been widely used for measuring physical parameters such as mass, acceleration, force, flow or pressure. The main advantage of a resonant sensor over other sensing principles is its quasi-digital output, which implies good resistance to noise or electrical interference and simple interface to a digital system. Resonant accelerometer has been produced and widely used for many years, for example, RBA500 of Honeywell Company. Most resonant accelerometers use quartz material for its excellent piezoelectric performance, but the quartz fabrication is

Z. Feng-Tian (✉), H. Xiao-Ping, S. Zhi-Gui, and Z. Wu
Lab 502, Institute of Electronic Engineering,
China Academy of Engineering Physics, Mianshan Road, Mianyang, Sichuan, China
e-mail: zftstuart@sohu.com; hxp@xleda.com;
szg@xleda.com; zhouwu916@yahoo.com.cn

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_4,

expensive and not compatible with IC (integrated circuits) fabrication technology, which makes it impossible to integrate the sensor and its interface circuits on one chip.

Resonant accelerometers based on the silicon micromachining technology become very attractive due to low cost, small size, compatibility with IC fabrication processes, and potential application in the fields where size and high precision are required. Most silicon resonant accelerometers detect resonant frequency of the vibrating beam subjected to axial loading which relates to inertial force on the proof mass. The vibrating beam is excited by alternating electrostatic or electrothermal force, and the frequency change is sensed by capacitors or piezoresistive resistors respectively [1–6]. In a relatively new concept of resonant accelerometer, inertial force of the proof mass pushes or pulls one electrode as one part of the proof mass, and changes the gap distance of the parallel-plate capacitor, then the electrostatic force between parallel plate electrodes changes with the gap distance, and equivalently changes the efficient mechanical rigidity of the vibrating beam [7, 8]. In the previous work of other research groups, the accelerometer was fabricated and showed excellent performances. But the movement and dynamic characteristics of the vibrating beam were not presented, which is very important for understanding the working principle and the sensor structure design. Therefore, in this study, the movement and dynamic characteristics are analyzed with mechanical dynamic theory. The relationship between resonant frequency of the vibrating beam and acceleration is obtained in theory. Finally, a resonant accelerometer with sensitivity of 60 Hz/g is designed and fabricated with bulk-silicon dissolved processes.

4.2 Operating Principle

The concept of the resonant accelerometer is shown in Fig. 4.1. It consists of double-ended tuning fork (DETF), driving capacitors, sensing capacitors, proof mass, and supporting springs. The fixed comb electrodes of driving capacitors are excited by an AC voltage with DC bias, DETF is connected to ground, and the proof mass including the electrode of sensing capacitor is connected to a DC voltage. In the horizontal plane, each clamped-clamped beam of DETF vibrates 180° out of phase with its natural frequency to cancel reaction forces at the ends when there is no acceleration. The sensing capacitors can detect the vibrating frequency of DETF through interface circuits. When there is acceleration, the proof mass moves near or away from the electrode of DETF under inertial force and changes the gap distance of sensing capacitor. The electrostatic force between sensing electrodes induces an additional electrostatic rigidity, which results in the variation of DETF resonant frequency with acceleration. Therefore, detecting the resonant frequency of DETF can measure acceleration.

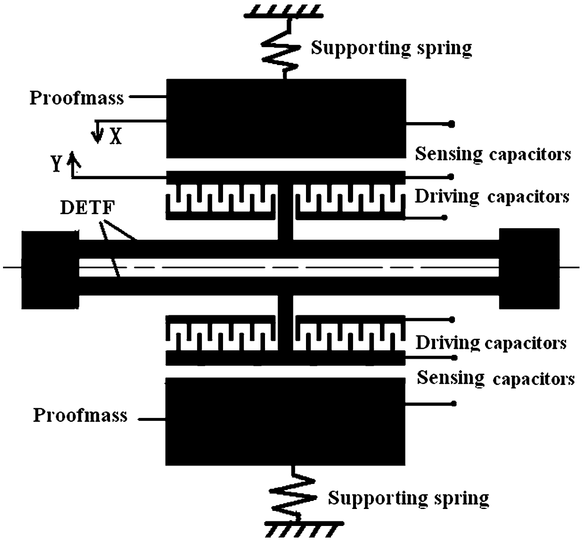


Fig. 4.1 Principle schematic of a resonant accelerometer

4.3 Theory Analysis

Due to the symmetry of sensor structure, one half of the structure shown as in Fig. 4.2 is analyzed. The supporting spring in Fig. 4.1 is realized with four folded beams which not only support the proof mass but also make it capable of moving freely along x axis. The DC voltage of the proof mass is $V_{\mathrm{d}2}$, the clamped-clamped beam of DETF is connected to GND, and the driving voltage is $V_{\mathrm{c}} \times \mathrm{Sin}\ (\omega t)$ with DC bias voltage $V_{\mathrm{d}1}$. The forces applied to the clamped-clamped beam include inertial force, damping force, elastic force, driving and sensing electrostatic forces. The mechanical dynamic equation of the vibrating beam can be written as

$$m\ddot{Y} + c\dot{Y} + \kappa Y = \frac{\varepsilon A {V_{d2}}^2}{2(g_0 - x - Y)^2} - \frac{N_1 \varepsilon h (V_{d1} + V_c \sin(\omega t))^2}{2g_0} \tag{4.1}$$

Where m is the effective mass of the vibrating beam, c damping coefficient, κ the effective spring constant of vibrating beam, ε the permittivity of free space (8.85×10^{-12} F/m), A the efficient area of the sensing capacitor, $V_{\mathrm{d}2}$ the sensing voltage, g_0 the gap distance of driving capacitor and sensing capacitor, x and Y are respectively the displacement of proof mass or vibrating beam relative to initial position when no acceleration is applied, and N_1 is number of driving comb finger pairs, h is the structure thickness, ω is the frequency of AC voltage. The first part on the right in (4.1) is sensing electrostatic forces, while the second part is driving electrostatic forces. In (4.1), it assumes that the damping force is proportional to velocity.

From (4.1), we can see that electrostatic forces of the vibrating beam include fixed and alternate parts. We can think that the vibrating beam moves

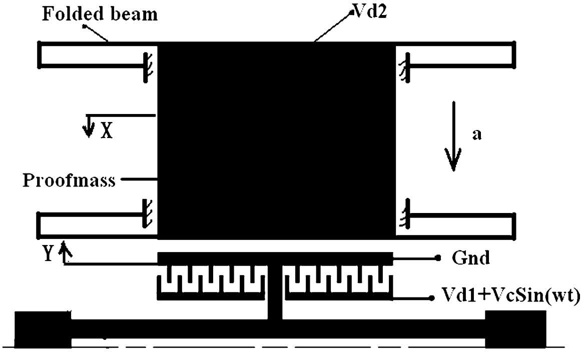


Fig. 4.2 Schematic of the half structure

to some position by one fixed electrostatic force and vibrates harmonically about this equilibrium position. So Y in (4.1) can be expressed as $y_1 + y$, and $y = y_0 \times \sin(\omega t + \phi)$, where y_1 is the displacement of vibrating beam by the fixed electrostatic force and y_0 is the displacement amplitude by harmonic force. If $V_{d1} >> V_c$, with Taylor expansion, (4.1) can be rewritten as

$$m\ddot{y} + c\dot{y} + \kappa(y_1 + y) \cong \frac{\varepsilon A V_{d2}{}^2}{2(g_0 - x - y_1)^2} + \frac{\varepsilon A V_{d2}{}^2}{(g_0 - x - y_1)^3}\, y - \frac{N_1 \varepsilon h V_{d1}{}^2}{2g_0} - \frac{N_1 \varepsilon h V_{d1} V_c \sin(\omega t)}{g_0} \tag{4.2}$$

Equivalently, we can obtain

$$\kappa y_1 - \frac{\varepsilon A V_{d2}{}^2}{2(g_0 - x - y_1)^2} + \frac{N_1 \varepsilon h V_{d1}{}^2}{2g_0} = 0 \tag{4.3}$$

$$m\ddot{y} + c\dot{y} + \left(\kappa - \frac{\varepsilon A V_{d2}{}^2}{(g_0 - x - y_1)^3}\right) y = -\frac{N_1 \varepsilon h V_{d1} V_c \sin(\omega t)}{g_0} \tag{4.4}$$

Equation (4.3) describes the displacement of vibrating beam by fixed force. We can see that y_1 is determined by the driving or sensing DC voltage in addition to the sensor structure and size. Equation (4.4) is the harmonic oscillation equation. It shows that there is an additional force proportional to the vibrating amplitude in (4.4), equivalently an additional rigidity caused by electrostatic force between parallel plates of sensing capacitor. The rigidity named electrostatic rigidity κ_e can be written as

$$\kappa_e = \frac{\varepsilon A V_{d2}{}^2}{(g_0 - x - y_1)^3} \tag{4.5}$$

Then, the resonant frequency of vibrating beam is as follows

$$f_n = \frac{1}{2\pi}\sqrt{\frac{\kappa - \kappa_e}{m}} \tag{4.6}$$

When there is acceleration which direction shown as in Fig. 4.2, the displacement of proof mass caused by inertial force is Δx. The electrostatic force between parallel plates of sensing capacitor will change and result in displacement variation Δy_1 of the electrode on the vibrating beam. Thus, the gap distance of sensing parallel-plate capacitor change $\Delta x + \Delta y_1$. The electrostatic rigidity induced by sensing parallel-plate capacitor will vary and alter resonant frequency of the vibrating beam shown as

$$f_n = \frac{1}{2\pi}\sqrt{\frac{\kappa - \dfrac{\varepsilon A {V_{d2}}^2}{(g_0 - (x - \Delta x) - (y_1 - \Delta y_1))^3}}{m}} \tag{4.7}$$

Due to that the natural frequency of vibrating beam is greatly higher than that of proof mass-spring system, the electrostatic force frequency applied to the proof mass by beam vibrating is also largely higher than that of proof mass-string system. So the displacement of proof mass caused by beam vibrating can be neglected. When there is no acceleration, the elastic force and electrostatic force applied to the proof mass are equal. We can obtain the force equilibrium equation of proof mass as follow

$$\kappa_s x - \frac{\varepsilon A {V_{d2}}^2}{2(g_0 - x - y_1)^2} = 0 \tag{4.8}$$

Where κ_s is spring constant of the folded supporting beams linked with proof mass.

When there is acceleration shown as in Fig. 4.2, the gap of sensing capacitor will change, and the elastic force plus inertial force is equal to electrostatic force by sensing capacitor, the force equilibrium equation can be written as

$$\frac{\varepsilon A {V_{d2}}^2}{2(g_0 - (x - \Delta x) - (y_1 - \Delta y_1))^2} - \kappa_s(x - \Delta x) - Ma = 0 \tag{4.9}$$

Here M refers to the mass of the proof mass. As for the vibrating beam, from (4.3), we can obtain

$$\frac{\varepsilon A {V_{d2}}^2}{2(g_0 - (x - \Delta x) - (y_1 - \Delta y_1))^2} - \frac{N_1 \varepsilon h {V_{d1}}^2}{2g_0} + ma - \kappa(y_1 - \Delta y_1) = 0 \tag{4.10}$$

When the sensor structure size, driving and sensing voltage are determined, from (4.3 and 4.8–4.10), we can calculate the value of $x, y_1, \Delta x, \Delta y_1$, and substitute them into (4.7), the relationship between the beam vibrating frequency and acceleration can be obtained, which is important for sensor structure design.

4.4 Structure Design

The accelerometer sensitivity directly affects its precision. The important aspect of sensor structure design is to determine proper structure size to improve the sensor sensitivity when the structure stability is guaranteed. The sensor sensitivity will be calculated at different structure size, driving or sensing voltage, with the above established equations. The vibrating beam is 500 μm long and 5 μm wide. The sensor structure is 25 μm in thickness. The driving AC amplitude is 1 V with 15 V DC bias. The folded supporting beams are U-shape.

Figure 4.3a shows the relationship between the sensor sensitivity and sensing voltage when the capacitor gap is 3 μm, the folded beam is 450 μm long and 6 μm width. Figure 4.3b presents the variation of sensitivity with the capacitor gap distance when the folded beam is 450 μm long and 6 μm wide and sensing voltage is 16 V. We can see that sensor sensitivity is 70 or 250 Hz/g when sensing voltage is 17 or 18 V respectively. Larger sensing voltage and small capacitor gap mean higher sensitivity. The reason is that the sensing capacitor electrostatic force increases when sensing voltage increases and capacitor gap gets narrow, then the

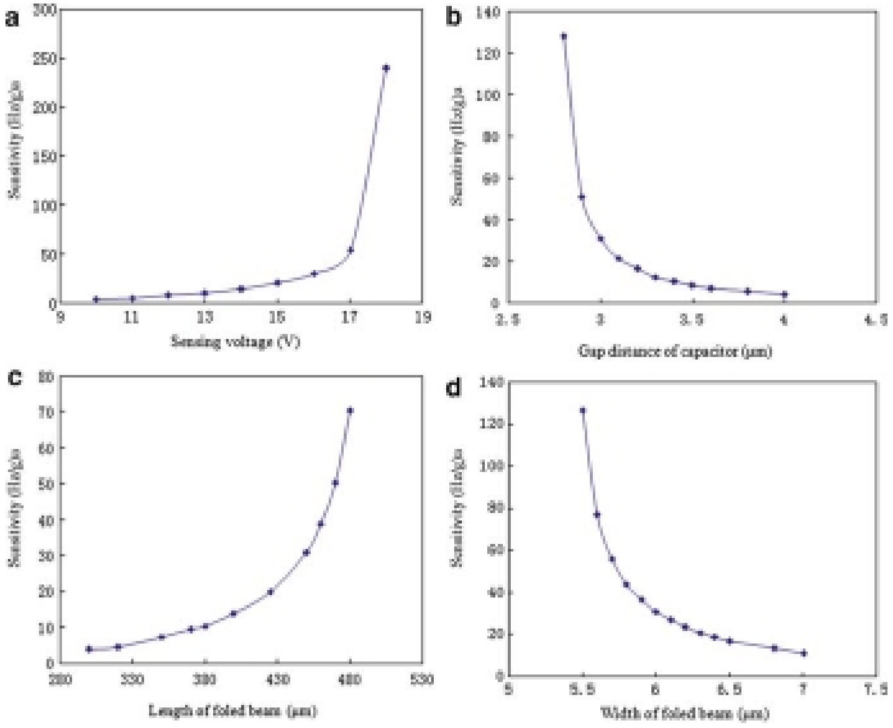


Fig. 4.3 Sensor sensitivity versus voltage and structure size

variation of electrostatic rigidity at same acceleration become larger and produce larger deviation of vibrating beam resonant frequency. But too large sensing voltage and small capacitor gap may lead to absorption of capacitor plates and structure stability will be ruined.

Figure 4.3c is the relationship of sensor sensitivity and folded supporting beam width, and Fig. 4.3d shows that of sensor sensitivity and supporting beam length. We can see that narrower and longer supporting beams can improve sensor sensitivity. This is due to that the capacitor gap variation caused by inertial force of the proof mass is larger when the folded beam is narrow and long, and electrostatic rigidity varies greatly. Also, too narrow or long folded beams will affect the strength and resistance to impact, even maybe result in absorption of capacitor plates and destroy the structure stability.

From above analysis, smaller capacitor gap, longer and narrower supporting beam, and larger sensing voltage are expected to improve the sensor sensitivity. But in the respect of stability and strength, they should be controlled properly. So sensitivity and structure stability should be considered at the same time when designing the sensor structure size. The sensor structure size in this study is finally designed as follows: folded supporting beam 450 μm long and 6 μm wide, capacitor gap 3 μm wide, and sensing voltage 17 V. The relationship between sensor resonant frequency and applied acceleration is shown in Fig. 4.4. We can see that the designed sensor sensitivity is at least 60 Hz/g for one clamped-clamped beam of DETF.

The designed sensor is fabricated with bulk-silicon dissolved processes, and SEM pictures of the sensor chip with are shown in Fig. 4.5.

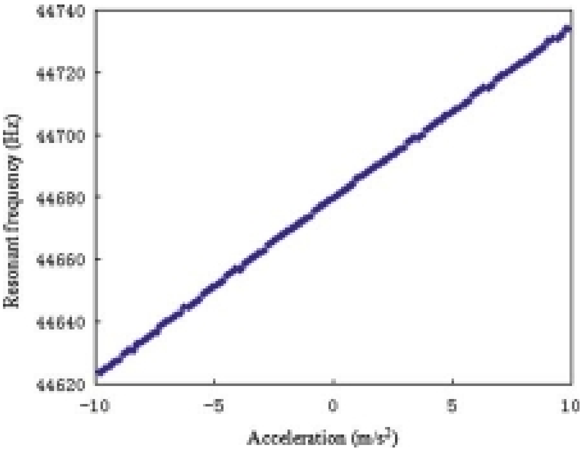


Fig. 4.4 Resonant frequency versus applied acceleration

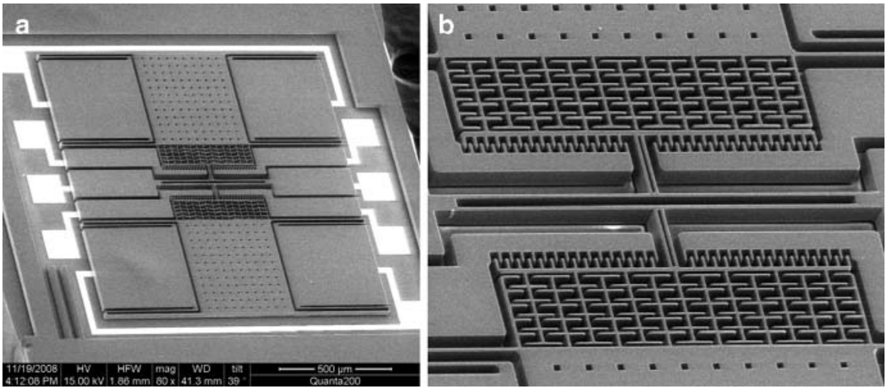


Fig. 4.5 The SEM pictures of the fabricated sensor

4.5 Conclusion

In the silicon resonant accelerometer based on electrostatic rigidity, the proof mass will not vibrate when DETF beams vibrate without acceleration. Meanwhile, the proof mass will move away the initial equilibrium position with acceleration, and apply an additional rigidity to the vibrating beam for change the resonant frequency. According to the dynamic equations of proof mass and DETF, the accelerometer resonant frequency can be determined at any acceleration. By analyzing effects of the structure size, sensing voltage on the sensor sensitivity, a sensor structure suitable for actual fabrication condition is designed and fabricated. Future research work will focus on the measurement of the sensor performance.

References

1. Burrer, C., Esteve, J.: A novel resonant silicon accelerometer in bulk-micromachining technology. Sensor. Actuator. A **46–47**, 185–189 (1995)
2. Susan, X.P.S., Yang, H.S., Agogino, A.M.: A resonant accelerometer with two-stage microleverage mechanisms fabricated by SOI-MEMS technology. IEEE Sensor. J. **5(6)**, 1214–1223 (2005)
3. Seshia, A.A., Palaniapan, M., Roessig, T.A., Howe, R.T., Gooch, R.W., Schimert, T.R., Montague, S.: A vacuum packaged surface micromachined resonant accelerometer. J. Microelectromech. Syst. **11(6)**, 784–793 (2002)
4. Helsel, M., Gassner, G., Robinson, M., Woodruff, J.: A navigation grade micro-machined silicon accelerometer. Position Location and Navigation Symposium, IEEE **94**, 51–58 (1994)
5. Yubin, J., Yilong, H., Rong, Z.: Bulk silicon resonant accelerometer. Chin. J. Semiconduct. **26**(2), 281–286 (2005)
6. Burns, W., Horning, R.D., Herb, W.R., Zook, J.D., Guckel, H.: Resonant microbeam accelerometers. The 8th International Conference on Solid-State Sensors and Actuators, and Eurosensors IX. Stockholm, Sweden, pp. 659–662, 25–29 June 1995

7. Lee, B.L., Oh, C.H., Oh, Y.S., Chun, K.: A novel resonant accelerometer: electrostatic stiffness type. The 10th International Conference on Solid State Sensors and Actuators (Transducer'99), Sendai, Japan, pp. 1546–1549, 7–10 June 1999
8. Seok, S., Kim, H., Chun, K.: An inertial-grade laterally-driven mems differential resonant accelerometer. IEEE 2004:654–657 (M. Young, The Techincal Writers Handbook). University Science, Mill Valley, CA (1989)

Chapter 5
Longer MEMS Switch Lifetime Using Novel Dual-Pulse Voltage Driver

Lai Chean Hung and Wallace S.H. Wong

Abstract A novel dual-pulse voltage driver has been proposed to reduce dielectric charging in micro-electromechanical system (MEMS) switch, leading to a longer switch lifetime. Mathematical and transient circuit models have been utilized to simulate dielectric charging in the RF MEMS switch, enabling the analysis of charge built-up at the switch dielectric and substrate brought about by the actuation voltage curve used. The proposed dual-pulse actuation signal has shown to improve the lifetime of the RF MEMS switch as it minimizes the charge built-up during its long continuous operation. Practical experiment on the commercial TeraVicta TT712-68CSP MEMS switch shows that the proposed actuation voltage can reduce the pull in/out voltage shift and therefore prolong the switch lifetime. The technique has also shown to reduce switching bounces.

Keywords Radio frequency (RF) · micro-electromechanical system (MEMS) · dielectric · charging · reliability · lifetime

5.1 Introduction

Microelectromechanical Systems or MEMS switch is becoming the preferred choice for RF switching due to its outstanding performance when compare to the conventional solid state RF switch such as p-i-n diodes or FET transistor. RF MEMS switch has very low insertion loss but high isolation and consumes minimal power in the microwatts rather than the milliwatts that solid state switches require. However, unlike its solid state counterparts, due to the electro-mechanical nature the MEMS switch suffers from shorter lifecycle ranging from 100 million to 10 billion cycles only [1].

L.C. Hung (✉) and W.S.H. Wong
School of Engineering & Science, Swinburne University of Technology (Sarawak campus),
Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia
e-mail: clai@swinburne.edu.my; wwong@swinburne.edu.my

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_5,

In capacitive membrane switches, the main life-limiting mechanism is dielectric charging trapped within the switch dielectric layer due to the high actuation voltage required to actuate the switch [2, 3]. When sufficient actuation voltage is applied to the electrodes beneath the membrane, the membrane is pulled down towards the dielectric layer by electrostatic force, creating a capacitive short. Over time and repeated ON-OFF cycles charges are trapped in the dielectric layer and the substrate of the switch, pulling the membrane to the dielectric even without any actuation voltage applied.

This chapter presents an analysis on the charge trapped in the switch using mathematical and transient circuit models of the charge built-up [5]. Then the implementation of a novel dual-pulse actuation voltage which reduces charge built-up and therefore extends the lifetime of the switch will be presented. The analysis is then corroborated through experiment where the commercially available TeraVicta (now defunct) TT712-68CSP MEMS switch is shown to experience less pull in/out voltage shift during long term operation, even when the switch is of metal-to-metal (ohmic) type.

5.2 Dielectric Charging

A typical capacitive membrane switches generally require 30–50 V of actuation voltage which will form a very high electric field in a region of 100 MV/m across the dielectric layer. In this condition, it is possible for charges to tunnel across the dielectric and become trapped within the dielectric layer through a process similar to that of Frenkel-Poole emissions in thin insulating films [2], where the charged trapped is exponentially related to the applied electric field. During switching ON, charges will be accumulated on the surface of the dielectric or even on the bulk of the substrate since the recombination time for these charges can be very long and there is a lack of conduction path to drain off the trapped charges. When the trapped charges build up to a level that is just enough to hold the membrane to the dielectric layer even without the presence of actuation voltage, the switch is stuck at the ON state.

The built-up charge also affects the pull-in voltage V_{pi} and pull-out voltage V_{po} of the switch. Pull-in voltage V_{pi} is the threshold voltage that the actuation voltage must exceeds so that the electrostatic force generated will be greater than the restoring force of the switch membrane and closes the switch. Once the switch closes, the electric field will be higher due to the smaller gap between the membrane and electrodes. Hence, the switch will only open when the applied voltage is reduced to below the pull-out voltage, V_{po}.

The trapped charges change the magnitude of the electric field presents in the dielectric layer and thus the generated electrostatic force as well. In case of a positive actuation voltage, positive charges will tunnel into the dielectric layer due to the high electric field across the gap. The trapped charges will generate electrostatic force itself and increase the net amount of electrostatic force between the membrane

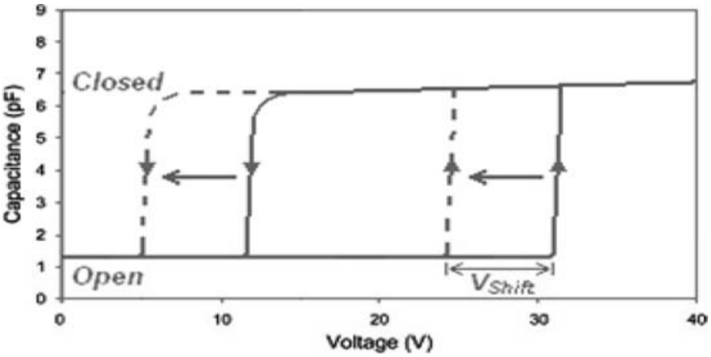


Fig. 5.1 C-V curve of capacitive RF MEMS switch before (*solid line*) and after continuous actuation (*dotted line*) [4]

and electrodes. This reduces the amount of external force/voltage needed to pull down the switch membrane. The net effect of injected positive charges is therefore a negative shift of the switch C-V curve as shown in Fig. 5.1, which in turn affects the pull-in and pull-out voltages where $V_{pi} = V_{pi} - V_{shift}$ and $V_{po} = V_{po} - V_{shift}$. Since V_{shift} is proportional to the amount of trapped charge, the longer the switch has been in operation, the more charges are accumulated and V$_{shift}$ increases. The switch will fail when V_{po} in positive region becomes negative. In that case, the switch will be in the closed state even at 0 V.

5.3 Modeling of Dielectric Charging

5.3.1 Mathematical Model

Dielectric charging for each ON time of the operating cycle can be modeled as [3]:

$$Q_C = \sum_{J=1,2} Q_J \times (1 - \exp(-(t_{on} + t_{eJ})/\tau_{CJ})) \tag{5.1}$$

where Q_J is the steady state charge density for Jth trap species (there are two types only, $J = 1$ and $J = 2$), t_{on} is the ON time duration for one cycle, τ_{CJ} is the charging time constant for Jth trap species and t_{eJ} is the equivalent time required to charge the dielectric to the value just before the present ON time. The steady state charge density Q_J for the Jth trapped species when absolute voltage V is applied is given by:

$$Q_J = Q_{0J} \times \exp(V/V_{0J}) \tag{5.2}$$

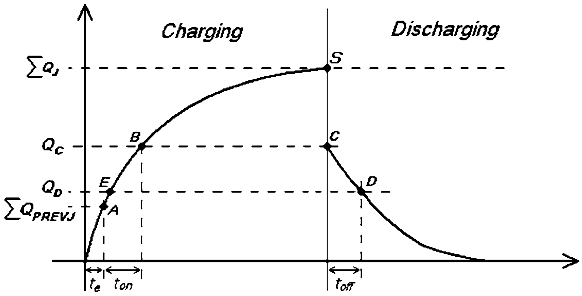


Fig. 5.2 Dielectric charging and discharging under square wave actuation voltage after one operating cycle, the charge density increases from the initial-state A to the end-state E

where Q_{0J} and V_{0J} are fitting parameters. Dielectric discharging can be model as:

$$Q_D = \sum_{J=1,2} Q_{PREVJ} \times \exp\left(-t_{off}/\tau_{DJ}\right) \tag{5.3}$$

where Q_D is the charge accumulated after the OFF time duration, Q_{PREVJ} is the amount of charges trapped in the dielectric just before the OFF time for Jth trap species, t_{off} is the OFF duration and τ_{DJ} is the discharging time constant for Jth trap species. By iterating Eqs. (5.1) and (5.3), the accumulation of charge over many operating cycles can be obtained. Figure 5.2 depicts the dielectric charging and discharging curve under a square wave ON-OFF actuating voltage [3]. The trapped charge at the beginning of each operating cycle can be somewhere between empty and full, represented by point A on the charging curve. When the switch is turned ON, the amount of charges increases to point B (Q_C) at the end of the ON time, which can be calculated by using Eq. (5.1). Once the switch is turned OFF, the charges trapped inside the dielectric start to discharge from point C on the discharging curve. At the end of the OFF time, the dielectric is discharged to point D(Q_D) which can be calculated by Eq. (5.3). For the next cycle, point D is mapped back onto point E on the charging curve. Thus, the net amount of charges accumulated on the dielectric after one operating cycle is equals to point E minus point A.

Figure 5.3 shows the accumulated charge simulated using the above model over a long operating cycle. The switch model used is similar to that in [3], using similar extracted parameters as shown in Table 5.1. The switch is operated using square wave of 100 Hz, 50% duty cycle. The voltage shift due to dielectric charging can then be calculated by using Eq. (5.4):

$$\Delta V = \frac{dQq}{\varepsilon_o \varepsilon_r} \tag{5.4}$$

where, ΔV is the voltage shift V_{shift} due to dielectric charging, d is the distance between the centroid of the charge sheet (dielectric) and the top electrode (switch membrane), Q is the charge density modeled by Eqs. (5.1) and (5.3), q is the

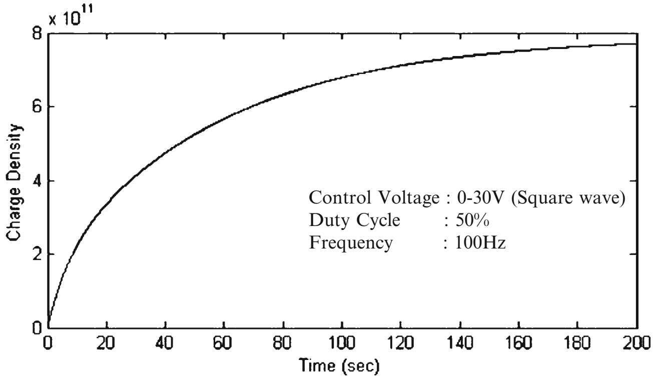


Fig. 5.3 Dielectric charge density accumulated over 200 s stressed by 0–30 V square wave actuation signal

Table 5.1 Extracted model parameters

J	τ_C (s)	τ_D (s)	Q_0 (q/cm^2)	V_0 (V)
1	6.6	6.8	3.1×10^{10}	13
2	54	62	1.6×10^{11}	15

electron charge, $\varepsilon_O\varepsilon_r$ is the permittivity of dielectric. This equation calculates how much the V_{pi} and V_{po} have been shifted based on the density trapped charge and therefore the switch failure due to dielectric charging can be modeled.

5.3.2 Equivalent Circuit Model

In order to analyze the dielectric charging under a more complex actuation voltage waveform, a SPICE model was implemented by Yuan et al. by using two RC subcircuis to simulate the charging and discharging behavior of the dielectric [5]. As shown in Fig. 5.4, the two sets of RC circuit represent the two trapped species with different charging and discharging time constants. Dielectric charging is represented by the charging of both capacitors, C_1 and C_2. Both capacitors were set to unity so that the resistances directly correspond to the charging and discharging time constants. R_{C1} and R_{D1} represent the charging and discharging time constants for $J = 1$, and R_{C2} and R_{D2} represent the charging and discharging time constants for $J = 2$. Diodes in the circuit were used to direct charge flow. The total charge trapped in the dielectric can be obtained by adding the charge accumulated on the unity capacitors (C_1 and C_2). Two voltage sources, V_1 and V_2 were implemented to represent the steady state charge density for different trap species. The value of the voltage sources are determined by Eq. (5.2).

The simulated result by using the equivalent circuit model is shown in Fig. 5.5. In this figure, it is shown that the result from the equivalent circuit model is very

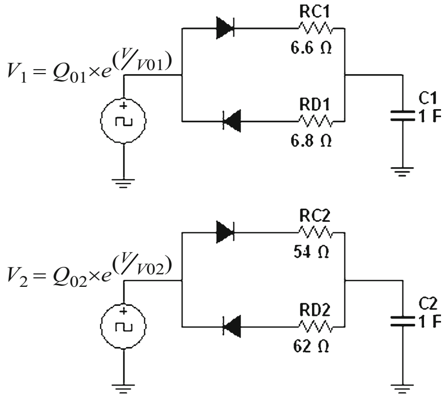

$V_1 = Q_{01} \times e^{(V/V01)}$
RC1
6.6 Ω
RD1
6.8 Ω
C1
1 F
$V_2 = Q_{02} \times e^{(V/V02)}$
RC2
54 Ω
RD2
62 Ω
C2
1 F

Fig. 5.4 Equivalent-circuit model for transient circuit simulation. Voltage dependence of the steady-state charge densities was implemented in the two voltage sources V_1 and V_2

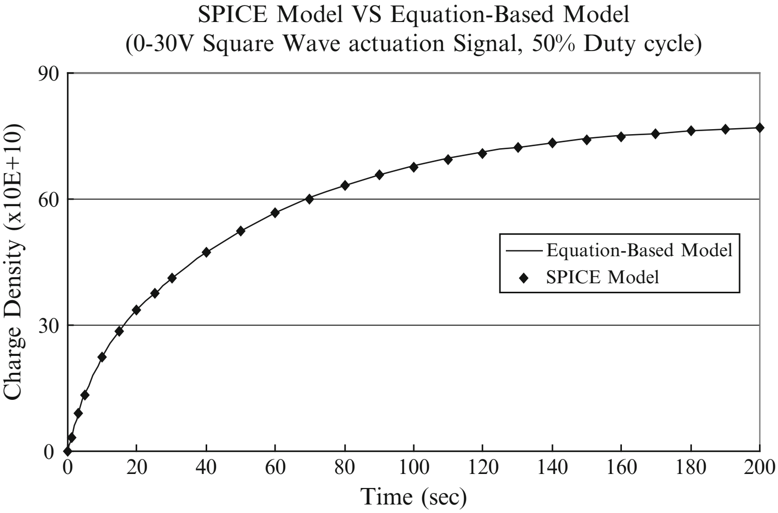


Fig. 5.5 Dielectric charging curve over 200 s at 100 Hz switching modeled by using equation-based model compared to the spice model

close to the one predicted by the equation-based model. Hence, dielectric charging effect for complex waveform such as dual-pulse actuation signal can be obtained using the Transient SPICE model by simply changing the shape of the two voltage sources V_1 and V_2.

5.4 Dual-Pulse Actuation Signal

Goldsmith et al. has proposed a dual-pulse actuation voltage to reduce dielectric charging [2]. The waveform comprises of a short high voltage pulse to pull down the switch membrane and a low-voltage pulse to hold the membrane at ON state.

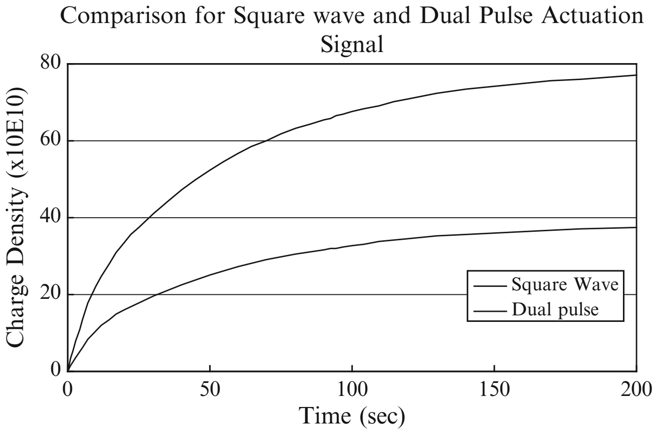


Fig. 5.6 Comparison of dielectric charging effect between a 100 Hz, 0–30 V square wave actuation signal and a 100 Hz dual pulse actuation signal

The initial high voltage must exceed the switch's pull-in voltage V_{pi}. Once the membrane has been pulled down, the increase in electrostatic force due to the smaller distance between the membrane and electrodes and the stiction force on the membrane surface reduce the required voltage to hold the switch at the ON state.

Dual-pulse actuation signal minimizes the time that high voltage is applied across the dielectric and hence reduces the dielectric charging. The effect of using dual-pulse actuation on dielectric charging is simulated using the aforementioned equivalent circuit model with some modification on the voltage sources in order to obtain the dual pulse signal. The voltage source V_1 is replaced by the summation of two voltage sources with different duty cycle. V_{P1} and V_{P2} represent the peak voltages for $J = 1$ and $J = 2$, and V_{h1} and V_{h2} represent the holding voltages for $J = 1$ and $J = 2$ respectively. Figure 5.6 shows the comparison of dielectric charging effect between a 100 Hz, 0–30 V square wave actuation signal and a 100 Hz dual-pulse actuation signal. From this figure, it is shown that the dual-pulse successfully reduces the charge trapped, leading to less V_{shift} and ultimately longer switch lifetime.

5.5 Novel Dual-Pulse Actuation Signal

In order to further improve the lifetime of the switch, a novel dual-pulse waveform as shown in Fig. 5.7 has been proposed. The novel actuation waveform gradually increases the actuation voltage at the beginning of the ON period, rather than a short constant pulse. This reduces the dielectric charging by effectively minimizing the time where high voltage is applied across the gap of two electrodes.

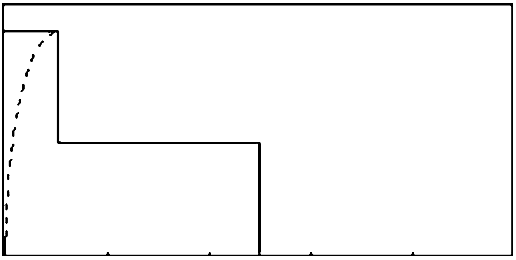


Fig. 5.7 A typical dual-pulse actuation signal (*solid line*) and the novel dual pulsed actuation signal (*dotted line*)

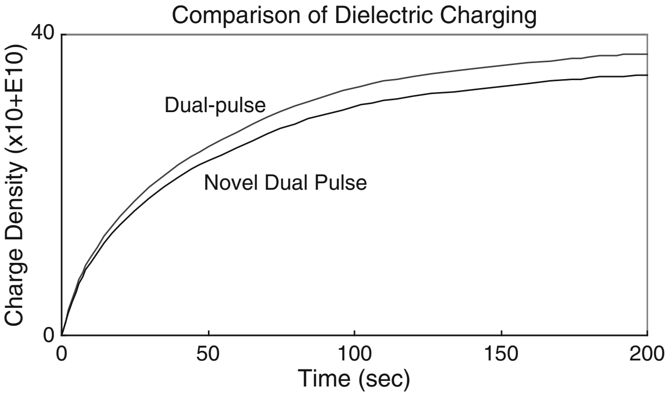


Fig. 5.8 Comparison of dielectric charging effect between a 100 Hz, typical dual pulse actuation signal and the modified dual pulse actuation signal

One main advantage of this technique is the simplicity in its implementation. The proposed actuation voltage curve can be simply driven by adding a simple RC low pass filter after the peak voltage source. This is a straightforward analog circuit implementation suitable for any MEMS switch rather than using active component such as microcontroller and sensing circuit. Admittedly, the RC value has to be properly tuned for a particular MEMS switch so that the switch operates correctly and does not compromise too much on the delay in the switching time. This novel actuation voltage has been shown to reduce the dielectric charging even further, as shown in Fig. 5.8.

5.6 Experimental Setup

To test the dielectric charging effect of the RF MEMS switch, the switch is stressed under a continuous ON-OFF actuation signal. The pull-in and pull-out voltages are then measured periodically. The MEMS switch used is the commercially produced

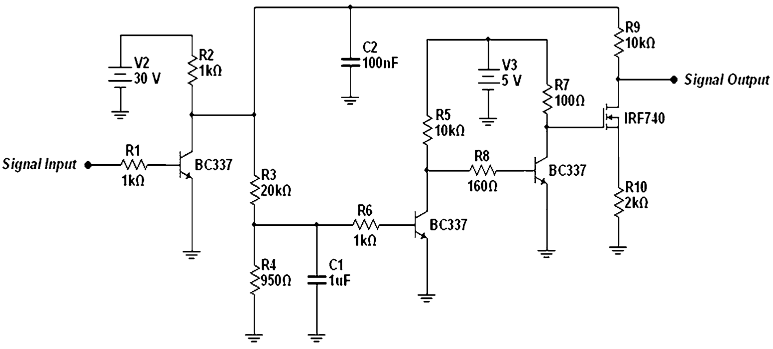


Fig. 5.9 Driver circuit for generating the novel dual-pulse actuation signal

TT712-68CSP SPDT RF MEMS switch fabricated by TeraVicta. The switch was actuated at 200 Hz, with 200 ms of high voltage at 68 V followed by a 55 V holding voltage for the rest of the ON time. A sine wave (1 V peak to peak, 100 kHz) was applied at the input pin of the MEMS switch to emulate RF signal, and the output of the switch is measured using oscilloscope. The MEMS switch is hot switched.

Two RF MEMS switches were tested in the experiment, labeled SW1 and SW4. Firstly, SW1 was tested with the normal dual-pulse actuation (DP) signal while SW4 was tested with the novel dual-pulse actuation (NDP) signal for 10 min. The pull-out voltage readings were taken in every 2 min. The switches were rested for days to allow the charges trapped in the dielectric to discharge before the next test, whereby SW1 was tested with the novel dual-pulse actuation signal and SW4 was tested with the normal dual-pulse actuation signal for 10 min. A 200 Hz, 0–2 V square wave signal was applied to the driver circuit shown in Fig. 5.9 to generate the desired actuation signal. The DP and NDP actuation voltage implementation differs only in the addition of capacitor C2 in the circuit.

The actuation voltage shift for both devices under test is shown in Fig. 5.10. The graph shows that both the variation and the rate of voltage shift are higher for the normal dual-pulse actuation signal (DP) compared to the proposed novel dual-pulse actuation signal (NDP). Therefore, the experimental results support the analysis done using simulation, whereby the dielectric charging effect can be reduced by applying a gradually build-up actuation voltage.

In addition, the proposed novel dual-pulse actuation signal is observed to generate less bouncing the moment the switch closes compared to the two-step dual-pulse actuation signal. Figure 5.11 depicts the switch contact bouncing for the two different actuation schemes observed during the experiment. Though reaching the contact faster, the conventional dual-pulse actuated switch suffers from bouncing which delay the switching time significantly. Mechanical impact and contact bouncing also lead the switch mechanical failures [7, 8] in the long run.

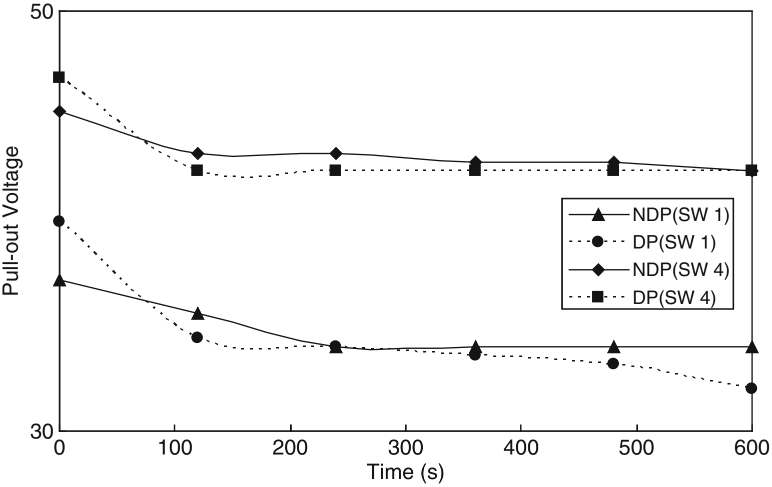


Fig. 5.10 Pull-out voltage shift for SW 1 and SW 4 over 10 min

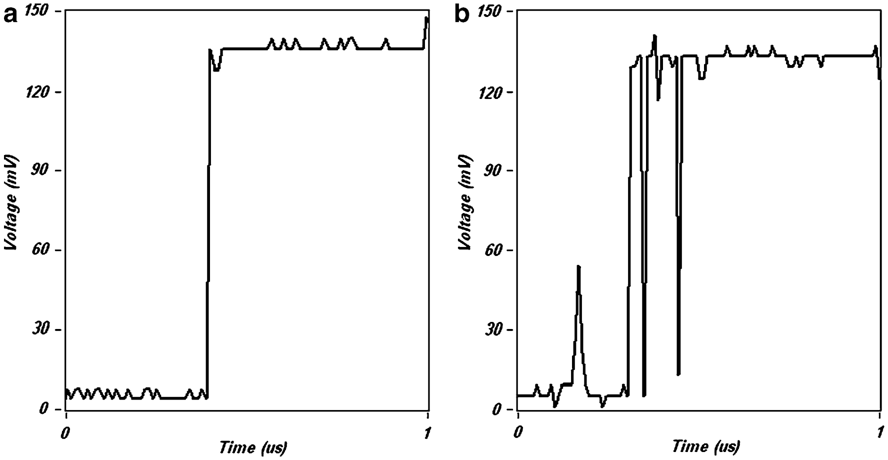


Fig. 5.11 Contact bouncing for (**a**) novel dual-pulse actuation signal and (**b**) normal dual-pulse actuation signal. A DC signal of 130 mV is applied to the switch input

5.7 Conclusion

Mathematical and transient circuit models of dielectric charging in a MEMS switch have been used to analyze the dielectric charging generated by different actuation voltage curves applied to the switch. From the analysis, a novel, simple to implement dual-pulse actuation voltage has been proposed to reduce the dielectric charging. Simulation using the model and experiment carried out using commercial MEMS switch have shown that the dielectric charging can be reduced, leading to less voltage shift and eventually increasing the lifetime of the MEMS switch.

References

1. Chan, R., Lesnick, R., Becher, D., Feng, M.: Low-actuation voltage RF mems shunt switch with cold switching lifetime of seven billion cycles. J. Microelectromech. Syst. **12**(5), (2003)
2. Goldsmith, C.L., Ehmke, J., Malczewski, A., Pillans, B., Eshelman S., Yao, Z., Brank, J., Eberly, M.: Lifetime characterization of capacitive RF MEMS switches. IEEE MTT-S Int. Microw. Symp. Dig. **1**, 227–230 (2001)
3. Yuan, X., Hwang, J.C.M., Forehand, D., Goldsmith, C.L.: Modeling and characterization of dielectric-charging effects in RF MEMS capacitive switches IEEE MTT-S Int. Microw. Symp. Dig. (June 2005), 753–756 (2005)
4. Herfst, R.W., Steeneken, P.G., Huizing, H.G.A., Schmitz, J.: Center-shift method for the characterization of dielectric charging in rf mems capacitive switches IEEE Trans Semiconduct Manufact **21**(2) (May 2008)
5. Yuan, X., Peng, Z., Hwang, J.C.M. Forehand, D., Goldsmith, C.L.: A transient SPICE model for dielectric-charging effects in RF MEMS capacitive switches IEEE Trans. Electr Device **53**(10), 2640–2648 (October 2006)
6. Merlijn van Spengen, W., Puers, R. Mertens, R., Wolf, I.D.: A comprehensive model to predict the charging and reliability of capacitive RF MEMS switches J. Micromech. Microeng. **14**(4):514–521 (January 2004)
7. Patton, S.T. Zabinski J.S.: Fundamental studies of Au contacts in MEMS RF switches Tribol Lett **18**(2) (February 2005)
8. Czaplewski, D.A., Dyck, C.W., Sumali, H., Massad, J.E., Kuppers, J.D., Reines, I., Cowan, W.D., Tigges, C.P.: A soft-landing waveform for actuation of a single-pole single-throw ohmic RF mems switch. J. Microelectromech. Syst. **15**(6) (December 2006)

Chapter 6
Optimal Control of Full Envelope Helicopter

Semuel Franko

Abstract Controlling rotary wing platforms, especially helicopters, is a difficult problem because of the nonlinearity of the structure and strong coupled motion dynamics. In this paper, linear quadratic regulator method is used to control the trajectory and mission paths of the autonomous helicopter. Nonlinear motion dynamics is trimmed and linearized about certain operating points and linear model is obtained by Taylor's expansion formula. This model is integrated into MATLAB/Simulink software. By using LQR methodology the attitude of the autonomous Puma helicopter is controlled and two simulations are realized. The results show that this approach can be effectively applied.

Keywords Auto pilot · optimal control · LQR · helicopter · trajectory control

6.1 Introduction

In recent years the concept of autonomous helicopter controlling has gained a big acceleration, because of their vertical take-off/landing advantages and hovering. Although the coupled and nonlinear dynamics of the helicopter makes the attitude control difficult, numerous control techniques are applied to perform missions like hovering, aggressive manoeuvring, course keeping etc. But conventional techniques like PD or PID becomes insufficient to control such a platform. Even for an experienced engineer it is hard to regulate considerable amount of parameters of the 6° of freedom helicopter. In the literature, among the control methods that are applied to helicopters LQR is a very efficient and relatively easy way to utilize.

Despite the fact that many researchers applied optimal control techniques to small scale helicopters [1], there is relatively few studies about full envelope

S. Franko (✉)
TUBITAK-Marmara Research Center, Information Technologies Institute,
Gebze, Kocaeli, Turkey
e-mail: semuel.franko@bte.mam.gov.tr

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_6,

helicopter control. Though, in war/tactical simulators it is necessary for the full envelope platforms have middle/high fidelity relative to real helicopters. The helicopter members of the simulator must hover, take-off and follow a path, etc. So this study aims to clarify the main points of modelling, trajectory/attitude control of the helicopter by LQR and contribute the literature about this problem. In future researchers will easily be able to integrate this model to their simulators.

6.2 Manual Control of Helicopter

Due to the strong coupling between the longitudinal and lateral motion of the helicopter, the work of the pilot is harder than an aircraft pilot. Pilot should simultaneously control three controllers, collective, cyclic, and tail pedals. Basically with collective controller pilot can adjust the altitude. By cyclic controllers pilot can change the angle of blades of main rotor so longitudinal and lateral motion can be performed. By feet pedals angle of the blades of the tail rotor is changed so yaw motion is performed. Any mistake can cause the collapse of the platform. In this paper, owing to the optimal control methods, the controller gains will be hold at the optimum values. These inputs will be defined in this study as $\theta_{0mr}, a_1/b_1, \theta_{0tr}$ respectively.

6.3 Mathematical Modelling of Helicopter

6.3.1 Coordinate Frames and Transformations

Basically two frames are needed to demonstrate the motion of the helicopter, body fixed and earth fixed frames. Force, moment and other effects act on the body frame. The origin of body fixed frame is center of gravity of the platform, and it moves with the motion of the fuselage. In this frame, x shows longitudinal, y shows lateral, and z shows up/down movement. In the latter coordinate system, x points the north, y points east, and z points the center of the earth. Earth frame notation is necessary for the calculation of the displacements (Fig. 6.1).

To transform between body and earth frames, orthonormal rotation matrix R is used. Motion equations are multiplied with R, which is the result of the rotation by Euler angles. Yaw, pitch, roll rotation order is the standard in aircraft modelling [2]. As cΘ shows cos(Θ) and sΘ shows sin(Θ) two rotation matrices can be shown as follows:

Body frame to earth frame:

$$R_{eb} = \begin{bmatrix} c\Theta c\Psi & s\Phi s\Theta c\Psi - c\Phi s\Psi & c\Phi s\Theta c\Psi + s\Phi s\Psi \\ c\Theta s\Psi & s\Phi s\Theta s\Psi + c\Phi c\Psi & c\Phi s\Theta s\Psi - s\Phi c\Psi \\ -s\Theta & s\Phi c\Theta & c\Phi c\Theta \end{bmatrix} \tag{6.1}$$

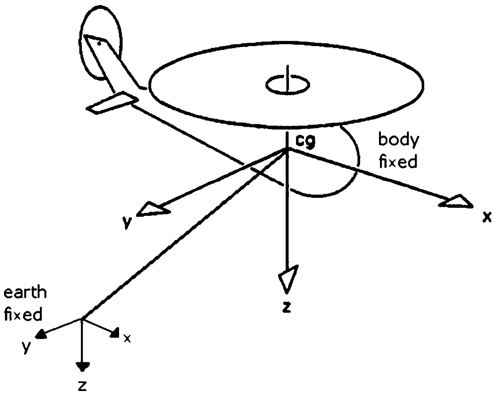


Fig. 6.1 Helicopter's two main frames

6.3.2 Dynamic Equations of Motion

By assuming that the platform as a rigid body, any two points on the helicopter doesn't change during the mission. The fuselage can make two types of movements: Translational and rotational. They define change in position and rotate around an axis respectively.

Translational motion, which is the motion of the center of gravity, can be defined by Newton's second law and Coriolis Effect. Linear accelerations along x, y, and z axes can be defined as:

$$\begin{bmatrix} \dot{u} = vr - qw + \dfrac{F_x}{m} \\ \dot{v} = pw - ur + \dfrac{F_y}{m} \\ \dot{w} = uq - pv + \dfrac{F_z}{m} \end{bmatrix} \tag{6.2}$$

Angular accelerations around x, y, and z axes can be defined as:

$$\begin{bmatrix} \dot{p} = qr\dfrac{I_{yy} - I_{zz}}{I_{xx}} + \dfrac{M_x}{I_{xx}} \\ \dot{q} = pr\dfrac{I_{zz} - I_{xx}}{I_{yy}} + \dfrac{M_y}{I_{yy}} \\ \dot{r} = pq\dfrac{I_{xx} - I_{yy}}{I_{zz}} + \dfrac{M_z}{I_{zz}} \end{bmatrix} \tag{6.3}$$

6.3.3 Kinematic Equations

To represent the motion of the helicopter with respect to earth fixed frame, kinematic equations must be used. For translational kinematics, relation between body and earth fixed frame is as follows, where x_E, y_E, z_E identifies position of the helicopter with respect to earth-fixed frame.

$$\frac{dy}{dx}\begin{bmatrix} x_E \\ y_E \\ z_E \end{bmatrix} = R_{eb}\begin{bmatrix} u_B \\ v_B \\ w_B \end{bmatrix} \tag{6.4}$$

Rotational kinematic equations of helicopter are as fallows, where ϕ, θ, ψ defines Euler angles of roll, pitch, and yaw respectively.

$$\begin{bmatrix} \dot{\phi} = p + \tan(\theta)\,[q\sin(\phi) + r\cos(\phi)] \\ \dot{\theta} = q\cos(\phi) - r\sin(\phi) \\ \dot{\psi} = [q\sin(\phi) + r\cos(\phi)]\sec(\theta) \end{bmatrix} \tag{6.5}$$

6.3.4 Force and Moments Acting on Helicopter

In order to represent the motion of the helicopter, force and moment effects must be taken into account.

Helicopter can be modeled by combining five subsystems: main-rotor, fuselage, empennage (consist of horizontal stabilizer and vertical fin), tail rotor and engine [4]. To define the force and moment effects originated from main rotor, tail rotor, gravity and drag on main rotor; mr, tr, g, and d subscripts are used respectively.

$$\begin{bmatrix} F_x = X_{mr} + X_{tr} + X_g \\ F_y = Y_{mr} + Y_{tr} + Y_g \\ F_z = Z_{mr} + Z_{tr} + Z_g \\ L = L_{mr} + L_{tr} + L_d \\ M = M_{mr} + M_{tr} + M_d \\ N = N_{mr} + N_{tr} + N_d \end{bmatrix} \tag{6.6}$$

As T_{mr} and T_{tr} shows main and tail rotor thrust, a1 and b1 shows longitudinal flapping angle and lateral flapping angle respectively, we obtain combined force equation matrix:

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \begin{bmatrix} -T_{mr}{}^*\sin(a_1) - \sin(\theta)^* mg \\ T_{mr}{}^*\sin(b_1) + T_{tr} + \sin(\phi)^*\cos(\theta)^* mg \\ -T_{mr}{}^*\cos(b_1)^*\cos(a_1) + \cos(\phi)^*\cos(\theta)^* mg \end{bmatrix} \tag{6.7}$$

As h_{mr}, h_{tr} represents distance between cog and main/tail rotor along z axis, l_{mr}, l_{tr} represents distance between cog and main/tail rotor along x axis, Q_{mr} defines counter torque that comes from the drag of main rotor, we can obtain combined torque equation matrix:

$$\begin{bmatrix} \tau_x \\ \tau_y \\ \tau_z \end{bmatrix} = \begin{bmatrix} Y_{mr}{}^* h_{mr} - Z_{mr}{}^* y_{mr} + Y_{tr}{}^* h_{tr} + Q_{mr}{}^* \sin(a_1) \\ -X_{mr}{}^* h_{mr} - Z_{mr}{}^* l_{mr} - Q_{mr}{}^* \sin(b_1) \\ X_{mr}{}^* y_{mr} + Y_{mr}{}^* l_{mr} - Y_{tr}{}^* l_{tr} + Q_{mr}{}^* \cos(a_1)^* \cos(b_1) \end{bmatrix} \tag{6.8}$$

By using variables above nonlinear mathematical model can be build. To apply a linear controller the model must be linearized about certain operating points which will be covered in next section.

6.3.5 *Trimming and Linearization*

Nonlinear motion equations must be linearized about certain operating points. To increase the fidelity of the model eight trim points (0, 20, 40, 60, 80, 100, 120 and 140 knots) have been used. First assuming that linear and angular accelerations are zero; trimming conditions are obtained. This algorithm changes conditions until ϕ, θ, a_1 and b_1 reaches to steady state value in our desired flight condition.

By using Taylor's series expansion, external forces acting on platform become linear functions of perturbed states. Total force along x axis, by the advantage of small perturbation theory ($x = x_e + \Delta x$), can be written as follows:

$$\left[F_x = X = X_e + \frac{\partial x}{\partial u}\Delta u + \frac{\partial x}{\partial w}\Delta w + \frac{\partial x}{\partial q}\Delta q + \frac{\partial x}{\partial \theta}\Delta\theta + \cdots \right] \tag{6.9}$$

$$\left[X = X_e + X_u\Delta u + X_w\Delta w + X_q\Delta q + X_\theta\Delta\theta + \cdots \right]$$

If we consider that the motion can be described nonlinearly as $\dot{x} = F(x, u, t)$, the linearized model can be defined as $\dot{x} = Ax + Bu$, where $x = \begin{bmatrix} u \; w \; q \; \theta \; v \; p \; \phi \; r \end{bmatrix}$ and $u = \begin{bmatrix} \theta_{0mr} \; a_1 \; b_1 \; \theta_{0tr} \end{bmatrix}$.

The coefficients like $X_u, X_w, \ldots$ are called stability derivatives in flight dynamics.

The result of these formulations can be found as A and B matrices.

6.3.6 *Obtaining the Stability Derivatives*

Stability derivatives can be calculated by using numerical or analytical methods. By using platform's main characteristics [5, 6] all derivatives can be figured. For example X_u can be found analytically by the equations below:

$$\begin{bmatrix} \dfrac{\partial X_{mr}}{\partial u} = \dfrac{\partial (Ta_1)}{\partial u} = \dfrac{\partial T}{\partial u} a_1 + T\dfrac{\partial a_1}{\partial u} \cdots \\ \dfrac{\partial T}{\partial u} = \rho(\Omega R)^2 \pi R^2 \dfrac{\partial C_T}{\partial u} \cdots \end{bmatrix} \tag{6.10}$$

Calculation of other derivatives was not mentioned in this paper. Further formulae can be found from [4]. For control study the stability derivatives of the PUMA type helicopter was used by author.

6.4 Controller Design

Helicopters which are open loop unstable must be stabilized and controlled carefully (Figs. 6.2 and 6.3).

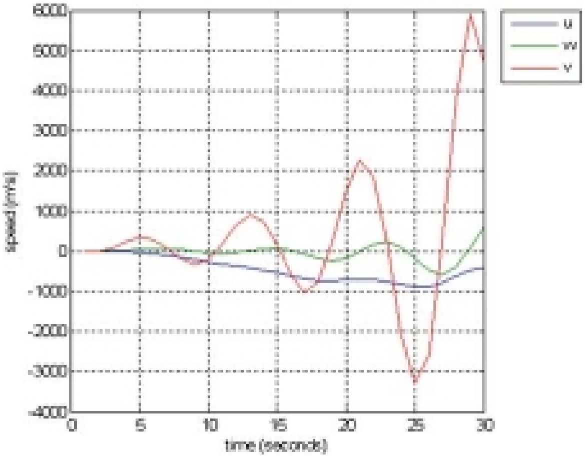


Fig. 6.2 Open loop step response

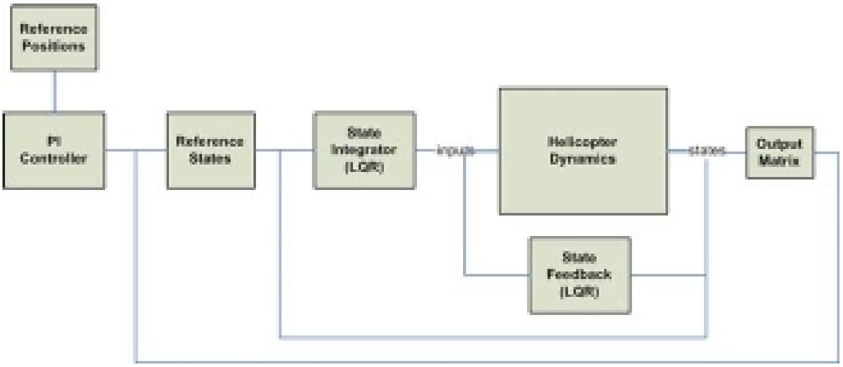


Fig. 6.3 Block diagram of controllers

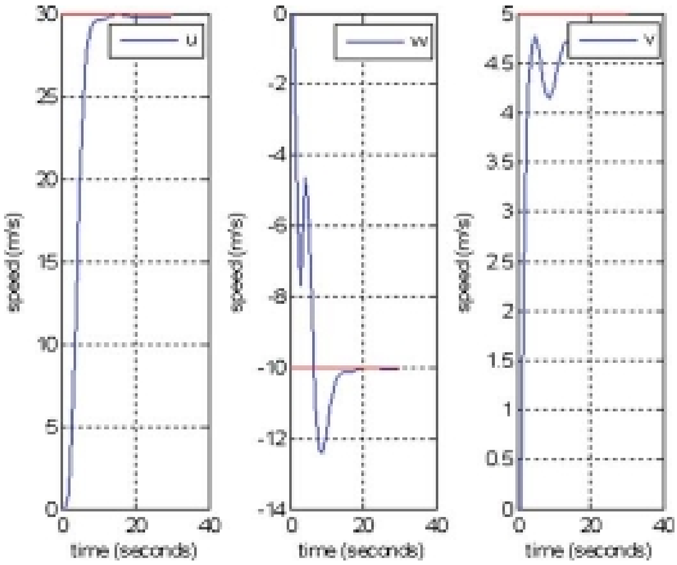


Fig. 6.4 Comparisons of u, w, and v

Open loop system's inputs are deflected up to full angle and a step input is put in the first second. The speed responses of the system on each axis can be seen in Fig. 6.4.

As seen in Fig. 6.5 proposed optimal controllers consist of three subsystems. State feedback controller, state integrator and PI controller. And gain scheduling is used to reflect the change in platform dynamics with respect to the forward velocity.

6.4.1 State Feedback Controller

Full-state feedback control algorithm tries to minimize the performance index (J), where "x" shows states, "u" inputs, Q and R are weighting matrices,

$$J = \frac{1}{2}\int_{0}^{\infty} (xQx^T + uRu^T)dt \tag{6.11}$$

To decide appropriate weighting matrices, Bryson's rule is used [8]. In this rule weights are calculated with the inverse of to the maximum available state/input value's square.

By checking the set/rise time, overshoot and controlling effort; weights are tuned. By solving steady state Riccati equation, K gain matrix calculated offline. In the simulation input is calculated in each time step by u = −Kx.

6.4.2 State Integrator

Full state feedback control gives adequate results. But controller will regulate the dynamical system state values to zero. For tracking control of the helicopter, error between reference and actual states must be taken into account [7]. So the error term is defined as $\dot{e} = \dot{x}_{desired} - \dot{x}_{actual}$. After integrating $\dot{e}$, e is obtained. Then new control input becomes $u = K_i{}^{*}e - K^{*}\dot{x}$.

Comparison of the reference and actual longitudinal, vertical and lateral speed can be seen from figure below respectively. Controller tracks reference speed values successfully.

6.4.3 PI Controller

After controlling the states and setting that values according to the reference states, for position control, which has slower dynamics than attitude control a proportional-integral feedback controller is used. Position error is calculated as $e_{pos} = [x, y, z]_{desired} - [x, y, z]_{actual}$. By using classical $K_P{}^{*}e_{pos} + K_I \int e_{pos}\, dt$ formula the trajectory control is realized.

6.5 Simulations

For testing the controllers following two scenarios are formed.

6.5.1 Movement to Point

Initial point = [x z y] = [0 0 0]
Target point = [500 −200 100] m
Reference velocities = [30 −10 5] m/s (Figs. 6.5–6.8)

6.5.2 Movement Through Waypoints

For testing the controllers a waypoint scenario is formed. Initial point is set as [x z y] = [0 0 0] m.

Four waypoints were selected as follows.
Waypoint 1 = [1000 400 0]
Waypoint 2 = [0 800 0]

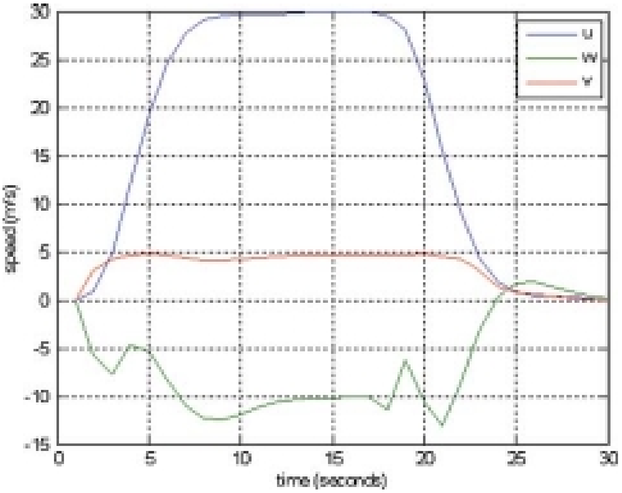


Fig. 6.5 Resultant velocities on three axes

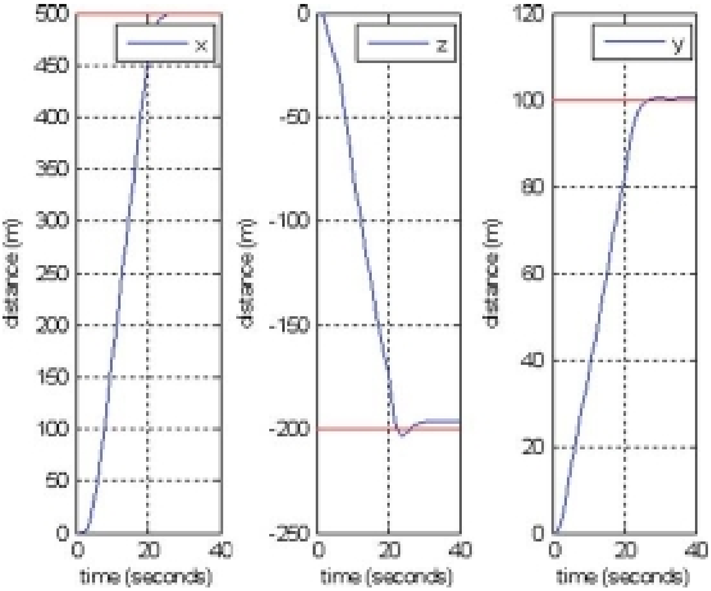


Fig. 6.6 Resultant distances on three axes

Waypoint 3 = $[-1000 \quad 400 \quad 0]$
Waypoint 4 = $[0 \quad 0 \quad 0]$
Commanded composite velocity (of u and v) is 20 m/s.

In Fig. 6.9, squares show waypoints, lines show the actual way of the helicopter.

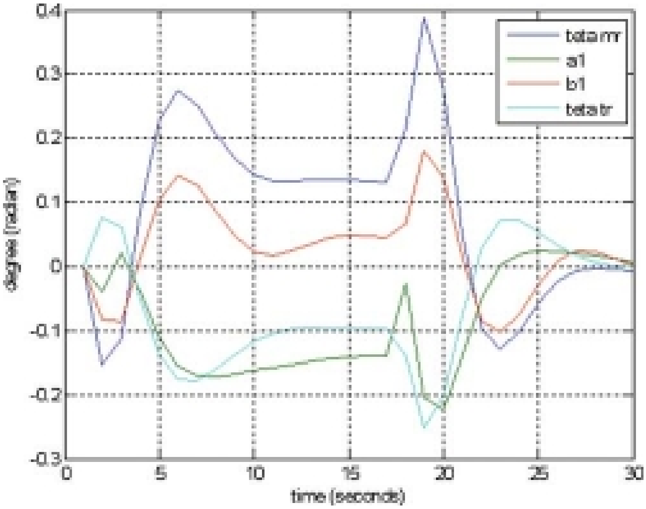


Fig. 6.7 Control efforts

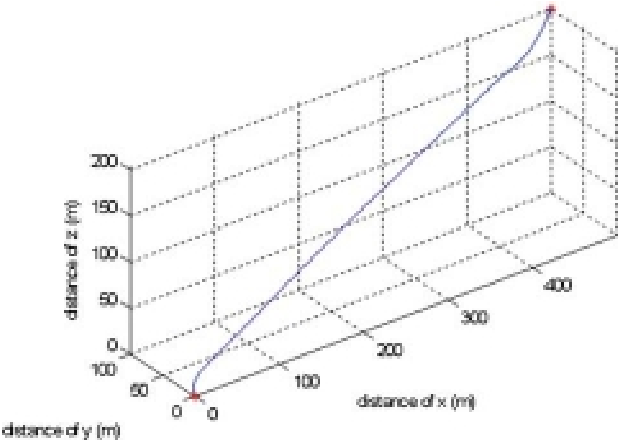


Fig. 6.8 3D view of the motion

6.6 Conclusion

To control an unmanned air vehicle, kinematics, dynamics and mathematical modelling of the platform was examined in detail. Optimal and classical control techniques are applied to achieve the missions.

Basic results of the study and future work can be summarized as follows:

- A platform which has strong coupling affects can effectively be controlled by LQR methods. Fast dynamics and control efforts can easily be optimized to reflect the real motion of the helicopter in simulators.

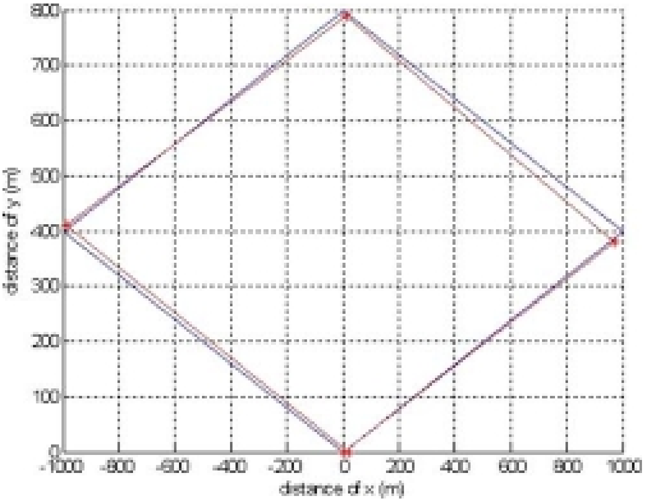


Fig. 6.9 2D view of motion through waypoints

- For full envelope platforms, PI control is sufficient to control slow dynamics like position control.
- In future obstacle avoidance algorithms can be integrated to these works to use in tactic environment.
- Disturbance scenarios can be added to the simulations to increase reality.
- Intelligent and/or model based controller techniques can be applied to helicopter. These controllers can be compared with the optimal techniques.

Acknowledgement The author gratefully acknowledges Mr. Mehmet Haklıdır and Mrs. Melike Gürbüz's helpful comments about this work.

References

1. Jiang, Z., Han, J., Wang, Y., Song, Q.: Enhanced LQR control for unmanned helicopter in hover, ISSCA-Systems and Control in Aerospace and Astronautics (2006)
2. Mettler, B.: Identification Modelling and Characteristics of Miniature Rotorcraft. Kluwer, Norwell, MA (2003)
3. Prouty, R.W.: Helicopter Performance, Stability and Control. PWS Publishers, Boston, MA (1986)
4. Padfield, G.D.: Helicopter Flight Dynamics: The Theory and Application of Flying Qualities and Simulation Modeling. AIAA Educational Series, Washington, DC (1995)
5. Jensen, R., Nielsen, A.: Robust control of an autonomous. M.Sc. thesis, Helicopter, Aalborg University (2005)
6. Munzinger, C.: Development of a real-time flight simulator for an experimental model helicopter. M.Sc. thesis, Georgia Institute of Technology, USA (1998)
7. Ogata, K.: Discrete-Time Control Systems. Prentice–Hall, Upper Saddle River, NJ (1995)
8. Bryson, A.E., Ho, Y.-C.: Applied Optimal Control: Optimization, Estimation, and Control. Taylor & Francis, London (1975)

Chapter 7
Determination of Consistent Induction Motor Parameters

Christian Kral, Anton Haumer, and Christian Grabner

Abstract From the rating plate data of an induction motor the nominal efficiency can be determined. Without detailed knowledge of equivalent circuit parameters, partial load behavior cannot be computed. Therefore, a combined calculation and estimation scheme is presented, where the consistent parameters of an equivalent circuit are elaborated, exactly matching the nominal operating point. From these parameters part load efficiencies can be determined.

Keywords Induction motor · consistent parameters · partial load efficiency

7.1 Introduction

The parameter identification and estimation of permanent magnet or [10] electric excited synchronous motors [8], switched reluctance motors [11] and induction motors [1] is essential for various applications. For a controlled drive the performance is very much related with the accuracy of the determined motor parameters [13], e.g., in energy efficient drives [3]. Another field of applications which may rely on properly determined parameters is condition monitoring and fault diagnosis of electric motors [7].

For an induction motor, operated at nominal voltage and frequency, and loaded by nominal mechanical power, the current and power factor indicated on the rating plate show certain deviations of the quantities obtained by measurements. The rating plate data are usually rounded and subject to certain tolerances. Deviations of the measured data from the data stated on the rating plate are therefore not surprising.

In this paper a method for the determination of the consistent parameters of the equivalent circuit of an induction motor is proposed. In this context it is very important to understand, that the proposed approach does not estimate the

C. Kral (✉), A. Haumer, and C. Grabner
Electric Drive Technologies, Austrian Institute of Technology, Giefinggasse 2,
1210 Vienna, Austria
e-mail: christian.kral@ait.ac.at; anton.haumer@ait.ac.at; christian.grabner@ait.ac.at

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_7,

Table 7.1 Rating plate data of induction motor

Quantity	Symbol	SI unit
Nominal mechanical output power	$P_{m,N}$	W
Nominal phase voltage	$V_{s,N}$	V
Nominal phase current	$I_{t,N}$	A
Nominal power factor	pf_N	–
Nominal frequency	$f_{s,N}$	Hz
Nominal speed	n_N	1/s

parameters of a real induction motor. In this sense, consistent parameters mean, that the determined parameters *exactly* model the specified operating point specified by the rating plate data summarized in Table 7.1.

The estimation of motor parameters from name plate data [2, 9, 12] is thus not sufficient, since the parameters are not implicitly consistent. This means, that the nominal efficiency computed by means of estimated parameters does *not* exactly compute the nominal operating point.

A calculation for the determination of consistent parameters is proposed. The applied model takes stator and rotor ohmic losses, core losses, friction losses and stray-load losses into account, while exactly modeling the specified nominal operating conditions. Some of the required parameters can either be measured, others may be estimated through growth relationships. Due to implicit consistency restrictions, the remaining parameters are solely determined by mathematical equations.

7.2 Power Balance

For the presented power balance and equivalent circuit the following assumptions apply:

- Only three phase induction motors (motor operation) are investigated.
- The motor and the voltage supply are fully symmetrical.
- The voltages and currents are solely steady state and of sinusoidal waveform.
- Only fundamental wave effects of the electromagnetic field are investigated.
- Non-linearities such as saturation and the deep bar effect are not considered.

The algorithm presented in this paper relies on the single phase equivalent circuit depicted in Fig. 7.1 and the power balance shown in Fig. 7.2. This equivalent circuit considers stator and rotor ohmic losses as well as core losses, which is why the phasors of the terminal current $\underline{I}_t$ and the stator current $\underline{I}_s$ have to be distinguished.

In order to simplify the explicit calculation of the equivalent circuit parameters, the conductor, representing core losses, is connected directly to the stator terminals. In this approach, stator and rotor core losses

$$P_c = 3G_c V_s^2 \tag{7.1}$$

are modeled together, where G_c represents the total core conductance.

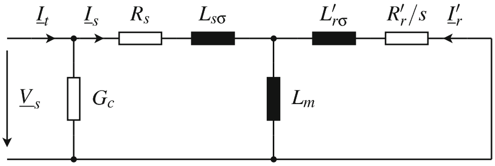


Fig. 7.1 Equivalent circuit of the induction motor

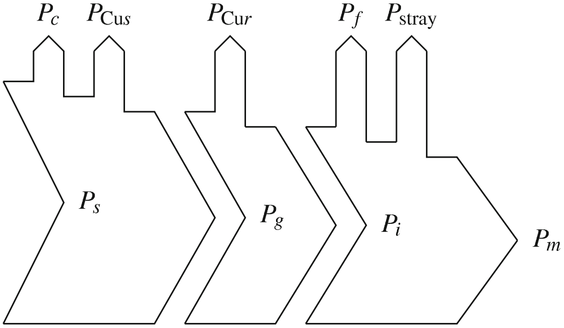


Fig. 7.2 Power balance of the induction motor

Due to the modeling topology of the equivalent circuit, stator and rotor ohmic losses are

$$P_{\text{Cu}s} = 3R_s I_s^2, \tag{7.2}$$

$$P_{\text{Cu}r} = 3R'_r I_r'^2. \tag{7.3}$$

The rotor parameters of the equivalent circuit are transformed to the stator side, and thus indicated by the superscript $'$.

The inner (electromagnetic) power of the motor is determined by

$$P_i = R'_r \frac{1-s}{s} I_r'^2. \tag{7.4}$$

Friction and stray-load losses are not considered in the power balance of the equivalent circuit. Nevertheless, these effects have to be taken into account independently.

In the total power balance of the motor, the total electrical input power is P_s, and the air gap power is

$$P_g = P_s - P_c - P_{\text{Cu}s}. \tag{7.5}$$

The inner power

$$P_i = P_g - P_{\text{Cu}r} \tag{7.6}$$

is computed from the inner power and the copper heat losses. Mechanical output (shaft) power

$$P_m = P_i - P_f - P_{\text{stray}} \tag{7.7}$$

is determined from inner power, P_i, friction losses, P_f, and stray-load losses, P_{stray}.

7.3 Determination of Parameters

In the following a combination of parameter calculations and estimations, based on empirical data, is presented. The presented results all refer to empirical data obtained from 50 Hz standard motors. Unfortunately the empiric data cannot be revealed in this paper. Nevertheless the applied mathematical approach for the estimation is presented. For motors with a certain mechanical power and number of pole pairs empirical data can be approximated with good accuracy.

The determination of the equivalent circuit parameters is, however, based on the rating plate data of the motor (Table 7.1). From these data the nominal electrical input power

$$P_{s,N} = 3V_{s,N} I_{t,N}\,\mathrm{pf}_N \tag{7.8}$$

and the nominal angular rotor velocity

$$\Omega_{m,N} = 2\pi n_N \tag{7.9}$$

can be computed.

7.3.1 Measurement or Estimation of Core Losses

If measurement results are obtained from a real motor, the core losses can be determined according to IEEE Standard 122 [4], by separating core and friction losses. The no load core losses, $P_{c,0}$, refer to the nominal voltage and nominal frequency. It is assumed that the no load test is performed at synchronous speed ($s = 0$),

$$\Omega_{m,0} = \frac{2\pi f_{s,N}}{p}. \tag{7.10}$$

In case measurement results are not available, the no load friction losses, for a motor with a certain nominal mechanical output power, $P_{m,N}$, and a certain number of pole pairs, p, can be estimated from empirical data by

$$\log_{10}\left(\frac{P_{c,0}}{P'_c}\right) = k_{c[p]}\log_{10}\left(\frac{P_{m,N}}{P'_m}\right) + d_{c[p]}. \tag{7.11}$$

In this equation, P'_c and P'_m are arbitrary reference power terms to normalize the argument of the logarithm. The parameters $k_{c[p]}$ and $d_{c[p]}$ are obtained from the empiric data and particularly refer to a specific number of pole pairs. From the estimated or measured no load core losses the reference core conductance can be derived,

$$G_{c,\mathrm{ref}} = \frac{P_{c,0}}{3V_{s,N}}. \tag{7.12}$$

7.3.2 Measurement or Estimation of Friction Losses

From a no load test ($s = 0$) the friction losses, $P_{f,0}$, of a particular real motor can also be determined. In case no measurements are available, the no load friction losses can be estimated from empirical data,

$$\log_{10}\left(\frac{P_{f,0}}{P'_f}\right) = k_{f[p]}\log_{10}\left(\frac{P_{m,N}}{P'_m}\right) + d_{f[p]}. \tag{7.13}$$

In this equation P'_f and P'_m are some arbitrary reference power terms, and $k_{f[p]}$ and $d_{f[p]}$ are empiric parameters corresponding for a certain number of pole pairs. The no load friction losses $P_{f,0} = P_{f,\text{ref}}$ are corresponding with $\Omega_{m,0} = \Omega_{m,\text{ref}}$.

7.3.3 Calculation of Stray-Load Losses

The nominal stray load losses can be derived from

$$P_{\text{stray},N} = a_{\text{stray}} P_{m,N}, \tag{7.14}$$

and parameter a_{stray} can be obtained by IEEE [4].

7.3.4 Estimation of Stator Resistance

The stator resistance can be estimated applying the power balance with respect to the nominal operation point. From the nominal mechanical output power, $P_{m,N}$, the nominal friction losses

$$P_{f,N} = P_{f,0}\left(\frac{\Omega_{m,N}}{\Omega_{m,0}}\right)^{a_f+1} \tag{7.15}$$

and the nominal stray-load losses (7.14), the nominal inner power

$$P_{i,N} = P_{m,N} - P_{f,N} - P_{\text{stray},N} \tag{7.16}$$

can be calculated. The air gap power can be determined from the nominal inner power and nominal slip, s_N [5]:

$$P_{g,N} = \frac{P_{i,N}}{1 - s_N} \tag{7.17}$$

The stator copper losses can then be obtained by the power balance (7.5), applied to the nominal operating point,

$$P_{\mathrm{Cu}s,N} = P_{s,N} - P_{g,N} - P_{c,N}. \tag{7.18}$$

Since the core loss conductor is connected to the terminals in Fig. 7.1, the core losses with respect to the nominal and no load operating point are equal,

$$P_{c,N} = P_{c,0}. \tag{7.19}$$

From the nominal stator phase current, and the nominal stator copper losses (7.18) consistent stator resistance are determined by

$$R_s = \frac{P_{\mathrm{Cu},s,N}}{3I_{s,N}^2}. \tag{7.20}$$

7.3.5 Measurement or Estimation of No Load Current

In the following space phasor calculations are applied to the nominal operating conditions. The phase angle φ_f of synchronous reference is chosen such way that the imaginary part of the stator voltage space phasor

$$\underline{V}_{s,N}^{f} = V_{sx,N} + \mathrm{j}V_{sy,N} \tag{7.21}$$

is zero,

$$V_{sx,N} = \sqrt{2}V_{s,N} \tag{7.22}$$

$$V_{sy,N} = 0. \tag{7.23}$$

If the no load terminal current

$$\underline{I}_{t,0}^{f} = I_{tx,0} + \mathrm{j}I_{ty,0} \tag{7.24}$$

has not been obtained by measurement results, the reactive component has to be estimated. This can be performed by

$$\log_{10}\left(\frac{I_{ty,0}}{I'_{ty,0}}\right) = k_{0[p]}\log_{10}\left(\frac{P_{m,N}}{P'_m}\right) + d_{0[p]}, \tag{7.25}$$

where $I'_{ty,0}$ and P'_m are an arbitrary reference current and power term. The parameters $k_{0[p]}$ and $d_{0[p]}$ are estimated from empiric data with respect to a certain number of pole pairs. Then the imaginary part of the stator current space phasor can be derived by

$$I_{sy,0} = I_{ty,0}, \tag{7.26}$$

according to (7.24). Once all the parameters of the motor are determined, the consistent no load current $I_{t,0}$ can be determined from the equivalent circuit in Fig. 7.1.

7.3.6 Determination of Stator Inductance

The rotor current space phasor diminishes under no load conditions, and thus the stator current equation of the equivalent circuit in Fig. 7.1 yields:

$$\sqrt{2}V_{s,N} = R_s I_{sx,0} - \omega_s L_s I_{yx,0} \tag{7.27}$$

$$0 = \omega_s L_s I_{sx,0} + R_s I_{sy,0} \tag{7.28}$$

By eliminating the real part, $I_{sx,0}$, in these equations, the remaining equation for the imaginary part, considering $I_{sy,0} < 0$, can be used to determine the stator inductance

$$L_s = -\frac{V_{sx,N} + \sqrt{V_{sx,N}^2 - 4R_s^2 I_{sy,0}^2}}{2\omega_s I_{sy,0}}. \tag{7.29}$$

7.3.7 Determination of Stray Factor and Rotor Time Constant

The components of the nominal terminal current space phasor are

$$\underline{I}_{t,N}^{f} = I_{tx,N} + \mathrm{j} I_{ty,N} \tag{7.30}$$

From the components of the terminal current

$$I_{tx,N} = +I_{t,N}\,\mathrm{pf}_N, \tag{7.31}$$

$$I_{ty,N} = -I_{t,N}\sqrt{1-\mathrm{pf}_N^2}, \tag{7.32}$$

the components of the stator current space phasor can be derived:

$$I_{sx,N} = I_{tx,N} - \sqrt{2}G_c V_{s,N}, \tag{7.33}$$

$$I_{sy,N} = I_{ty,N}. \tag{7.34}$$

The real and the imaginary part of the voltage equation of Fig. 7.1 yields:

$$\frac{V_{sx,N}}{R_s} = (1 - \sigma a_s a_r) I_{sx,N} - (a_s + a_r) I_{sy,N} \tag{7.35}$$

$$\frac{a_r}{a_s}\frac{V_{sx,N}}{R_s} = (a_s + a_r) I_{sx,N} + (1 - \sigma a_s a_r) I_{sy,N} \tag{7.36}$$

From these two equations the two unknown parameters

$$a_r = \frac{a_s R_s I_{s,N}^2 + I_{sy,N} V_{sx,N}}{I_{sx,N} V_{sx,N} - R_s I_{s,N}^2}, \tag{7.37}$$

$$\sigma = \frac{(2I_{sx,N} - a_s I_{sy,N}) V_{sx,N} - R_s I_{s,N}^2 - \frac{V_{sx,N}^2}{R_s}}{a_s (a_s R_s I_{s,N}^2 + I_{sy,N} V_{sx,N})}, \tag{7.38}$$

can be obtained. The rotor time constant is the ratio of the rotor inductance and the rotor resistance and can be expressed in terms of known parameters,

$$T_r = \frac{a_r}{\omega_{r,N}}, \tag{7.39}$$

where $\omega_{r,N}$ refers to the nominal operating point. From (7.37)–(7.39) it can be seen that the rotor time constant and the leakage factor are independent of the particular choice of σ_{sr}.

7.3.8 Determining Magnetizing Inductance, Rotor Inductance and Rotor Resistance

The rotor leakage and main field inductances of the equivalent circuit in Fig. 7.1 and the rotor resistance cannot be determined independently [6]. It is thus useful to introduce the stray factor

$$\sigma = 1 - \frac{L_m^2}{L_s L_r'}, \tag{7.40}$$

and the ratio of stator to rotor inductance,

$$\sigma_{sr} = \frac{L_s}{L_r'}. \tag{7.41}$$

From the leakage factors σ and σ_{sr}, and the stator inductance L_s, the following equations can be derived:

$$L_m = L_s \frac{\sqrt{1 - \sigma}}{\sqrt{\sigma_{sr}}} \tag{7.42}$$

$$L'_r = \frac{L_s}{\sigma_{sr}} \tag{7.43}$$

From (7.39) and (7.43) the rotor resistance can be computed.

$$R'_r = \frac{L'_r}{T_r}. \tag{7.44}$$

The quantities determined in this subsection are fully dependent on the particular choice of σ_{sr}.

7.4 Measurement and Calculation Results

A comparison of calculation and measurement results is presented for a 18.5 kW four pole induction motor in Figs. 7.3 and 7.4. The investigations are performed for the motor operated at nominal voltage and frequency. Core and friction losses are known from a no load test. The stray load losses are determined according to IEEE [4].

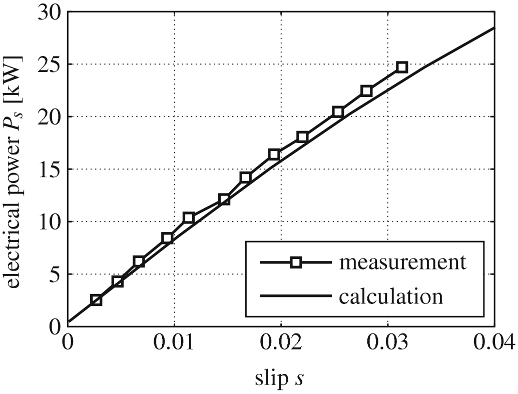


Fig. 7.3 Electrical input power versus slip of a four pole induction motor with 18.5 kW; measurement and calculation

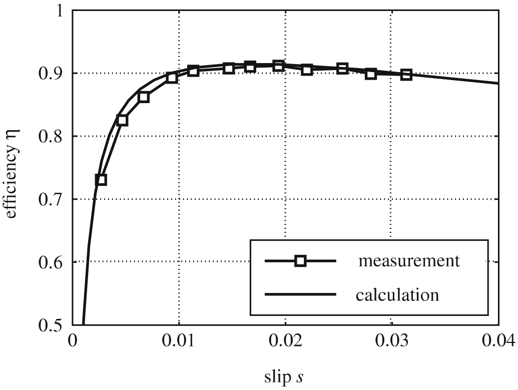


Fig. 7.4 Efficiency versus slip of a four pole induction motor with 18.5 kW; measurement and calculation

7.5 Conclusions

The intention of this paper was the calculation of the consistent parameters of an induction motor. In this context it is important to know that the proposed approach does not model the real motor behavior, but exactly models the nominal operating point, specified by the rating plate. In the presented model ohmic losses, core losses, friction losses and stray-load losses are considered. Mathematical algorithms and estimations – where necessary – are presented. For an investigated four pole 18.5 kW induction motor, measurement and calculation results are compared, revealing well matching results and thus proving the applicability of the presented calculations.

References

1. Babau, R., Boldea, I., Miller, T.J.E., Muntean, N.: Complete parameter identification of large induction machines from no-load accelerationdeceleration tests. IEEE Trans. Ind. Electron. **54**(4), 1962–1972 (2007)
2. Davey, K.: Predicting induction motor circuit parameters. IEEE Trans. Magn. **38**(4), 1774–1779 (2002)
3. de Almeida Souza, D., de Aragao Filho, W.C.P., Sousa, G.C.D.: Adaptive fuzzy controller for efficiency optimization of induction motors. IEEE Trans. Ind. Electron. **54**(4), 2157–2164 (2007)
4. IEEE: Standard test procedures for polyphase induction motors and generators. IEEE Standard 112, 2004
5. Jordan, H., Klima, V., Kovacs, K.P.: Asynchronmaschinen. F. Vieweg & Sohn Verlag, Braunschweig, 1975
6. Kleinrath, H.: Ersatzschaltbilder für Transformatoren und Asynchronmaschinen. *e&i*, **110**, 68–74 (1993)
7. Kral, C., Wieser, R.S., Pirker, F., Schagginger, M.: Sequences of field-oriented control for the detection of faulty rotor bars in induction machines – the vienna monitoring method. IEEE Trans. Ind. Electron. **47**(5), 1042–1050 (2000)
8. Kyriakides, E., Heydt, G.T., Vittal, V.: Online parameter estimation of round rotor synchronous generators including magnetic saturation. IEEE Trans. Energy Conver. **20**(3), 529–537 (2005)
9. König, H.: Ermittlung der Parameter der Drehstrom-Asynchronmaschine vorwiegend aus den Typenschildangaben. Elektrie **6**, 220–220 (1988)
10. Lee, J.-Y., Lee, S.-H., Lee, G.-H., Hong, J.-P., Hur, J.: Determination of parameters considering magnetic nonlinearity in an interior permanent magnet synchronous motor. IEEE Trans. Magn. **42**(4), 1303–1306 (2006)
11. Niazi, P., Toliyat, H.A.: Online parameter estimation of permanent-magnet assisted synchronous reluctance motor. IEEE Trans. Ind. Appl. **43**(2), 609–615 (2007)
12. Pedra, J., Corcoles, F.: Estimation of induction motor double-cage model parameters from manufacturer data. IEEE Trans. Energy Conver. **19**(2), 310–317 (2004)
13. Toliyat, H.A., Levi, E., Raina, M.: A review of rfo induction motor parameter estimation techniques. IEEE Trans. Energy Conver. **18**(2), 271–283 (2003)

Chapter 8
Broken Rotor Bars in Squirrel Cage Induction Machines – Modeling and Simulation

Christian Kral, Anton Haumer, and Christian Grabner

Abstract This paper presents a physical model of a squirrel cage induction machine with rotor topology. The machine is modeled in Modelica, an object oriented multi physical modeling system. The presented machine model is used to investigate electrical rotor asymmetries in induction machines. For the case of a single broken rotor bar a simulation model is compared with measurement results.

Keywords Modelica · induction machines · squirrel cage · rotor asymmetries · broken rotor bar

8.1 Introduction

The squirrel cage of an induction machine can be seen as a multiphase winding with shorted turns. Instead of a wound structure the rotor consists of N_r bars and two end rings, connecting the bars on both ends, as depicted in Fig. 8.1. At the end rings fins are located to force the circulation of air in the inner region of the machine.

In the production plant it is intended to fabricate a fully symmetrical squirrel cage. Yet, manufacturing tolerances and technological uncertainties give rise to inhomogeneous material densities. These inhomogeneities cause unequal rotor bar and end ring resistances – the so called electrical rotor asymmetries. The causes for such rotor electrical asymmetries are:

- Shrink holes and voids in the material of the bars or end rings
- Improper junctions of the bars and end rings
- Heavy duty start-ups that the machine is not designed for
- Thermal overloading of the machine
- High temperature gradients, causing cracks

C. Kral (✉), A. Haumer, and C. Grabner
Electric Drive Technologies, Austrian Institute of Technology,
Giefinggasse 2, 1210 Vienna, Austria
e-mail: christian.kral@ait.ac.at; anton.haumer@ait.ac.at;
christian.grabner@ait.ac.at

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_8,

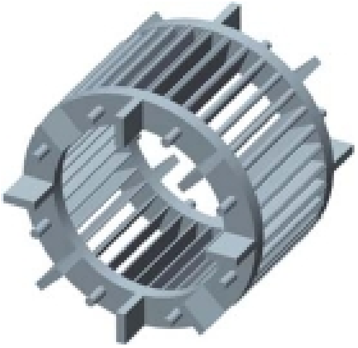


Fig. 8.1 Scheme of a rotor cage of an induction machine

The presented machine topology is modeled in *Modelica*, an acausal object oriented modeling language for multi physical systems [3]. In the *Modelica Standard Library* (MSL) a standard set of physical packages for electric, mechanic, thermal, control and logic components is collected. The *Machines* package provides models of electric machines based on text book equations. The modeled three phase induction machines rely on space phasor theory and a full symmetry of the stator and rotor winding. Electrical rotor asymmetries can therefore not be modeled using the MSL.

In order to model electrical rotor asymmetries the full topology of the rotor cage, respectively has to be taken into account. Such models are developed in the *ExtendedMachines* Library [5].

8.2 Model of Stator Winding

It is assumed in the following that the stator winding is fully symmetrical. Additionally, the number of stator phases is restricted to three. In this case the stator voltage equation can be written as

$$V_{s[i]} = R_s I_{s[i]} + L_{s\sigma}\frac{\mathrm{d}I_{s[i]}}{\mathrm{d}t} + \sum_{j=1}^{3} L_{sm[i,j]}\frac{\mathrm{d}I_{s[j]}}{\mathrm{d}t} + \sum_{j=1}^{N_r}\frac{\mathrm{d}L_{sr[i,j]}I_{r[j]}}{\mathrm{d}t}, \tag{8.1}$$

where $V_{s[i]}$ and $I_{s[i]}$ and $I_{r[i]}$ are the stator voltages and stator currents and the rotor currents, respectively. The stator resistance R_s and the stator stray inductance $L_{s\sigma}$ are symmetrical, due to the symmetry of the stator winding. The coupling of the stator windings is represented by the main field inductance matrix

$$L_{sm[i,j]} = L_0 w_s^2 \xi_s^2 \cos\left[\frac{(i-j)2\pi}{3}\right]. \tag{8.2}$$

The mutual coupling between the stator and rotor is represented by the matrix

$$L_{sr[i,j]} = L_0 w \xi_s \xi_r \cos\left[\frac{(i-1)2\pi}{3} - \frac{(j-1)2\pi}{N_r} - \gamma_m\right]. \tag{8.3}$$

Both these matrices are fully symmetrical, since it is assumed that the coupling over the magnetic main field is not influenced by the any electric asymmetry. The term L_0 indicates the base inductance of a coil without chording, i.e., the coil width equal to the pole pitch. The number of series connected turns and the winding factor of the stator winding are represented by the parameters w_s and ξ_s. The product $w_s \xi_s$ is thus the *effective number of turns.* The winding factor of the rotor winding

$$\xi_r = \sin\left(\frac{p\pi}{N_r}\right) \tag{8.4}$$

is a pure geometric factor, which is derived from the mesh width of two adjacent rotor bars. In this equation, however, skewing is not considered. The quantity γ_m represents the electric displacement angle of the rotor with respect to the stator.

The effective number of turns, $w_s \xi_s$, can be determined from a winding topology, which is indicated by the begin and end location and the number of turns of the stator winding coils. Alternatively, a symmetric stator winding may be parametrized by entering the effective number of turns.

8.3 Model of Rotor Winding

The winding topology of the squirrel cage rotor with N_r rotor bars can be seen as a winding with an effective number of turns equal to one. The matrix of the main rotor field can be expressed as

$$L_{rr[i,j]} = L_0 \xi_r^2 \cos\left[\frac{(i-j)2\pi}{N_r}\right]. \tag{8.5}$$

For the bars and the end rings on both sides (index a = drive end side, DE; index b = non drive end, NDE) constant leakage inductances $L_{b[i]}$ and $L_{ea[i]}$ and $L_{eb[i]}$ are considered. The rotor voltage equations can be derived from the equivalent circuit of a rotor mesh as shown in Fig. 8.2:

$$\begin{aligned} 0 = {} & (R_{ea[i]} + R_{eb[i]} + R_{b[i]} + R_{b[i+1]}) I_{r[i]} - R_{b[i]} I_{r[i-1]} - R_{b[i+1]} I_{r[i]} \\ & + R_{eb[i]} I_{eb} + (L_{ea[i]} + L_{eb[i]} + L_{b[i]} + L_{b[i+1]}) \frac{\mathrm{d}I_{r[i]}}{\mathrm{d}t} \\ & - \frac{\mathrm{d}}{\mathrm{d}t}(L_{b[i]} I_{r[i-1]} + L_{b[i+1]} I_{r[i]} - L_{eb[i]} I_{eb}) \\ & + \sum_{j=1}^{3} \frac{\mathrm{d}L_{sr[j,i]} I_{s[j]}}{\mathrm{d}t} + \sum L_{rr[i,j]} \frac{\mathrm{d}I_{r[j]}}{\mathrm{d}t} \end{aligned} \tag{8.6}$$

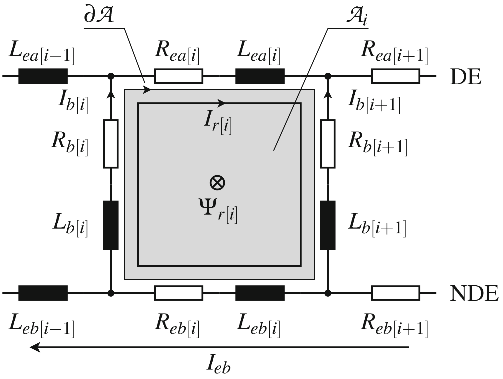


Fig. 8.2 Rotor cage topology (DE = drive end, NDE = non drive end)

In this equation $R_{b[i]}$ are the bar resistances, and $R_{ea[i]}$ and $R_{eb[i]}$ are the resistances of the end ring segments on both sides. Due to the topology of the rotor cage (Fig. 8.2), $N_r + 1$ linearly independent meshes have to be taken into account. The mesh current I_{eb} is thus introduced and the additional voltage equation

$$0 = \sum_{i=1}^{N_r} R_{eb[i]}(I_{r[i]} + I_{eb}) + \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{N_r} L_{eb[i]}(I_{r[i]} + I_{eb}) \tag{8.7}$$

has to be considered, accordingly.

As the air gap of the induction machine is modeled with smooth surface, the main field inductances $L_{ss[i,j]}$ and $L_{rr[i,j]}$ are constant and only the mutual inductances (8.3) are a function of the rotor angle γ_m.

The rotor cage can be parametrized in the *ExtendedMachines* library in two different ways. First, a symmetric rotor cage can be indicated by the rotor resistance R'_r and the rotor leakage inductance $L'_{r\sigma}$, equivalently transformed to the stator side. These are the typical parameters as they appear in the an equivalent circuit of the induction machine. The same parameters are also used for the Machines package of the MSL. Second, each resistance and leakage inductance of the rotor bars and the end ring segments on both sides can be parametrized. The relationship between the symmetric rotor bar and end ring resistance and the rotor resistance with respect to the stator side is determined by

$$R'_r = 2\frac{3w_s^2\xi_s^2}{N_r\xi_r^2}\left\{R_{e,\mathrm{sym}} + R_{b,\mathrm{sym}}\left[1-\cos\left(\frac{2\pi p}{N_r}\right)\right]\right\}, \tag{8.8}$$

where p is the number of pole pairs. A similar equation can be obtained for the rotor leakage inductance with respect to the stator side,

$$L'_{r\sigma} = 2\frac{3w_s^2\xi_s^2}{N_r\xi_r^2}\left\{L_{e\sigma,\mathrm{sym}} + L_{b\sigma,\mathrm{sym}}\left[1-\cos\left(\frac{2\pi p}{N_r}\right)\right]\right\}. \tag{8.9}$$

For the symmetric cage the ratios of the resistances, ρ_r, and leakage inductances, ρ_l, each with respect to the rotor bars over the end ring segments, can be specified,

$$\rho_r = \frac{R_{b,\mathrm{sym}}}{R_{e,\mathrm{sym}}}, \tag{8.10}$$

$$\rho_l = \frac{L_{b\sigma,\mathrm{sym}}}{L_{e\sigma,\mathrm{sym}}}. \tag{8.11}$$

This way, the symmetric cage resistance and leakage inductance parameters can be determined from R'_r, $L'_{r\sigma}$, ρ_r and ρ_l.

An electric rotor asymmetry can be modeled by increasing the resistance of a rotor bar or and end ring segment, respectively. Alternatively the equation structure could be adopted by removing a completely broken bar or end ring segment accordingly – this case is, however, not considered in this paper.

8.4 Torque Equation

The inner electromagnetic torque of the machine is determined by

$$T_{\mathrm{el}} = \sum_{i=1}^{3} \sum_{j=1}^{N_r} \frac{\mathrm{d}L_{sr[i,j]}}{\mathrm{d}\gamma_m} I_{s[i]} I_{r[j]}, \tag{8.12}$$

where $\frac{\mathrm{d}L_{sr[i,j]}}{\mathrm{d}\gamma_m}$ can be expressed analytically from (8.3). Even if friction, ventilation losses and stray load losses are not taken into account in the presented paper, they could be considered as a breaking torque in the angular momentum equation.

8.5 Theoretical Background of Rotor Faults

An electrical rotor asymmetry gives rise to a distortion of the rotor bar currents and the fundamental wave of the rotor magneto motive force (MMF). Therefore, the fundamental rotor MMF can be decomposed into a forward and backward traveling wave with respect to the rotor fixed reference frame. The forward travelling wave represents the main magnetic field and the backward traveling wave is due to the electrical rotor asymmetry. For infinite inertia drives the backward travelling wave induces a stator voltage harmonic component at the frequency

$$f_l = (1 - 2s) f_s. \tag{8.13}$$

In this equation f_s is the stator supply frequency and

$$s = \frac{f_s - pn}{f_s} \tag{8.14}$$

is slip, expressed in terms of rotor speed n and the number of pole pairs. The impedance of the machine (including supply) leads to a stator current harmonic with the same frequency as the induced stator voltage harmonic. Since the frequency of this harmonic component is less than the frequency of the fundamental wave this stator current harmonic component is called the *lower side band* harmonic. A finite inertia of the drive causes an additional upper side band harmonic current at frequency

$$f_u = (1 + 2s) f_s, \tag{8.15}$$

the *upper side band harmonic* [2]. Due to the interaction of these side band currents with flux and the inertia specific speed ripple, additional harmonics arise,

$$f_{l[k]} = (1 - 2ks) f_s, \tag{8.16}$$

$$f_{u[k]} = (1 + 2ks) f_s, \tag{8.17}$$

where k is an integer order number. Between no load and rated operating conditions slip varies between zero and some per cent.

The side band currents of stator currents are also reflected in the flux linkages of the main field (air gap) [9] and the stray flux [4]. The interaction of the fundamental and harmonic waves of the currents and flux linkages gives rise to double slip frequency oscillations

$$f_t = 2s f_s \tag{8.18}$$

of the electrical power and torque. The magnitudes of these fault specific oscillations are much smaller the average values of the electrical power and torque, respectively.

8.6 Investigated Machine

The investigated simulation and measurement results refer to a 18.5 kW, four pole induction machine with 40 rotor bars (Fig. 8.3). The experiments were performed for nominal load torque, nominal line-to-line voltage (400 V) and nominal frequency (50 Hz). For experimentally investigating electric asymmetries one rotor bar was fully broken by drilling a hole into the aluminum part of the rotor as shown in Fig. 8.3b.

For the investigated machine and design geometry the parameters of the rotor bar and end ring segments, $R_{b,\mathrm{sym}}$, $L_{b\sigma,\mathrm{sym}}$, $R_{e,\mathrm{sym}}$ and $L_{e\sigma,\mathrm{sym}}$ are estimated according to (8.8)–(8.11) for a given rotor resistance R_r' and a leakage inductance $L_{r\sigma}'$. In the

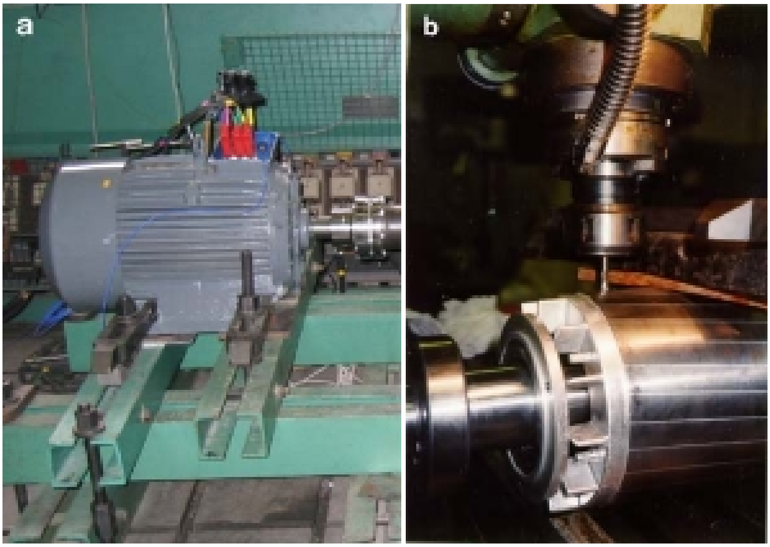


Fig. 8.3 (**a**) 18.5 kW four pole induction machine in the lab, (**b**) drilling a hole into the aluminum part of the squirrel cage rotor

simulation the broken bar was modeled by setting the faulty bar resistance with index 1 to

$$R_{b[1]} = 100R_{b,\mathrm{sym}}. \tag{8.19}$$

This is enough resistance increase for the respective bar current to sufficiently vanish.

8.7 Simulation Results

The distortion of the current distribution of the rotor bars and end ring segments due to an electrical rotor asymmetry is evident. The rotor bar currents can be computed from the rotor mesh currents according to Fig. 8.2,

$$I_{b[i]} = I_{r[i]} - I_{r[i-1]}. \tag{8.20}$$

Under nominal and steady state operating conditions the peak values of the sinusoidal currents of the rotor bars are depicted in Fig. 8.4. The current of the broken rotor bar (index 1) is almost zero. Additionally, an interesting phenomenon can be observed. The currents of the adjacently located rotor bars (e.g. index 40 and 2) are significantly larger than the currents of the remaining healthy rotor bars. Due to increased currents in the adjacent rotor bars the associated heat losses increase. If the adjacent bars fail as a result of the increased thermal stress, the extent of the

Fig. 8.4 Peak values of the rotor bar currents; broken rotor bar with index 1; simulation results

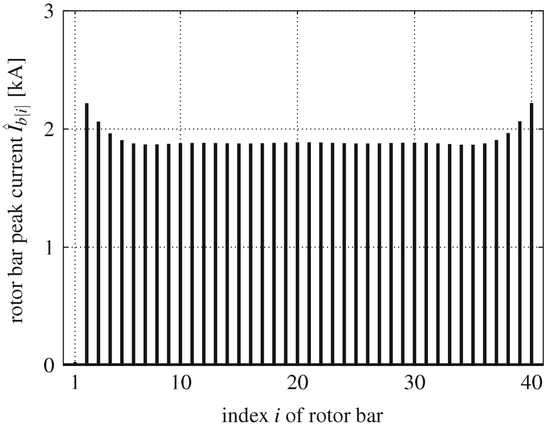
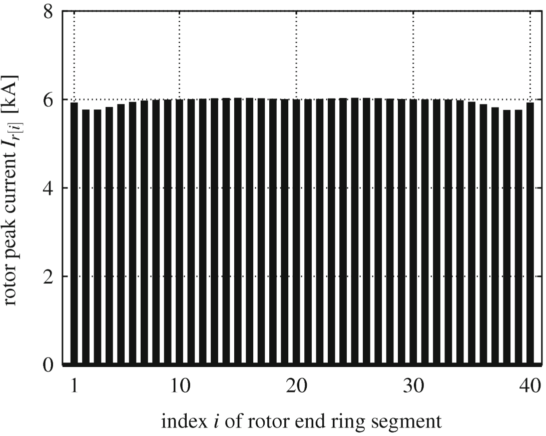


Fig. 8.5 Peak values of the currents of the end ring segments; broken rotor bar with index 1; simulation results



fault may spread in an avalanche-like way. Electrical rotor asymmetries spread relatively slow compared to other machine faults. It will thus take weeks, months or even years for a rotor failure to significantly increase.

Caused by the distortion of the current distribution in the rotor bars, the end ring current are distorted, too. Mathematically, the rotor end ring currents of the A- and B-side can be expressed by

$$I_{ea[i]} = I_{r[i]}, \tag{8.21}$$

$$I_{eb[i]} = I_{r[i]} + I_{eb}. \tag{8.22}$$

Without any asymmetry on either side of the end rings, the mesh current $I_{eb} = 0$ and thus the currents of the end ring segments of the A- and B-side are equal. The peak values of the currents of the rotor end ring segments are depicted in Fig. 8.5 for the case of one broken rotor bar.

The rotor asymmetry specific lower and upper side band harmonics of the current arise close to the fundamental wave according to (8.13) and (8.15). For the investigated 50 Hz machine the Fourier spectrum of a stator current (phase 1) is depicted in Fig. 8.6. The lower and upper side band harmonics clearly appear at 48.6 Hz and 51.4 Hz. Since the magnitudes of these side band components are much smaller than the magnitude of the fundamental, electrical rotor asymmetries are difficult to detect.

8.8 Measurement Results

In Fig. 8.7 the measured Fourier spectrum of a stator phase current is depicted. This plot reveals compared to the simulation results of Fig. 8.6 the same frequencies of the side band harmonics, but slightly deviating magnitudes. The differences of the magnitudes are mainly due to the inertia of the drive which is not exactly estimated [7]. Some additional deviation of the modeled rotor bar and end ring resistance ratio

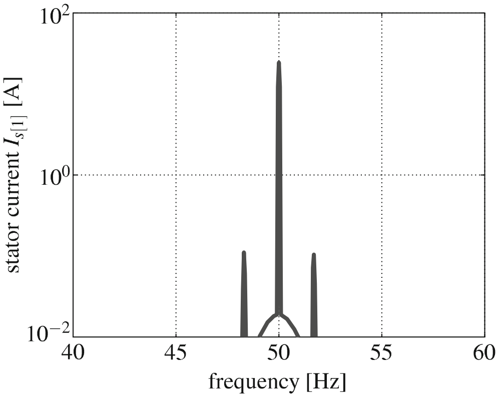


Fig. 8.6 Fourier spectrum of the stator current $I_{s[1]}$; simulation results

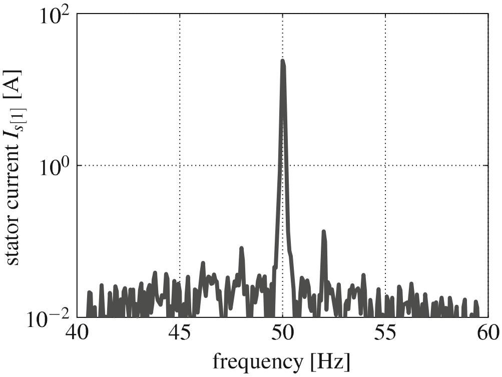


Fig. 8.7 Fourier spectrum of the stator current $I_{s[1]}$; measurement results

from the real machine cage, also have an impact on the magnitudes of the side band currents. With respect to the comparison of measurement and simulation results it should also be noted that in a real motor with die-cast rotor interbar currents arise [10].

8.9 Rotor Fault Detection Methods

Since the magnitudes of the fault specific current harmonics are much smaller than the fundamental wave current, only severe rotor asymmetries can be detected through visual observations of fluctuations of the current or power pointer instruments. For the detection of upcoming electric rotor asymmetries require some more sophisticated detection methods have to be used. The most common rotor fault detection methods are based on the measurement of one stator current – this class of methods is called *current signature analysis* (CSA) methods [1]. Then a fast Fourier transform or a wavelet transform or some other signal processing techniques are performed in order to determine the fault specific current harmonic side bands.

Another class of methods uses *power signature analysis* (PSA), evaluating either total or phase power [8]. For the assessment of the fault severity – in combination with either CSA or PSA techniques – neural networks or Fuzzy based methods may be applied.

A third class of fault detection methods uses model based techniques for the detection of rotor faults. One model based technique is the Vienna Monitoring Method (VMM) which was introduced in 1997 [11]. The VMM compares the calculated torques of a voltage and a current model to derive a fault indicator. If the models are well tuned, both models compute the same torque in case of a fully symmetrical rotor cage. An electrical rotor asymmetry excites double slip frequency torque oscillations in both models – with different magnitude and phase shift, however. The magnitude of the derived torque difference is directly proportional to the average load torque. This relationship can be used to calculate a robust fault indicator. For this purpose the torque difference is divided by the estimated load torque which leads a new quantity: the relative torque difference. A spacial data clustering technique, applied to the relative torque difference, then eliminates any harmonic components which is not an integer multiple of slip frequency. The magnitude of second harmonic of the relative torque difference can the be determined by applying a discrete Fourier analysis of the clustered data values, and serves as fault indicator for the VMM. The particular advantage this technique is that it provides a reliable fault indicator independent of load torque, speed, supply and inertia of the drive [6, 7].

The VMM is applied to both the simulation and measurement results. The fault indicator determined from the simulation and measurement results is 0.0093 and 0.0105, respectively. The deviation of these two quantities is about 11%.

8.10 Conclusions

For induction machines with squirrel cage rotors the background of electrical rotor asymmetries is introduced and discussed. A rotor topology model as it is implemented in the *ExtendedMachines*, is presented.

For a 18.5 kW induction machine with one broken rotor bar – out of 40 bars – simulated and measured results are compared and shown good coherence. The results refer to the Fourier spectra of the stator currents and a model based rotor fault detection method – the Vienna Monitoring Method. The comparison of the fault indicators determined by the Vienna Monitoring Method for the simulation and measurement case, reveals a deviation of about 11%. Considering that no parameter tuning of the simulation model has been performed, this is a satisfactory matching result.

References

1. Didier, G., Ternisien, E., Caspary, O., Razik, H.: Fault detection of broken rotor bars in induction motor using a global fault index. IEEE Trans. Ind. Appl. **42**(1), 79–88 (2006)
2. Filippetti, F., Franceschini, G., Tassoni, C., Vas, P.: Impact of speed riple on rotor fault diagnosis of induction machines. Proceedings of the International Conference on Electrical Machines, ICEM, pp. 452–457, 1996
3. Fritzson, P.: Principles of Object-Oriented Modeling and Simulation with Modelica 2.1. IEEE Press, Piscataway, NJ, 2004
4. Henao, H., Demian, C., Capolino, G.-A.: A frequency-domain detection of stator winding faults in induction machines using an external flux sensor. IEEE Trans. Ind. Appl. **39**, 1272–1279 (2003)
5. Kral, C., Haumer, A., Pirker, F.: A modelica library for the simulation of electrical asymmetries in multiphase machines – the extended machines library. IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives, The 6th, SDEMPED 2007, Cracow, Poland, pp. 255–260, 2007
6. Kral, C., Pirker, F., Pascoli, G.: Model-based detection of rotor faults without rotor position sensor – the sensorless vienna monitoring method. IEEE Trans. Ind. Appl. **41**(3), 784–789 (2005)
7. Kral, C., Pirker, F., Pascoli, G.: The impact of inertia on rotor fault effects – theoretical aspects of the vienna monitoring method. IEEE Trans. Power Electron. **23**(4), 2136–2142 (2008)
8. Liu, Z., Yin, X., Zhang, Z., Chen, D., Chen, W.: Online rotor mixed fault diagnosis way based on spectrum analysis of instantaneous power in squirrel cage induction motors. IEEE Trans. Energy Conver. **19**(3), 485–490 (2004)
9. Schagginger, M.: Luftspaltfelduntersuchungen an umrichtergespeisten Asynchronmaschinen im Hinblick auf elektrische Unsymmetrien im Rotorkäfig. Master's thesis, Technische Universität Wien, Vienna, 1997
10. Walliser, R.: The influence of interbar currents on the detection of broken rotor bars. ICEM, pp. 1246–1250, 1992
11. Wieser, R., Kral, C., Pirker, F., Schagginger, M.: On-line rotor cage monitoring of inverter fed induction machines, experimental results. Conference Proceedings of the First International IEEE Symposium on Diagnostics of Electrical Machines, Power Electronics and Drives, SDEMPED, pp. 15–22, 1997

Chapter 9
Different Designs of Large Chipper Drives

Hansjörg Kapeller, Anton Haumer, Christian Kral, and Christian Grabner

Abstract This paper presents two simulation models for two variants of a large chipper drive used in a paper mill. If a slip ring induction motor is used, the impact of a rotor circuit rheostat with respect to starting behavior and heavy duty load impulses can be examined. Furthermore a speed controlled squirrel cage induction machine will be investigated. The modeling language for both drives is Modelica. The simulation results for both drives are compared and discussed.

Keywords Induction motor · squirrel cage · slip ring · speed control · load impulses

9.1 Introduction

Paper mills use chippers for crushing trunks and making wood chips. Drives used for such applications are rated from several 100 kW up to 2 MW. Chipper drives are usually not operated continuously, because load impulse-arise only if a trunk is shredded. After that, a period of no-load operation follows until the next trunk arrives. The heavy duty load impulses which can even exceed double the nominal torque give rise to large motor currents which in turn cause large voltage drops at the mains terminals. Certain voltage drops may not be exceeded during impulse load or starting operation according to the regulations, depending on the actual configuration. If the chipper drive has been set to stand still, leaving some remaining parts of a trunk, re-starting is a critical condition for the whole drive. In order to take appropriate measures to avoid these mains reactions, two drive configurations are investigated. First, an induction machine with slip ring rotor (see Fig. 9.1a) and rheostat in the rotor circuit is presented. Second, a speed controlled induction machine with squirrel cage rotor (see Fig. 9.1b) is compared.

H. Kapeller (✉), A. Haumer, C. Kral, and C. Grabner
AIT Austrian Institute of Technology, Mobility Department,
Electric Drive Technologies Giefinggasse 2, 1210 Vienna, Austria
e-mail: hansjoerg.kapeller@ait.ac.at; anton.haumer@ait.ac.at;
christian.kral@ait.ac.at; christian.grabner@ait.ac.at

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_9,

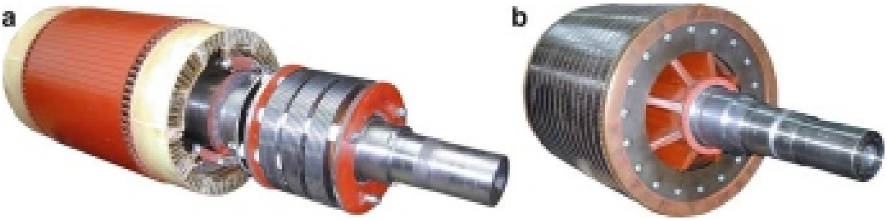


Fig. 9.1 (**a**) Slip ring rotor and (**b**) squirrel cage rotor

9.1.1 Slip Ring Motor with Rotor Rheostat

A motor not being supplied by an inverter causes high starting currents due to the low locked rotor impedance of the induction motor [1]. For a slip ring motor (see Fig. 9.1a), simple measures can be taken to reduce starting currents and to improve the torque speed characteristic. The improved torque speed characteristic (see Fig. 9.4a) diminishes the reactions of the load impulses on the motor currents and voltage sags, since part of the energy of a load impulse is absorbed by kinetic energy. After the load impulse, speed is increased and the resulting load current is homogenized.

The disadvantages of a slip ring motor, especially with additional resistances in the rotor circuit, is the deterioration of efficiency due to additional losses in the external rotor resistances and the high abrasion of the brushes, which causes increased deposit of brush dust in the motor. This brush dust subsequently increases the risk of isolation breakdown and causes higher costs of maintenance.

Instead of a slip ring motor with additional rotor resistances, a speed controlled inverter drive with squirrel cage motor can be used (Fig. 9.2).

9.1.2 Speed Controlled Squirrel Cage Motor with Inverter

A speed controlled drive consists of the electric motor, the power converter, cascaded current and speed controller, the power supply, the mechanical load, the current and speed sensors respectively [2]. The speed controller determines the reference stator current of the machine according to the deviation of the actual speed from the reference speed. The current controller observes limitations of the stator current to avoid overloading the drive. Both, controlling the speed as well as limiting the currents, lead to efficiency savings over a wide operating range and indicates an advantage compared to the slip ring motor. Investment costs are higher due to the additional inverter, but maintenance costs are lower than for the slip ring motor drive, however.

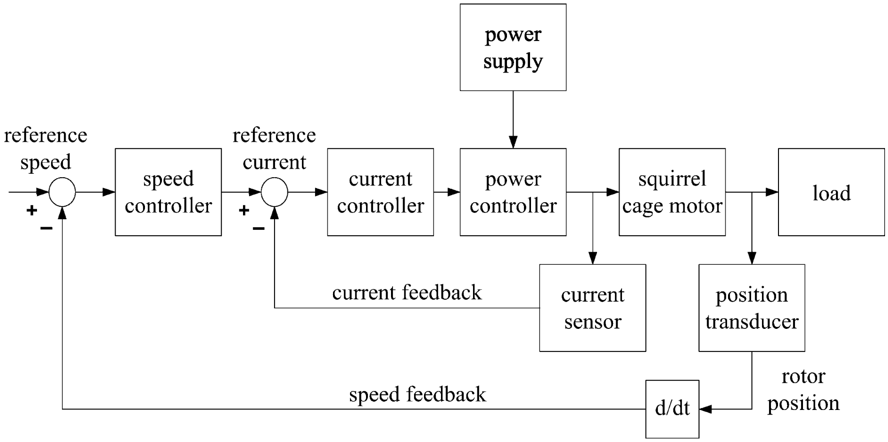


Fig. 9.2 Layout of a speed controlled drive

Table 9.1 Grid data

Mains supply	
Frequency	50 Hz
RMS voltage, line-to-line	6,000 V
Short-circuit apparent power	50 MVA
Short-circuit power factor	0.05
Transformer's nominal apparent power	1.8 MVA
Transformer's short-circuit p.u. voltage	0.06
Transformer's copper losses	17.5 kW

9.1.3 Technical Data

The technical parameters of the power supply and the slip ring motor as well as the squirrel cage induction machine are summarized in Tables 9.1–9.3.

9.2 Simulation Models

For performing simulations the software tool Dymola is used. Dymola is based on the modeling language Modelica [3]. The Modelica Association not only develops the language specification but also provides the comprehensive *Modelica Standard Library* (MSL). Except for the inverter and its control all models can be taken from this library. Based on the *Machines* library [4] the *SmartElectricDrives* (SED) library [5] provides models of electric drives using different control structures and strategies.

Table 9.2 Parameters of the investigated chipper drive with slip ring motor

Chipper drive with slip ring motor	
Frequency	50 Hz
Number of pole pairs	2
RMS stator voltage, line-to-line	6,000 V
RMS stator current	161.1 A
RMS rotor voltage, line-to-line	1,500 V
RMS rotor current	595.3 A
Warm stator resistance per phase	129.0E-3 Ω
Stray stator inductance per phase	6.845E-3 H
Main inductance per phase	273.8E-3 H
Stray rotor inductance per phase	0.4631E-3 H
Warm rotor resistance per phase	8.729E-3 Ω
Motor rated power	1.5 MW
Motor rated rpm	1,490.8 min^{-1}
Motor inertia	120 kg m^2
Load inertia	20,000 kg m^2
Gear unit	1,500 : 300 min^{-1}

Table 9.3 Parameters of the investigated chipper drive with low voltage inverter supplied squirrel cage motor

Chipper drive with squirrel cage motor	
Frequency	50 Hz
Number of pole pairs	2
RMS stator voltage, line-to-line	690 V
RMS stator current	1,404.5 A
Warm stator resistance per phase	1.702E-3 Ω
Stray stator inductance per phase	0.10835E-3 H
Main inductance per phase	4.063E-3 H
Stray rotor inductance per phase	0.10835E-3 H
Warm rotor resistance per phase	1.135E-3 Ω
Motor rated power	1.5 MW
Motor rated rpm	1,493.9 min^{-1}
Motor inertia	80 kg m^2
Load inertia	20,000 kg m^2
Gear unit	1,500 : 300 min^{-1}

The SED library contains models of the components used in recent electric drives: sources (batteries and a PEM fuel cell), converters (ideal and power balanced), electric loads, process controllers, sensors, etc. Two classes of drive models are provided:

For fast simulations focused on energy consumption or the efficiency of a drive configuration, the models of the *QuasiStationaryDrives* can be used. These models neglect all electrical transients in the machines, i.e. they calculate quasi-stationary points of operation, but mechanical transients are considered, however. This enables a remarkably shorter simulation time due to the simpler controller configuration and the neglect of switching effects. For the analysis of electric transients the *TransientDrives* have to be used. By choosing this higher level of detail the user has to provide more parameters, besides the fact that these models consume more computing time. Additionally to all elementary components that give the user the freedom to design an entire controlled drive, 'ready to use' models are provided. These models can be used to conveniently and quickly arrange simulations [5]. The '*ready to use*' models contain the machine, the converter, measurement devices and a *field oriented control* (FOC).

9.2.1 Slip Ring Motor with Additional Rotor Resistances

An external rotor resistance can be used to increase the impedance of a slip ring motor (Fig. 9.3). This measure allows to reduce starting currents and improves the torque speed characteristic (Fig. 9.4a).

The total rotor circuit resistance R_r^* consists of the internal rotor winding resistance R_r and the external resistance, which in turn is built up from a variable rheostat R_v and an external constant resistance R_c:

$$R_r^* = R_r + R_v + R_c \tag{9.1}$$

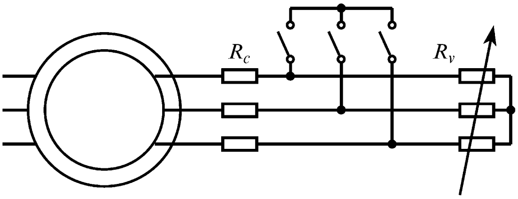


Fig. 9.3 Three phase rheostat of a slip ring induction motor

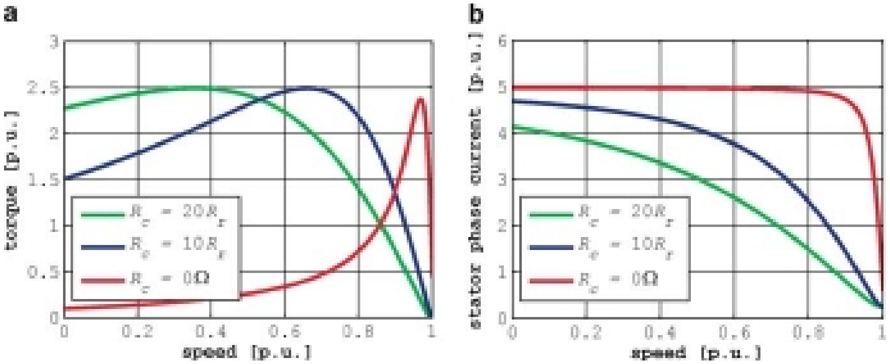


Fig. 9.4 (**a**) Stationary torque versus speed characteristic of a slip ring motor with external rotor resistance and (**b**) stationary current versus speed characteristic of a slip ring motor with external rotor resistance

Starting from stand still, the resistance of the rheostat R_v is reduced along a linear time dependent ramp. The duration of the ramp has to be chosen according to the drive configuration to meet the actual starting time. After reaching nominal speed, the rheostat R_v is shortened.

If the motor is not loaded, the influence of the constant resistance R_c on the motor current and speed can be neglected. Applying a constant load torque the stationary speed depends on the actual resistance R_c according to the following equation

$$\frac{R_r}{s} = \frac{R_r + R_c}{s_c} \tag{9.2}$$

where s denotes the slip for the case without external rotor circuit resistance, and s_c is the slip for the case with the external rotor circuit resistance R_c [6]. The impact of the external rotor resistance R_c on the torque-slip characteristic is shown in Fig. 9.4a: The same torque is delivered at higher slip, i.e. lower speed. Therefore, load impulses can be absorbed temporarily by the stored energy of all rotating masses. However, the speed dip is compensated after the load pulse.

The characteristic of stator current versus speed is shown in Fig. 9.4b. For a shortened slip ring rotor ($R_r^* = R_r$) the torque speed characteristic shows a very low starting torque and a starting current of approximately five times the nominal current. For higher external rotor resistances, e.g. $R_r^* = 21R_r$, the stationary characteristics show a much higher starting torque – close to the breakdown torque – and the locked rotor current is less than 4.5 times the nominal current, decreasing rapidly with increasing speed.

Figure 9.5 depicts the Modelica model of the chipper drive realized with a slip ring motor. The 6 kV/50 Hz voltage supply is modeled as three sinusoidal supply voltages (`sineVoltage`) which are star connected. The overall mains impedance including all transmission lines and transformers, is taken into account by a series connection of a three phase resistor (`resistorG`) and a three phase inductor (`inductorG`). In order to be able to show *root mean square* (RMS) values of the voltages and currents in the simulation results, an RMS voltmeter and amperemeter are connected. Additionally, an electrical power sensor measures the active and reactive power consumption of the drive. The stator winding of the slip ring induction motor (`AIM`) is star connected, connecting the stator terminals to the mains impedances. The rotor windings of the slip ring rotor are star connected, too. The terminals of the slip rings are series connected to the external constant resistor (`Rc`) and the variable rheostat (`Rv`). For simplifying the simulation model depicted in Fig. 9.5, the variable resistor is controlled by a ramp during the start-up of the motor. The signal inputs of the variable resistor model, however, could be controlled by any other strategy as well, e.g. dependent on current or speed.

A torque and a speed sensor provide signals for the simulation result. The mechanical power of the induction motor is transmitted through a gear (`idealGear`) to the load inertia and a simplified load torque model (`loadTorque`). The signal input of the load torque model is commanded by a time dependent table (`loadTable`) where the cycle of load impulses is stored.

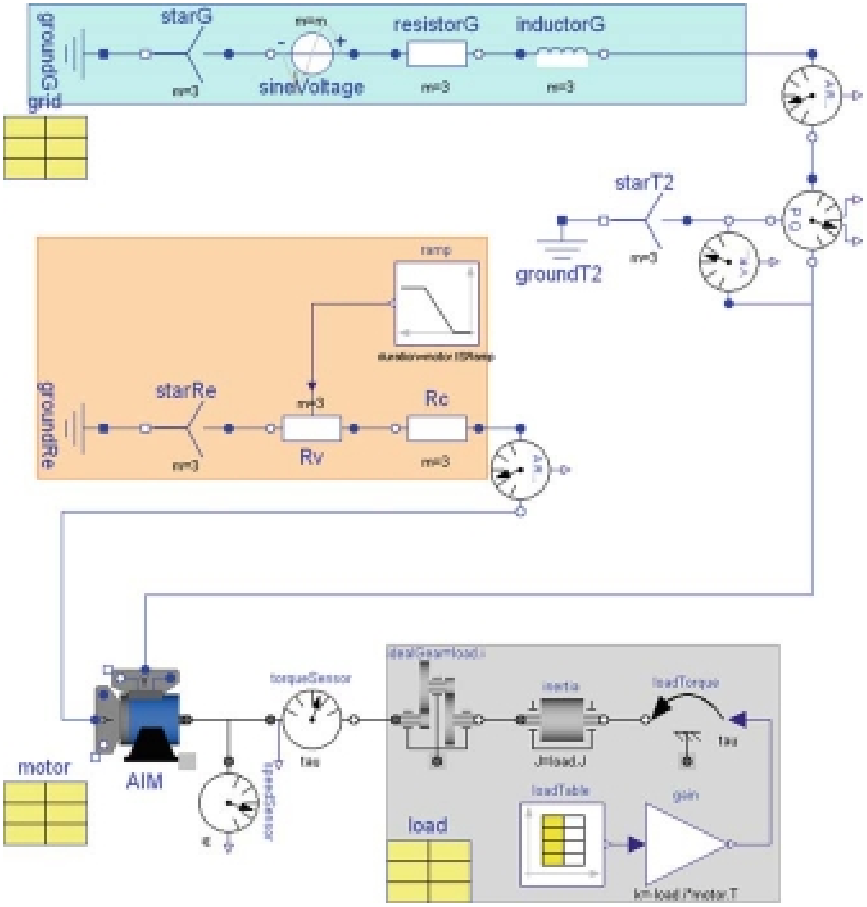


Fig. 9.5 Modelica model of the chipper drive with slip ring motor

9.2.2 *Speed Controller Squirrel Cage Motor with Inverter*

Figure 9.6a shows a squirrel cage induction motor with a speed controller, using components from the SED library. The voltage supply and electric measurement is the same as in the slip ring motor model, except that a transformer is used to provide 690 V from the 6 kV grid to the low voltage drive. The transformed supply voltage (`transformer`) is rectified (`diodeBridge`) and provides the intermediate circuit voltage for the inverter. The rectifier model does not take into consideration switching effects, i.e. the typical non-sinusoidal waveform of a diode bridge. Therefore the supply current is rather comparable to that of an IGBT rectifier. The DC/AC-inverter which is implemented in the field oriented controlled *QuasyStationaryDrive* model (`AIMfoc`) feeds the squirrel cage motor model. The mechanical load model is the same, however.

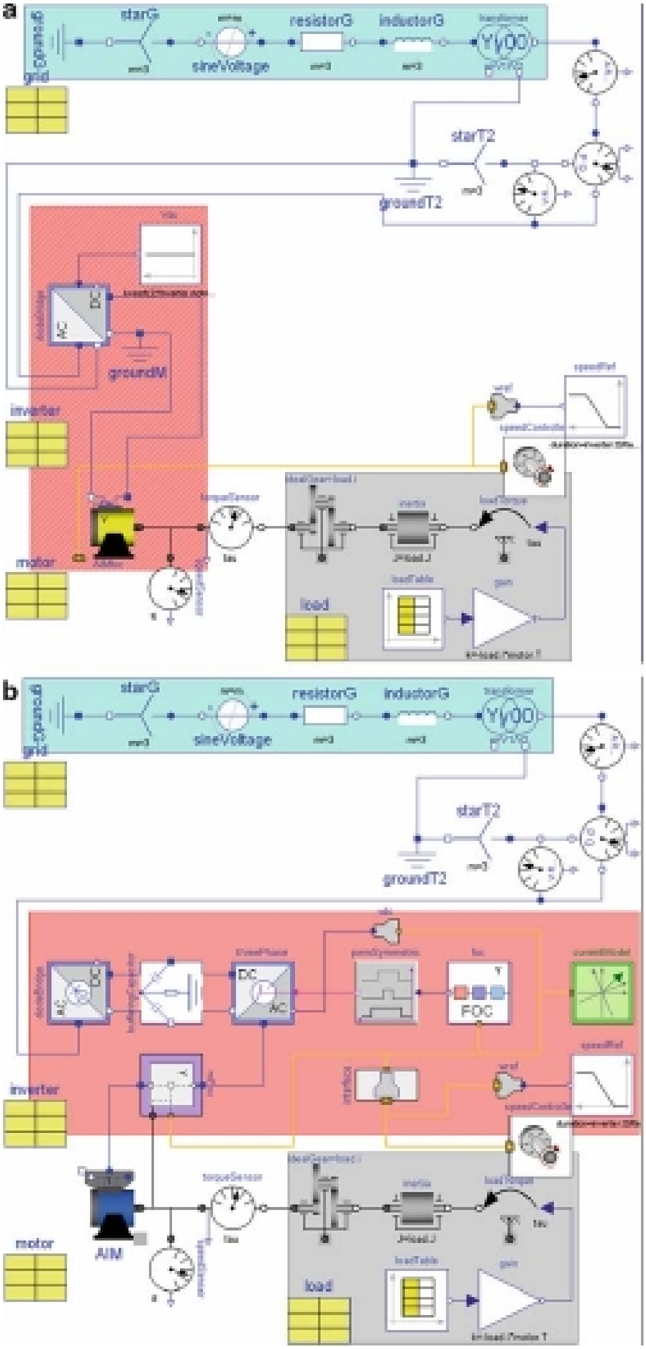


Fig. 9.6 (**a**) Modelica model of a speed controlled chipper drive with a squirrel cage motor and (**b**) full transient Modelica model of a speed controlled chipper drive with a squirrel cage motor

The cascade control system shown in Fig. 9.2 can be parameterized separately [7] for the speed and the current controller. Starting from the current control loop to the speed control loop, various parameterization methods can be applied to achieve the desired dynamic behavior [8].

Figure 9.6b shows the full transient Modelica model of a speed controlled chipper drive with a squirrel cage induction motor. Again, this drive uses components from the SED library. The voltage supply and electric measurement is the same as before, but an ideal switching diode rectifier is utilized. Additionally, the machine inverter is modeled in detail, not being integrated in the drive model.

The model of the ideal switching inverter leads to a high number of switching events during simulation, and therefore causes a significantly longer simulation time. The mechanical time constants are significantly greater than the electrical time constants. If current peaks due to inverter switching are not of interest for this investigation, the *QuasiStationaryDrive* model can be used to save a significant amount of simulation time.

9.3 Simulation Results

In order to have a fair basis for comparison, each load impulse has the same duration (2 s), equal rise and fall times (0.1 s) and the torque amplitude is twice the nominal torque. The first load impulse starts after complete start up of the drive at $t = 30$ s, the subsequent load impulses are applied in a 10 s cycle, i.e. the second load impulse starts at $t = 40$ s (Figs. 9.7a and 9.9a).

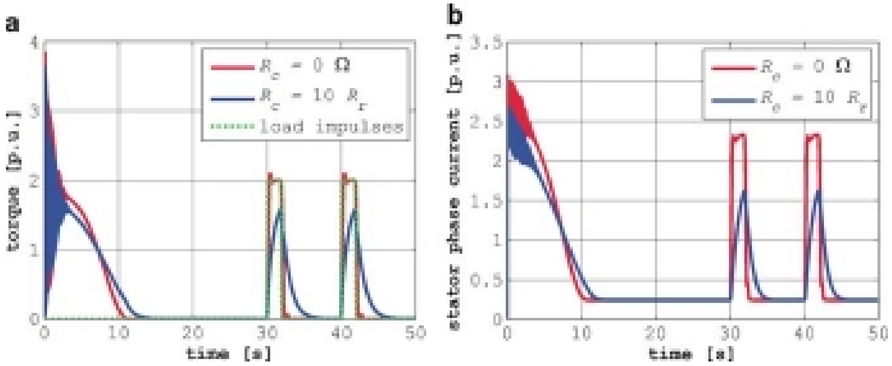


Fig. 9.7 (**a**) Load response of the slip ring motor and wave form of the modeled load impulses and (**b**) stator phase current during start-up and during periodic loading of the chipper drive with slip ring motor

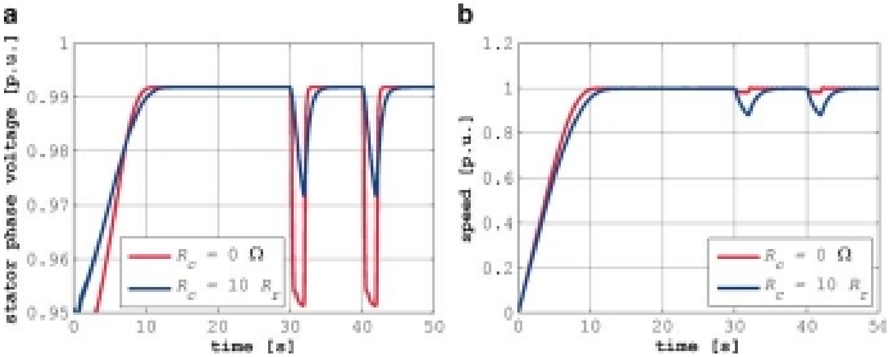


Fig. 9.8 (**a**) Stator phase voltage during start-up and during periodic loading of the chipper drive with slip ring motor and (**b**) speed during start-up and during periodic loading of the chipper drive with slip ring motor

9.3.1 Slip Ring Motor

Simulation results for the chipper drive with slip ring motor are shown in Figs. 9.7 and 9.8. $R_c = 0\,\Omega$ and $R_c = 10R_r$ were set for two consecutive simulations. The duration of the linear ramp for decreasing the variable resistor ($R_{v,max} = 40R_r$) was set to 10 s. Figure 9.7a and b show that the torque and current get reduced due to the additional resistances in the rotor circuit. From Fig. 9.8a it is evident, that furthermore the maximum voltage sags at the motor terminals get reduced. Figure 9.8b shows that a higher external rotor resistance leads to larger speed drop during the load impulses, and it takes some time after the load pulse until the motor has re-accelerated the whole drive to full speed again.

9.3.2 Squirrel Cage Motor

Simulation results for the chipper drive with squirrel cage motor are shown in Figs. 9.9 and 9.10. The current controller limits overloading of the machine, which can be seen in Fig. 9.9a and b. Additionally, during the start up procedure the current limitation of the speed controlled drive is more effective than of the slip ring drive. The voltage drops during load impulses are less than 2% (cp. Fig. 9.10a). Figure 9.10b illustrates, that the speed drop during the load impulses shows a similar dynamic characteristic as the chipper drive with slip ring motor.

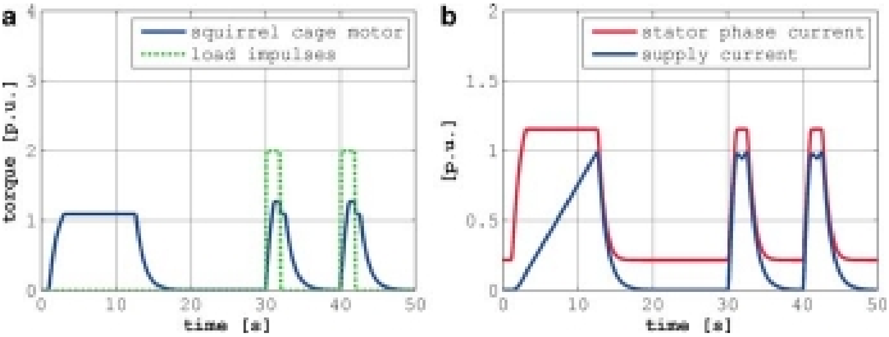


Fig. 9.9 (**a**) Load response of the speed controlled squirrel cage motor and wave form of the modeled load impulses and (**b**) stator phase current during start-up and during periodic loading of the speed controlled squirrel cage motor

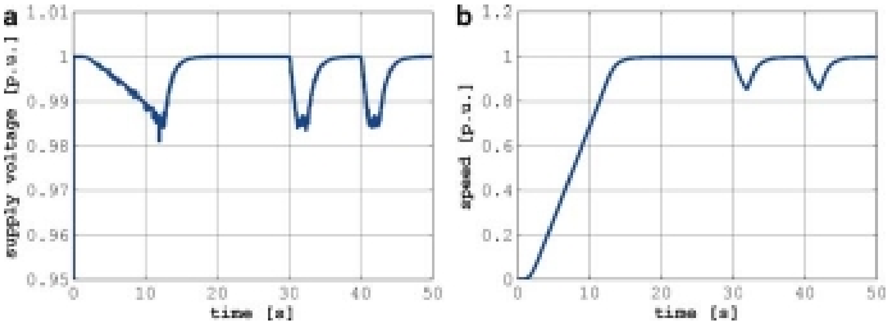


Fig. 9.10 (**a**) Stator phase voltage during start-up and during periodic loading of the speed controlled squirrel cage motor and (**b**) speed during start-up and during periodic loading of the chipper drive with speed controlled squirrel cage motor

9.4 Conclusions

Two different designs of a large chipper drive were modeled in Modelica. The first drive consists of a slip ring motor with external rotor resistances, the second drive is a speed controlled squirrel cage motor fed by an inverter. Both drives provide a reduction of current peaks and voltage dips during the start-up and the load pulses.

Using a slip ring induction motor, the investment costs for the motor are higher than for an induction motor with squirrel cage, but the rather simple additional equipment leads to low total investment costs. Disadvantages of the slip ring motor drive are the deterioration of efficiency due to additional losses in the external rotor resistance and the abrasion of the brushes, which gives rise to an increased deposit of brush dust in the motor.

Contrarily, the inverter drive has lower costs of maintenance but higher investment costs. Although the squirrel cage induction motor is cheaper, more expensive additional equipment is needed: a transformer as well as a frequency inverter and the entire speed control unit with all necessary sensors. However, the efficiency savings over a wide operating range and the absence of brushes are the advantages compared to the slip ring motor solution.

References

1. Alger, P.L.: Induction Machines. Gordon and Breach Science Verlag, New York (1970)
2. Nasar, S., Boldea, I.: Electric Machines; Dynamics and Control 2000, 1st edn. CRC Press, Boca Raton, FL (1993)
3. Fritzson, P.: Principles of Object-Oriented Modeling and Simulation with Modelica 2.1. IEEE Press, Piscataway, NJ (2004)
4. Kral, C., Haumer, A.: Modelica libraries for Dc machines, three phase and polyphase machines. Modelica Conference, pp. 549–558 (2005)
5. Giuliani, H., Kral, C., Gragger, J.V., Pirker, F.: Modelica Simulation of Electric Drives for Vehicular Applications. The Smart Electric Drives Library. ASIM (2005)
6. Fischer, R.: Elektrische Maschinen, 5th edn. C. Hanser Verlag, München (1983)
7. Föllinger, O.: Regelungstechnik, 8th edn. Hüthig Verlag, Heidelberg (1994)
8. Lutz, H., Wendt, W.: Taschenbuch der Regelungstechnik, 5th edn. Wissenschaftlicher Verlag Harri Deutsch, Frankfurt am Main (2003)

Chapter 10
Macro Cell Placement: Based on a Force Directed Flow

Meththa Samaranayake and Helen Ji

Abstract Macro cells are used more and more in current designs as they provide the benefit of reusability directly resulting in the decrease in design time and cost. However, there lies a gap in the EDA industry for macro-cell placement tools. This chapter looks at the implementation of a force-directed macro-cell placement tool that is developed to target the gap in industry.

Keywords Macro-cell · placement · force directed · EDA · graph drawing

10.1 Introduction

The past few years have seen an exponential rise in the growth rate of the semiconductor industry. The increase in usage and demand of electronic devices among consumers has resulted in the need to provide better and faster design methods. The designers are pushed to their limits in meeting these demands whilst juggling the constraints of power and performance of ever shrinking circuits. To help designers meet their targets, EDA (Electronic Design Automation) tools are used to help fully or partially automate the design processes. One of such important backend processes is the placement component.

The placement problem simply is the problem of finding the ideal locations for each cell in a circuit achieving as many or all of the placement objectives. The two main objectives that every placement tool has to achieve for today's fixed die design are,

- Overlap free layout
- Fit in the given placement area

M. Samaranayake (✉) and H. Ji
Department of Engineering and Technology, Manchester Metropolitan University,
Chester Street, Manchester M1 5GD UK
e-mail: meththa.t.samaranayake@stu.mmu.ac.uk; h.ji@mmu.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_10,

Other objectives may include minimization of wirelength, congestion, area, run time etc. The optimal solution will be one that satisfies all of the given criteria. Achieving such a placement solution is far from possible and even the simplest of cell placement problems are defined to be NP-hard. The consequence of falling short of a good placement could result in an unroutable design or a slower and/or larger chip. This will cost time and money to either correct the placement at a later stage or re-do the design.

In the past, designs mainly carried standard cells but as the complexity and components increased, Macros were utilized to reduce the complexity of circuits. Macros can mainly be seen as black boxes that are designed to carry out specific tasks such as implementation of a logic function (e.g. an IP block). It can also be on-chip memories that are common in SoC (System on Chip) designs. Increased use of Macro cells help designer's reuse of their designs which in turn helps reduce design time and cost.

Previous work [1, 2] has looked at the possibility of using graph-drawing algorithms as the basis of a Macro-cell placement tool. In this work, the authors have extended the capabilities of the algorithms in order to achieve better placements of cells. The two algorithms used in the work are those authored by Kamada and Kawai [3] and by Fruchterman and Reingold [4] (referred to as KK and FR respectively). They were chosen for their ability to handle undirected graphs, their simplicity in implementation, their speed as well as the criteria they follow to produce aesthetically pleasing graphs. In many cases, these criteria are shared by good placements.

10.2 Placement Tools

There are many standard cell placement tools currently available both academically and commercially within the EDA community. Several of them are capable of mixed-mode cell placement i.e. designs that contain both standard cells and modules, but there are only a few placement tools specifically for macro cells. This is because standard cells used to govern most of the circuit designs up till recent times. Recent changes have seen designs to contain Macro-based designs such as memory blocks and IP blocks (Intellectual Property) and furthermore, the hierarchical design methodology intended to tackle design complexity has resulted in macro-dominated designs at the top level. Even though mixed mode placement tools can handle macros, for designs that contain a majority of macros these tools may not place the cells in the best interest of the macros giving the need for a dedicated macro cell placer.

Some leading edge mixed-mode placement tools identified are Capo [5], Dragon [6], FastPlace [7] and APlace [8]. The Capo tool is based on a combination of simulated annealing and recursive bisection techniques and handles macro cells mainly by the help of the floorplanner Parquet [9]. Capo placement tool has a secondary method of placing macro cells where it shreds them to smaller sub-cells. These sub-cells are connected by two pin nets ensuring that they are placed close to one another. The design is then solved as a standard cell placement problem. FastPlace and APlace tools are based on analytical techniques and incorporates macro-cell placement within its normal placement flow. In FastPlace, the macro-cells are given priority during legalization stage where overlaps are resolved between macros before standard cells. Dragon is a hybrid placement tool that combines the use of simulated annealing with min-cut partitioning. To accommodate macro cells, it modifies the min-cut algorithm so that the partitions can be of different sizes. All these placement tools were designed for standard cells and as a result, with the exception of the Capo placer, these tools do not consider factors such as macro pin locations, cell orientation, soft macro cells and fixed cell support etc. Due to the capabilities of the Parquet floorplanner, Capo can handle all factors affecting macro cells aside from pin handling.

A Java based macro-cell placer [10] based on a force directed placement algorithm has been identified to be different from traditional force directed algorithms. In this work, the cell shapes and sizes have been considered when developing the force equation. A limitation of this tool is that it does not handle placement on a fixed placement area and instead treats the chip as a soft cell with a variable size and aspect ratio. The pads of the chip are also not fixed; therefore, the positions are found with the use of the force directed algorithm at a later stage.

A macro-cell placement method based on net clustering and force directed method is proposed in literature [11]. This approach is unique such that, it treats the nets as the placement components. In the graphs they draw, the nodes represent the nets whilst an edge only exists for the nets that share one or more cells. The forces on the nets determine the initial locations for the cells. Pin locations are determined last, suggesting that this placement tool is mainly focused on soft cell macros. This work reiterates the importance of the pin locations and cell orientation in macro cell placement. Another limitation seen is that the tool only handles connected graphs, again limiting the type of designs that can be processed.

Looking at both macro-cell placers identified above, a common disadvantage recognized is that both tools are not standardized – inputs are not of industry recognized LEF/DEF format [12] but formats limited to the tool. This has limited the tools from reading in standard designs currently available, therefore disabling measuring their quality of placement. The same is true for outputs where they are not given out in any standard format so that the placement can be processed by a routing tool.

10.3 Standard Cells Versus Macro-cells

Macro-cell placement is not as straightforward as standard cell placement. In standard cell placement, the cells are of uniform height and are restricted to rows in which they must sit in. These restrictions allow the placement tools to be more precise in choosing locations for the standard cells and to allocate routing resources. Macro cells on the other hand do not have such restrictions. They can be of any height, width and shape (L, T and U shapes though the most common is rectangular) and are not restricted to a specific location of the placement area. As a result, choosing a good placement for macro-cells can be much harder as the permutations of locations they can be placed-at are unbounded. Similarly, the different shapes of the cells can bring in unwanted limitations on finding placements. This can bring negative results such as more expensive computations and longer runtimes.

It has been found that the size of macro-cells can sometimes be a considerable amount of the total area; sometimes even up to half of the placement area. This can have a significant impact on the finalizing the positions of cells with respect to others. Therefore, it is necessary to give due consideration to the magnitude of cells and the impact they can have on other cells.

As well as cell position, cell size has a significant impact on the position of pins. Unlike in standard cell placement, pin locations can have a significant impact on wirelength, routability and congestion of the chip. To overcome this, the placement tool will need to handle extra features such as cell mirroring and cell rotation to consider the best possible cell orientation in order to minimize the above-mentioned costs and to place the pins in the best locations possible.

Fixed cells are also an important factor that needs to be looked at during cell placement. There are times when one or more components of the design need to be placed in a fixed position within the placement area. For macro-cells, these fixed cells will create a blockage on the area on which cells are to be placed and will need to be given due consideration during the placement process.

It is seen that there are important differences between standard cell designs and macro-cell designs and these differences need to be given appropriate priority during placement. It further reaffirms the need for macro-cell placement tools that are separate from standard cell placement tools. Not doing so will result in poor placements and increased design costs in terms of wirelength, congestion and routing resources etc. that is detrimental for both designers and manufacturers alike.

10.4 Force Directed Graph Drawing Algorithms

Graph drawing algorithms are mainly concerned about nodes lacking any size or shape, whereas for cell placement cell sizes need to be given due consideration. A recent published work introduces methods of successfully modifying graph-drawing algorithms to incorporate dimensions to nodes [13]. This work is mainly aimed towards general graph drawing algorithms and the criteria they use for graph drawing include,

- Vertices are not to overlap
- Edges are not to cross vertices

For this work, the first criterion directly applies, as the objective of the placement tool is to produce a non-overlapping placement. The second criterion also applies as it tends to place directly connected cells together, but it could be too conservative if routing is allowed to be over-the-cell. One of the limitations of this work [13] is that the node orientation is fixed and cannot be mirrored or rotated.

Force directed graph-drawing algorithms generally tend to be analogous to the classic problem of Hookes law for a spring system. Most of the current force directed algorithms follow the footsteps of Eades' spring embedded algorithm [14]. Hooke's law simply stated that the force exerted by an extended spring is proportional to the length of the spring. Eades modeled the graph as a system of rings in place of the nodes and springs for edges. His formula for the forces exerted by the springs differed Hooke's law by the former taking both attraction and repulsion forces in to consideration. The aim of all the force directed algorithms is to find zero-force locations for all nodes – i.e. state of equilibrium for that system.

A comparison [15] of several force-directed algorithms has been carried out where KK and FR algorithms were the two main contenders. It was identified that KK is successful in achieving high computation speed, minimizing the computation time. Even though FR is quick in giving aesthetically pleasing layouts, it is said to suffer from long run times when the number of nodes/edges exceeds 60. One cannot declare a certain algorithm to be the best where each has its pros and cons and how relevant each algorithm is depends on the application [15].

KK Algorithm [3] is concerned about general undirected, connected graphs. It has the ability to handle weighted graphs such that edges with higher weighting are longer than those with a lower weighting. One advantage in this algorithm is that it introduces a "graph theoretic distance" which defines a minimum edge length in order to minimize node overlaps. The main objective of the algorithm is to find a balanced formulation of the spring forces within the system. The graph drawing criteria followed by KK [3] are,

- Reduce number of edge crossings.
- Distribute the vertices and edges uniformly.

Comparing these criteria with those of the macro-cell placement tool, it can be seen that both are related to the 'good placement criteria'. Reducing number of edge crossings results in directly connected cells being placed close to each other. The second criterion allows the nodes to be evenly distributed within the placement area as well as show any symmetry within the layout. This not only is an advantage for graph drawing where the aesthetics are improved, but for cell placement, by illustrating the cell connections in an uncomplicated manner. It is worth pointing out that symmetry is a very important heuristic for placement. While most of placement tools have difficulty in incorporating it into their algorithms, the KK algorithm handles it neatly.

The main objectives of the FR algorithm are to achieve a visually pleasing graph with increased speed and simplicity. Following Eades work, the FR algorithm also makes use of both attraction and repulsion forces, but takes it one-step further by defining that the attraction forces only to be calculated for neighboring nodes whilst repulsion forces are calculated for all nodes within the graph.

Looking at the criteria followed by FR [4] when drawing graphs, it is seen that two main points are considered.

- Vertices connected by an edge should be drawn near each other.
- Vertices should not be drawn too close to each other.

The first criteria does apply for the cell placement tool as the cells connected to one another will need to be close to each other in order to minimize wirelength. This can be further enhanced by edge weights to ensure that cells connected to edges with higher weights are as close as possible. Unfortunately, the current implementation of the FR algorithm does not contain support for edge weights. The second criterion is set quite vaguely and according to literature [4] it depends on the number of nodes and the placement area. Literally, this should mean that the nodes do not overlap each other, which is directly applicable to the objectives of the placement tool.

FR algorithm uses a method similar to simulated annealing to control the 'cooling schedule' of the algorithm, which controls the number of sweeps it goes through in optimizing the layouts. This can be both advantageous and disadvantageous. It is advantageous such that it helps limit the displacement prohibiting the algorithm to be trapped in local minima. It is disadvantageous such that the number of sweeps is kept at a constant so that the algorithm does no check on the quality of placement before ending the sequence.

The main difference between the FR and KK algorithm is that the FR algorithm can handle disconnected graphs. Even though this is not an absolute requirement compared to the objectives of the placement tool, it does give an advantage as to the type of designs the algorithm will be able to handle. Authors of KK [3] points out that even though KK algorithm does not support disconnected graphs, it can be easily extended to do so without a significant delay in time as follows.

> Partition the graph to its connected components giving each component a region of area proportional to its size, with each component laid out independently.

FR algorithm puts this theory into practice in its technique in handling disconnected graphs. Authors of FR names this technique as the *"grid variant option"* where the placement area is divided into a grid and nodes are given locations within the grid. Changes are made to the calculation of the repulsion forces; for each node, the repulsion forces are calculated from the nodes within the current grid as well as those in neighboring grids, unlike the basic algorithm which calculated repulsion forces for all nodes.

Another difference between the two algorithms is that KK does not specify a clear placement area for the graph whereas FR implements support for a customizable placement area. Whilst for graph drawing this may not be very important, it does carry greater significance in cell placement where the cells are expected to be placed within the given placement area in order for the placement to be legal. It is believed that limitation on placing components within the placement area can be imposed upon in later stages when being used in the placement tool. Table 10.1 summarizes the basic capabilities of the two algorithms.

10.5 Implementation Details

Initial work carried out [1] has proved that both KK and FR are good candidates as a basis for a macro-cell placement tool. The basic algorithms of both KK and FR were while sufficient as graph drawing algorithms, lacked the necessary functionalities to

Table 10.1 Comparison of Kamada-Kawai and Fruchterman-Reingold algorithms

KK algorithm	FR algorithm
Undirected graphs	Undirected graphs
Cannot handle disconnected graphs	Can handle disconnected graphs
Objective is to evenly distribute nodes and edges	Objective is to place nodes close to each other
Does not respect boundary conditions	Nodes are placed within the given boundary
Supports the use of edge weights	Current implementation cannot handle edge weights

be used as a module placement algorithm. Further modifications were introduced to the algorithms in-order to function better. The basic implementations of the two algorithms were taken from the peer reviewed Boost [16] library.

10.5.1 Non-zero Size Vertices Implementation

Traditional force directed algorithms tend to treat the cells as points that do not posses any size or shape. The edges do not connect to any pins but to the nodes that represent the cells. This method may be acceptable for standard cell design [10] but in Macro cell placement it can cause inaccuracies of positions, wirelength, area, congestion etc. due to the cell dimensions. Recent literature [13, 17, 18] has been found to carry out work regarding the implementation of different size nodes for graph drawing.

The simplest method of representing a cell is to consider the node to be circular [2]. However, previous work showed that for macro-cell placement, circular nodes could introduce inaccuracies of the actual dimensions and shapes of the cells. Cells with high aspect ratios can overlap one another requiring further work towards legalizing the placement. To tackle this, the elliptic spring method [13] was applied to FR algorithm. The attraction and repulsion forces for the cells are calculated such that assuming the nodes are elliptical in shape. The values of the forces are selected based on the condition if the source and target nodes are overlapping or not. All the modules are assumed rectangular shaped and the width and height of each cell is used to calculate the radii of the elliptical vertex that will represent the module.

For KK, this method could not be applied. During the flow of the algorithm, KK constantly calculates the distance between the boundaries of the connecting nodes. Whilst considering the nodes as circular, this was easily achievable. However, if the nodes were to be considered elliptical, this would cause much complexity in the algorithm increasing the runtime significantly; therefore this was not applied to KK.

10.5.2 Fixed Node Support

Another feature that was lacking within the graph drawing algorithms was support to handle fixed nodes. This is especially useful when designers may specify locations for some of the cells to be fixed or for the placement of the IO (Input Output) pads, which communicate with the external world. In force directed algorithms, since there are attraction and repulsion forces that affects all cells, it was needed to ensure that the forces emitted by the fixed cells were still being taken into account whilst the forces felt upon the fixed cells do not cause the fixed cell to displace as is illustrated in Fig. 10.3.

The algorithm of FR was altered so that the fixed nodes are treated equally as movable cells during force calculation. During displacement calculations, fixed nodes are ignored and for added measure ignored during positional updating of the cells as well. This has shown to be a more accurate method of force calculation for the algorithm when containing fixed cells.

The KK algorithm was modified to filter out the fixed nodes during the energy minimization calculations. This was accomplished in a manner that, whilst minimizing the energy function for the movable cells, the affect made from fixed nodes are still felt. Again, this has been proven successful in implementation.

10.5.3 Input/Output Format

Not all placement tools follow a single format for input and output. This hinders benchmarking and comparison of placement tools. It is with this in mind that it was thought best to use the industry standard formats; the Cadence LEF/DEF file format. The LEF/DEF format is written in ASCII format and can be easily understood. The DEF file contains all the information relevant to the design, constraints, layout and netlist information whilst the LEF file contains the library information of all the cells and modules within the design as well as information regarding layers, vias and core area information. In order to read in the necessary information for the placement tool, a parser was developed. The parser reads data in from the two files, extracts the necessary data and saves it into a text file, which can then be read in by the placement algorithms. It is hoped that in the future, this will be integrated within the placement algorithm itself so that the data input will be a one-step process.

Once the algorithms have generated a placement, it will output the summary in text format and plot the placement in order to inspect the results achieved. In future, it is hoped to output the data into a DEF file such that the final placements then can be routed. Routing congestion is another important quality measurement of the placement.

10.6 Experimentation and Results

With the implementation of different features to the algorithms, they were simulated under different conditions to identify their strengths and weaknesses. To start, the two algorithms were subjected to a selection of graphs, some with known golden topologies. Cells of different dimensions were used to observe the impact the changes described in this chapter. The simulations were run on an Intel Pentium IV PC running at 3.2 GHz and with 2 GB of RAM.

Table 10.2 compares the results obtained through this exercise. The runtime and HPWL (Half-perimeter wirelength) are the cost factors looked at during the experimentation to evaluate the performance of the algorithms. The first two columns show previous results of KK and FR whilst using circular nodes. The results shown for the FR are those obtained for the grid option, which allows the use of disconnected graphs and with elliptical nodes. As it was not possible to use elliptical nodes for KK, the placement of KK was taken as an initial placement and further enhanced by FR signified as the KK_FR flow. Figure 10.1 gives the placement of FR with circular nodes, FR with elliptical nodes and that of KK_FR for graph5 and graph6.

It is noted that the runtime of FR algorithm has increased due to the increased complexity of the algorithm by the changeover from circular nodes to elliptical nodes. Wirelength has seen in increase as well, however, this can be said to be due to the reduction in overlap where the elliptical nodes has allowed the cells to be better placed. This is further apparent in Fig. 10.1a where we see in the previous

Table 10.2 Comparison of runtime (ms) and wirelength (μm) results from previous work (marked as old) and from current work

	KK (old)		FR (old)		FR		KK_FR	
	Time	WL	Time	WL	Time	WL	Time	WL
G1	0	360	15	369	15	416	0	415
G2	15	390	15	364	15	412	0	366
G3	15	509	16	522	15	648	15	588
G4	46	1,281	15	1,163	2,187	1,489	1,296	1,461
G5	78	1,072	31	1,011	3,812	1,060	2,750	1,051
graph1	0	429	0	446	1,312	576	484	631
graph2	15	231	0	235	1,125	318	343	257
graph3	15	619	15	624	93	843	15	820
graph4	15	23	15	32	15	27	0	45
graph5	15	321	0	366	1,390	371	15	361
graph6	16	887	15	819	2,171	919	1,281	982
graph7	343	975	140	751	125	877	156	723
graph8	46	763	46	786	2,625	1,042	31	1,080
graph9	296	980	78	999	93	1,276	171	1,082
graph10	31	157	15	147	15	127	578	131
graph11	219	480	93	459	109	461	124	461

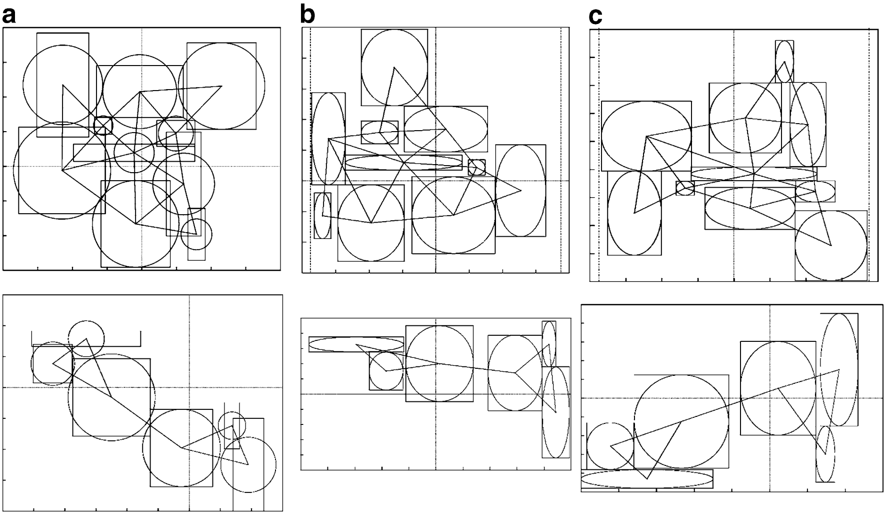


Fig. 10.1 Placement results of graph5 (*above*) and graph6 (*below*) as achieved by (**a**) FR_old, (**b**) FR with elliptical nodes, and (**c**) KK_FR placement flow

work FR has overlaps between the actual cells. However in Fig. 10.1b the reduction in overlaps is visible due to the fact that long rectangles are better represented by the elliptical nodes.

Looking at the results achieved by KK_FR, it is seen that for the same designs, runtime is less than when FR is applied on its own. It is believed that the initial placement given by KK has helped FR to reduce the complexity of the placement problem and therefore achieves better placements in a lower amount of time.

10.7 Future Work and Conclusion

In conclusion it can be said that the use of elliptical nodes has helped the FR graph drawing algorithm find more accurate placements than when using circular nodes. Even though it was not possible to apply elliptical nodes to KK algorithm, initial work carried out here has suggested that the algorithm will be successful in generating initial layouts which can then be improved by a secondary algorithm. Future work will focus on optimizing those added features and in developing the idea of having multiple algorithms within the placement flow to reduce the runtime and produce higher quality placements. In addition, further work will include establishing techniques to use pin locations to optimize wirelength by means of rotating and/or mirroring of cells. The experiments carried out so far have given positive results in achieving good layouts even with the presence of cell dimensions.

References

1. Samaranayake, M., Ji, H., Ainscough, J.: A force directed macro cell placement tool. The 2009 International Conference of Electrical and Electronics Engineering, London, 1–3 July 2009
2. Samaranayake, M., Ji, H., Ainscough, J.: Force directed graph drawing algorithms for Macro cell placement. World Congress on Engineering, London, 2–4 July 2008
3. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Inform. Process. Lett. **31**:15 (1989)
4. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Softw.-Pract. Exp. **21**:1129–1164 (November 1991)
5. Adya, S., Chaturvedi, S., Roy, J., Papa, D.A., Markov, I.L.: Unification of partitioning, placement and floorplanning. International Conference of Computer Aided Design, pp. 550–557, November 2004
6. Cong, J., Kong, T., Shinnerl, J.R., Xie, M., Yuan, X.: Large-scale circuit placement: gap and promise. IEEE/ACM International Conference on Computer-Aided Design, pp. 883–890, 2003
7. Viswanathan, N., Pan, M., Chu, C.: FastPlace 3.0: a fast multilevel quadratic placement algorithm with placement congestion control. Asia and South Pacific Design Automation Conference, pp. 135–140, 23–26, January 2007
8. Kahng, A.B., Reda, S., Wang, Q.: APlace: A general analytic placement framework. International Symposium of Physical Design, pp. 233–235, San Francisco, CA, April 2005
9. Adya, S., Markov, I.L.: Fixed-outline floorplanning: enabling hierarchical design. IEEE Trans.VLSI Syst. **11**:1120–1135 (December 2003)
10. Mo, F., Tabbara, A., Brayton, R.K.: A force-directed macro-cell placer. International Conference on Computer-Aided Design, pp. 177–180, San Jose, CA, November 2000
11. Alupoaei, S., Katkoori, S.: Net-based force-directed macro cell placement for wirelength optimization. IEEE Trans. VLSI Syst. **10**:824–835 (December 2002)
12. Sanrarini, M.: Open source website offers LEF/DEF formats. EE Times (2000)
13. Harel, D., Koren, Y.: Drawing graphs with non-uniform vertices. Proceedings of Working Conference on Advanced Visual Interfaces, pp. 157–166 (2002)
14. Eades, P.: A heuristic for graph drawing. Congressus Numerantium, pp. 149–160 (1984)
15. Brandenburg, F.J., Himsholt, M., Rohrer, C.: An experimental comparison of force-directed and randomized graph drawing algorithms. Symposium on Graph Drawing, pp. 76–87, 20–22 September 1995
16. Boost. http://www.boost.org/. Accessed September 2007
17. Wang, X., Miyamoto, I.: Generating customized layouts. In: Brandenburg, F.J. (ed.) Graph Drawing, vol. 1027, pp. 504–515. Springer, Berlin (1996)
18. Gansner, E., North, S.: Improved force-directed layouts. In: Whitesides, S.H. (ed.) Graph Drawing, vol. 1547, pp. 364–373. Springer, Berlin (1998)

Chapter 11
Surface Roughness Scattering in MOS Structures

Raheel Shah and Merlyne DeSouza

Abstract The comprehensive Ando's surface roughness (SR) model examined for nMOSFETs. Four distinct source terms contribute in SR scattering. Relative strength of these contributing source terms are evaluated and compared. The most influential term turned out to be due to scattering with the "physical steps" at the interface. Remote SR scattering is also significant in ultra-thin MOS structures. The proposed model of Gámiz et al. for remote SR scattering is studied. It is shown that modification to the Gámiz model is necessary in order to observe the full impact of rms height of the abrupt "steps".

Keywords Surface roughness · ultra-thin MOSFET · mobility · modelling and simulations

11.1 Introduction

Apart from electron-phonon scattering the most damaging effect to charge carrier mobility in MOS structures is scattering at the rough insulator/substrate interface. This scattering is particularly dominant at high inversion densities, however, due to its nature, it weakly depends on lattice temperature variations. Unlike phonon scattering it is not intrinsic in nature, since with technological advancement the insulator can be grown on the substrate with relative smoothness.

Theoretical models of interface scattering date back to 1968, when Prang and Nee performed simulations to quantify the irregularities of a rough surface [1]. This was followed by a model with explicit mobility dependence on the transverse effective

R. Shah (✉)
Emerging Technologies Research Centre, De Montfort University LE1 9BH, UK
e-mail: raheel@dmu.ac.uk

M. DeSouza
EEE Department, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK
e-mail: m.desouza@sheffield.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_11,

field by Matsumoto and Uemura, which is still adopted today for its simplicity [2]. But a more complete and comprehensive theory is by T. Ando which is by far the best available model regarded by the researchers [3,4]. Ando treated various source terms contributing towards surface roughness in detail. Recently Jin and co-workers have extended Prang and Ando's model to SOI (silicon on insulator) structures where the severity of surface roughness has been predicted to hinder ballistic transport in ultra thin SOI MOSFETS [5].

This chapter is organised as follows: an introduction to the underlying physics of the problem and the statistical measure to compute the "surface randomness" is given. Next various scattering potentials associated with surface roughness are presented. Surface roughness limited mobilities are simulated, incorporating the source terms presented in the previous section and their relative strengths are also compared.

Remote surface roughness model presented by Gámiz et al. and appropriate modification to it is also discussed along with the explanation to the observed trend in remote SR limited mobility.

11.2 Physics of the Problem

Surface Roughness (SR) in the context of MOS structures is associated with random fluctuations of the boundary between the insulator and the substrate. This roughness appears as atomic "steps" at the interface between the two materials. Deviation from the ideal flat surface introduces electric potentials from number of sources e.g. dipoles created at the interface, associated image charges, etc.

Moreover, wavefunctions of charge carriers are also physically perturbed from their normal states and consequently the potential of the system changes as well (wavefunctions and electrostatic potential are linked together via Schrödinger-Poisson coupled equations). Thus, charge carriers interact with the rough surface via these potentials and their momentum dissipates in the process.

Surface topology is usually unknown, thus it has to be modeled appropriately. Quantitatively roughness is measured via a 2D roughness function, $\Delta(\mathbf{r})$, which describes the fluctuations from an assumed ideal flat boundary. The two- dimensional vector $\mathbf{r}$ is measured along the interface plane.

The autocovariance function $C(\mathbf{r})$ determines the statistical properties of the roughness function $\Delta(\mathbf{r})$, which depends on two parameters viz.: Δ – the rms height of the steps and Λ – the average width of the fluctuation (see Fig. 11.1 for illustration). Mathematically it is given by:

$$C(\mathbf{r}) = \left\langle \Delta(\mathbf{r})\, \Delta(\mathbf{r} - \mathbf{r}') \right\rangle \tag{11.1}$$

which simply gives the probability of a random step at position $\mathbf{r}$ to be repeated at $(\mathbf{r} - \mathbf{r}')$, the notation in (11.1) $\langle \bullet \rangle$ represents the average value of a function.

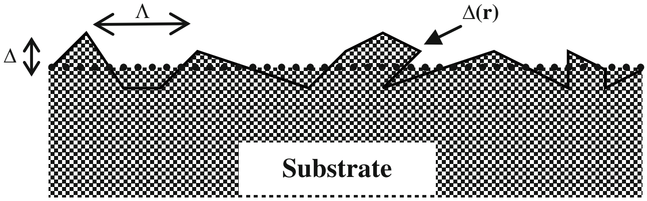


Fig. 11.1 Roughness function with the two parameters, Λ and Δ, illustrated

In relation to surface roughness, two forms of the autocovariance function are widely accepted – The Gaussian form, first proposed by Prang and Nee [1] and the exponential function [6]. Mathematically they are given as:

$$\begin{aligned} C(\mathbf{r}) = \left\langle \Delta(\mathbf{r})\, \Delta(\mathbf{r}-\mathbf{r}') \right\rangle &= \Delta^2 e^{-r^2/\Lambda^2} \qquad (\mathit{Gaussian}) \\ &= \Delta^2 e^{-r/\Lambda} \qquad (\mathit{Exponential}) \end{aligned} \tag{11.2}$$

Fourier transformation is needed to convert the autocovariance function from real space into **k**-space. Also according to Wiener–Khinchin theorem the Fourier transform of the autocovaraince function is in fact the "Power Spectrum Density", $|S(q)|^2$ of the function. Gaussian form mimics the nature ("Normal distribution") and has the other advantage that its Fourier transform is again Gaussian and thus easy to manipulate. However, Goodnick et al. [6] showed that the exponential form of the autocovariance fits more accurately to experimental data, measured through High Resolution Transmission Electron Microscope (HRTEM). The exponential form is expected for a stochastic Markov process [7]. The power spectrums for the two forms are given by [6]:

$$\begin{aligned} |S(q)|^2 &= \pi\Lambda^2\Delta^2 e^{(-q^2\Lambda^2/4)} \qquad (\mathit{Gaussian}) \\ &= \pi\Lambda^2\Delta^2(1+q^2\Lambda^2/2)^{-\frac{3}{2}} \qquad (\mathit{Exponential}) \end{aligned} \tag{11.3}$$

In simulations the exponential form will be utilized, which is, as stated, more consistent with measurement. Surface roughness limited mobility is greatly affected by the choice of Λ and Δ which are obviously device process dependent. In literature the range of these two parameters widely varies as: $\Lambda = 0.5 - 2.0\,\mathrm{nm}$ and $\Delta = 0.2 - 0.7\,\mathrm{nm}$ [7–10].

11.3 Associated Scattering Potentials

According to Ando's argument there are two main sources of surface roughness scattering affecting the charge carrier's motion, viz. [3, 4]:

- Fluctuation of *wavefunctions* due to physical "steps" at the interface ($\Gamma^{(1)}$)
- Fluctuation in *potential energy* due to Coulomb interactions

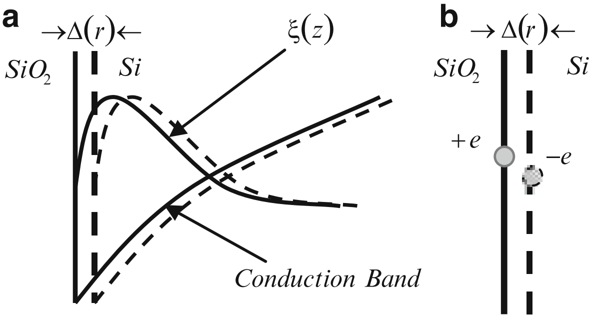


Fig. 11.2 *Solid lines* depicting the "normal" situation whereas *dashed lines* are presenting the distortions due to the interfacial "steps". (**a**) Wavefunctions and the conduction band edge are perturbed at the interface. (**b**) Image charge and a dipole are illustrated

The effects of "change in the potential energy" are further classified as

- Change in image potential ($\Gamma^{(2)}$)
- Creation of interface polarization charges ($\Gamma^{(3)}$)
- Fluctuation in charge carrier densities ($\Gamma^{(4)}$)

The first main source of scattering is obvious, since the "steps" at the interface perturbs the surface potential which consequently affects the wavefunctions, eigenvalues, etc. (see Fig. 11.2a), or simply the wavefunctions originate from the two surfaces: perturbed and unperturbed. This change in wavefunctions is propagated all along the depth of the substrate.

In order to compute the channel mobility the matrix element arising from the scattering potentials is required. Matrix element associated with change in wavefunctions is given by [5, 11]:

$$\Gamma_{i,j}^{(1)} = \int_{0}^{\infty} \left\{ \xi_i(z) \frac{\partial V(z)}{\partial z} \xi_j(z) + E_i \frac{\partial \xi_j(z)}{\partial z} \xi_i(z) + E_j \frac{\partial \xi_i(z)}{\partial z} \xi_j(z) \right\} dz \qquad (11.4a)$$

where i and j stand for initial and final subbands of a conduction valley, respectively, E_i is the eigenvalue corresponding to the wavefunction ξ_i. The potential well in the substrate is denoted by V (in units of energy e.g. *eV*).

Next among the "Coulomb pieces" i.e. SR induced charge fluctuations: the scattering potential associated with the second term ($\Gamma^{(2)}$) appears due to the mismatch of the dielectric constants of the two materials (Si and SiO_2) across the interface. SR deforms the image potential developed at the surface. M. Saitoh studied 2D electrons on the surface of ^{4}He fluid and calculated the potential resulting from such image charges, this idea and expression was utilized in the context of 2DEG in inversion layer by Ando [4, 12].

Scattering matrix for the change in image potential is given by:

$$\Gamma_{i,j}^{(2)}(q) = \frac{e^2\tilde{\varepsilon}q^2}{16\pi\varepsilon_s}\int\limits_0^\infty \xi_i(z)\left\{\frac{K_1(qz)}{qz} - \frac{\tilde{\varepsilon}}{2}K_0(qz\right\}\xi_j(z)\,dz \tag{11.4b}$$

where $\tilde{\varepsilon} = (\varepsilon_s - \varepsilon_{ox})/(\varepsilon_s + \varepsilon_{ox})$ with ε_s and ε_{ox} as the dielectric constants of the substrate and the oxide, respectively. K_0 *and* K_1 are modified Bessel functions of the second kind and of order zero and one, respectively.

The third Coulomb interaction ($\Gamma^{(3)}$) is also related to the difference in dielectric constants of the adjacent materials. An extra polarization charge is formed which changes the electric field distribution [13]. The arising electrostatic potential is calculated by considering an effective dipole moment at $z = 0$. The matrix element pertaining to $\Gamma^{(3)}$ is given by [5, 13]:

$$\Gamma_{i,j}^{(3)}(q) = e\tilde{\varepsilon}\int\limits_0^\infty \xi_i(z)\left\{E_{eff}e^{-qz}\right\}\xi_j(z)\,dz \tag{11.4c}$$

where e is the electronic charge and E_{eff} is the effective field at the substrate side of the interface.

Lastly, the fluctuations at the interface also affect the electron distribution normal to the interface ($\Gamma^{(4)}$ term). The redistributed electron charges give rise to an additional scattering potential whose matrix element is [5, 14]:

$$\Gamma_{i,j}^{(4)}(q) = -\frac{e^2}{2q\varepsilon_s}\int\limits_0^\infty \xi_i(z)\xi_j(z)dz\int\limits_0^\infty\left\{e^{-q|z-z'|} + \tilde{\varepsilon}e^{-q|z+z'|}\right\}\frac{\partial n(z')}{\partial z'}dz' \tag{11.4d}$$

where $n(z)$ is the volume density of electrons along the z-direction (normal to the interface).

All the three "Columbic potentials" are wave vector $q = \left|\mathbf{k}_j - \mathbf{k}_i\right|$ dependent while potential due to change in wavefunctions is only electron energy dependent. It is to be noted that matrix elements given above are under the assumption of infinite potential barrier at the interface and vanishing wavefunctions deep into the thick substrate.

With four different source terms contributing in SR scattering, an "effective" matrix element is required whose squared value could be plugged in the scattering rate computed through Fermi golden rule [11].

It is well known that the scattering potential is effectively "screened" by the sheet of electrons present between the source of the electric potential and the scattered charge carriers, within certain specific distance called as the Debye length [7, 13]. Screening has a profound effect on the SR limited mobility; it is thus inevitable to include screening effects in simulations. SR limited mobilities are calculated using appropriate expressions in this context [14].

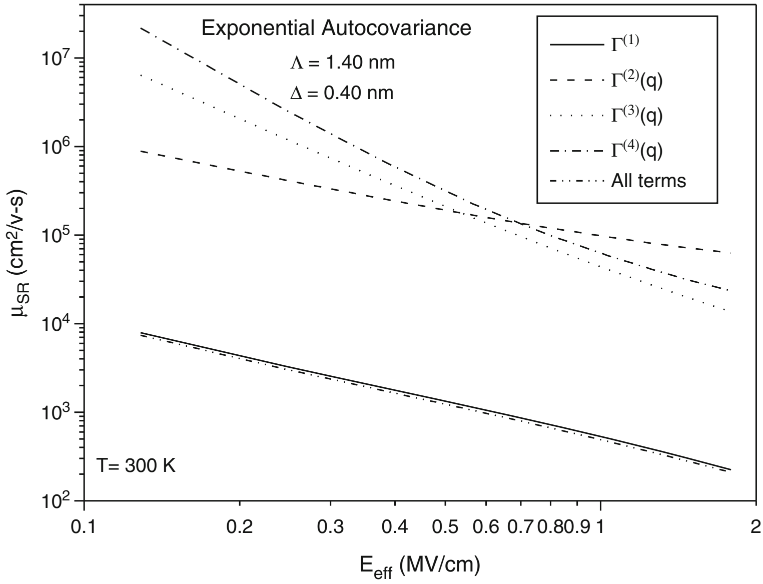


Fig. 11.3 Relative strengths of four sources of SR scattering are compared. The most damaging source for channel mobility is the $\Gamma_{i,j}^{(1)}$ term

11.4 Relative Strength of the Scattering Potentials

Once the screening mechanism is formulated the SR limited mobilities can be evaluated using the above mentioned four scattering potentials. Next, the relative importance of each scattering potential is evaluated. Figure 11.3 is a compilation of the results for SR mobility with the effects of individual scattering terms and with their combined influence.

It is clearly evident that the most dominant scattering source is due to the perturbations in electron wavefunctions ($\Gamma^{(1)}$). The weakest source is the variation of electron density due to physical "steps" introduced i.e., $\Gamma^{(4)}(q)$, though computationally it is a most time consuming term to evaluate. The percentage difference between mobility computed with all terms and then with $\Gamma^{(1)} + \Gamma^{(2)}(q) + \Gamma^{(3)}(q)$ terms is around 1% at $E_{eff} = 1\,\text{MV/cm}$. Thus these three terms are sufficient to account for SR scattering and $\Gamma^{(4)}(q)$ can be safely ignored.

11.5 Remote Surface Roughness Scattering

Another scattering mechanism, closely related to SR scattering, is the "Remote Surface Roughness" (RSR) scattering. For ultra thin oxide layered MOS structures, charge carriers in the channel can significantly dissipate their momentum by remotely interacting with the gate/insulator interface. Similar to oxide/substrate interface the second interface i.e. gate/oxide interface is not smooth and deviates from the ideal plane. Degree of roughness at the two interfaces is uncorrelated and depends on the device processing mechanism.

Extending the concept first presented by Li [15], Gámiz et al. proposed a simple scattering model for the remote surface roughness mechanism [16, 17]. In their proposed model the Hamiltonian of the system is given by:

$$H' = H_0 + \frac{\Delta V(z)}{\Delta_m}\Delta(\mathbf{r}) \tag{11.5}$$

where:

$$\Delta V(z) = V_{t_{ox}+\Delta_m}(z) - V_0(z) \tag{11.6}$$

with $V_{t_{ox}+\Delta_m}$ as the perturbed potential in the presence of "steps" at the gate/insulator interface while V_0 is the unperturbed potential i.e. in the case of ideal boundary. Surface topology is measured via 2D roughness function, $\Delta(\mathbf{r})$, which describes the fluctuations from an assumed ideal flat boundary. The two-dimensional vector $\mathbf{r}$ is measured along the interface plane. H_0 is the initial unperturbed Hamiltonian and the final Hamiltonian H' arising from the change in potential energy along the z-direction. The rms value of the step height at the second interface is denoted by Δ_m. Using the Hamiltonian (11.5) the matrix element was constructed as [16]:

$$\Gamma_{i,j}^{RSR} = \int_0^\infty \xi_i(z)\frac{\Delta V(z)}{\Delta_m}\xi_j(z)dz \tag{11.7}$$

The matrix element (11.7) modulated by dielectric function $\varepsilon(q)$ is utilized to compute the RSR scattering rate. However, Hamiltonian (11.5) of the present system can also be used to construct a relatively better RSR matrix element, following the approach described below:

Consider the change in the Hamiltonian of the system due to the presence of a random "step" at the interface, given by:

$$\Delta H = H' - H_0 \tag{11.8}$$

Next the matrix element for the changed Hamiltonian is generated from Eq. (11.8):

$$\Gamma_{i,j}^0(\mathbf{r}) = \int_0^\infty \xi_i(z+\Delta(\mathbf{r}))[H']\xi_j(z+\Delta(\mathbf{r}))dz - \int_0^\infty \xi_i(z)[H_0]\xi_j(z)dz \tag{11.9}$$

Note that for the final Hamiltonian H', the final perturbed wavefunctions are used. This perturbation is caused due to the potential difference arising at the gate/oxide interface. Substituting Eq. (11.5) in (11.9) to get:

$$\Gamma^{0}_{i,j}(\mathbf{r}) = \int_{0}^{\infty} \xi_i(z + \Delta(\mathbf{r}))\left[H_0 + \frac{\Delta V(z)}{\Delta_m}\Delta(\mathbf{r})\right]\xi_j(z + \Delta(\mathbf{r}))dz - \int_{0}^{\infty} \xi_i(z)[H_0]\xi_j(z)\,dz \tag{11.10}$$

Next using Taylor's theorem for the expansion of the wavefunctions to the lowest order reveals:

$$\Gamma^{0}_{i,j}(\mathbf{r}) = \int_{0}^{\infty} \left[\xi_i(z) + \frac{\partial \xi_i(z)}{\partial z}\Delta(\mathbf{r})\right]\left[H_0 + \frac{\Delta V(z)}{\Delta_m}\Delta(\mathbf{r})\right] \left[\xi_j(z) + \frac{\partial \xi_j(z)}{\partial z}\Delta(\mathbf{r})\right] dz - \int_{0}^{\infty} \xi_i(z)\,[H_0]\,\xi_j(z)\,dz \tag{11.11}$$

Ignoring the product terms involving $\Delta(\mathbf{r})^2$:

$$\Gamma^{0}_{i,j}(\mathbf{r}) = \int_{0}^{\infty} \begin{bmatrix} \xi_i(z)H_0\xi_j(z) + H_0\xi_i(z)\dfrac{\partial \xi_j(z)}{\partial z}\Delta(\mathbf{r}) + \\ \xi_i(z)\,\dfrac{\Delta V(z)}{\Delta_m}\Delta(\mathbf{r})\,\xi_j(z) + H_0\xi_j(z)\dfrac{\partial \xi_i(z)}{\partial z}\Delta(\mathbf{r}) \end{bmatrix} dz - \int_{0}^{\infty} \xi_i(z)\,[H_0]\,\xi_j(z)\,dz \tag{11.12}$$

Now, from the time independent Schrödinger equation:

$$H_0\xi = E\xi \tag{11.13}$$

Equation (11.12) is modified using (11.13) and after simplification the net result is:

$$\Gamma^{0}_{i,j}(\mathbf{r}) = \Delta(\mathbf{r}) \int_{0}^{\infty} dz \begin{bmatrix} \xi_i(z)\dfrac{\Delta V(z)}{\Delta_m}\xi_j(z) + E_j\dfrac{\partial \xi_i(z)}{\partial z}\xi_j(z) \\ +E_i\dfrac{\partial \xi_j(z)}{\partial z}\xi_i(z) \end{bmatrix} \tag{11.14}$$

$$\Gamma^{0}_{i,j}(\mathbf{r}) = \Delta(\mathbf{r})\,\Gamma^{RSR}_{i,j} \tag{11.15}$$

where,

$$\Gamma_{i,j}^{RSR} = \int_0^\infty dz \left[\begin{array}{c} \xi_i(z)\dfrac{\Delta V(z)}{\Delta_m}\xi_j(z) + E_j \dfrac{\partial \xi_i(z)}{\partial z}\xi_j(z) \\ +E_i \dfrac{\partial \xi_j(z)}{\partial z}\xi_i(z) \end{array} \right] \tag{11.16}$$

is the modified form of the matrix element given earlier in Eq. (11.7).

In (11.16) the derivative of the electron wavefunctions appears, which essentially is a characteristic of surface roughness scattering [1]. The matrix element given in (11.6) over estimates the RSR limited mobility as compared to one in computed using (11.14). Another shortcoming of the Gámiz model is that the effect of rms Δ_m is not present explicitly (it cancels out in the scattering rate when squared matrix element is multiplied with power spectrum $|S(q)|^2$). The only weak dependence of Δ_m appears in Gámiz model is through simulated value of $\Delta V(z)$ (via the coupled Schrödinger Poisson solver) in (11.6), while the modification presented here to the transport model includes Δ_m explicitly and thus its effect is realistically observed.

For comparison, results obtained using the two model equations are shown in Fig. 11.4. In this study, two different theoretical values of Δ_m (0.5 and 0.3 nm) are used for a fixed oxide thickness of 1.0 nm. Potential $\Delta V(z)$ in (11.6) with wavefunctions and eigenvalues are computed using UT-Quant Schrödinger-Poisson solver [18]. With $\Delta_m = 0.5$ nm, the drop from the "Universal mobility" is around 5% at $E_{\text{eff}} \sim 1\text{MV/cm}$.

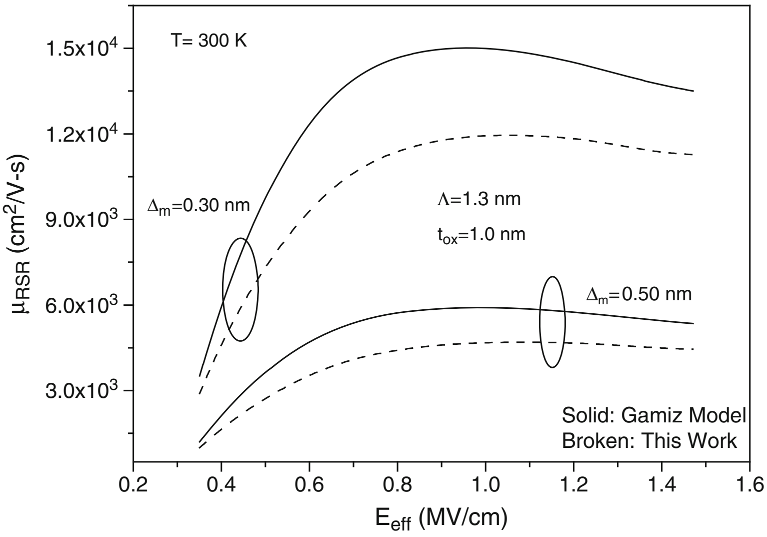


Fig. 11.4 RSR limited mobility computed using Gámiz et al. model [16] and the modified model proposed in this work. Gámiz model overestimates the mobility as compared to the modified model

11.5.1 Observed Trend in RSR Mobility

From Fig. 11.4 two observations can be clearly made, first: smoother the surface (small Δ_m) better is the RSR limited mobility. The reason for this behavior is obvious. Secondly: Initially RSR mobility increases with increasing sheet density, N_s, after reaching to an absolute maximum, mobility then starts declining. Possible reasons for this trend are explored below.

The scattering potential, which is in fact the difference in perturbed and unperturbed potentials $(V_{t_{ox}+\Delta_m}(z) - V_0(z))$, decreases with evolving sheet density, N_s. Figure 11.5 illustrates this fact graphically. Additionally, screening also contributes towards mobility enhancement. On the other hand, with increasing transverse field, the wavefunctions are more squeezed towards the interface and thus magnitude of the matrix element increases, consequently lowering the mobilities (see (11.7) or (11.16)). At the maximum ("breakeven point"), observed in Fig. 11.4, the two effects i.e. $\Delta V(z)$ and the "squeezing wavefunctions" just balance each other. Beyond this point, a further increase in gate voltage favors the impact of squeezed wavefunctions and thus mobility starts dropping, similar to the trend observed in "normal" SR mobility.

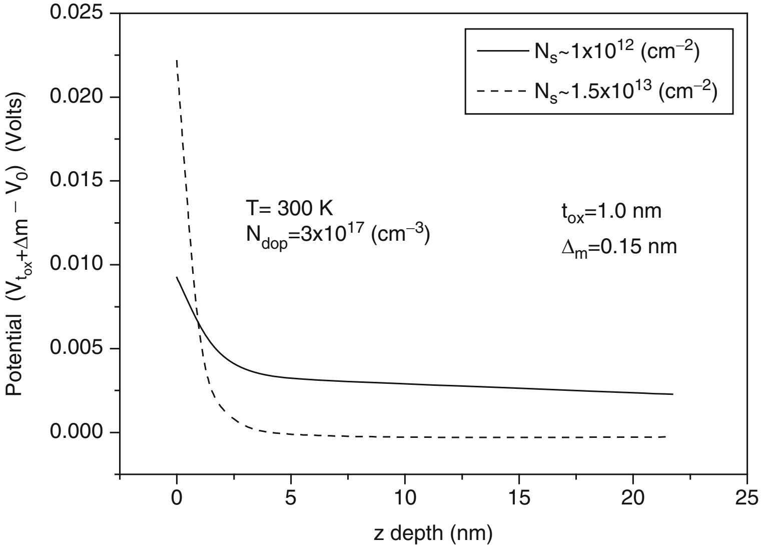


Fig. 11.5 Difference in the unperturbed and perturbed potentials are plotted for two different sheet concentrations. For strong inversion the difference in potentials ΔV drops sharply

References

1. Prange, R.E., Nee, T.W.: Quantum spectroscopy of the low-field oscillations in the surface impedance. Phys. Rev. **168**, 779–786 (1968)
2. Matsumoto, Y., Uemura, Y.: Scattering mechanism and low temperature mobility of MOS inversion layers. Jpn. J. Appl. Phys. (suppl.2, pt.2), 367–370. Kyoto, Japan (1974)
3. Ando, T.: Screening effect and quantum transport in a silicon inversion layer in strong magnetic fields. J. Phys. Soc. Jpn. **43**, 1616–1626 (1977)
4. Ando, T., Fowler, A.B., Stern, F.: Electronic properties of two-dimensional systems. Rev. Mod. Phys. **54**, 437–672 (1982)
5. Seonghoon, J., Fischetti, M.V., Ting-Wei, T.: Modeling of surface-roughness scattering in ultrathin-body SOI MOSFETs. IEEE Trans. Electron Device **54**, 2191–2203 (2007)
6. Goodnick, S.M., Ferry, D.K., Wilmsen, C.W., Liliental, Z., Fathy, D., Krivanek, O.L.: Surface roughness at the Si(100)-SiO$_2$ interface. Phys. Rev. B **32**, 8171 (1985)
7. Ferry, D.K., Goodnick, S.M.: Transport in Nanostructures. Cambridge University Press, Cambridge (1999)
8. Jungemann, C., Emunds, A., Engl, W.L.: Simulation of linear and nonlinear electron transport in homogeneous silicon inversion layers. Solid-St. Electron. **36**, 1529–1540 (1993)
9. Pirovano, A., Lacaita, A.L., Zandler, G., Oberhuber, R.A.O.R.: Explaining the dependences of the hole and electron mobilities in Si inversion layers. IEEE Trans. Electron Devices **47**, 718–724 (2000)
10. Low, T., Li, M.F., Fan, W.J., Ng, S.T., Yeo, Y.C., Zhu, C., Chin, A., Chan, L., Kwong, D.L.: Impact of surface roughness on silicon and Germanium ultra-thin-body MOSFETs. IEEE international electron devices meeting 2004 (IEDM), pp. 151–154, 13–15 Dec 2004. San Francisco, CA.
11. Esseni, D.: On the modeling of surface roughness limited mobility in SOI MOSFETs and its correlation to the transistor effective field. IEEE Trans. Electron Devices **51**, 394–401 (2004)
12. Saitoh, M.: Warm electrons on the liquid ^{4}He surface. J. Phys. Soc. Jpn. **42**, 201–209 (1977)
13. Fischetti, M.V., Laux, S.E.: Monte-Carlo study of electron-transport in silicon inversion-layers. Phys. Rev. B **48**, 2244–2274 (1993)
14. Esseni, D., Abramo, A.: Modeling of electron mobility degradation by remote Coulomb scattering in ultrathin oxide MOSFETs. IEEE Trans. Electron Devices **50**, 1665–1674 (2003)
15. Jia, L., Ma, T.P.: Scattering of silicon inversion layer electrons by metal-oxide interface roughness. J. Appl. Phys. **62**, 4212–4215, Nov 15 1987
16. Gámiz, F., Roldan, J.B.: Scattering of electrons in silicon inversion layers by remote surface roughness. J. Appl. Phys. **94**, 392–399 (2003)
17. Gámiz, F., Godoy, A., Jimenez-Molinos, F., Cartujo-Cassinello, P.A.C.-C.P., Roldan, J.B.A.R.J.B.: Remote surface roughness scattering in ultrathin-oxide MOSFETs. 33rd Conference on European Solid-State Device Research, 2003 (ESSDERC ’03), pp. 403–406 (2003)
18. Shih, S.J.W.-K., Chindalore, G.: UTQUANT 2.0 User’s Guide. University of Texas Press, Austin (1997)

Chapter 12
A Novel Transform Domain Based Hybrid Recurrent Neural Equaliser for Digital Communication Channel

Susmita Das

Abstract Efficient neural network based adaptive equalisations for digital communication channels have been suggested in recent past. Recurrent neural network (RNN) exhibits better performance in nonlinear channel equalization problem. In this present work a hybrid model of recurrent neural equaliser configuration has been proposed where a Discrete Cosine Transform (DCT) block is embedded within the framework of a conventional RNN structure. The heterogeneous configuration on the RNN framework needs training and involves updation of the connection weights using the standard RTRL algorithm, which necessitates the determination of errors at the nodes of the RNN module. To circumvent this difficulty, an adhoc solution has been suggested to back propagate the output error through this heterogeneous configuration. Simulation study and bit-error-rate performance analysis of the proposed Recurrent Transform Cascaded (RTCS) equaliser for standard communication channel models show encouraging results.

Keywords Recurrent neural network · equaliser · bit error rate · discrete cosine transform · normalization

12.1 Introduction

Channel equalization is a powerful technique for compensating intersymbol interference in a dispersive communication channel, the nonlinearities introduced by the modulation/demodulation processes and the noise generated in the system. However, linear equalisers do not perform well on channels with deep spectral nulls or with nonlinear distortions. Researchers have shown that nonlinear equalisers based nonlinear theory exhibit better performance than linear equalisers in applications

S. Das (✉)
Department of Electrical Engineering, National Institute of Technology,
Rourkela – 769008 Orissa, India
e-mail: sdas@nitrkl.ac.in

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_12,

where the channel nonlinear distortions exist [1, 2]. When the channel itself has nonlinear characteristics or nonlinear channel distortions are too severe to ignore, even the Decision Feedback Equaliser cannot recover the corrupted signals effectively. Since neural networks (NN) [3] can perform complex mapping between its input and output space, and are capable of forming complex decision regions with nonlinear decision boundaries, many types of NNs have successfully applied in channel nonlinear equalization problem [2].The use of NN's is justified by noting that in most cases, the boundaries of the optimal decision regions are highly nonlinear, thus requiring the use of nonlinear classifiers, even with linear channels. Efficient neural network based adaptive equalisations for digital communication channels have been suggested in recent past. Different ANN architectures such as multilayer perceptron (MLP), radial basis function (RBF) etc. and many novel architectures and efficient training algorithms have been proposed in the literature [4]. Moreover structure selection for an ANN equaliser has always been a point of concern because a less complex structure is much easier to implement in real-time using VLSI, DSP chips etc. and also more suitable for typical applications like time varying channels in mobile communication system, optical recording media [5] etc.

Among the techniques based NN, Recurrent Neural Network (RNN) [6, 7] equalisers are proposed to solve the nonlinear channel equalization problem. RNN has shown better performance than feed forward neural network, because it approximates infinite impulse response (IIR) filter while feed forward neural network approximates FIR filter, which makes it attractive in the presence of channels with deep spectral nulls. In addition, RNN is more attractive for their small size [8]. Results from the simulations show that the RNE with simple size can yield a significant improvement in performance relative to the equalisers with linear filter, and outperform MLP equalisers of larger computational complexity in no minimum phase, partial response, and nonlinear channel equalizations cases. Complex versions of the RNE based on a real time current learning (RTRL) algorithm are developed to process complex signals [9]. Although various algorithms and hybrid structures [10, 11] have improved the performance of RNE, the computational burdens would become greater. In summary, the heavy computational load and low convergence speed have limited the practical applications of RNE.

In this paper, a hybrid configuration has been proposed where a Discrete Cosine Transform (DCT) block is embedded within the framework of a conventional RNE structure. A signal vector is mapped from a given domain to another when fed to a transform block and basically the transform block performs a fixed filtering operation. The basic difference between the transform block and the neural block is that while adaptive weights are associated with the later, fixed weights are inherent in the former. Hence, this cascaded network representing a heterogeneous configuration has been proposed to solve the conventional RNE problem keeping the complexity of the weight adaptation less. It is obvious that the transform block does not require any weight adaptation, but the RNN module needs updation of the connection weights using the standard RTRL algorithm, which necessitates the determination of errors at the nodes of the RNN module. To circumvent this difficulty, an adhoc solution has been suggested. The primary objective of the proposed work is to design cascaded RNE on reduced structural framework with faster convergence keeping in mind real-time implementation issue.

The organization of this paper is as follows. In Section 12.2, cascaded RNE equaliser based on the hybrid technique utilizing the modified version of the RTRL algorithm used to train it are described in detail. In Section 12.3, the performances of the proposed equaliser through various simulations for linear and nonlinear channels are illustrated. Finally, Section 12.4 summarizes the research work.

12.2 Proposed Hybrid Recurrent Neural Equaliser

A real-valued discrete cosine transform block followed by power normalization block is cascaded with an RNN module at the output end as given in Fig. 12.1. Power normalisation technique [9] is applied to the transformed signals and the final output of the proposed structure is evaluated as a weighted sum of all normalised signals. In order to update the connection weights of this cascaded framework, a novel idea has been developed based on propagation of the output error through the network in the light of the conventional BP algorithm. The transform block does not require any weight adaptation as it consists of fixed weights, but the RNN module needs updation of the connection weights using the standard RTRL algorithm, which necessitates the determination of errors at the nodes of the RNN module. But this estimate cannot be accomplished directly by using BP algorithm due to positioning of the transform block close to the output end, so problem is encountered here in propagating the final output error back into the network. To circumvent this difficulty, an adhoc solution has been evolved and error estimation at the input end of the transform block is done from the knowledge of the error at its output by considering its inverse transform. The mathematical expressions governing this concept are described in subsequent section.

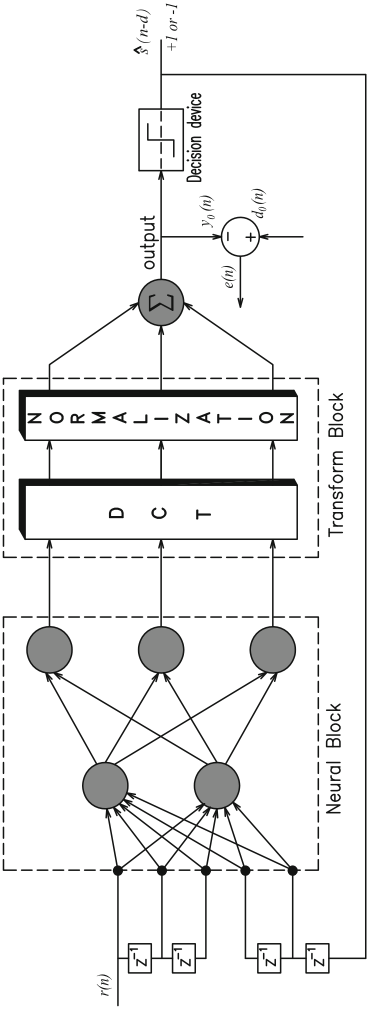
12.2.1 Training Algorithm of Hybrid Neural Structure

The proposed structure shown in Fig. 12.1 consists of nr processing units in the RNN module with nx external inputs and a transform block. A step by step procedure has been adopted to update the weights of the neural network as mentioned below. Sensitivity parameters $\left\{p\,_{kl}^{\,j}\right\}$ of all RNN nodes are intialised to zero. The input signal to the proposed equaliser structure is represented by a *m x 1* vector $\boldsymbol{x}(n) = [r(n), r(n-1), \ldots\ \ldots, r(n-m+1)]^T$.

Input signal vector to the RNN module is defined as $u(n)$, lth element of which is

$$u_l(n) = \begin{cases} y_j(n), & 1 \le j \le nr \\ x_i(n), & 1 \le i \le nx \end{cases} \quad \text{for} \quad 1 \le l \le (nr + nx) \tag{12.1}$$

Fig. 12.1 Hybrid neural – transform equaliser structure



The output of jth neuron of the RNN module at time index n is given by

$$y_j(n) = \frac{1 - e^{-\phi . c_j(n)}}{1 + e^{-\phi . c_j(n)}} \tag{12.2}$$

Where the net internal activity is described by

$$c_j(n) = \sum_{l=1}^{nx+nr} w_{kl}(n) \cdot u_l(n), \quad 1 \le k \le nr \tag{12.3}$$

where W denotes *nr* by *(nx + nr)* weight matrix of the RNN module. Sigmoid activation functions ($\mathcal{F}$) with slope parameter ϕ for neurons of the RNN module have been considered. Input signal vector to the transform block can be expressed as *z(n)*, whose jth element is denoted as,

$$z_j(n) = y_j(n), j = nr \tag{12.4}$$

Here all the processing units of the RNN module act as visible units giving externally reachable outputs. The jth element of the output from the transform block (DCT) is defined as

$$z_{\mathcal{T}j}(n) = DCT\left\{z_j(n)\right\} = \mathcal{T}z_j(n) \tag{12.5}$$

The $\mathcal{T}_{\mathrm{pq}}$th element of the N X N transforms matrix $\mathcal{T}$ is defined as

$$\mathcal{T}_{pq} = \begin{cases} \dfrac{1}{\sqrt{N}}, & p = 0;\ q = 0, 1, \ldots\ldots, N-1 \\ \left(\sqrt{\dfrac{2}{N}}\right)\cos\dfrac{\pi(2q+1)p}{2N}, & p = 1, 2, \ldots.N-1;\ q = 0, 1, \ldots., N-1 \end{cases} \tag{12.6}$$

Transformed signal $y_{Tk}(n)$ is then normalised by the square root of their power $\mathcal{B}_j(n)$ which can be estimated by filtering the signal an exponentially decaying window of scaling parameter $\gamma \in [0, 1]$ as derived in the literature [12, 13] and shown below.

The jth element of the normalized signal becomes

$$z_{\mathcal{N}j}(n) = \frac{z_j(n)}{\sqrt{\boldsymbol{\mathcal{B}}_j(n)\ +\varepsilon}} \tag{12.7}$$

and

$$B_j(n) = \gamma\ B_j(n-1) + (1-\gamma)z^2_{\mathcal{T}j}(n) \tag{12.8}$$

The small constant ε is introduced to avoid numerical instabilities when signal power $B_j(n)$ is close to zero.

The final output of the hybrid structure at time index n, y_o(n) is expressed as the weighted sum of all normalized signals from the transform blocks.

$$y_o(n) = \sum_{j=1}^{nr} g_j(n) \cdot z_{\mathcal{N}j}(n) \tag{12.9}$$

Where g denotes the weight matrix at the output end of the proposed network.

The error at the equaliser output at time index n is en by,

$$e(n) = d_o(n) - y_o(n) \tag{12.10}$$

With the knowledge of the output error, the errors at all the nodes of RNN module can be evaluated in order to facilitate the updation of weights using RTRL algorithm. But this is not possible directly as already explained before and hence a technique has been employed to tackle the situation.

At first the error e(n) is back propagated through various connection paths. Then the error at the jth output of normalization block is computed as given by

$$e_{\mathcal{N}_j}(n) = e(n) \cdot g_j(n),\ 1, \quad 1 \leq j \leq nr \tag{12.11}$$

The error terms at the output of the transform block $\delta_{\mathcal{T}j}(n)$ can be calculated using the following approach. The power normalisation can be considered as a process, whose operation is quite similar to the nonlinear transformation produced by sigmoid activation function of a neuron. This concept helps to calculate the error terms (i.e., local gradients) at the output of the transform block using the following equation

$$\begin{aligned}\delta_{\mathcal{T}j}(n) &= e_{\mathcal{N}j}(n)\ \frac{\partial y_{\mathcal{N}k}(n)}{\partial y_{\mathcal{T}k}(n)}\\ &= e_{\mathcal{N}j}(n)(y_{\mathcal{N}k}(n)/y_{\mathcal{T}k}(n))\left\{1-(1-\gamma)y^2_{\mathcal{T}k}(n)\right\}\end{aligned} \tag{12.12}$$

Further, to propagate the error back through the transform block and to estimate the error magnitudes at the input side of the transform block, Inverse Discrete Cosine Transform (IDCT) is applied. This provides an estimate of the error at the input end of the transform block.

The error at the jth processing unit of the RNN module at time index n is given by

$$err_{rnn-node_j}(n) = IDCT\left\{\delta_{\mathcal{T}j}(n)\right\} \tag{12.13}$$

Application of RTRL algorithm involves primarily the evaluation of sensitivity parameter, a triply indexed set of variables $\left\{p\ ^j_{kl}\right\}$ defined in literature [06].

$$p^j_{kl}(n) = \frac{\partial y_j(n)}{\partial w_{kl}(n)}, \quad k \in \mathbf{A} \text{ and } l \in \mathbf{A} \cup \mathbf{B}$$

where, A $= \{1, 2, \ldots, nr\}$ and B $= \{1, 2, \ldots, nf\}$.

The sensitivity parameters $\left\{p\ ^j_{kl}\right\}$ are updated as follows

$$p^j_{kl}(n+1) = \mathsf{F}'\left\{c_j(n)\right\}\left[\sum_{i=1}^{nr} w_{ji}(n) \cdot p^i_{kl}(n) + \partial_{kj} u_l(n)\right] \tag{12.14}$$

where, $1 \leq j \leq nr$, *and* $1 \leq l \leq (nr + nx)$

$$\mathsf{F}'\{c_j(n)\} = \{1 - y_j(n+1)^2\}(\phi/2) \text{ and}$$

∂_{kj} is termed as *Kronecker delta* as given by,

$$\partial_{kj} = 1 \text{ for } j = k \text{ and zero otherwise.}$$

While the incremental weight change $\Delta g_j(n)$ is calculated using BP algorithm, RTRL algorithm computes the incremental weight change $\Delta w_{kl}(n)$.

$$\Delta g_j(n) = \theta \cdot e(n) \cdot z_{Nj}(n), \quad 1 \leq j \leq nr \tag{12.15}$$

$$\Delta w_{kl}(n) = \lambda \cdot \sum_{j=1}^{nr} err_{rnn-node_j}(n) \cdot p_{kl}^{j}(n), \quad 1 \leq k \leq nr \text{ and } 1 \leq l \leq (nr + nx) \tag{12.16}$$

where, λ and θ are learning-rate parameters of the RNN module and the output layer respectively.

The connection weights are updated as given below.

$$g_j(n+1) = g_j(n) + \Delta g_j(n) \tag{12.17}$$

$$w_{kl}(n+1) = w_{kl}(n) + \Delta w_{kl}(n) \tag{12.18}$$

The objective here is to minimise the cost function i.e. to change the weights in the direction that minimizes $J(n)$. The recursion process of updating weights of the cascaded network continues till a this predefined condition is achieved.

12.3 Simulation Study and Discussions

An exhaustive computer simulation study has been undertaken for evaluating the performance of all the proposed neural equaliser structures based on FNN topologies for a variety of linear and non-linear real communication channels models. The simulation model of an adaptive equaliser considered is illustrated in Fig. 12.2. In the simulation study the channel under investigation is excited with a 2-PAM signal, where the symbols are extracted from uniformly distributed bipolar random numbers $\{-1, 1\}$. The channel output is then contaminated by an AWGN (Additive White Gaussian Noise). The pseudo-random input and noise sequences are generated with different seeds for the random number generators. For mathematical convenience, the received signal power is normalised to unity. Thus the received

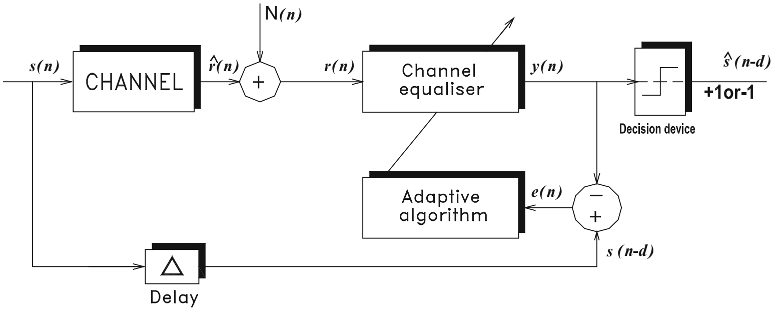


Fig. 12.2 Simulation model of channel equaliser in training phase

signal to noise ratio (SNR) is simply the reciprocal of the noise variance at the input of the equaliser. The power of additive noise has been taken as 0.01, representing a SNR of 20 dB.

Equalisation of different types of channel models (both linear and non-linear type) are attempted in order to establish the efficacy of the proposed equaliser structures based on RNN topology and to prove their robustness. It has been already reported in the literatures [6, 7], that a two-unit, one input, one output RNN is a non-linear IIR model which is sufficient to model many communication channels. Considering this aspect, all the proposed cascaded equalisers in RNN framework are compared with a conventional RNN equaliser (CRNN) with two recurrent units and one external input sample from the channel output. Further the TDRNN structure has two nodes in RNN module followed by a 2 × 2 DCT block with power normalisation and a summing unit at the output end.

For a comparative study and analysis purpose the number of training samples presented to the proposed equaliser considered here are restricted to 200 samples only as it is observed that their performances are quite satisfactory. The BER performance comparison of the proposed equaliser structures based on RNN topology has been carried out after all the structures has undergone a training phase (200 samples) The weight vectors of the equalisers are frozen after the training stage is over and then the performance test is continued. The BER performances for each SNR are evaluated, based on 10^7 more received symbols (test samples) and averaged over 20 independent realizations. All the proposed equalisers in RNN domain require fewer samples in training phase for satisfactory BER performance. Simulation results demonstrate this advantages offered by these structures. For the RTCS structure, the number of processing units remains the same as the CRNN equaliser. After the input signal is preprocessed in the RNN module, it is fed to the DCT transform block for further processing. As expected, such a proposed structure performs better than a CRNN due to the further signal de-correlation in the transform block followed by power normalization.

An example of a three tap channel characterized by

$$H_1(z) = 0.407 - 0.815z^{-1} - 0.407z^{-2} \tag{12.19}$$

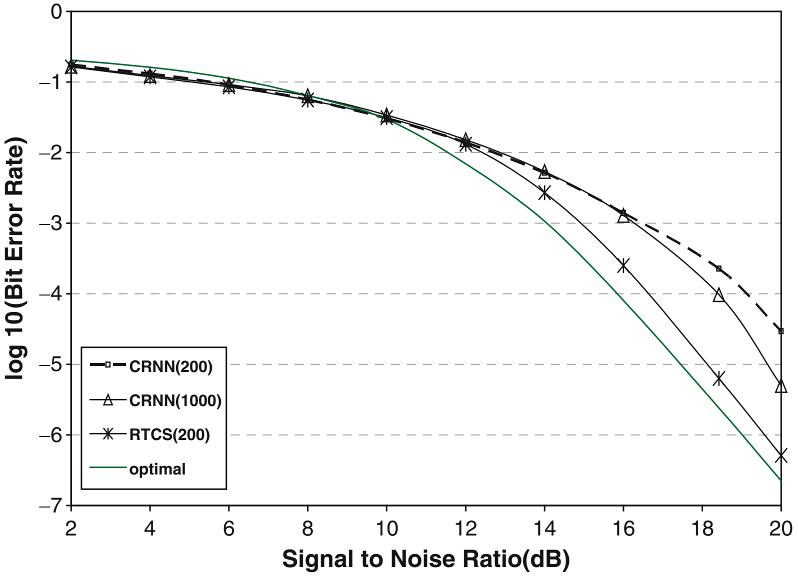


Fig. 12.3 BER performance of the proposed hybrid equaliser for channel $H_1(z)$

RTCS equaliser show distinct SNR gains of about 4.4 dB at a prefixed BER level of 10^{-4} over a conventional RNN equaliser which is quite encouraging (Fig. 12.3).

In order to prove the robustness and consistency in performance of all the proposed neural structures, equalisation of nonlinear channels is simulated. Such nonlinear channels are frequently encountered in several places like the telephone channel, in data transmission over digital satellite links, especially when the signal amplifiers operate in their high gain limits and in mobile communication where the signal may become non-linear because of atmospheric nonlinearities. These typical channels encountered in real scenario and commonly referred to in technical literatures [4, 6] are described by the following transfer functions.

$$H_2(z) = (1 + 0.5\,z^{-1}) - 0.9(1 + 0.5\,z^{-1})^3 \tag{12.20}$$

$$\begin{aligned} H_3(z) &= (0.3482 + 0.8704z^{-1} + 0.3482\,z^{-2}) \\ &\quad +.2(0.3482 + 0.8704z^{-1} + 0.3482\,z^{-2})^2 \end{aligned} \tag{12.21}$$

For the nonlinear channel $H_2(z)$, the proposed RTCS equaliser results a significant 2 dB gain in SNR level at a prefixed BER of 10^{-4} over the CRNN equaliser in Fig. 12.4 which clearly justifies their application for such type of channel. RTCS equaliser in Fig. 12.5 shows distinct SNR gains of about 4 dB at a prefixed BER level of 10^{-4} over a conventional RNN equaliser in channel $H_3(z)$. For all the examples proposed structure performance is approaching the optimal Bayesian equaliser. Further it is noticed that increasing the number of training samples of the conventional RNN equaliser to 1,000 samples does not yield comparable performance.

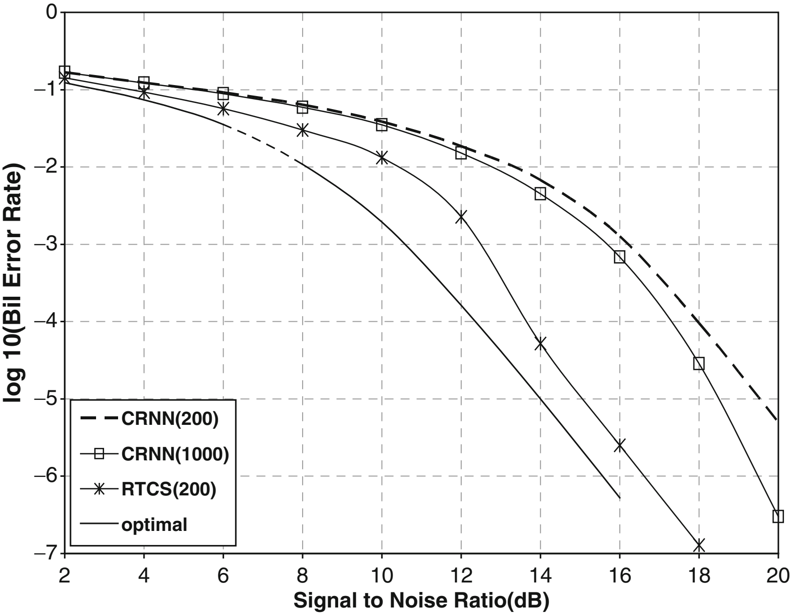


Fig. 12.4 BER performance of the proposed hybrid equaliser for channel $H_2(z)$

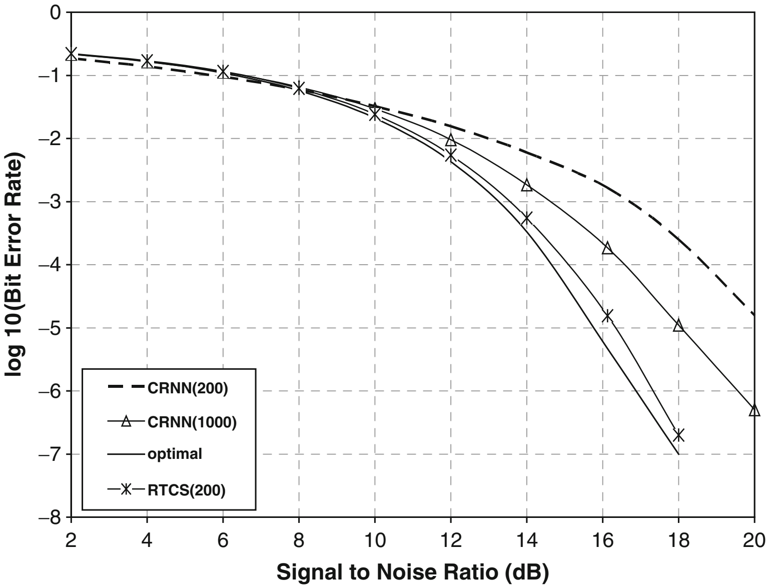


Fig. 12.5 BER performance of the proposed hybrid equaliser for channel $H_3(z)$

12.4 Conclusion

A real-valued transform is a powerful signal decorrelator which performs whitening of the signal by causing the eigen value spread of an auto-correlation matrix to reduce. The proposed neural equalisers with hybrid structures have outperformed their conventional counterparts to a large limit and require less number of samples in training phase simultaneously. The basic objective of this research of developing reduced network configurations remains and hence, while cascading is employed, it is ensured that under no circumstances this main purpose be defeated. It is interesting to note that recurrent neural structure with two nodes cascaded with a 2×2 DCT block with power normalization can outperform the conventional equaliser. As Bit Error Rate performance is a significant measure of channel equalization and proposed hybrid neural structure has an edge over conventional ones and even it is observed that it is close to the theoretically optimal Bayesian equalisers. Further a reduced structure has low computational complexity. Hence this hybrid ANN architecture has opened up new directions in designing efficient adaptive nonlinear equalisers and can be implemented in DSP processors for real – time applications.

References

1. Gibson, G.J., Siu, S., Chen, S., Cowan, C.F.N., Grant, P.M.: The application of nonlinear architectures to adaptive channel equalisation. Proceeding ICC'90 (Atlanta, GA), pp. 312.8.1–312.8.5, April 1990
2. Theodoridis, S., Cowan, C.F.N., Callender, C.P., See, C.M.S.: Schemes for equalisation of communication channels with non-linear impairments. IEEE Proc. Commun. **142**(3), 165–171 (June 1995)
3. Haykin, S.: Neural Networks – A Comprehensive Foundation. Macmillan, New York (1994)
4. Chen, S., Gibson, G.J., Cowan, C.F.N.: Adaptive channel equalisation using a polynomial perceptron structure. IEEE Proc. I Commun. Speech Vision **137**(5), 257–264 (October 1990)
5. Chen, C.N., Chen, K.H., Chiueh, T.D.: Algorithm and architecture design for a low complexity adaptive equaliser. IEEE Int. Symp. Circ. Syst. ISCAS'03 **2**, 304–307 (May 2003)
6. Kechriotis, G., Zervas, E., Manolakos, E.S.: Using recurrent neural networks for adaptive communication channel equalisation. IEEE Trans. Neural Networks **5**(2), 267–278 (1994)
7. Choi, J., Bouchard, M., Yeap, T.H.: Decision feedback recurrent neural equalization with fast convergence rate. IEEE Trans. Neural Network **16**(3), 699–708 (2005)
8. Parisi, R., Di Claudio, E.D., Orlandi, G., Rao, B.D.: Fast adaptive digital equalisation by recurrent neural networks. IEEE Trans. Signal Process **45**(11), 2731–2739 (November 1997)
9. Ortiz-Fuentes, J.D., Forcada, M.L.: A comparison between recurrent neural network architectures for digital equalisation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97) **4**, 3281–3284, 21–24, April 1997
10. Jiang, H.R., Kwak, K.S.: On modified complex recurrent neural network adaptive equaliser. J. Circ. Syst. Comput. **11**(1), 93–101 (2002)
11. Wang, X., Lin, H., Lu, J., Yahagi, T.: Combining recurrent neural networks with self-organizing map for channel equalisation. IEICE Trans. Commun. **E85-B**, 2227–2235 (October 2002)
12. Beaufays, F.: Transform-domain adaptive filters: an analytical approach. IEEE Trans. Signal Process. **43**(2), 422–431 (February 1995)
13. Marshall, D.F., Jenkins, W.K., Murphy, J.J.: The use of orthogonal transforms for improving performance of adaptive filters. IEEE Trans. Circ. Syst. **36**(4), 474–484 (April 1989)

Chapter 13
Embedding Interconnection Networks in Crossed Cubes

Emad Abuelrub

Abstract The hypercube parallel architecture is one of the most popular interconnection networks due to many of its attractive properties and its suitability for general purpose parallel processing. An attractive version of the hypercube is the crossed cube. It preserves the important properties of the hypercube and most importantly reduces the diameter by a factor of 2. In this chapter, we show the ability of the crossed cube as a versatile architecture to simulate other interconnection networks efficiently. We present new schemes to embed complete binary trees, complete quad trees, and cycles into crossed cubes.

Keywords Binary trees · crossed cubes · cycles · embedding · interconnection networks · quad trees

13.1 Introduction

Hypercube architectures are loosely coupled parallel processors based on the binary cube network. Parallel computers based on the hypercube topology have gained widespread acceptance in parallel computing. Recently, many machines based on the hypercube have been designed and made commercially available. The hypercube offers a rich interconnection topology with large bandwidth, logarithmic diameter, simple routing and broadcasting of data, recursive structure that is naturally suited to divide and conquer applications, and the ability to simulate other interconnection networks with minimum overhead [3, 13, 19]. Due to the popularity of the hypercube, many variations of the hypercube topology have been proposed to improve on its properties and computational power [7, 13, 17, 20]. Efe [6] proposed an attractive version of the hypercube called the crossed cube, where preliminary studies proved

E. Abuelrub (✉)
Department of Computer Science, Faculty of Science and IT,
Zarqa Private University, Zarqa 13132, Jordan
e-mail: abuelrub@zpu.edu.jo

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_13,

that the crossed cube preserves many of the attractive properties of the hypercube and more importantly reduces the diameter by a factor of 2 [1, 2, 4, 6, 8, 11, 15]. This implies that the crossed cube has an advantage over the hypercube when data communication is of major concern. It is well known that for parallel architectures, data communication cost dominates computation cost. Therefore, it is worthwhile to make comparative studies on crossed cubes and other interconnection networks, and explore the advantages provided by them. The problem of embedding one interconnection network into another is very important in the area of parallel computing for portability of algorithms across various architectures, layout of circuits in VLSI, and mapping logical data structures into computer memories.

The importance of binary trees comes from the fact that this class of structures is useful in the solution of banded and sparse systems, by direct elimination, and captures the essence of divide and conquers algorithms. Embedding binary trees into other interconnection networks attracted the attention of many researchers. Barasch et al. have considered embedding complete binary trees into generalized hypercubes [3]. Dingle and Sudborough considered simulation of binary trees and X-trees on pyramid networks [5]. Quad trees are becoming an important technique in the domain of image processing, computer graphics, robotics, computational geometry, and geographic information systems [16, 18]. This hierarchical structure is based on the principle of recursive decomposition, which is similar to divide and conquer methods. Mapping quad trees into other interconnection networks attracted the attention of many researchers [1]. The problem of embedding rings or cycles into other interconnection networks has been studied by many researchers. It is well known that rings can be embedded into hypercubes using cyclic Gray codes [19]. Latifi and Zheng [11] generalized the cyclic Gray code method to embed rings into twisted cubes. Many researchers have addressed the problem of embedding rings into hypercubes in the presence of faults. On the other hand, other researchers addressed the problem of embedding rings into fault-free and faulty topologies [8, 11, 12, 14] or the Hamiltonicity of such structures in fault-free and faulty environments [9, 10, 15].

The remainder of this chapter is organized as follows. In Section 13.2, we establish a few preliminary definitions and notations. Section 13.3 explains straight forward scheme to embed complete binary trees into the crossed cubes. Section 13.4 presents a recursive technique to embed complete quad trees. In Section 13.5, we extend the Gray code scheme to embed cycles into the crossed cube. Finally, Section 13.6 concludes the chapter and discusses some future possible work.

13.2 Definitions and Notations

In this chapter, we use undirected graphs to model interconnection networks. Each vertex represents a processor and each edge a communication link between processors. The embedding of a guest graph $\mathrm{G} = (\mathrm{V_G}, \mathrm{E_G})$ into a host graph $\mathrm{H} = (\mathrm{V_H}, \mathrm{E_H})$ is an injective mapping f from $\mathrm{V_G}$ to $\mathrm{V_H}$, where $\mathrm{V_G}$, $\mathrm{E_G}$ and $\mathrm{V_H}$, $\mathrm{E_H}$ are the vertex and edge sets of G and H, respectively, and where $|V_H| \geq |V_G|$. We consider a complete binary tree of height $n-1$, a complete quad tree of height $n-1$, and a cycle of size n, denoted CB_n, CQ_n, and C_n, respectively, as guest graphs and a crossed cube of dimension n, denoted XQ_n, as a host graph. Two cost functions, dilation and expansion often measure the quality of an embedding. If u and v are two nodes in G, then the distance from u to v, $d = (u, v)$, is the length of the shortest path from u to v. The *Dilation* (D) is the maximum distance in H between the images of adjacent vertices of G, $D = \max\{d(f(u), f(v))$, where $u\text{–}v \in E_G\}$. The *Expansion* (E) is the ratio of the cardinality of the host vertex set to the cardinality of the guest vertex set, $E = |V_H|/|V_G|$. Minimizing each of these measurements has a direct implication on the quality of the simulation of the guest network by the corresponding host network. The dilation of an embedding measures how far apart neighboring guest processors are placed in the host network. Clearly, if adjacent guest processors are placed far apart in the host network, then there will be a significant degradation in simulation due to the long length of the communication path between them. The expansion of an embedding measures how much larger is the host network than the guest network during the simulation. We want to minimize expansion, as we want to use the smallest possible host network that has at least as many processors as in the guest network. In reality, we usually have a fixed size host network and we may have to consider many-to-one embedding for larger guest networks. When the size of the guest network is not equal to the size of the host network in terms of the number of processors, then we try to find the smallest host network that has at least as many processors as the guest network. Such a host network is referred to as the *optimal host network*. There is a trade off between dilation, which measures the communication delay, and expansion, which measures processor utilization, such that one can achieve lower expansion at a cost of greater dilation and vice versa.

A *hypercube* of dimension n, denoted by Q_n, is an undirected graph consisting of 2^n vertices labeled from 0 to $2^n - 1$ and such that there is an edge between any two vertices if and only if the binary representation of their labels differs in exactly one bit position. A *complete binary tree* of height $n-1$, denoted by CB_n, is an undirected graph consisting of $2^n - 1$ vertices and such that every vertex of depth less than $n-1$ has exactly two sons and every vertex of depth $n-1$ is a leaf. A *complete quad tree* of height $n-1$, denoted by CQ_n, is an undirected graph consisting of $(4^n - 1)/3$ vertices and such that every vertex of depth less than $n-1$ has exactly four sons and every vertex of depth $n-1$ is a leaf. A *cycle* of size n, denoted C_n, is an undirected graph consisting of n vertices labeled from v_1 to v_n, such that node v_i is a neighbor with node $v_{(I+1)\bmod n}$, $1 \leq i \leq n$. A *path* $(v_o, v_1, v_2, \ldots, v_{n-1})$ is a sequence of nodes such that each two consecutive nodes are adjacent. A path in a graph G is a *Hamiltonian path* if all its nodes are distinct and they span G. A cycle or a circuit is called a *Hamiltonian circuit* if it traverses every node of G exactly once.

The crossed cube is defined recursively as follows. Let G be any undirected labeled graph, then G^b is obtained from G by prefixing every vertex label with b. Two binary strings $x = x_1x_0$ and $y = y_1y_0$, each of length two, are *pair-related* if and only if $(x, y) \in \{(00, 00), (10, 10), (01, 11), (11, 01)\}$. Now, we define a *crossed*

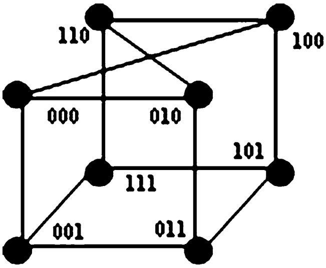


Fig. 13.1 The crossed cube XQ_3

cube of dimension n, denoted XQ_n, as an undirected graph consisting of 2^n vertices labeled from 0 to $2^n - 1$ and defined recursively as following:

1. XQ_1 is the complete graph on two vertices with labels 0 and 1.
2. For $n > 1$, XQ_n consists of two copies of XQ_{n-1} one prefixed by 0, $XQ^0{}_{n-1}$, and the other by 1, $XQ^1{}_{n-1}$. Two vertices $u = 0u_{n-2}\ldots u_0 \in XQ^0{}_{n-1}$ and $v = 1v_{n-2}\ldots v_0 \in XQ^1{}_{n-1}$ are adjacent, if and only if:

   (a) $u_{n-2} = v_{n-2}$, if n is even.
   (b) For $0 \leq i \leq \lfloor (n-1)/2 \rfloor$, $u_{2i+1}u_{2i}$ and $v_{2i+1}v_{2i}$ are pair-related.

Figure 13.1 shows crossed cubes of dimension 3. XQ_n is constructed recursively based on the construction of XQ_{n-1} by pasting together a copy of $XQ^0{}_{n-1}$ and the mirror image of $XQ^1{}_{n-1}$, then adding the appropriate links between the two copies according to the pair-related relationship.

13.3 Embedding Complete Binary Trees

This section describes our scheme to embed a complete binary tree CB_n into a crossed cube XQ_n with dilation two and unit expansion. Our scheme is based on the inorder labeling to embed CB_n into XQ_n in a straight forward way. The inorder embedding is constructed by Algorithm Embedding Complete Binary Tree (ECBT).

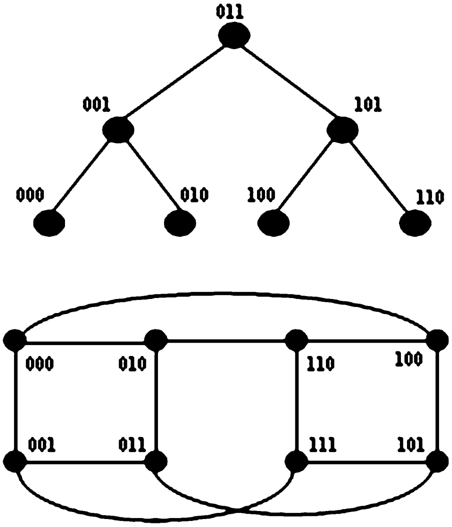
Algorithm ECBT

```
1. Begin
2. Label the nodes of the complete binary tree based on the inorder traversal using
   binary representation
3. Map each node of the complete binary tree to the node in the crossed cube with
   the corresponding binary representation
4. End
```

Theorem 13.1. *For all n, the inorder labeling of the complete binary tree embeds* CB_n *within the crossed cube* XQ_n *with dilation two.*

Prove of the theorems presented in this chapter is omitted due to the limited space. However, the reader can refer to [1] for more details. As an illustration to

Fig. 13.2 The inorder embedding of CB_3 into XQ_3



the resulted embedding, in the lowest level, each edge from a left child to its parent is mapped to the corresponding crossed cube edge between the images of the two nodes, while the edge between a right child to its parent is mapped to a path of length two, from the right child to the left child and from the left child to the parent. In the higher level, each edge from a left child, or a right child, to its parent is mapped to the corresponding crossed cube edge between the images of the two nodes. In all higher levels, each edge from a left child, or a right child, to its parent is mapped to a path of length two. Notice that the inorder embedding is very simple and straight forward, as shown in Fig. 13.2.

13.4 Embedding Complete Quad Trees

This section describes our recursive scheme to embed a complete quad tree CQ_n into its optimal crossed cube XQ_{2n-1} with dilation two and unit expansion. We proceed in four steps. In the first step, CQ_n is decomposed into a four complete quad sub trees; a left complete quad sub tree ACQ_{n-1} with root a, a left middle complete quad sub tree BCQ_{n-1} with root b, a right middle complete quad sub tree CCQ_{n-1} with root c, a right complete quad sub tree DCQ_{n-1} with root d, and a root r. In the second step, XQ_{2n-1} is decomposed into four sub cubes $XQ^{00}{}_{2n-3}$, $XQ^{01}{}_{2n-3}$, $XQ^{11}{}_{2n-3}$, and $XQ^{10}{}_{2n-3}$. In the third step, ACQ_{n-1} is embedded into $XQ^{00}{}_{2n-3}$, BCQ_{n-1} is embedded into $XQ^{01}{}_{2n-3}$, CCQ_{n-1} is embedded into $XQ^{11}{}_{2n-3}$, DCQ_{n-1} is embedded into $XQ^{10}{}_{2n-3}$, and the root r is embedded into one of the unused nodes in $XQ^{00}{}_{2n-3}$. In the last step, we construct CQ_n by finding the paths $r \sim a$, $r \sim b$, $r \sim c$, and $r \sim d$, each of at most length two. The embedding process is continued recursively by decomposing the complete quad sub trees and the crossed sub cubes, repeating the above steps, until we reach the leaves of the complete quad tree. At the bottom level of the complete quad tree, each complete quad sub tree with five nodes is mapped into a crossed sub cube of dimension 3. Next, we present Algorithm Embed Complete Quad Tree (ECQT) that uses a recursive divide and conquers technique to embed a complete quad tree $\mathrm{CQ_n}$ into its optimal crossed cube $\mathrm{XQ_{2n-1}}$.

Algorithm ECQT

Let δ_i be the binary string of length n with a 1 *in position i and* 0 *in all other positions, θ_k be the binary string of length k with 0 in all positions, and $\oplus$ be the xor operator.*

1. *Begin*
2. Decompose $\mathrm{CQ_n}$ to $\mathrm{ACQ_{n-1}}$, $\mathrm{BCQ_{n-1}}$, $\mathrm{CCQ_{n-1}}$, $\mathrm{DCQ_{n-1}}$, and r
3. Decompose $\mathrm{XQ_{2n-1}}$ *to* $\mathrm{XQ^{00}{}_{2n-3}}$, $\mathrm{XQ^{01}{}_{2n-3}}$, $\mathrm{XQ^{11}{}_{2n-3}}$, and $\mathrm{XQ^{10}{}_{2n-3}}$
4. Map the quad sub trees into the crossed sub cubes as follows:
   (a) *Embed ACQ_{n-1} into $XQ^{00}{}_{2n-3}$, BCQ_{n-1} into $XQ^{01}{}_{2n-3}$, CCQ_{n-1} into $XQ^{11}{}_{2n-3}$, and DCQ_{n-1} into $XQ^{10}{}_{2n-3}$. a, b, c, and d will appear at addresses* $000\theta_{2n-4}$, $010\theta_{2n-4}$, $110\theta_{2n-4}$, *and* $100\theta_{2n-4}$, *respectively*
   (b) *Translate the embeddings in $XQ^{00}{}_{2n-3}$ and $XQ^{10}{}_{2n-3}$ by complementing the* $(2n-3)$th *bit of each node. Formally, if a tree node was mapped to address x then after the translation it will appear at address* $\mathrm{x} \oplus \delta_{2n-3}$. *After the translation, the left root a and the right root d will appear at addresses* $001\theta_{2n-3}$ *and* $101\theta_{2n-4}$, *respectively. Therefore, the final position of a, b, c, and d are* $001\theta_{2n-4}$, $010\theta_{2n-4}$, $110\theta_{2n-4}$, *and* $101\theta_{2n-4}$, *respectively*
   (c) *Map the root r into the node with label* 0 *in* $XQ^{00}{}_{2n-3}$
5. *Construct CQ_n from ACQ_{n-1}, BCQ_{n-1}, CCQ_{n-1}, DCQ_{n-1}, CQ_n, and r by finding the four paths $r \sim a, r \sim b, r \sim c$, and $r \sim d$. The edges r-a and r-b of CQ_n are mapped to paths of length one in XQ_{2n-1}, while the edges r-c and r-d are mapped to paths of length two. The shortest paths from r to c and from r to d are* $000\theta_{2n-4} - 010\theta_{2n-4} - 110\theta_{2n-4}$ *and* $000\theta_{2n-4} - 100\theta_{2n-4} - 101\theta_{2n-4}$, *respectively*
6. *End*

Theorem 13.2. *For all n, Algorithm ECQT maps the complete quad tree* $\mathrm{CQ_n}$ *within the crossed cube* $\mathrm{XQ_{2n-1}}$ *with dilation two and unit expansion (Figs. 13.3 and 13.4).*

13.5 Embedding Cycles

Given a cycle $\mathrm{C_{2^n}}$ with 2^{n} nodes, consider the problem of assigning the cycle nodes to the nodes of the crossed cube $\mathrm{XQ_n}$ such that adjacency is preserved. Now, given any two adjacent nodes in the cycle, their images by this embedding should be neighbors in the crossed cube through some dimension i, where $1 \leq \mathrm{i} \leq \mathrm{n}$. We can view such an embedding as a sequence of dimensions crossed

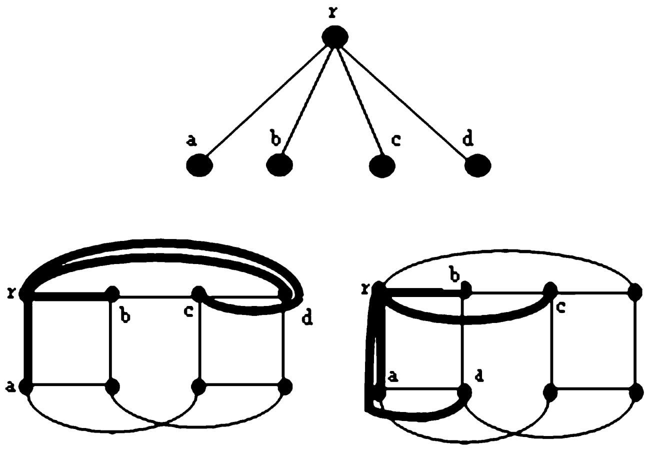


Fig. 13.3 Embedding CQ_1 into the sub cube XQ_3

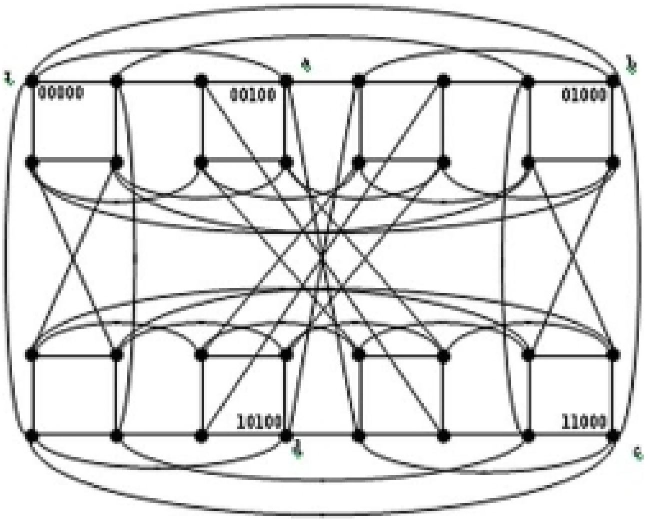


Fig. 13.4 Embedding CQ_3 into XQ_5

by adjacent nodes. We call such a sequence the *embedding sequence*, denoted by $ES = (d_1, d_2, \ldots, d_{2^n})$, where $d_i \in \{1, \ldots, n\}$ for all $1 \leq i \leq 2^n$. Figure 13.5 shows two different embeddings of the cycle C_{2^3} into the crossed cube XQ_3. It is more convenient to view the embedded cycle as will as the crossed cube in the way shown in Fig. 13.5. The embedding sequence of C_{2^3} is $ES = (1, 3, 1, 2, 1, 3, 1, 2)$. For example, in the first part of Fig. 13.5, notice that nodes 000 and 001 are connected by a link through dimension 1, 001 and 111 are connected by a link through

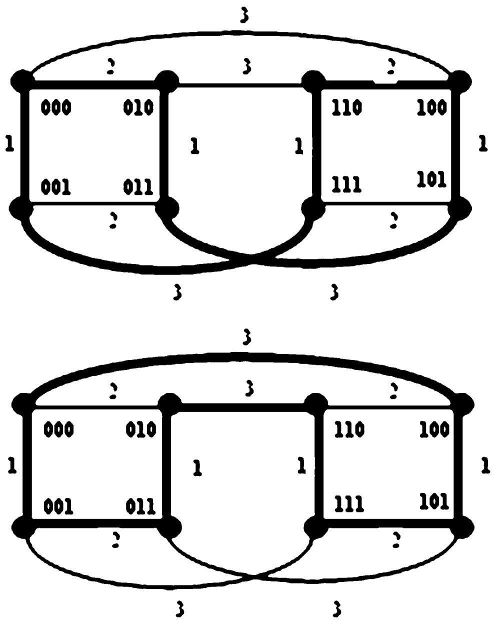


Fig. 13.5 The embedding sequence

dimension 3, 111 and 110 are connected by a link through dimension 1, 110 and 100 are connected by a link through dimension 2, and so on. The embedding sequence ES can be generated using Algorithm ES.

Algorithm ES

Let n be the dimension of the crossed cube and let the vertical bar be the concatenation operator.

```
1. Begin
2. $ES \leftarrow 1$
3. For $i \leftarrow 3$ to $n$ do
4. $ES \leftarrow ES|i|ES$
5. $ES \leftarrow ES|2|ES|2$
6. End
```

The embedding sequence is generated by applying Algorithm ES on n, where n is the dimension of the crossed cube. The number of nodes in the crossed cube is equal to the number of nodes in the embedded cycle, which is 2^n nodes. Thus, the embedding sequence of the cycle C_{2^4} is ES $= (1, 3, 1, 4, 1, 3, 1, 2, 1, 3, 1, 4, 1, 3, 1, 2)$ and the embedding sequence of the cycle C_{2^5} is ES $= (1, 3, 1, 4, 1, 3, 1, 5, 1, 3, 1, 4, 1, 3, 1, 2, 1, 3, 1, 4, 1, 3, 1, 5, 1, 3, 1, 4, 1, 3, 1, 2)$. The embedding sequence corresponds to the extended binary-reflected Gray code embedding of a cycle into a crossed cube. The binary-reflected Gray code is the most common technique to embed a cycle into a fault-free hypercube. Notice that the same embedding sequence may result in different embeddings of C_2n into XQ_n depending on the crossed cube node that initiates the cycle construction. Among all different embeddings, we are interested in one kind. The embedding when the node that initiates the cycle construction in the crossed cube is the upper leftmost node, node with label 0. This will not violate the generalization of the technique since the crossed cube is node and vertex symmetric [9], which means that we can relabel the nodes, where any node can be labeled as node 0, and hence initiates the construction of the cycle. In Fig. 13.5, the cycle is initiated by node 000 in the first part, while it is initiated by node 001 in the second part.

Theorem 13.3. *For every n, Algorithm ES will generate the embedding sequence to construct a cycle of size* 2^{n} *in a crossed cube of dimension n.*

Next, we present Algorithm Hamiltonian Circuit (HC) that uses a recursive divide and conquers technique to embed a cycle $\mathrm{C}_{2^{\mathrm{n}}}$ into a crossed cube XQ_{n}.

Algorithm HC

```
1. Begin
2. Partition $XQ_n$ into $2^{n-3}$ disjoint crossed cubes, each of dimension 3
3. Embed the cycle $C_{2^3}$ into each sub cube using the embedding sequence $ES =$
   $(1, 3, 1, 2, 1, 3, 1, 2)$
4. Connect the $2^{n-3}$ cycles, each of size 8, through the upper, or lower, links to come
   up with a cycle of size $C_{2^n}$
5. End
```

Theorem 13.4. *For every n, Algorithm HC will embed a Hamiltonian cycle of size* 2^{n} *in a crossed cube of dimension n.*

Note the use of the upper links of dimension n when the embedding sequence is generated by node 0, while the lower crossed links of dimension n are used when the embedding sequence is generated by node 1, as shown in Fig. 13.6.

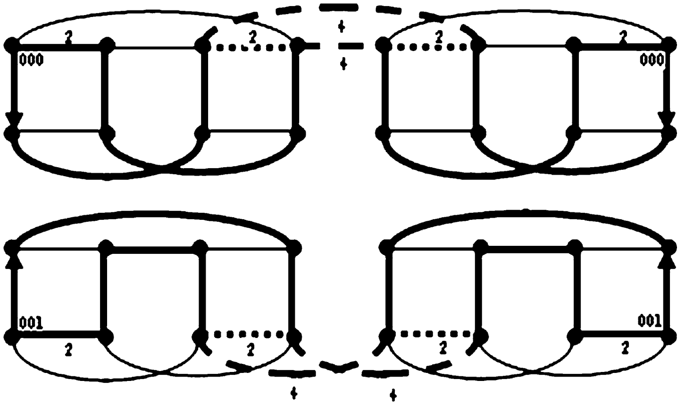


Fig. 13.6 The recursive construction of the cycle C_{2^n} in a fault-free environment

13.6 Conclusions and Future Work

This chapter has presented different schemes to show the ability of the crossed cube as a versatile architecture to simulate other interconnection networks efficiently. We present new schemes to embed complete binary trees, complete quad trees, and cycles into crossed cubes. A good problem will be to improve the dilation on embedding trees into crossed cubes. Another interesting problem is to generalize the schemes to embed trees and cycles in the presence of faults.

References

1. Abuelrub, E. Embeddings into crossed cubes. Proceedings of the World Congress on Engineering (WCE'2009), **1**, London (2009)
2. Abuelrub, E. The hamiltonicity of crossed cubes in the presence of faults. Eng. Lett. **16**(3), 453–459 (2008)
3. Barasch, L., Lakshmivarahan, S., Dhall, S.: Embedding arbitrary meshes and complete binary trees in generalized hypercubes. Proceedings of the 1st IEEE Symposium on Parallel and Distributed Processing, pp. 202–209, 1989
4. Chang, C., Sung, T., Hsu, L.: Edge congestion and topological properties of crossed cubes. IEEE Trans. Parall. Distrib. Syst. **11**(1), 64–80 (2006)
5. Dingle, A., Sudborough, I.: Simulating binary trees and x-trees on pyramid networks. Proceedings of the 1st IEEE Symposium on Parallel and Distributed Processing, pp. 210–219 (1989)
6. Efe, K.: The crossed cube architecture for parallel computation. IEEE Trans. Parall. Distrib. Syst. **3**(5), 513–524 (1992)
7. El-Amaway, A., Latifi, S.: Properties and performance of folded hypercubes. IEEE Trans. Parall. Distrib. Syst. **2**(1), 31–42 (1991)
8. Fan, J., Lin, X., Jia, X.: Optimal path embedding in crossed cubes. IEEE Trans. Parall. Distrib. Syst. **16**(12), 1190–1200 (2005)
9. Fu, J., Chen, G.: Hamiltonicity of the hierarchical cubic network. Theory Comput. Syst. **35**, 59–79 (2008)
10. Keh, H., Lin, J.: On fault-tolerance embedding of hamiltonian cycles, linear arrays, and rings in a flexible hypercube. Parall. Comput. **26**(6), 769–781 (2000)
11. Latifi, S., Zheng, S.: Optimal simulation of linear array and ring architectures on multiply-twisted hypercube. In: Proceedings of the 11th International IEEE Conference on Computers and Communications, 1992
12. Lee, S., Ho, H.: A 1.5 approximation algorithm for embedding hyperedges in a cycle. IEEE Trans. Parall. Distrib. Syst. **16**(6), 481–487 (June 2005)
13. Leighton, T.: Introduction to Parallel Algorithms and Architecture: Arrays, Trees, Hypercubes. Morgan Kaufmann, San Mateo, CA (1992)
14. Lin, J.: Embedding hamiltonian cycles, linear arrays, and rings in a faulty supercube. Int. J. High Speed Comput. **11**(3), 189–201 (2000)
15. Ma, M., Xu, J.: Panconnectivity of locally twisted cubes. Appl. Math. Lett. **17**(7), 674–677 (2006)
16. Markas, T., Reif, J.: Quad tree structures for image compression applications. Inform. Process. Lett. **28**(6), 707–722 (1992)
17. Preparata, F., Vuillemin, J.: The cube-connected cycles: a versatile network for parallel computation. Commun. ACM. **24**(5), 3000–3309 (1981)
18. Topi, L., Parisi, R., Uncini, A.: Spline recurrent neural network for quad-tree video coding. Proceedings of WIRN'2002, pp. 90–98 (2002)

19. Saad, Y., Schultz, M.: Topological properties of the hypercube. IEEE Trans. Comput. **37**(7), 867–872 (July 1988)
20. Youyao, L.: A hypercube-based scalable interconnection network for massively parallel computing. J. Comput. **3**(10) (October 2008)

Chapter 14
Software Fault Tolerance: An Aspect Oriented Approach

Kashif Hameed, Rob Williams, and Jim Smith

Abstract Software fault tolerance demands additional tasks like error detection and recovery through executable assertions, exception handling, diversity and redundancy based mechanisms. These mechanisms do not come for free; rather they introduce additional complexity to the core functionality. This paper presents light weight error detection and recovery mechanisms based on the rate of change in signal or data values. Maximum instantaneous and mean rates are used as plausibility checks to detect erroneous states and recover. These plausibility checks are exercised in a novel aspect oriented software fault tolerant design framework that reduces the additional logical complexity. A Lego NXT Robot based case study has been completed to demonstrate the effectiveness of the proposed design framework.

Keywords Aspect oriented design and programming · executable assertions · exception handling · fault tolerance · plausibility checks

14.1 Introduction

Adding fault tolerance measures to safety critical and mission critical applications introduces additional complexity to the core application. By incorporating handler code, for error detection, checkpointing, exception handling, and redundancy/diversity management, the additional complexity may adversely affect the dependability of a safety critical or mission critical system.

One of the solutions to reduce this complexity is to separate and modularize the extra, cross-cutting concerns from the true functionality.

At the level of design and programming, several approaches have been utilized that aim at separating functional and non-functional aspects. Component level

K. Hameed (✉), R. Williams, and J. Smith
University of the West of England, Bristol Institute of Technology, BS16 1QY, UK
e-mail: Kashif3.Hameed@uwe.ac.uk; Rob.Williams@uwe.ac.uk; james.smith@uwe.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_14,

approach like IFTC [1], computational reflection and meta-object protocol based MOP [2] have shown that dependability issues can be implemented independently of functional requirements.

The evolving area of Aspect-Oriented Programming & Design (AOP&D) presents the same level of independence by supporting the modularized implementation of crosscutting concerns.

Aspect-oriented language extensions, like AspectJ [3] and AspectC++ [4] provide mechanisms like *Advice* (behavioural and structural changes) that may be applied by a pre-processor at specific locations in the program called *join point.* These are designated by *pointcut* expressions. In addition to that, static and dynamic modifications to a program are incorporated by *slices* which can affect the static structure of classes and functions.

In the context of fault tolerance, an induced fault can activate an error that changes the behaviour of the program and may lead to system failure. In order to tolerate a fault, abnormal behaviour must be detected and transformed back by introducing additional behaviour changes (Exception Handler) or alternate structure adoption (Recovery Blocks, N-Version Programming) strategies.

The rate of change (ROC) of signals or data can be used to detect erroneous conditions that can help in tolerating faults and avoiding failures by triggering appropriate recovery mechanisms. ROC-based plausibility checks for error detection and recovery in the form of executable assertions have been addressed by Hiller in [5, 6] . In [7] the author utilizes dynamic signal values for modeling and predicting future sensor values. Unfortunately, these mechanisms will add to the complexity of the true functionality that could affect the overall dependability of the system. None of the previous studies propose the separation of these error handling concerns from true functionality. However Aspect Oriented Design and Programming approaches may be used to separate out these concerns from the true functionality of a computer based system.

In this paper the rate of change based executable assertions have been extended with more refined time bounded instantaneous and mean rate checks that reduce false positives and false negatives. Secondly an empirical method for determining the maximum instantaneous and mean rates of change has been devised.

The current work also proposes generalized aspect-oriented software fault tolerance design patterns. These design solutions provide an implementation framework to incorporate and validate the proposed ROC-based checks.

14.2 ROC Plausibility Based Error Detection and Recovery

Error detection is the basic step in deploying any fault tolerance strategy. Executable assertions are often utilized as an error detection mechanism. Rate of change (ROC) based plausibility checks on input and output data may be used to detect some erroneous conditions that could lead to failure. Although ROC based executable assertions have been addressed by Hiller in [5], these constraints are based on changes

Table 14.1 ROC assertions and recovery

ROC assertion (PC)	Recovery mechanism
Case: $\mathbf{y_i} > \mathbf{y_{i-1}}$ (increasing) PC1: $\frac{\mathbf{y_i - y_i - 1}}{\mathbf{t_i - t_{i-1}}} \le \mathbf{r}_{\max-\mathbf{incr}}$	$\mathbf{y_r} = \mathbf{y_{i-1}} + \mathbf{r}_{\max-\mathbf{incr}} \Delta \mathbf{T_{i-1 \to i}}$
Case: (PC1 & $\mathbf{y_i} < \mathbf{y}_{\max}$) PC2: $\frac{\mathbf{y_i - y_{i-1}}}{\mathbf{t_i - t_{i-1}}} \ge \mathbf{r}_{\min-\mathbf{incr}}$	If $\mathbf{y}_{\max} - \mathbf{y_{i-1}} \ge \mathbf{r}_{\min-\mathbf{incr}} \Delta \mathbf{T_{i-1 \to i}}$ then $\mathbf{y_r} = \mathbf{y_{i-1}} + \mathbf{r}_{\min-\mathbf{incr}} \Delta \mathbf{T_{i-1 \to i}}$ else $\mathbf{y_r} = \mathbf{y}_{\max}$
Case: $\mathbf{y_i} < \mathbf{y_{i-1}}$ (decreasing) PC3: $\frac{\mathbf{y_{i-1}} - \mathbf{y}_i}{\mathbf{t_i - t_{i-1}}} \le \mathbf{r}_{\max-\mathbf{decr}}$	$\mathbf{y_r} = \mathbf{y_{i-1}} - \mathbf{r}_{\max-\mathbf{decr}} \Delta \mathbf{T_{i-1 \to i}}$
Case: **PC**3 & $\mathbf{y_i} > \mathbf{y}_{\min}$ PC4: $\frac{\mathbf{y_{i-1} - y_i}}{\mathbf{t_i - t_{i-1}}} \ge \mathbf{r}_{\min-\mathbf{decr}}$	If $\mathbf{y}_{\min} - \mathbf{y_{i-1}} \ge \mathbf{r}_{\min-\mathbf{decr}} \Delta \mathbf{T_{i-1 \to i}}$ then $\mathbf{y_r} = \mathbf{y_{i-1}} - \mathbf{r}_{\min-\mathbf{decr}} \Delta \mathbf{T_{i-1 \to i}}$ else $\mathbf{y}_\mathrm{r} = \mathbf{y}_{\min}$

in variable values but without any bound on time. However true rate of change should employ the change in variable values in a specified time interval as asserted by Clegg [7]. Without considering a time boundary, there are more chances to have false positives and false negatives.

In order to apply various plausibility checks, it is first necessary to determine the characteristic range of values for key variables/signals. The characteristic parameters of variables assigned here are $y_{\max}$(maximum value), $y_{\min}$(minimum value), $r_{\max\text{-incr}}$(maximum increase/sample time), $r_{\min\text{-incr}}$(minimum increase/sample time), $r_{\max\text{-decr}}$(maximum decrease/sample time), $r_{\min\text{-decr}}$(minimum decrease/sample time).

When an error is detected a recovery mechanism is brought into service to avoid a failure and so tolerate the fault. The recovery mechanisms employed here are managed on the basis of running trends. The faulty data is replaced by computed values derived from past values and some increment based on the maximum and minimum rates of change. However, the forcefully assigned values are kept within the maximum and minimum data ranges as tabulated in Table 14.1.

14.3 Aspect Oriented Exception Handling Patterns

Exception handling has been deployed as a key mechanism in implementing software fault tolerance through forward and backward error recovery mechanisms. It provides a convenient means of structuring software that has to deal with erroneous conditions [8].

In [9], the authors addresses the weaknesses of exception handling mechanisms provided by mainstream programming languages like Java, Ada,C + +, C#. In their experience exception handling code is inter-twined with the normal code. This hinders maintenance and reuse of both normal and exception handling code.

Moreover as argued by Romanovsky in [10], exception handling is difficult to develop and has not been well understood. This is due to the fact that it introduces additional complexity and has been misused when applied to a novel application domain. This has further increased the ratio of system failures due to poorly designed fault tolerance strategies.

Thus fault tolerance measures using exception handling should make it possible to produce software where (a) error handling code and normal code are separated logically and physically; (b) the impact of complexity on the overall system is minimized; and (c) the fault tolerance strategy may be maintainable and evolvable with increasing demands of dependability.

In this respect, Garcia et al. [2] have proposed an architectural pattern for exception handling. They address the issues like specification and signaling of exceptions, specification and invocation of handlers and searching of handlers. These architectural and design patterns have been influenced by computational reflection and meta-object protocol.

However, most meta-programming languages suffer performance penalties due to the increase in meta-level computation at run-time. This is because most of the decisions about semantics are made at run-time by the meta-objects, and the overhead to invoke the meta-objects reduces the system performance [11].

Therefore we propose generalized aspect based patterns for monitoring, error detection, exception raising and exception handling using a static aspect weaver. These patterns would lead to integration towards a robust and dependable aspect based software fault tolerance. The following design notations have been used to express aspect-oriented design patterns (Fig. 14.1).

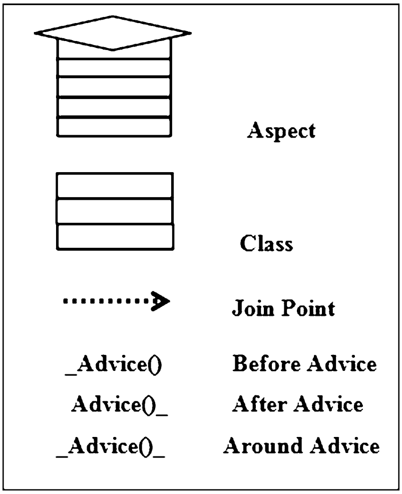
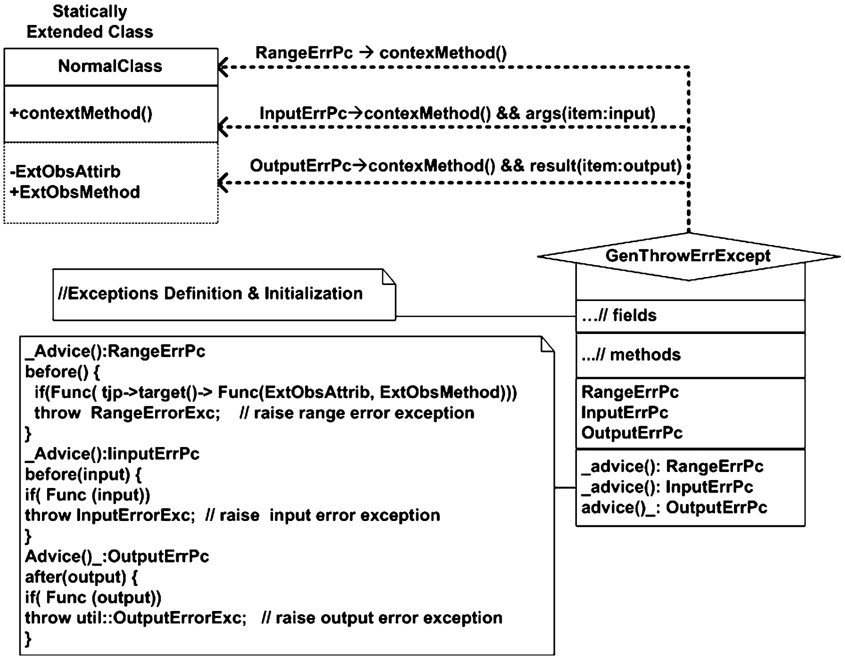


Fig. 14.1 Design notations



```
_Advice():RangeErrPc
before() {
  if(Func( tjp->target()-> Func(ExtObsAttrib, ExtObsMethod)))
  throw  RangeErrorExc;    // raise range error exception
}
_Advice():IinputErrPc
before(input) {
if( Func (input))
throw InputErrorExc; // raise  input error exception
}
Advice()_:OutputErrPc
after(output) {
if( Func (output))
throw util::OutputErrorExc;   // raise output error exception
}
```

Fig. 14.2 Error detection, exception throwing aspect pattern

14.3.1 Error Detection and Exception Throwing Aspect

Error detection and throwing exceptions has been an anchor in implementing any fault tolerance strategy. This aspect detects faults and throws range, input and output type of exceptions. The overall structure of this aspect is shown in Fig. 14.2.

The *GenThrowErrExcept* join points the *NormalClass* via three pointcut expressions for each type of fault tolerance case.

RangeErrPc: this join points the *contexMethod*() only. It initiates a before advice to check the range type errors before executing the *contextMethod*(). Incase the assertions don't remain valid or acceptable behavior constraints are not met, *RaneErrExc* exception is raised.

InputErrPc: this join points the *contextMethod*() further scoped down with input arguments of the *contextMethod*(). It initiates a before advice to check the valid input before the execution of the context method. Incase the input is not valid it raises *InputErrExc*.

OutputErrPc: this join points the *contextMethod*() further scoped down with results as output of the *contextMethod*(). It initiates an after advice to check the valid output after the execution of the context method. Incase the output is not valid it raises **OutputErrExc**.

14.3.2 ROC Plausibility Check Aspect

This aspect is responsible for checking the erroneous state of the system based on the rate of change in critical signal/data values. Once an erroneous state is detected, the respective exception is raised. Various exceptions are also defined and initialized in this aspect. The *pointcut GetSensorData* defines the location where error checking plausibility checks are weaved whenever a critical data/sensor reading function is called. The light weight ROC-based plausibility assertions are executed in the *advice* part of this aspect.

14.3.3 Catcher Handler Aspect

The *CatcherHandler* aspect as shown below is responsible for identifying and invoking the appropriate handler. This pattern addresses two run-time handling strategies.

The first strategy is designated by an *exit_main* pointcut expression. It checks the run-time *main*() function for various fatal error exceptions and finally aborts or exits the main program upon error detection. This aspect may be used to implement safe shut-down or restart mechanisms in safety critical systems to ensure safety, if a fatal error occurs or safety is breached.

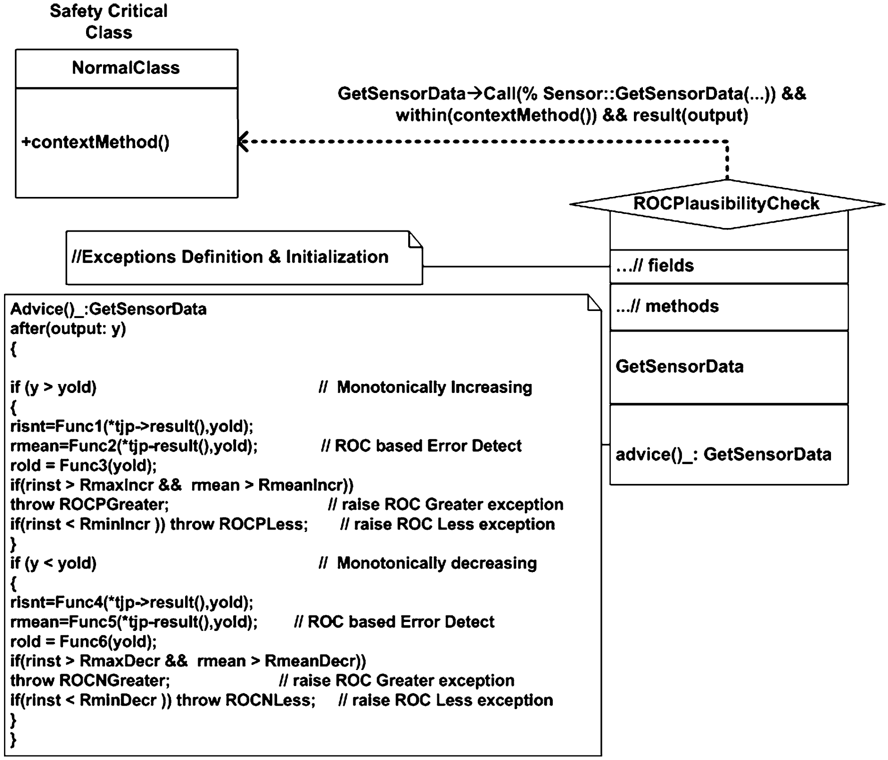
The second strategy returns from the called function as soon as the error is detected. The raised exception is caught after giving warning or doing some effective action in the catch block. This can help in preventing error propagation. Using this aspect, every call to critical functions is secured under a try/catch block to ensure effective fault tolerance against an erroneous state.

It can be seen in the diagram below that *exit_main* pointcut expression join points the main() run-time function. Whereas *caller_return* pointcut expression join points every call to the *contextMethod*(). Moreover *exit_main* and *caller_return* pointcut expressions are associated with an around advice to implement error handling. The tjp→proceed() allows the execution run-time main() and called functions in the try block.

The **advice** block of the catcher handler identifies the exception raised as a result of in-appropriate changes in the rate of signal or data. Once the exception is identified, the recovery mechanism is initiated that assign new values to signal or data variables based on previous trends or history of the variable (Figs. 14.3 and 14.4).

14.3.4 Dynamics of Cather Handler Aspect

This scenario shows an error handling aspect. It simulates two error handling strategies. In the first case, control is returned from the caller to stop the propagation of errors along with a system warning. In the second case the program exits due to a


Safety Critical Class
NormalClass
+contextMethod()
GetSensorData→Call(% Sensor::GetSensorData(...)) && within(contextMethod()) && result(output)
ROCPlausibilityCheck
...// fields
...// methods
GetSensorData
advice()_: GetSensorData
//Exceptions Definition & Initialization

```
Advice()_:GetSensorData
after(output: y)
{

if (y > yold)                              //  Monotonically Increasing
{
risnt=Func1(*tjp->result(),yold);
rmean=Func2(*tjp-result(),yold);           // ROC based Error Detect
rold = Func3(yold);
if(rinst > RmaxIncr &&  rmean > RmeanIncr))
throw ROCPGreater;                         // raise ROC Greater exception
if(rinst < RminIncr )) throw ROCPLess;     // raise ROC Less exception
}
if (y < yold)                              //  Monotonically decreasing
{
risnt=Func4(*tjp->result(),yold);
rmean=Func5(*tjp-result(),yold);       // ROC based Error Detect
rold = Func6(yold);
if(rinst > RmaxDecr &&  rmean > RmeanDecr))
throw ROCNGreater;                     // raise ROC Greater exception
if(rinst < RminDecr )) throw ROCNLess;    // raise ROC Less exception
}
}
```

Fig. 14.3 ROC plausibility aspect pattern structure

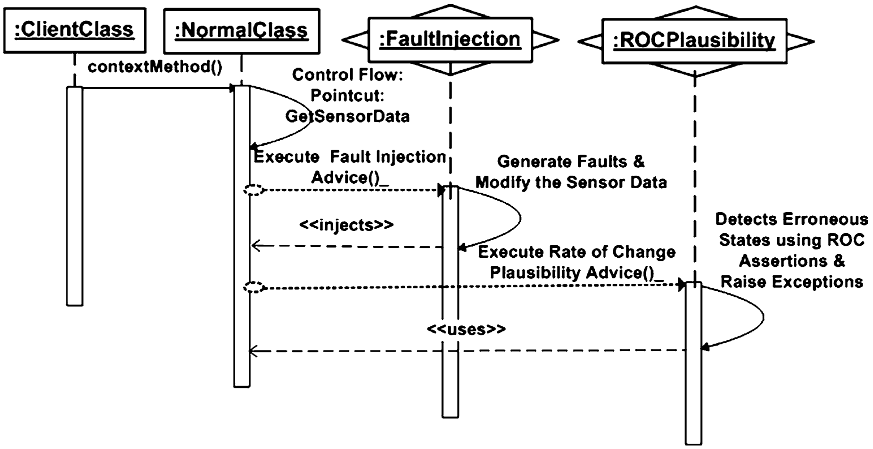

:ClientClass
:NormalClass
:FaultInjection
:ROCPlausibility
contextMethod()
Control Flow: Pointcut: GetSensorData
Execute Fault Injection Advice()_
Generate Faults & Modify the Sensor Data
<<injects>>
Execute Rate of Change Plausibility Advice()_
Detects Erroneous States using ROC Assertions & Raise Exceptions
<<uses>>

Fig. 14.4 ROC aspect pattern dynamics

fatal error. This may be used to implement shutdown or restart scenarios. Moreover the extension of a class member function with a *try* block is also explained.

1. A client object invokes the *contextMethod*() on an instance of *NormalClass*.
2. The control is transferred to *CatcherHandler* aspect that extends the *contextMethod()* by wrapping it in a *try* block and executes the normal code.
3. In case an exception is raised by previous aspect, the exception is caught by the *CatcherHandler* aspect. This is shown by the catch message. The condition shows the type of exception *e* to be handled by the handler aspect.
4. *CatcherHandler* aspect handles the exception e. the caller_return strategy warns or signals the client about the exception and returns from the caller. The client may invoke the *contextMethod2*() as appropriate. In exit_main strategy, the control is retuned to client that exits the current instances as shown by the life line end status (Fig. 14.5).

14.4 Case Study

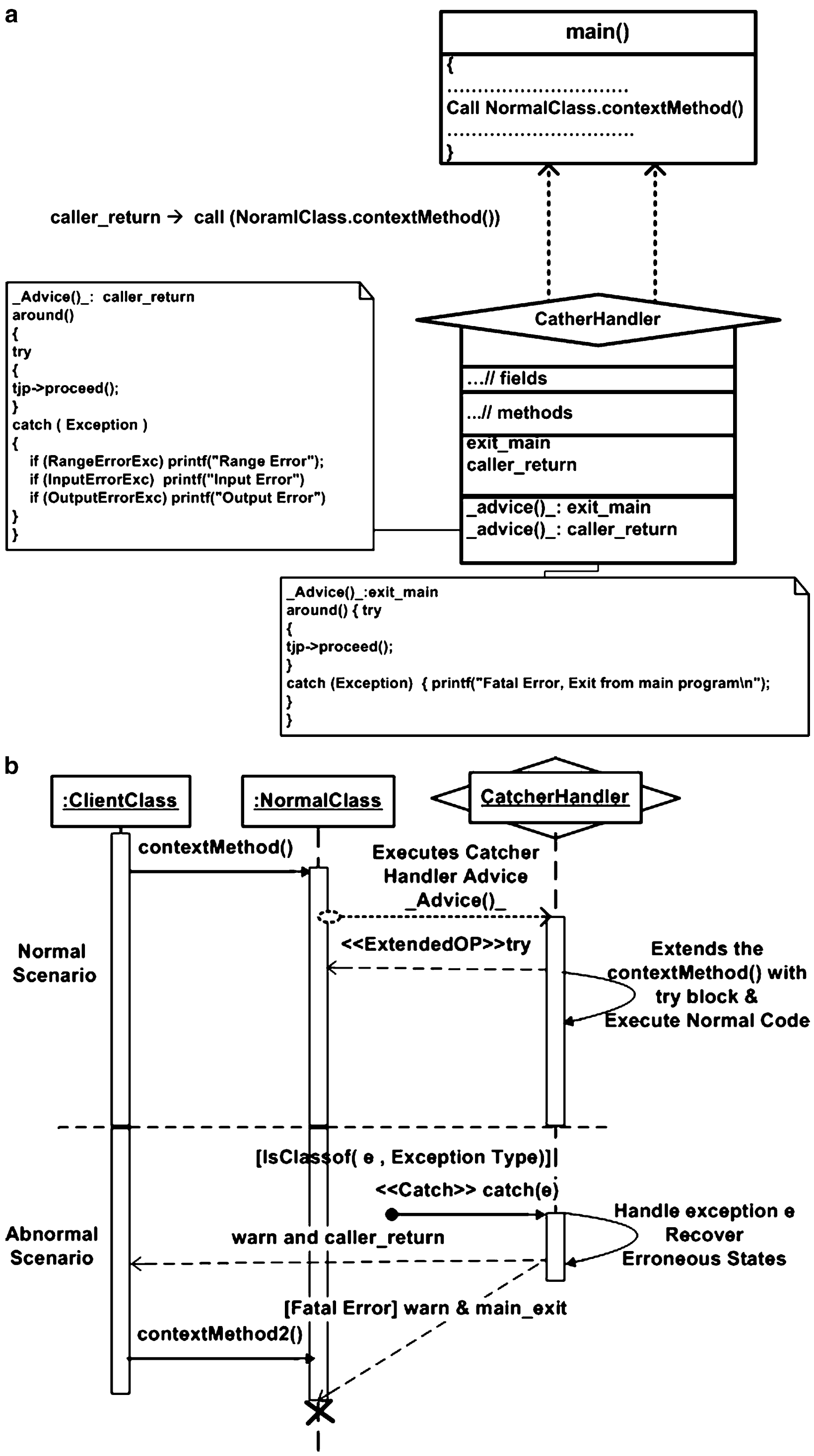
In order to validate aspect oriented fault tolerance patterns for exception handling and executable assertions as proposed earlier, a case study has been carried out using a LEGO NXT Robot (Tribot). This uses an Atmel 32-bit ARM processor running at 48 MHz. Our development environment utilizes AspectC++ 1.0pre3 as aspect weaver [4].

The Tribot has been built consisting of two front wheels driven by servo motors, a small rear wheel and an arm holding a hockey stick with the help of some standard Lego parts. Ultrasonic and light sensors are also available for navigation and guidance purposes.

An interesting task has been chosen to validate our design. In this example Tribot hits a red ball with its hockey stick avoiding the blue ball placed on the same ball stand. It makes use of the ultrasonic and light sensors to complete this task. Any deviation in full-filling the OR goals and corresponding AND sub-goals in fulfilling the overall task are considered as a mission failure.

14.5 Result and Discussion

The dependability assessment of the proposed scheme has been done via fault injection. All the faults are injected into the most critical functionality of the system, which is reading the ultrasonic sensor, light sensor, motor speed sensor and writing motor servo commands. The faults are injected by supplementary code in an aspect oriented way using AspectC++ [4]. The faults injected are permanent stuck, noise bursts and random spikes at pre-defined or random locations. These faulty data


a
main()
b
:ClientClass
:NormalClass
CatcherHandler
contextMethod()
Executes Catcher Handler Advice _Advice()_
<<ExtendedOP>>try
Extends the contextMethod() with try block & Execute Normal Code
Normal Scenario
[IsClassof(e , Exception Type)]
<<Catch>> catch(e)
Handle exception e Recover Erroneous States
Abnormal Scenario
warn and caller_return
[Fatal Error] warn & main_exit
contextMethod2()

```
{
..............................
Call NormalClass.contextMethod()
..............................
}
```

caller_return → call (NoramlClass.contextMethod())

```
_Advice()_:  caller_return
around()
{
try
{
tjp->proceed();
}
catch ( Exception )
{
   if (RangeErrorExc) printf("Range Error");
   if (InputErrorExc)  printf("Input Error")
   if (OutputErrorExc) printf("Output Error")
}
}
```

```
CatherHandler
...// fields
...// methods
exit_main
caller_return
_advice()_: exit_main
_advice()_: caller_return
```

```
_Advice()_:exit_main
around() { try
{
tjp->proceed();
}
catch (Exception)  { printf("Fatal Error, Exit from main program\n");
}
}
```

Fig. 14.5 Catcher handler aspect. (a) Structure. (b) Dynamics

scenarios may simulate both permanent and transient faults originating in a faulty hardware, software or corrupted environment within or outside a computer-based system.

Although ROC-based plausibility checks are very effective in detecting faulty data values, yet a number of false positives and false negatives were generated. The proposed recovery mechanism deviates if faults persist for a longer duration. Thus maximum instantaneous ROC assertions are augmented with mean rate base check to reduce these fault positives and negatives. Mean rate is measured in a moving average window of size m. The choice of window size m is a trade off between avoiding faulty data and reducing too much estimation bias if fault bandwidth is large. For the ultra sonic sensor of Lego NXT case study, a moving window of size (m = 4, 5) provides optimal results.

In order to provide better test coverage, the ultrasonic sensor data has been injected with periodic noisy bursts and random spikes. The frequency of these noisy spikes is controlled by modulo-n of a random number. It has been observed that mission critical failures are avoided using the proposed strategy with much higher confidence level (Figs. 14.6 and 14.7).

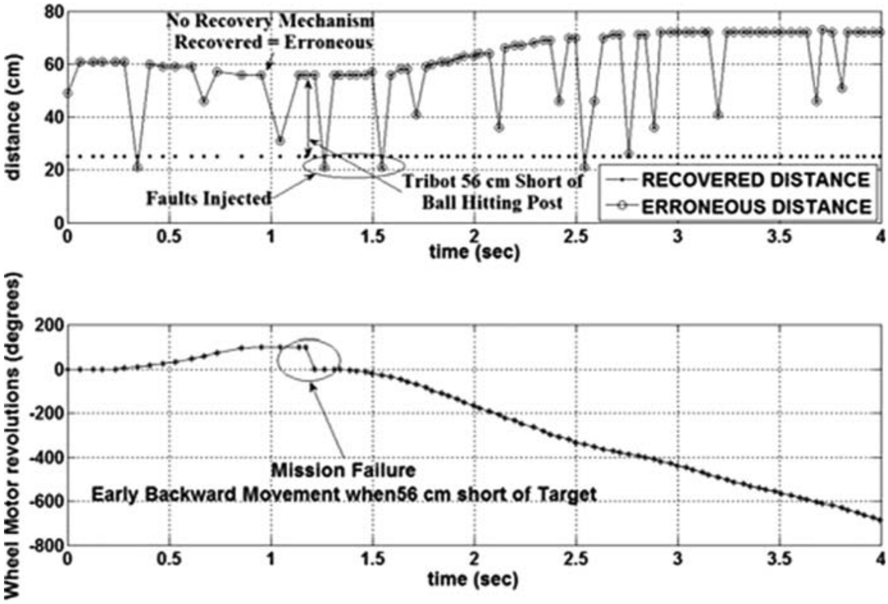


Fig. 14.6 Mission failure without recovery aspect in place

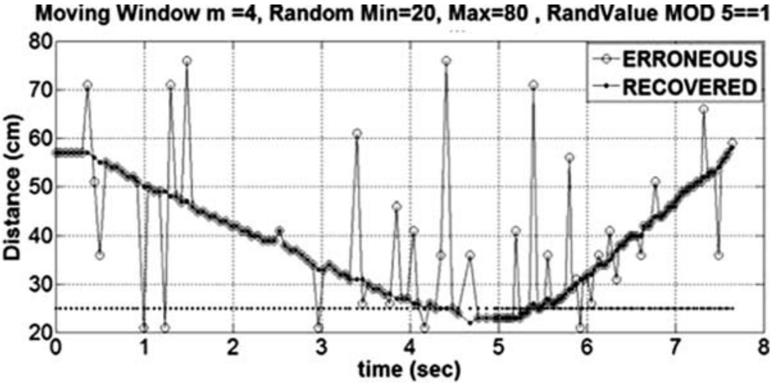


Fig. 14.7 Random spikes with error recovery

14.6 Conclusions and Future Work

The current work proposes an aspect oriented error detection and exception handling design framework. The aspect oriented design patterns under this framework bring additional benefits like the localization of error handling code in terms of definitions, initializations and implementation. Thus error handling code is not duplicated since the same error detection and handling aspect is responsible for all the calling contexts of a safety critical function. Reusability has also been improved because different error handling strategies can be plugged in separately. In this way, aspect and functional code may both be ported more easily to new systems.

The current work also investigated the use of maximum instantaneous and mean rate plausibility checks to detect and recover from erroneous states. It has been observed that mission critical variables which have monotonically increasing or decreasing trends can be augmented with carefully designed maximum instantaneous and mean rate plausibility checks to detect and recover from erroneous states.

The feedback from this initial case-study has led us to apply the same strategy to more complex applications involving the university's Merlin 521 Flight simulator. The intention is now to design and implement an aspect oriented protective wrapper that will allow students to experience physical motion within the flight simulator, under the control of their own designed autopilot, with much reduced physical risk.

This further probes the need for incorporating an error masking strategy like Recovery Blocks and N-Version Programming. An aspect oriented design version of these strategies is also under consideration.

References

1. Asterio, P., et al.: Structuring exception handling for dependable component-based software systems. Proceedings of the 30th EUROMICRO Conference (EUROMICRO'04), 2004
2. Garcia, A.F., Beder, D.M., Rubira, C.M.F.: An exception handling software architecture for developing fault-tolerant software. Proceedings of the 5th IEEE HASE USA, pp. 311–32, November 2000
3. AspectJ project homepage. http://eclipse.org/aspectj/
4. AspectC++ project homepage. http://www.aspectc.org
5. Hiller, M., et al.: Executable assertions for detecting data errors in embedded control systems. Proceedings of the International Conference on Dependable Systems & Networks, 2000
6. Hiller, M.: Error recovery using forced validity assisted by executable assertions for error detection: an experimental evaluation. 25th EUROMICRO, Milan, Italy, 1999
7. Clegg, M., Marzullo, K.: Predicting physical processes in the presence of faulty sensor readings. Proceedings of 27th International Symposium on Fault Tolerant Computing, pp. 373–378, 1996
8. Pullum, L.L: Software fault tolerance techniques and implementation. Artech House Inc., Boston, MA (2001)
9. Filho, F.C., et al.: Error handling as an aspect. Workshop BPAOSD 2007, Vancouver, BC, Canada, 12–13 March 2007
10. Romanovsky, A.: A looming fault tolerance software crisis. ACM SIGSOFT Software Engineering Notes **32**(2), 1 (March 2007)
11. Murata, K., Nigel Horspool, R., Manning, E.G., Yokote, Y., Tokoro, M.: Unification of compile-time and run-time metaobject protocol. ECOOP Workshop in Advances in Meta Object Protocols and Reflection (Meta'95), August 1995

Chapter 15
An Adaptive Multibiometric System for Uncertain Audio Condition

Dzati Athiar Ramli, Salina Abdul Samad, and Aini Hussain

Abstract Performances of speaker verification systems are superb in clean noise-free conditions but the reliability of the systems drop severely in noisy environments. In this study, we propose a novel approach by introducing Support Vector Machine (SVM) as indicator system for audio reliability estimation. This approach directly validate the quality of the incoming (claimant) speech signal so as to adaptively change the weighting factor for fusion of both subsystem scores. The effectiveness of this approach has been experimented to a multibiometric verification system that employs lipreading images as visual features. This verification system uses SVM as a classifier for both subsystems. Principle Component Analysis (PCA) technique is executed for visual features extraction while for the audio feature extraction; Linear Predictive Coding (LPC) technique has been utilized. In this study, we found that the SVM indicator system is able to determine the quality of the speech signal up to 99.66%. At 10 dB SNR, EER performances are observed as 51.13%, 9.3%, and 0.27% for audio only system, fixed weighting system and adaptive weighting system, respectively.

Keywords Biometric verification system · reliability estimation · support vector machine

15.1 Introduction

Biometric speaker verification is a technology that utilizes behavioral and physiological information of speech signal for the purpose of authentication of individual for identity claim. According to [1, 2], the advantages of using speech signal trait for biometric systems are that the signal is natural and easy to produce, requiring

D.A. Ramli (✉), S.A. Samad, and A. Hussain
Department of Electrical, Electronic & System Engineering, Engineering Faculty,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
e-mail: dzati@vlsi.eng.ukm.my; salina@vlsi.eng.ukm.my; aini@vlsi.eng.ukm.my

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_15,

little custom hardware, has low computation requirement and is highly accurate in clean noise-free conditions. However, in uncontrolled conditions, the reliability of the system drops severely as the signal to noise ratio (SNR) of the speech signal decreases. One of the solutions to overcome these limitations is by implementing fusion approach to the biometric system [3].

Research on fixed weighting fusion approach can be found in [4]. This study reported the fusion of scores produced independently by speaker recognition system and face recognition system using a weighted merged score. The optimal weight was found by maximizing the performance of the integrated system on one of the available training sets. In another case, a weighted product approach to fuse two voice features i.e. static and dynamic and three face features i.e. eye, noise and mouth was evaluated [5]. The tan-estimators were used for score normalization and weighted geometric average was used for score combination. Reference [5] combined different biometric cues i.e. voice, lip motion and face image. Reference [6, 7] integrated the scores of speech and lip modality using weighted summation fusion. In another experiment, information from speaker verification system (SVS) and profile verification system (PVS) using a weighted summation fusion was combined [7]. In [9], fuse-HMM that integrates the audio and visual features of speech were reported. In this method, the learning algorithm maximizes the two HMMs separately and consequently fuse the HMM by Bayesian fusion method. The experimental results showed that the fuse-HMMs constantly performed better than the unimodal method under clean and low noise conditions. But under stronger noise level, the performance of the fusion systems is worse compared to the speech only system. Multistage information fusion by taking both feature fusion and decision fusion approach was implemented in [10]. The study observed that the multistage system achieves significant improvement over both feature fusion and decision fusion system at different SNR levels.

For the adaptive weighting fusion approach, the reliability estimation of the current speech signal is performed either relying on the statistic based measure or directly based on the quality of the speech signal. The weight for fusion scheme is adapted correspondingly to the quality of the current input (claimant) speech signal instead of using the optimum weight that is estimated from the available training set. This approach is more advantageous especially when the system is implemented in uncertain environment conditions. Two methods have been proposed for the statistics based reliability measures i.e. entropy of a posteriori probabilities and dispersion of a posteriori probabilities. The reliability information can be obtained by the shape of a posteriori probabilities distribution of HMM states, GMM and MLP as studied in [11–13], respectively. A high entropy interprets low confidence hence signifies very unreliable input. Consequently, a mapping function between the entropies and the corresponding weight is calculated. On the other hand, study based on the quality of the speech signal was reported in [13]. This study described the use of voicing index as audio reliability measure. Implementation of the degree of voicing index as reliability measure is also reported in [14].

In this study, we propose a novel approach by introducing Support Vector Machine as indicator system for audio reliability measure. The development of this system is made up of three modules i.e. an audio front-end module, a visual front-end module and a fusion module. For audio front-end module, a vector of LPC coefficients is computed from the autocorrelation vector using Durbin recursion method. For the visual front-end module, lipreading features are employed to the system. Lipreading features are the sequence of lip images while the speaker utters the words for example, zero to nine. Several researches using lip information as features to recognition systems have been reported. As in [15], shape and intensity information from a person's lip were used in a speaker recognition system. The utilization of geometric dimension such as height, width and angle of speaker's mouth as features was also investigated [16]. Apart from lip contour-based features, pixel-based features i.e. Discrete Cosine Transform (DCT) has also been experimented as features for person recognition in [17]. The overall architecture of the proposed adaptive weighting fusion system is illustrated in Fig. 15.1.

The database used in this study is the Audio-Visual Digit Database (2001) [18]. The database consists of video and the corresponding audio recording of people reciting digits zero to nine. The video recording of each person is stored as a sequence of JPEG images with a resolution of 512 × 384 pixels while the corresponding audio recording provided is a monophonic, 16 bit, 32 kHz, WAV format. For the purpose of evaluating the systems in noisy conditions, the clean testing audio data are corrupted into 30, 20 and 10 dB SNR data by using the simulated additive white Gaussian noise (AWGN). Due to the objective of this research is to investigate the biometric system in uncertain audio condition, no artificial degradation was imposed to the visual data. However, some natural challenges such as facial expression, pose and illumination invariant are occurred within the sequence of the images and from session to session.

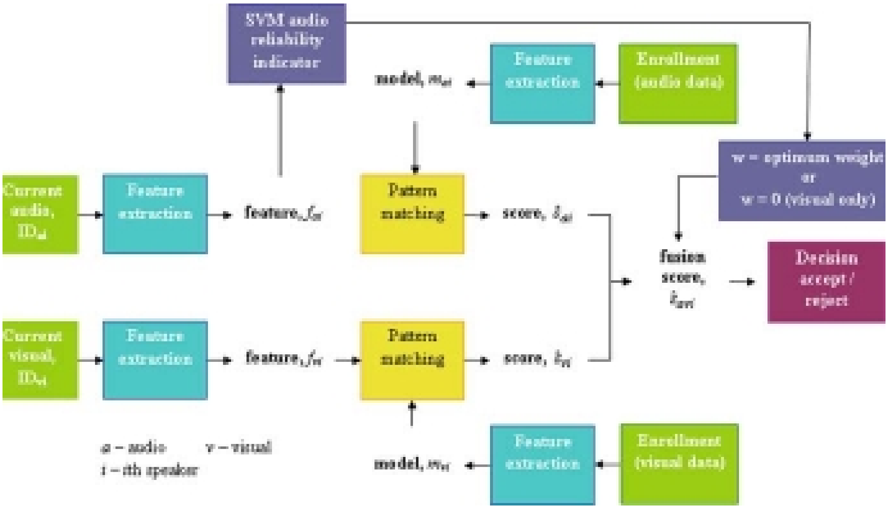


Fig. 15.1 An adaptive multibiometric system

15.2 Support Vector Machine Classifier

Support vector machine (SVM) classifier in its simplest form, linear and separable case is the optimal hyper plane that maximizes the distance of the separating hyper plane from the closest training data point called the support vectors [19, 20].

From [19], the solution of a linearly separable case is given as follows. Consider a problem of separating the set of training vectors belonging to two separate classes,

$$D = \left\{\left(x^1, y^1\right), \ldots \left(x^L, y^L\right)\right\}, \ldots x \in \Re^n, y \in \{-1, -1\} \tag{15.1}$$

with a hyperplane, $\langle w, x\rangle + b = 0$. The hyperplane that optimally separates the data is the one that minimizes

$$\phi(w) = \frac{1}{2}\|w\|^2 \tag{15.2}$$

which is equivalent to minimizing an upper bound on VC dimension. The solution to the optimization problem, Eq. (15.2) is given by the saddle point of the Lagrange functional (Lagrangian)

$$\phi(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i \left(y^i\left[\left\langle w, x^i\right\rangle + b\right] - 1\right) \tag{15.3}$$

where α are the Lagrange multipliers. The Lagrangian has to be minimized with respect to w, b and maximized with respect to $\alpha \geq 0$. Eq. (15.3) is then transformed to its dual problem. Hence, the solution of the linearly separable case is given by,

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{L}\sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j\right\rangle - \sum_{k=1}^{L} \alpha_k \tag{15.4}$$

with constrains, $\alpha_i \geq 0, \quad i = 1, \ldots, L$ and $\sum_{j=1}^{L} \alpha_j y_j = 0.$

Subsequently, consider a SVM as a non-linear and non-separable case. Non-separable case is considered by adding an upper bound to the Lagrange multipliers and non-linear case is considered by replacing the inner product by a kernel function. The solution of the non-linear and non-separable case is given as:

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{L}\sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{k=1}^{L} \alpha_k \tag{15.5}$$

with constrains, $0 \leq \alpha_i \leq C, \quad i = 1, \ldots, L$ and $\sum_{j=1}^{L} \alpha_j y_j = 0.$

Non-linear mappings (kernel functions) that can be employed are polynomials, radial basis functions and certain sigmoid functions. In this study, polynomial kernel has been employed.

15.3 Visual Front-End Subsystem

In order to locate the lips on a face, techniques for face detection and lip localization have been used in this study [21, 22]. In the first task, we implement a color-based technique and template matching algorithm to segment human skin regions from non-skin color. For the lip localization task, hue/saturation color thresholding has been employed in order to differentiate the lip area from the face [21,22]. As demonstrated in [23], the detection of the lip in hue/saturation color is much easier owing to its robustness under a wide range of lip colors and varying illumination condition. Our lipreading database has 22,200 images in total size 64 × 64 pixels from 37 persons. For each person, 60 sequences of images (with ten images per sequence) have been utilized.

Consequently, Principle component analysis (PCA) technique or also known as Karhunen–Loeve method is used for dimensionality reduction. This statistical method aims to obtain an optimum linear subspace from a covariance matrix of a set of samples [24]. This technique executes linear projection on the original samples that maximizes the scatter (variance) of all projected samples. This technique is beneficial for reducing storage capacity because the projected features are presented in a lower dimensionality space compared to the original sample space. Theory of PCA technique for feature extraction can be simply stated as follows. Given a set of N sample images $x_i, i = 1, 2, \ldots, M$ where each image in the set is lithographically re-ordered in L^2 dimensional space and belongs to one of the c classes $\{C_1, C_2, \ldots, C_c\}$. By considering a linear transformation mapping, the original sample in L^2 dimensional space are then transformed into a P-dimensional feature space, where $P \ll M \ll L^2$. The new transformed features $y_i, i = 1, 2, \ldots, M$ is known as subspace and the process of transforming is called projection. In PCA, the transformation process is executed by the following linear transformation:

$$y_i = U^T x_i, \ldots i = 1, 2, \ldots, M \tag{15.6}$$

where $U \in \Re^{L^2 \times P}$ represents matrix of Eigen pictures in $L^2 \times P$ and P corresponding to the P largest Eigen values.

The transformed lip features are then used for the verification process using SVM as classifier. Clean training visual data and clean testing visual data are used for this purpose. In order to model the classifier discriminatively, each speaker model is trained using 3, 6, 10 and 20 client data as well as with 108, 216, 360 and 720 imposter data, respectively. Thus, four types of speaker models are developed for each speaker. During testing, each type of speaker model from each speaker is tested on 40 client data and 1,440 (40 × 36) imposter data from the other 36 persons.

15.4 Audio Front-End Subsystem

Linear Predictive Coding is a time domain analysis that approximates a speech sample as a linear combination of past speech samples. A unique set of predictor coefficients are determined by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones [25, 26]. The parameter values that have been used at each stage of the experiment are also indicated. A set of feature vector computed from each frame consists of 14 cepstrum coefficients. Let assume the relation between the present sample $x(n)$ and first-order linear combination of the previous p samples as in Eq. (15.7). Consequently, LPC cepstrum coefficients can be derived through the LPC model.

$$x(n) \approx \alpha_1 x(n-1) + \cdots + \alpha_p x(n-p) \tag{15.7}$$

For a time sequence $x(n)$, complex cepstrums $\hat{c}_n$ are represented as below:

$$\hat{c}_1 = -\alpha_1 \tag{15.8}$$

$$\hat{c}_n = -\alpha_n - \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \alpha_k \hat{c}_{n-k}, \ldots 1 < n \leq p \tag{15.9}$$

$$\hat{c}_n = -\sum_{k=1}^{p} \left(1 - \frac{k}{n}\right) \alpha_k \hat{c}_{n-k}, \ldots n > p \tag{15.10}$$

Experiments for audio only biometric system are divided into two categories. Firstly, to evaluate the system using different numbers of training data and secondly, to evaluate the system based on different SNR levels. Clean training speech data and clean testing speech data are used for this purpose. In order to model the classifier discriminatively, each speaker model is trained using 3, 6, 10 and 20 client data as well as with 108, 216, 360 and 720 imposter data, respectively. Thus, four types of speaker models are developed for each speaker. During testing, each type of speaker model from each speaker is tested on 40 client data and 1,440 (40 × 36) imposter data from the other 36 persons.

Consequently, three experiments are conducted in order to evaluate the system based on different SNR levels. For this purpose, clean data are used for speaker modeling. Each speaker model is trained using 20 client training data and 720 (20 × 36) imposter training data. Three levels of corrupted testing data i.e. 30, 20 and 10 dB SNR data are used. During testing, speaker model from each speaker is tested on 40 client data and 1,440 (40 × 36) imposter data from the other 36 persons for each level of the corrupted signals.

15.5 Fusion System Implementation

Speech quality measurement is done by developing an indicator system which is based on SVM classification technique. By modeling the clean data features as sample type $+1$ and the noisy data features as sample type -1, the system is used to discriminate the incoming speech signal either as high quality or low quality speech signals. Modeling data are taken from the training data set. The indicator system is constructed to differentiate clean speech signal (high quality) from 30, 20 and 10 dB SNR speech signal (low quality). We have used 2,960 training data and 5,920 testing data for this task. This system is capable to achieve 99.66% accuracy.

In fixed weighting approach, the weight to be used for the fusion system is estimated by first running the audio subsystem and visual subsystem separately using the enrollment data. The fusion system is a soft fusion system that uses raw scores from audio and visual subsystems. The scores from each subsystem are then normalized by using min-max normalization technique. In this case, the minimum and maximum scores are transformed to 0 and 1, respectively. The normalized values are then combined by using a weighted summation fusion. In order to calculate the optimum weight, w_{opt}, w is varied from 0 to 1 in steps of 0.2. The overall performance in each step is then evaluated and the optimum weight, w_{opt} is defined at which the weight, w give the highest performance. The optimum weight is calculated at $w = 0.4$ in this experiment. For the adaptive weighting approach, each audio testing data is first checked for its quality by the audio reliability indicator system. After the speech quality measurement process is completed, the system will decide the weight for the fusion process. If the indicator system determines the current speaker's speech signal as clean speech signal, the optimum weight is employed for the fusion system otherwise the visual only system is executed.

15.6 Results and Discussions

Figure 15.2 shows the performances of the visual only system using 3, 6, 10 and 20 training data. By increasing the numbers of training data to the system, a great improvement in the GAR is observed. At FAR of 0.1%, the GAR of the 3 training data system is 85%. By using 6 and 10 training data, the percentage of GAR increases to 96%. Subsequently, the system reaches to 98% GAR when 20 training data are used. Almost 100% GAR is found for 3, 6, 10 and 20 training data systems when the FAR is equal to 35%, 4%, 3% and 0.2%, respectively. EER performances are observed as 0.27%, 0.94%, 1.15% and 2.70% for system using 20, 10, 6 and 3 training data, respectively.

The audio only system performances using 3, 6, 10 and 20 training data are showed in Fig. 15.3. The increment of the numbers of training data increases the performance of GAR. At FAR of 1%, the GAR of the 3, 6, 10 and 20 training data systems are 83%, 93%, 94% and 96%, respectively. This study observes that the performance of 20 training data system is capable to accomplish almost 100% GAR

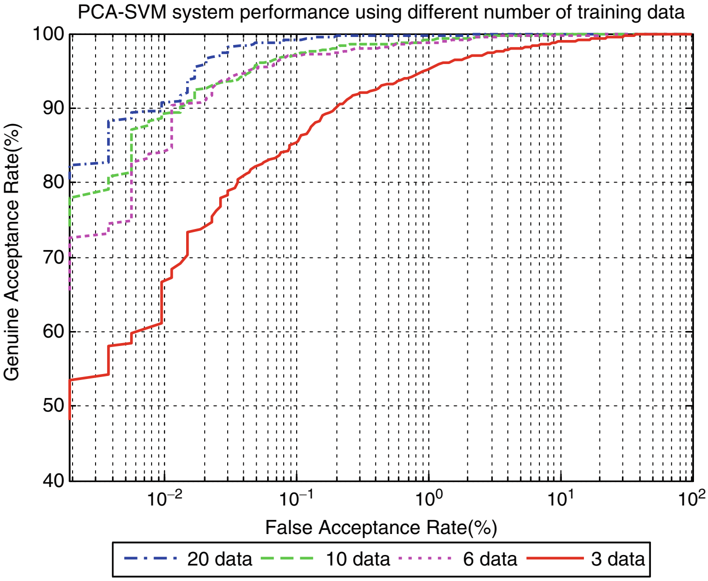


Fig. 15.2 ROC curve for visual only system

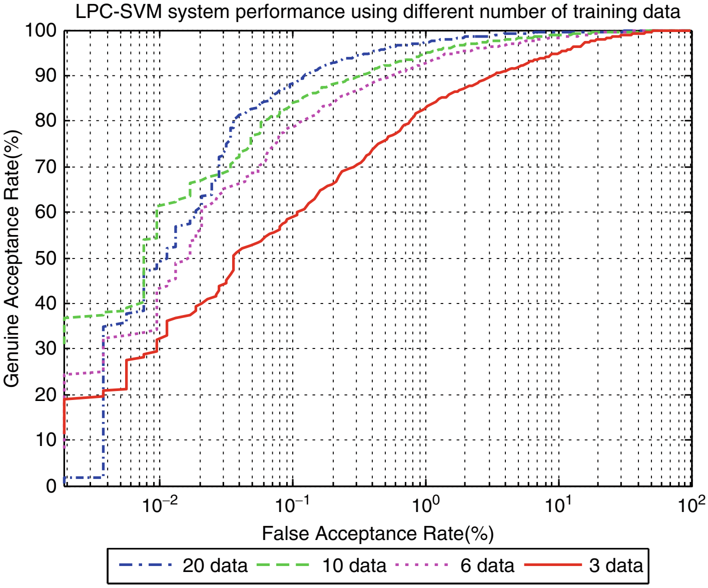


Fig. 15.3 ROC curve for visual only system

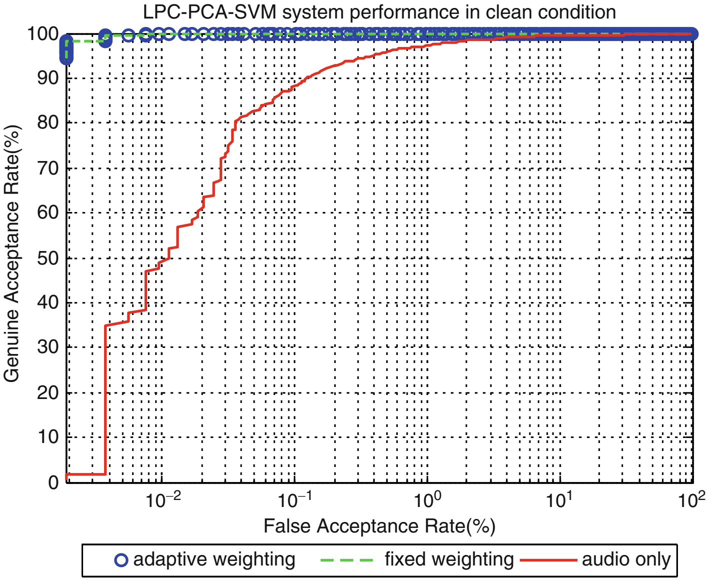


Fig. 15.4 System performances in clean condition

at FAR of 7% compared to FAR of 20%, 25% and 50% for the 10, 6 and 3 training data systems, respectively. System performances based on EER are observed as 1.79%, 2.70%, 3.50% and 6.60% for system using 20, 10, 6 and 3 training data, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system in clean condition are given in Fig. 15.4. The 100% GAR performance is evaluated at FAR is equal to 0.004% for the adaptive weighting and fixed weighting system compared to 35% GAR performance for the audio only system at the same percentage of FAR. On the other hand, it is observed that the audio only system reaches nearly 100% GAR at FAR of 7%. System performances based on EER are observed as 0.067%, 0.067% and 1.79% for adaptive weighting system, fixed weighting system and audio only system, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system based on 30 dB SNR data are given in Fig. 15.5.

Hundred percent of GAR performance is evaluated for the adaptive weighting system at FAR equal to 0.3% compared to FAR equal to 1% for fixed weighting system. The GAR performance at the FAR equal to 0.3% for fixed weighting system is 96% meanwhile GAR performance for the audio only system is 18% at the same FAR percentage. System performances based on EER are observed as 0.29, 0.87 and 17.02% for adaptive weighting system, fixed weighting system and audio only system, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system based on 20 dB SNR data are given in Fig. 15.6. The 100% GAR is evaluated at

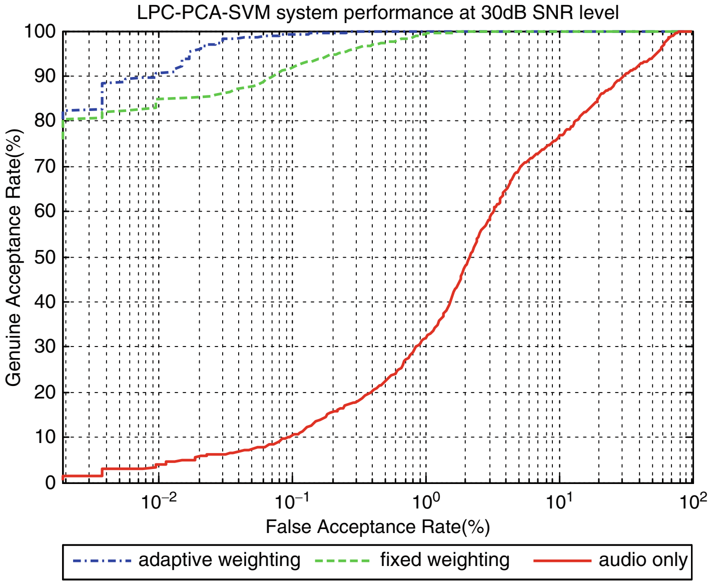


Fig. 15.5 System performances at 30 dB SNR level

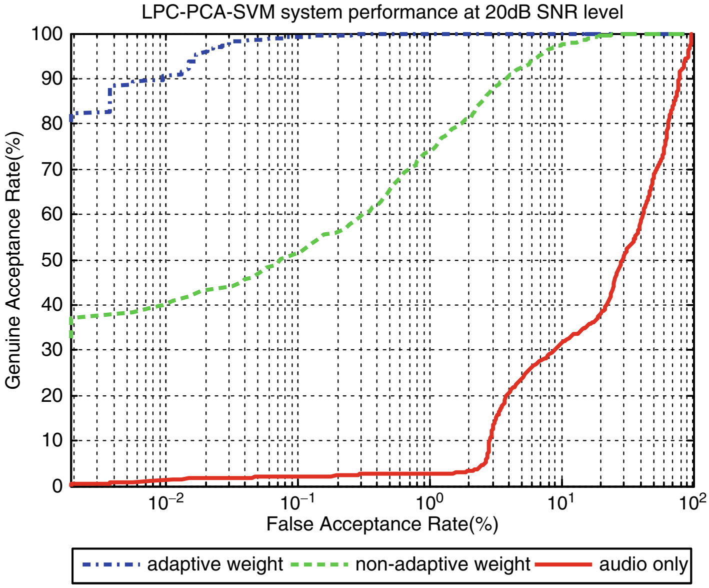


Fig. 15.6 System performances at 20 dB SNR level

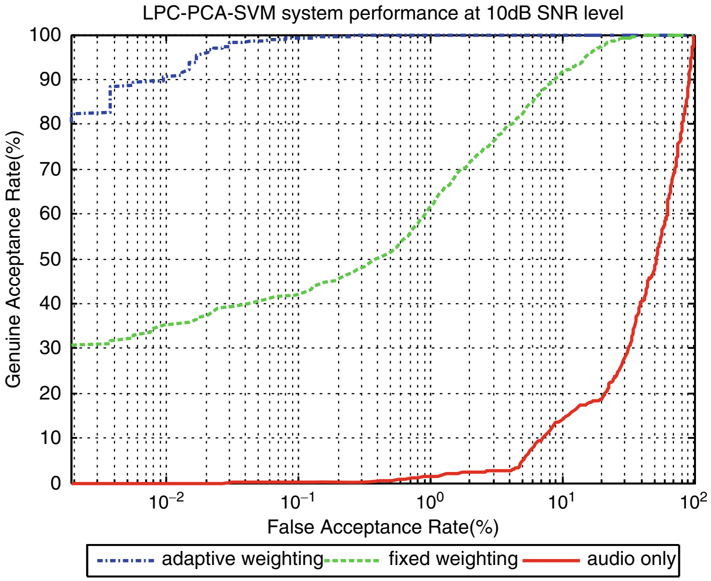


Fig. 15.7 System performances at 10 dB SNR level

FAR is equal to 0.3% for the adaptive weighting system compared to 60% GAR for fixed weighting system at the same FAR percentage. In contrast, the fixed weighting system reaches to 100% GAR at FAR equal to 11%. At the FAR 0.3%, the audio only system simply attains 4% GAR. System performances based on EER are observed as 0.27%, 6% and 40.75% for adaptive weighting system, fixed weighting system and audio only system, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system based on 10 dB SNR data are given in Fig. 15.7.

The 100% GAR performance for adaptive weighting system is evaluated at FAR equal to 0.3% while for fixed weighting system is found at FAR of 35%. In contrast, the performance at FAR of 0.3% for fixed weighting system is 48% GAR meanwhile GAR performance for audio only system is evaluated as 0% at the same FAR percentage. System performances based on EER are observed as 0.27%, 9.3% and 51.14% for adaptive weighting system, fixed weighting system and audio only system, respectively.

15.7 Conclusions

The performances of the adaptive weighting, fixed weighting and audio only system at different SNR levels have been reported for comparison. This study proved that the proposed SVM indicator system is viable for estimating the quality of speech signal and the implementation of the adaptive weighting approach is imperative for uncertain audio condition. The advantage of using the adaptive weighting instead of fixed weighting is to avoid unreliable scores to be fused together in fusion systems that can spoil the accuracy of the total scores. By using the adaptive weight fusion approach, the performances of the verification systems can be further enhanced when high quality speech signal is obtained. Besides, in corrupted speech signal environment, the system performances can still be maintained by adjusting the fusion weight by using the visual only systems. However, the effectiveness of this approach depends on the performance of the audio indicator system and visual verification system. Future work will be devoted on all SNR levels and different types of noises. Noise eradication techniques will also be experimented to the audio and visual subsystems so as to enhance the system performance.

References

1. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE **85**, 1437–1462 (1997)
2. Reynolds, D.A.: An overview of automatic speaker recognition technology. Proc. IEEE Acoustics Speech Signal Processing **4**, 4072–4075 (2002)
3. Ramli, D.A., Samad, S.A., Hussain, A.: In: Corchado, E., et al. (ed.) Score Information Decision Fusion using Support Vector Machine for a Correlation Filter Based Speaker Authentication System, vol **53**, pp. 235–242. Springer, Berlin, Heidelberg (2008)
4. Brunelli, R., Falavigna, D., Stringa, L., Poggio, T.: Automatic person recognition by using acoustic and geometric. Mach. Vis. Appl. **8**, 317–325 (1995)
5. Brunelli, R., Falavigna, D.: Personal identification using multiple cue. IEEE Trans. Pattern Anal. Mach. Int. **17**(3), 955–966 (1995)
6. Dieckmann, U., Plankensteiner, P., Wagner, T.: SESAM: A biometric person identification system using sensor. Pattern Recog. Lett. **18**(9), 827–833 (1997)
7. Jourlin, P., Luettin, J., Genoud, D., Wassner, H.: Integrating acoustic and labial information for speaker identification and verification. Proc. 5th European Conf. Speech, Commun. Technol. **3**, 1603–1606 (1997)
8. Sanderson, C., Paliwal, K.K.: Multi-modal person verification system based on face profile and speech. Fifth International Symposium on Signal Processing and Its Applications, pp. 947–950 (1999)
9. Pan, H., Liang, Z.P., Huang, T.S.: Fusing audio and visual features of speech. Proc. IEEE Int. Conf. Image Processing **3**, 214–217 (2000)
10. Chu, S.M., Marcheret, V.L.E., Neti, C., Potamianos, G.: Multistage information fusion for audio-visual speech recognition. Proc. IEEE Int. Conf. Multimedia Expo, pp. 1651–1654 (2004)
11. Gurban, M., Thiran, J.P.: Using entropy as a stream reliability estimate for audio-visual speech. 16th European Signal Processing Conference (2008, in press)
12. Potamianos, G., Neti, C.: Stream confidence estimation for audio-visual speech. Proc. Int. Conf. Spoken Language **3**, 746–749 (2000)
13. Heckmann, M., Berthommier, F., Kroschel, K.: Noise adaptive stream weighting in audio-visual speech. EURASIP J. Appl. Signal Process. **2002**(11), 1260–1273 (2002)
14. Chetty, G., Wagner, M.: Robust face-voice based speaker verification using multilevel. Image Vision Comput. **26**(9), 1249–1260 (2008)
15. Wark, T., Sridharan, S.: A syntactic approach to automatic lip feature extraction for speaker identification. IEEE Int. Conf. Acoustics Speech Signal Processing **6**, 3693–3696 (1998)

16. Broun, C.C., Zhang, X., Mersereau, R.M., Clements, M.: Automatic speechreading with application to speaker verification. IEEE Int. Conf. Acoustics Speech Signal Processing **1**, 685–688 (2002)
17. Fox, N.A., Reilly, R.B.: Robust multi-modal person identification with tolerance of facial expression. Proc. IEEE Int. Conf. System Man Cybernetics **1**, 580–585 (2004)
18. Sanderson, C., Paliwal, K.K.: Noise compensation in a multi-modal verification system. Proc. Int. Conf. Acoustics, Speech Signal Processing **1**, 157–160 (2001)
19. Gunn, S.R.: Support vector machine for classification and regression. Technical Report, University of Southampton (2005)
20. Wan, V., Campbell, W.M.: Support vector machines for speaker verification and identification. Proc. Neural Networks Signal Processing **2**, 775–784 (2000)
21. Chetty, G., Wagner, M.: Liveness verification in audio-video speaker authentication. Proc. Int. Conf. Spoken Language Processing ICSLP 04, pp. 2509–2512 (2004)
22. Chetty, G., Wagner, M.: Automated lip feature extraction for liveness verification in audio-video authentication. Proc. Image Vision Comput., pp. 17–22 (2004)
23. Matthews, I., Cootes, J., Bangham, J., Cox, S., Harvey, R.: Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213 (2002)
24. Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve procedure for the characterisation of human. IEEE Trans. Pattern Anal. Mach. **12**(1), 103–108 (1990)
25. Rabiner, L.R., Juang, B.H.: Fundamental of Speech Recognition. Prentice-Hall, New York (1993)
26. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust Speech Signal Process. **29**(2), 254–272 (1981)

Chapter 16
Analysis of Performance Impact due to Hardware Virtualization Using a Purely Hardware-Assisted VMM

Saidalavi Kalady, Dileep PG, Krishanu Sikdar, Sreejith BS, Vinaya Surya, and Ezudheen P

Abstract This article presents a discussion on the impact of Hardware-assisted Virtualization for the ×86–64 architecture. A quantitative performance analysis has been done using a simple prototype of a purely Hardware-assisted Virtual Machine Monitor. The performance impact is measured in terms of the CPU time consumed by certain critical sections of Virtualization specific code. The analysis reveals that Hardware Virtualization at its current stage suffers from considerable performance overhead, but can be improves considerably with better hardware.

Keywords Hardware virtualization · virtual machine monitor · performance analysis · ×86 architecture · AMD secure virtual machine

16.1 Introduction

Virtualization is a technique for the efficient utilization of hardware resources. A Virtual Machine Monitor (VMM), also known as a Hypervisor, is a piece of software, which is used to implement Virtualization. A VMM controls the concurrent execution of multiple Operating Systems on a single physical machine. It is a thin software layer that provides a Virtual Machine (VM) abstraction. The abstraction resembles the real hardware to an extent that is sufficient to enable the software written for the real machine to run without change in the VM [1]. The VMM is referred to as the Host and each Operating System (OS) that runs atop the VMM a Guest. The VMM provides each Guest OS, the appearance of full control over a complete physical machine (processor, memory, and all peripheral devices).

S. Kalady (✉), P.G. Dileep, K. Sikdar, B.S. Sreejith, V. Surya, and P. Ezudheen
Department of Computer Science and Engineering, National Institute of Technology Calicut, Kerala, India 673 601
e-mail: said@nitc.ac.in; dileeppg.nitc@gmail.com; Krishanu.6@gmail.com; reejithbs.nitc@gmail.com; Vinayasurya@gmail.com; Ezhudheen@gmail.com

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_16,

Until the recent wave of processor Virtualization assistance from hardware manufacturers, ×86 – the world's most popular architecture was very hostile towards Virtualization. Techniques such as Full Virtualization and Paravirtualization resorted to complicated software workarounds to virtualizes ×86 machines, but not without the associated overhead [2]. Hardware-assisted Virtualization, though not fully mature at present, represents the future of Virtualization [3].

This article describes (1) The recent architectural extensions for ×86 Virtualization; (2) Design and implementation of a simple prototype of a purely Hardware-assisted Virtual Machine Monitor (HVMM); (3) A quantitative study of the performance impact associated with Hardware Virtualization.

A comparison of running times with the VMM running identical Virtual Machines atop real machines with differing processor speeds reveals that the performance overhead diminishes with increasing processing power.

16.2 Hardware Virtualization

The ×86 processor architecture failed to meet classical Virtualization standards until major hardware vendors came up with the recent architectural extensions in support of Virtualization [4]. The respective technologies from Intel and AMD in this domain are Intel VT codenamed Vanderpool and AMD Secure Virtual Machine (SVM) codenamed Pacifica.

16.2.1 ×86 Architectural Limitations

In its native form, the ×86 architecture doesn't meet the Popek & Goldberg's formal requirements for Virtualization [4]. In a virtualized environment, those instructions that require control to be handed over to the VMM are called sensitive instructions. Privileged instructions within a VM can be trapped to handover control to the VMM. If all the *sensitive instructions* are privileged then the processor architecture is said to be virtualizable. There are 17 instructions in ×86 which violate this basic requirement [5].

Moreover, the ×86 architecture has privilege levels called *Rings*. The OS normally runs at ring level 0. So, to control a Guest OS within a VM, the VMM must run at a higher privilege level than the OS. But the highest privilege available is at level 0. Even in that case the Guest OS must execute at a level numerically greater than or equal to 1. But most proprietary Operating Systems are designed to run at level 0 or they will fail to operate. One solution to this problem is modifying the kernel of the Guest OS. But for some commercial Operating Systems it may violate their Licensing terms and conditions. Hence in such cases implementing a virtualized environment requires expensive changes in Operating Systems and it might also cause software incompatibility issues [5].

Different software techniques have been in use for ×86 virtualization. These include *Full Virtualization*, which completely emulates the underlying hardware, and *Paravirtualization* which necessitates modifications to the OS kernel. Both these techniques devise expensive software techniques to overcome the inherent limitations of the underlying hardware.

The recent ×86 hardware Virtualization extensions from Intel and AMD are the answer to many of the aforesaid problems. The two technologies are similar but incompatible in the sense that a VMM designed to work on one cannot automatically run on the other. The AMD SVM has been identified for the study and analysis related to the present work.

16.2.2 AMD SVM Architectural Extensions

The AMD SVM architecture is a set of hardware extensions to the ×86 architecture specifically designed to enable effective implementation of Virtual Machine systems [6].The AMD SVM is designed to provide quick mechanisms for World Switch between Guest Operating Systems and the Host VMM. World Switch refers to the operation of switching between the Host and Guest [7, 8].

Important features of the SVM include the ability to intercept selected instructions or events in the Guest, external access memory protection, assistance for interrupt handling, virtual interrupt support, a Guest/Host tagged Translation Lookaside Buffer (TLB) and Nested Paging to reduce Virtualization overhead [7].

AMD SVM introduces several new instructions and modifies many existing ones to facilitate simpler and yet robust implementations of Virtual Machine systems on the ×86 architecture or more specifically the AMD64 architecture. The newly introduced instructions are VMRUN, VMLOAD, VMSAVE, VMMCALL, STGI, CLGI, SKINIT, and INVLPGA [6].

Another feature provided by the architecture is a new processor mode called Guest Mode entered through the VMRUN instruction. While executing in Guest Mode, subtle changes are introduced in the behavior of some ×86 instructions in order to facilitate Virtualization [6].

There is also a new memory resident data structure called Virtual Machine Control Block (VMCB) for each running Guest OS. The VMCB is divided into the Control area and the State area. Control area contains various control bits including the intercept vector with settings that determine what actions cause #VMEXIT (transfer of control from the Guest to Host). The CPU state for each Guest is saved in the state area. In addition, information about the intercepted event is put into the VMCB on #VMEXIT [7].

The SVM also includes Nested Paging facility to allow two levels of address translation in hardware, thus eliminating the need for the VMM to maintain the so called *shadow page table* structures that are involved in software Virtualization techniques [6].With nested paging enabled, the processor applies two levels of address translation. A Guest page table (gPT) maps Guest virtual addresses to Guest physical addresses located in Guest physical space. Each Guest also has a Host page table (hPT) maps Host virtual addresses to Host physical addresses located in Host physical space. Both Host and Guest levels have their own copy of the CR3 register, referred to as hCR3 and gCR3, respectively. The complete translation from Guest virtual to Host physical address is cached in the TLB and used on subsequent Guest accesses [6].

In addition, the TLB entries are tagged with an Address Space IDentifier (ASID) distinguishing Host-space entries and different Guest-space entries from each other. The ASID Tag specifies to which Virtual Machine, the corresponding memory page is assigned. This eliminates the need for TLB flushes each time a different Virtual Machine is scheduled [6].

16.3 The Hardware-Assisted Virtual Machine Monitor

The HVMM, which we use for performance analysis, completely, relies on the AMD SVM Hardware Virtualization support. It is based on an open source prototype VMM called Tiny Virtual Machine Monitor (TVMM) by Kaneda [9].

The TVMM is a simple VMM developed for the purpose of education and verification and has the following functionalities [9]. It performs the basic tasks necessary to initialize the AMD SVM processor extensions. It successfully creates a single VM and then boots a skeletal Guest OS within the VM.

Our HVMM is an extended version of the TVMM with the following additional capabilities. It can handle multiple Guest Operating Systems. It creates multiple Virtual Machines each of which runs different Guest Operating Systems. It also has the functionality to schedule the Guests one after the other in a Round Robin fashion.

16.3.1 HVMM Design

Getting the HVMM running requires a few important steps.

The first step is to get the SVM initialized which is done by INITIALIZE_SVM(). This invokes ENABLE_SVM() and SETUP_HYPERVISOR(). ENABLE_SVM() does the job of initializing the SVM specific flags such as EFER.SVME with the needed values. SETUP_HYPERVISOR() allocates and sets up the basic VMM data structures such as Host Save Area {HOST_SAVE_PA} which is a protected memory that can be accessed by the VMM alone.

The second step is to get the HVMM running which is done by the function HYPERVISOR_CORE().

HYPERVISOR_CORE() does the following. First of all it creates a different virtual machine (VM) using VM_CREATE(gos) {where gos stands for guest operating system} for each guest OS and adds them to the list of active virtual machines by invoking ADD_ACTIVE (vm). This list is considered for the round robin scheduling of the virtual machines. VM_CREATE(gos) does the job of setting up the VMCB

(Virtual Memory Control Block) using SETUP_VMCB(). The image of the guest OS is then loaded into memory using LOAD_GUEST_IMAGE(gos). SETUP_VMCB() does the task of allocating the VMCB using ALLOCATE_VMCB(), which allocates a region of physically contiguous page-aligned memory to the VMCB. The VMCB is partitioned into control area and state area using SETUP_CONTROL_AREA() and SETUP_STATE_AREA(). SETUP_CONTROL_AREA() initializes the control area of the VMCB with conditions that shall control the guest OS execution and transfer of control to VMM. SETUP_STATE_AREA() initializes the state area of the VMCB with the initial machine state of the guest OS. Next, HYPERVISOR_CORE() enters an infinite loop wherein each iteration the next VM from the list is obtained following a simple round robin rule using GET_NEXT_VM(), and booted using VM_BOOT(vm); if no active virtual machines are found the VMM terminates. One iteration of the infinite loop terminates when the virtual machine exits with a #VMEXIT or a VMMCALL {which are AMD SVM specific instructions}. The exit handler HANDLE_VMEXIT(vm) then deals with the problem that causes the virtual machine to exit (in other words transferred control from guest OS to VMM). In this implementation, HANDLE_VMEXIT(vm) simply displays the information associated with #VMEXIT. After handling the #VMEXIT, the next loop begins and the next VM is scheduled and so on.

The Following Pseudo Code describes the control flow behind the working of the HVMM

```
HVMM {
    INITIALIZE_SVM();
    HYPERVISOR_CORE();
}
INITIALIZE_SVM() {
    ENABLE_SVM();
    SETUP_ HYPERVISOR();
}
ENABLE_SVM() {
    // Initialize SVM specific flags such as EFER,SVME
    // with appropriate value;
}
SETUP_ HYPERVISOR(){
    //Allocate and setup basic VMM Data structures such as Host State
    Save Area
}
HYPERVISOR_CORE() {
    for (each Guest OS gos) {
        vm = VM_CREATE(gos);
        ADD_ACTIVE(vm);
}
```

```
   while (1) {
       vm = GET_NEXT_VM();
       i f (no active vm) break;
       VM_BOOT(vm);
       HANDLE_VMEXIT(VM);
   }
}
VMCREATE(gos) {
       vm->VMCB = SETUP_VMCB();
       LOAD_GUEST_IMAGE(gos);
}
SETUP_VMCB() {
       ALLOCATE_VMCB();
       SETUP_CONTROL_AREA();
       SETUP_STATE_AREA();
}
ALLOCATE_VMCB() {
       // Al locate a region of physically contiguous,
       // page-aligned memory
}
SETUP_CONTROL_AREA() {
  // Initialize the control area of the VMCB
       // with conditions that will control the
       // guest OS execution
}
SETUP_STATE_AREA() {
       // Initialize the state area of the VMCB with the
       // initial machine state of the guest OS
}
LOAD_GUEST_IMAGE(gos){
       // Load the image of the guest OS gos into memory
}
ADD_ACTIVE(vm){
       // Adds the Virtual Machine vm to the list of VMs
       // considered for Round Robin execution
}
GET_NEXT_VM() {
       //Return one from the active list of VMs following
       // a simple Round Robin execution
}
VM_BOOT(vm){
       // Transfer control to the Guest Os within the vm
       // using special SVM Instructions
```

```
}

HANDLE_VMEXIT(vm) {
        // A naive #VMEXIT handle which just displays
        // the information associated with the #VMEXIT
}
```

16.3.2 HVMM Implementation

The HVMM is designed to run on AMD64 machines with the 64-bit Long mode and SVM extensions enabled. The bulk of the HVMM code is written in ‘C’ language. The startup code needs to be in assembly and is done using the GNU Assembler. The development environment is described below. Here the object files from AS and GCC are linked together to form a single ×86–64 Executable and Linkable Format (ELF) binary file ‘hvmm’ by the linker ‘ld’ using a separate custom linker script (Fig. 16.1).

A disk image is created using the dd utility to emulate a real physical disk. The virtual disk created is mounted on a 64 bit Linux machine installed on AMD SVM supported hardware. The binary files of one or more Guest kernels, HVMM along with GRUB boot files and settings are copied to the disk. Then the disk is loaded in AMD SimNow and the kernel is run from within the HVMM.

We use simple skeletal 64-bit kernels in ELF format for our purpose. The OS kernels complete the minimal initialization tasks and then simply loop infinitely printing dummy messages on to the screen [11, 12]. We can load any number of OS kernels provided as separate GRUB modules along with the HVMM kernel. The OS kernels are copied to disjoint locations in memory from within our HVMM. Then Virtual Machines are created for each OS, which involves populating SVM specific data structures such as the VMCB to control the VM execution. The VMM finally boots the kernels within the corresponding Virtual Machines in a Round Robin fashion with a specific time slice. Once an OS gains control, it continues to execute until

Fig. 16.1 HVMM development environment

CPU	:	AMD x2 3800+
Base OS	:	64-bit openSUSE 10.3
Simulator	:	AMD simnow 4.4.2
'C' Compiler	:	GCC 4.3
Assembler	:	GNU AS
Linker	:	GNU ld
Image Tool	:	dd
Bootloader	:	GRUB
Other	:	GNU Make utility

certain specific operations such as, the completion of a time slice which in turn will cause a #VMEXIT forcing transfer of control back to the VMM. The OS can also transfer the control explicitly to the VMM using a special VMMCALL instruction.

16.4 Performance Analysis

In this section, e present a quantitative study of the ormance overheads associated with Hardware-assisted Virtualization using the AMD SVM. The overhead is measured in terms of CPU time consumed by certain critical sections of HVMM code.

16.4.1 Test Setup

The HVMM is made to boot and run two Guest Operating Systems one after the other. It is possible to setup break points at critical points in the OS code with the help of the SimNow debugger [10]. We setup break points at the following two points (1) Immediately before the #VMEXIT in the Guest OS 1 and (2) Just before start of Guest OS 2 (Fig. 16.2).

The time at the two instances are noted as the value given by the Host seconds field in the SimNow main window. The difference between the two values gives the time taken to execute minimal VMM specific tasks before the next OS gets control. We call this value "Switch Lag". Here, we neglect the time taken for specific #VMEXIT handling code and just consider the minimal time taken for a blind transfer of control back to the next Guest OS.

The test is performed on processors with three different clock speeds – AMD x2 Turion, AMD x2 3800+, AMD x2 4400+, all of which have AMD SVM extensions, and a comparison is made between the values of Switch Lag obtained.

16.4.2 Analysis

The following are the results obtained from the aforementioned test repeated on different hardware (Table 16.1).

CPU	:	AMD x2 Turion, AMD x2 3800+, 4400+
Base OS	:	OpenSUSE 10.3 64 bit
Simulator	:	AMD SimNow V 4.4.2
Host	:	HVMM
Guests	:	64 bit skeletal OS kernels

Fig. 16.2 Test environment

Table 16.1 Comparisons of switch lag values

Processor	Clock speed (GHz)	Switch lag (ms)
AMD x2 Turion	1.6	89
AMD x2 3800+	2.0	79
AMD x2 4400+	2.2	75

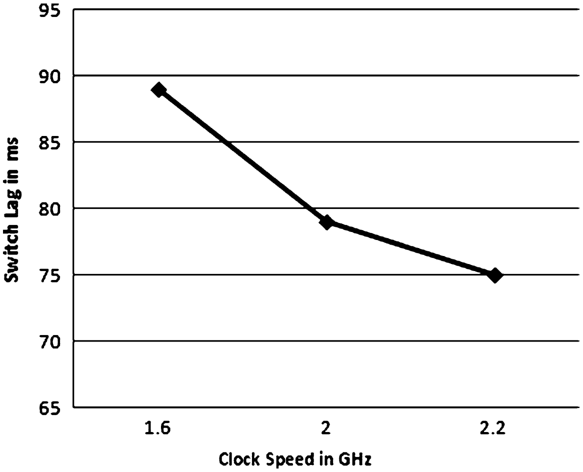


Fig. 16.3 Plot of CPU speed versus switch lag

A graph is plotted with the Switch Lag on the vertical axis against the CPU speed on the horizontal axis for the above three machines (Fig. 16.3).

From the graph, we can infer that the Switch Lag comes down considerably with ascending clock speeds and increasing processing power.

16.5 Conclusion

Comparison of running times with the HVMM running identical Virtual Machines atop real machines having differing processor speeds reveals that the performance overhead diminishes with increasing processing power. With evolving hardware enhancement techniques, the performance impact due to Virtualization is expected to come down. Hardware Virtualization is still at a nascent stage. The results obtained from the test with different processors show that there is a long way still to go in reducing the performance overhead associated with Hardware-assisted Virtualization.

16.6 Future Works

A multithreaded or multi-process implementation of VMM scheduler, which enables the HVMM to execute multiple guest operating systems in round robin fashion, will bring down the switch lag into the order of microseconds. It will necessitate additional inter process communication among Host OS and Guest OS for exchanging virtualization events among them. This will result in higher response time for Guest OS and efficient CPU utilization.

Hardware vendors are keen on replacing single core processors with multi-core processors. The future of virtual machine depends on how the VMM performs on multi-core machines and how HVMM uses the available hardware resources to maximize the VMM performance.

A real performance gain can be achieved by reducing the size of both data and code of individual guest OS. Digging more into this idea will reveal the atomic difference between different operating systems which is nothing but processing the same set of hardware and operating system information (stored in different format by different operating systems) getting processed by slightly different algorithms. The codes for hardware drivers mainly depend on hardware manufacturers and they are almost independent of the Operating Systems. Reducing these redundant codes will optimize guest operating systems performance. Another technique for improving the performance is to standardize all the systems and OS dependent information at large and keeping them in a single file system format such that all the guest OS can reuse it and improve the locality of reference in the data cache.

References

1. Rosenblum, M., Garfinkel, T.: Virtual machine monitors: current technology and future trends. IEEE Comput. **38**(5) (May 2005)
2. Intel. Intel virtualization technology. Intel Tech. J. **10**, 3 (August 2006)
3. Adams, K., Agesen, O.: A comparison of software and hardware techniques for ×86 virtualization. International Conference on Architectural Support for Programming Languages and Operating Systems, ACM (2006)
4. Popek, G.J., Goldberg, R.P.: Formal requirements for virtualizable third generation architectures. Commun. ACM, July 1974
5. Robin, J., Irvine, C.: Analysis of the Intel pentium's ability to support a secure virtual machine monitor. Proceedings of the 9th USENIX Security Symposium, August 2000
6. AMD. AMD 64 virtualization codenamed "Pacifica" technology secure virtual machine architecture reference manual. Publication no. 33047 Revision 3.01, May 2005
7. AMD. AMD 64 architecture programmer's manual vol. 2: system programming. Publication no.24593 Revision 3.13, July 2007
8. AMD. AMD64 architecture programmer's manual vol. 3: general-purpose and system instructions. Publication no.24594 Revision 3.13, July 2007
9. Tiny Virtual Machine Monitor. http://web.yl.is.s.u-tokyo.ac.jp/∼kaneda/tvmm/
10. AMD. SimNow v4.4.2 Simulator Users Manual
11. An introduction to OS development. http://osdever.net/
12. Advanced OS development. http://www.osdev.org/

Chapter 17
Metrics-Driven Software Quality Prediction Without Prior Fault Data

Cagatay Catal, Ugur Sevim, and Banu Diri

Abstract Software quality assessment models are quantitative analytical models that are more reliable compared to qualitative models based on personal judgment. These assessment models are classified into two groups: generalized and product-specific models. Measurement-driven predictive models, a subgroup of product-specific models, assume that there is a predictive relationship between software measurements and quality. In recent years, greater attention in quality assessment models has been devoted to measurement-driven predictive models and the field of software fault prediction modeling has become established within the product-specific model category. Most of the software fault prediction studies focused on developing fault predictors by using previous fault data. However, there are cases when previous fault data are not available. In this study, we propose a novel software fault prediction approach that can be used in the absence of fault data. This fully automated technique does not require an expert during the prediction process and it does not require identifying the number of clusters before the clustering phase, as required by the K-means clustering method. Software metrics thresholds are used to remove the need for an expert. Our technique first applies the X-means clustering method to cluster modules and identifies the best cluster number. After this step, the mean vector of each cluster is checked against the metrics thresholds vector. A cluster is predicted as fault-prone if at least one metric of the mean vector is higher than the threshold value of that metric. Three datasets, collected from a Turkish white-goods manufacturer developing embedded controller software, have been used during experimental studies. Experiments revealed that unsupervised software fault prediction can be automated fully and effective results can be achieved by using the X-means clustering method and software metrics thresholds.

Keywords Software metrics · X-means algorithm · quality prediction · unlabeled program modules · metrics thresholds · clustering

C. Catal (✉) and U. Sevim
TUBITAK-Marmara Research Center, Information Technologies Institute, Gebze, Kocaeli, Turkey
e-mail: cagatay.catal@bte.mam.gov.tr; ugur.sevim@bte.mam.gov.tr

B. Diri
Yildiz Technical University, Department of Computer Engineering, Besiktas, Istanbul, Turkey
e-mail: banu@ce.yildiz.edu.tr

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_17,

17.1 Introduction

Today's software systems are becoming more and more complex and their lines of code will reach from millions to billions of lines of code in the near future. These kinds of large-scale systems pose extraordinary challenges in software quality. Software quality can be defined differently based on different views of quality [11]:

- *Transcendental view*: Quality can be described with abstract terms instead of measurable characteristics and users can recognize quality if it exists in a software product.
- *Value-based view*: Quality is defined with respect to the value it provides and customers decide to pay for the software if the perceived value is desirable.
- *User view*: Quality is the satisfaction level of the user in terms of his/her needs.
- *Manufacturing view*: Process standards improve the product quality and these standards must be applied.
- *Product view*: This view of quality focuses on internal quality characteristics.

As software systems are becoming more complex and people's quality expectations are increasing, we need to manage these quality expectations in a quality engineering process. A quality engineering process includes three groups of activities, as described below [11]:

- *Quality planning*: Quality goals are identified, quality assurance (QA) activities are selected, and quality measurements are chosen for quality improvement.
- *Execution of QA activities and management of faults*: Selected QA activities are performed and discovered faults are removed.
- *Quality measurement, assessment and improvement*: Quality measurement activities start parallel to the QA activities.

Among the software quality assurance (QA) activities, software testing is the most known QA activity. In addition to software testing, there are many alternative QA approaches such as inspection, formal verification, fault tolerance, and defect prevention [11]. Objective assessment of product quality is performed based on measurement data collected during quality engineering processes. Quality assessment models are quantitative analytical models and more reliable compared to qualitative models based on personal judgment. They can be classified into two categories, which are generalized and product-specific models [11]:

- *Generalized models*: Project or product-specific data are not used in generalized models and industrial averages help to estimate the product quality roughly. They are grouped into three sub-categories:
    - *Overall models*: A rough estimate of quality is provided.
    - *Segmented models*: Different quality estimates are applied for different product segments.
    - *Dynamic models*: Dynamic model graphs depict the quality trend over time for all products.

- *Product-specific models*: Product-specific data are used in these models in contrast to the generalized models.
    - *Semi-customized models*: These models apply historical data from previous releases of the same product instead of a profile for all products.
    - *Observation-based models*: These models only use data from the current project and quality is estimated based on observations from the current project.
    - *Measurement-driven predictive models*: These models assume that there is a predictive relation between software measurements and quality.

Software fault prediction models are in the *measurement-driven predictive models* group and the objective of our research is to develop tools, methods or techniques for software fault prediction models in the absence of fault data. The quality of software components should be tracked continuously during the development of high-assurance systems such as telecommunication infrastructure, medical devices, and avionic systems. The quality assurance group can improve product quality by allocating the necessary budget and human resources to low-quality modules identified with different quality estimation models. Software metrics are independent variables and a dependent variable is fault data for software fault prediction models. The aim of building this kind of model is to predict the fault labels (fault-prone or not fault-prone) of the modules for the next release of the software. A typical software fault prediction process includes two steps. First, a fault prediction model is built using previous software metrics and fault data belonging to each software module (class or method level). After this training phase, fault labels of program modules can be estimated using this model. Most of the software fault prediction studies focused on developing fault predictors using previous fault data. From a machine learning perspective, these studies are *supervised learning* approaches. However, there are cases when previous fault data are not available. For example, a software company might start to work on a new project domain or might plan on building fault predictors for the first time in their development cycle. In addition, current software version's fault data might not be collected and therefore, there might not exist in any previous fault data for the next release of the software. In these cases, supervised learning approaches cannot be developed because of the absence of class labels.

There are a few studies that have tried to build a fault prediction model when the fault labels for modules are unavailable. Zhong et al. [13] used K-means and Neural-Gas clustering methods to cluster modules, and then an expert labeled each cluster as fault-prone or not fault-prone. To remove the obligation of expert assistance, we developed a prediction model in our previous work [4]. But there is one drawback in our previous study. Since the K-means clustering method is used in the first stage, the K number should be selected heuristically and this number may affect the overall performance of the model. In this new study, we propose a novel fault prediction model that does not require the selection of the K number heuristically. Instead, the K number is calculated automatically with the X-means clustering algorithm. After the identification of the K number and the clusters, metrics thresholds are used again as was done in our previous study. The main contribution of this paper is the development of an automated way of assigning fault-proneness labels to the modules and removing the subjective expert opinion. Subjective human interaction directly affects the quality of the software fault prediction model and adds unnecessary complexity. We explored our new approach on three datasets, collected from a Turkish white-goods manufacturer developing embedded controller software for washing machines, dish washers, and refrigerators. (These datasets, AR3, AR4, and AR5 are available at http://promisedata.org.) Six method-level metrics were used for the development of this model. The metrics used in our experiments are lines of code, cyclomatic complexity, unique operator, unique operand, total operand, and total operator. Threshold vector [LoC, CC, UOp, UOpnd, TOp, and TOpnd] was chosen as [65, 10, 25, 40, 125 and 70]. We started the analysis with the metrics thresholds proposed by Integrated Software Metrics, Inc. (ISM). Later, values were calibrated according to our experiments in order to achieve high-performance prediction models. There are several approaches to calculate software metrics thresholds and Shatnawi et al. [10] developed a ROC (Receiver Operating Characteristic) curve-based threshold calculation technique. Companies may use this approach on previous projects or projects from different domains to find metrics thresholds. The remainder of this paper is structured as follows. Section 17.2 presents related work and Section 17.3 introduces clustering methods. Section 17.4 presents an empirical case study using real-world data from embedded controller software and Section 17.5 explains the conclusion and future work.

17.2 Related Work

Software fault prediction models have been studied since the 1990s and fault-prone program modules can be identified prior to system tests by using these prediction models. According to recent studies, the probability of detection (PD) of fault prediction models (71%) may be higher than PD of software reviews (60%) if a robust model is built [7]. There is an increasing interest in the development of software fault prediction models and numerous techniques have been proposed. Menzies et al. [7] investigated several data mining algorithms for software fault prediction on public NASA datasets and the Naive Bayes algorithm with logNum filter achieved the best performance. Arisholma et al. [1] reported that modeling technique has a limited affect on the prediction accuracy, process metrics are very useful for fault prediction, and the best model is highly dependent on the performance evaluation parameter. However, there are a few software fault prediction studies that do not use prior fault data for modeling. Zhong et al. [13] used the K-means and Neural-Gas clustering methods to cluster modules. Then, one expert labeled each cluster as fault-prone or not fault-prone by examining not only the representative of each cluster, but also some statistical data such as the global mean. Mean squared error (MSE), average purity, and time parameters were used to evaluate the clustering quality. False positive rate (FPR), false negative rate (FNR), and overall misclassification rate parameters were used to evaluate the performance of the expert. According to the expert's opinion, it was simpler to label modules by using Neural-Gas clustering. K-means clustering worked faster than Neural-Gas and Neural Gas's overall misclassification rate was better than K-means clustering's misclassification rate. Neural-Gas performed much better than K-means according to the MSE parameter and its average purity was slightly better than K-means clustering's purity value. However, this approach is dependent on the capability of the expert who should be specialized in the areas of machine learning and software engineering. Furthermore, the selection of the cluster number, K, is done heuristically when the K-means clustering method is chosen and this process can affect the model's performance drastically. Seliya and Khoshgoftaar [9] proposed a constraint-based semi-supervised clustering scheme that uses the K-means clustering method to predict the fault-proneness of program modules when the defect labels for modules are unavailable. The expert labels clusters mainly basing his predictions on his knowledge and some statistical information. On the other hand, the enlargement of the dataset causes an increase of the number of clusters and iterations, which will require the expert to spend much more time on this approach. The need for an expert prevents this method from being performed automatically, presenting one of the major drawbacks of this study. Seliya and Khoshgoftaar [9] reported that the semi-supervised clustering scheme provided better performance than traditional clustering methods and half of the modules that could not be labeled were noisy. Bingbing et al. [3] used the Affinity Propagation clustering algorithm on two datasets and compared the performance of it with the performance of the K-means clustering method. They assumed that there are only two clusters for the dataset and they did not use an automated approach to find the number of clusters.

17.3 Clustering

This section provides introductory information about clustering analysis and clustering methods used during our experimental studies.

17.3.1 Clustering Basics

Clustering is an unsupervised learning approach. While classification uses class labels for training, clustering does not use class labels and tries to discover relationships between the features [5]. Clustering methods can be used to group the modules that have similar metrics by using similarity measures or distances. Cluster analysis has four basic steps: feature selection, clustering algorithm selection, cluster validation, and results interpretation [12]. The classification of clustering algorithms is not easy. A categorization was created by Berkhin [2]. Another categorization was shown in Gan et al.'s [5] study.

17.3.2 Clustering Algorithms

17.3.2.1 K-Means

K-means is a centroid-based clustering algorithm and centroid-based algorithms are more efficient than similarity-based clustering algorithms [13]. In the initialization phase, clusters are initialized with random instances and in the iteration phase, instances are assigned to clusters according to the distances computed between the centroid of the cluster and the instance. This iteration phase goes on until changes do not occur in the clusters.

17.3.2.2 X-Means

One drawback of the K-means algorithm is the selection of the number of clusters, k, as an input parameter. Pelleg and Moore [8] developed an algorithm to solve this problem and used the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) measure for optimization [5]. Rather than choosing the specific number of clusters, k, x-means needs $\mathrm{k}_{\mathrm{min}}$ and $\mathrm{k}_{\mathrm{max}}$ values. The algorithm starts with $\mathrm{k}_{\mathrm{min}}$ value and adds centroids if needed. The BIC or Schwarz criterion is applied to split some centroids into two and hence, new centroids are created. The final centroid set is the one that has the best score.

Given n objects in a dataset $\mathrm{D} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\}$ in a d-dimensional space and a set of alternative models $\mathrm{M}_\mathrm{j} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{k}\}$, scoring of these alternative models, identified with different k values, is done by using the posterior probabilities $\mathrm{P}(\mathrm{M}_\mathrm{j}|\mathrm{D})$ [5]. The Schwarz criterion is shown in Eq. (17.1)

$$BIC(M_j) = \widehat{I}_j(D) - \frac{p_j}{2} \log n \tag{17.1}$$

$\widehat{I}_j(D)$ is the loglikelihood of the jth model and M_j's number of parameters are represented with p_j. The largest score reflects the true model and it is selected as the final model. The maximum likelihood estimate of variance is calculated using Eq. (17.2) under the identical spherical Gaussian distribution and μ_i is the centroid that is closest to the object x_i.

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (x_i - \mu_i)^2 \tag{17.2}$$

The point probabilities are calculated using Eq. (17.3).

$$\hat{P}\,(x_i) = \frac{|C_i|}{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}^d} \exp\left(-\frac{1}{2\hat{\sigma}^2} \|x_i - \mu\|^2\right) \tag{17.3}$$

The loglikelihood of the data is calculated using Eq. (17.4). “The Schwarz criterion is used in X-means globally to choose the best model it encounters and locally to guide all centroid splits” [5].

$$1(\mathrm{D}) = \prod_{i=1}^{n} P(x_i) = \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}\hat{\sigma}^{d}} - \frac{1}{2\hat{\sigma}^{2}} \left\| x_i - \mu_i \right\|^{2} + \log \frac{|C_i|}{n} \right) \tag{17.4}$$

17.4 Empirical Case Studies

This section presents the results of experimental studies.

17.4.1 Performance Evaluation Parameters

We used FPR, FNR and error parameters to evaluate the models. Error is the percentage of mislabeled modules, false positive rate (FPR) is the percentage of not faulty modules labeled as fault-prone by the model, and false negative rate (FNR) is the percentage of faulty modules labeled as not fault-prone. The confusion matrix used to calculate evaluation parameters is shown in Table 17.1. Equations (17.5)–(17.7) are used to calculate FPR, FNR and error values, respectively. All of them must be minimized, but there is a trade-off between FPR and FNR values. FNR value is much more critical than FPR value because high FNR value means that a large amount of fault-prone modules cannot be detected prior to the system testing.

$$FPR = \frac{FP}{FP + TN} \tag{17.5}$$

$$FNR = \frac{FN}{FN + TP} \tag{17.6}$$

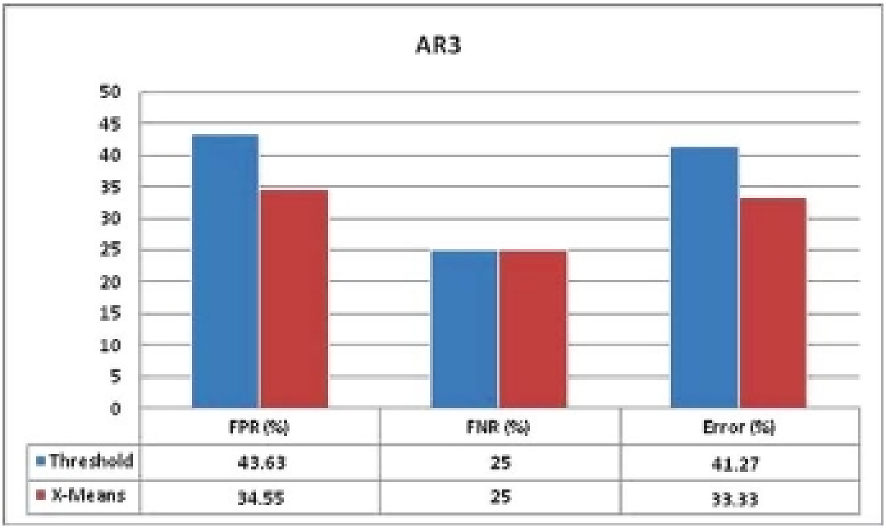
$$Error = \frac{FN + FP}{TP + FP + FN + TN} \tag{17.7}$$

Table 17.1 Confusion matrix

Predicted Labels	Actual labels		
		Yes	No
	Yes	True-positive (TP)	False-positive (FP)
	No	False-negative (FN)	True-negative (TN)

17.4.2 Results and Analysis

We used the X-means clustering-based unsupervised software fault predictor and pure metrics thresholds-based fault predictor for experimental studies. Since K-means clustering requires the selection of the cluster number, we applied X-means clustering method in three datasets and calculated the K values for each of these datasets. The X-means algorithm requires an interval to calculate the best K value. We chose the minimum K value as 2 and maximum K value as the number of data points in that dataset. The X-means algorithm identified the K value as 2 for AR5 and calculated the K value as 3 for AR3 and AR4 datasets. Experimental results are shown in Figs. 17.1–17.3. In order to evaluate the performance of our fully automated approach based on the X-means clustering method, we compared it with our metrics thresholds-based approach. Figures 17.1 and 17.2 show that FPR values decreased for AR3 (from 43.63% to 34.55%) and AR5 datasets (from 32.14% to 14.29%) when the X-means based approach is used. As shown in Fig. 17.3, even though FPR value increased for the AR4 dataset, its FNR value decreased from 20 to 5. As explained in Section 17.4.1, FNR value is much more critical for our models. While our pure metrics thresholds-based approach (Threshold) detects fault-prone modules according to the metrics thresholds, the X-means-based approach first calculates the best K value, divides data points into K clusters and then the mean vector of each cluster is checked against the metrics thresholds vector. Our clustering-based approaches have two stages and the second step for both of them is similar to the pure metrics thresholds-based approach. According to our pure metrics thresholds-based approach, a module is predicted as fault-prone if at least one metric of the module is higher than the specified value of that metric. According



	FPR (%)	FNR (%)	Error (%)
■ Threshold	43.63	25	41.27
■ X-Means	34.55	25	33.33

Fig. 17.1 Experimental results on AR3

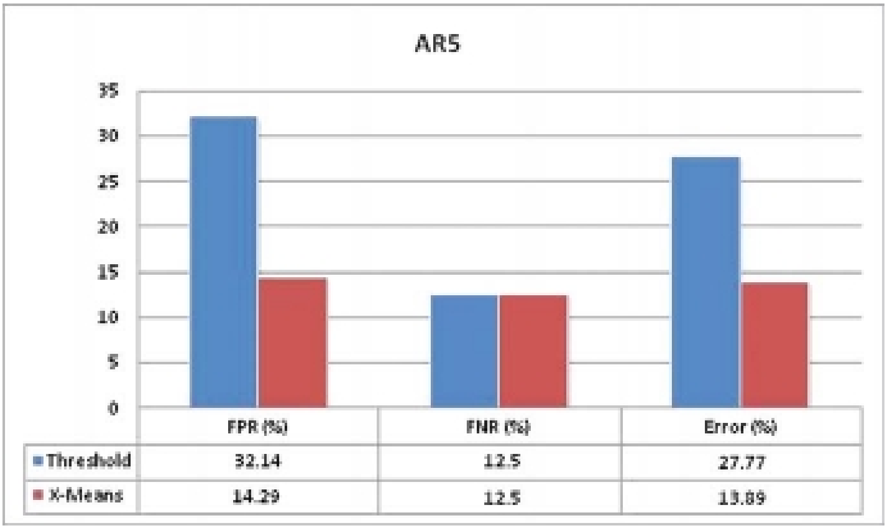


Fig. 17.2 Experimental results on AR5

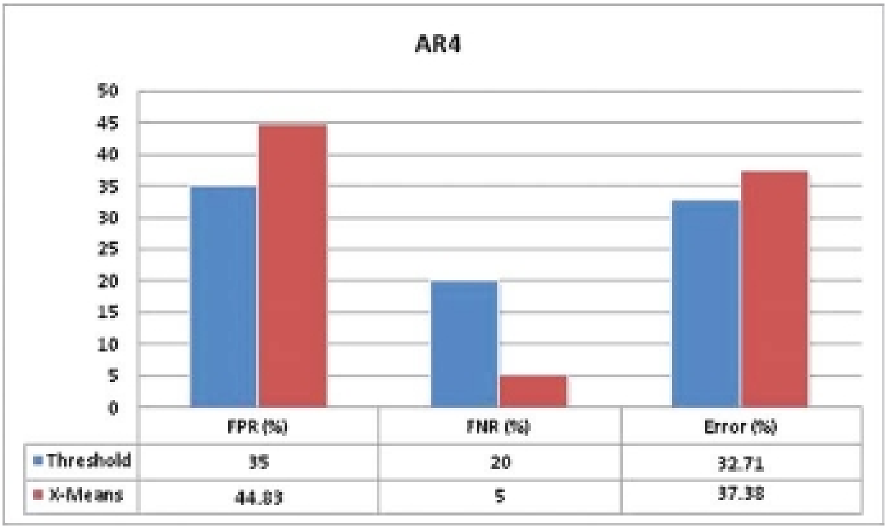


Fig. 17.3 Experimental results on AR4

to our clustering-based approach, a cluster is predicted as fault-prone if at least one metric of the mean vector is higher than the specified threshold value of that metric. Datasets include class labels, but we ignored this column for modeling because our purpose was to develop models for software fault prediction without prior fault data. Class labels were used to calculate the evaluation parameters. We suggest using the X-means-based prediction model for software fault prediction in the absence of class labels. If previous fault data of projects exist, supervised learning algorithms such as Naive Bayes and Random Forests can be applied. However, our research focus was to build fault prediction models that can be used when the fault labels for modules are unavailable. Experiments reveal that unsupervised software fault prediction can be automated fully and effective results can be achieved using X-means clustering with software metrics thresholds. In order to generalize the results of an empirical study outside of the experimental setting, threats to the external validity should be discussed. Datasets used in this study were collected from an industrial environment in Turkey, systems were developed by professional developer groups, and these systems are real industry projects. These features satisfy the requirements explained in Khoshgoftaar et al.'s [6] study. An important point for our models is the effect of noisy instances. Since we calculate the mean vector of each cluster, noisy instances can change the mean vector drastically and hence, the performance of our models may be affected negatively. However, projects used in these experiments were middle-sized and the dataset collection process was done carefully. Therefore, we assumed that there are no noisy instances in these datasets.

17.5 Conclusion and Future Work

Most of the software fault prediction studies in literature assume that there is enough fault data to build the prediction models. However, there are cases when previous fault data are not available. This research problem can be known as *software fault prediction of unlabeled program modules*. This study proposed a novel unsupervised software fault prediction approach. Experiments revealed that unsupervised software fault prediction can be automated fully and effective results can be achieved by using X-means clustering and software metrics thresholds. The main contribution of this paper is the development of an automated way of assigning fault-proneness labels to the modules and removing the subjective expert opinion. There is not a heuristic step in our model to estimate the k number as needed in the K-means clustering method. We studied three publicly available datasets collected from a Turkish white-goods manufacturer. Results are promising and our model can be used when there is not any prior fault data. Our models are not dependent on threshold vector and each company can identify its threshold vector with different approaches such as Shatnawi et al.'s [10] ROC (Receiver Operating Characteristic) curve-based threshold calculation technique. Future work will consider evaluating our model for datasets that have noisy instances, such as the JM1 dataset in the PROMISE repository. A pre-processing step is necessary to remove noisy instances before our prediction model is applied or we need to develop a new, unsupervised fault prediction model that is not sensitive to noisy instances. In addition, additional clustering algorithms and larger datasets can be used for an in-depth analysis of our model.

Acknowledgement This project is supported by the Scientific and Technological Research Council of TURKEY (TUBITAK) under Grant 107E213.

References

1. Arisholma, E., Briand, L.C., Johannessen, E.B.: A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. J. Syst. Softw. **83**(1):2–17 (January 2010)
2. Berkhin, P.: Survey of clustering data mining techniques. Technical Report, Accrue Software, San Jose, CA (2002)
3. Bingbing, Y., Qian, Y., Shengyong, X., Ping, G.: Software quality prediction using affinity propagation algorithm. IJCNN – International Joint Conference on Neural Networks, pp. 1891–1896 (2008)
4. Catal, C., Sevim, U., Diri, B.: Clustering and metrics thresholds based software fault prediction of unlabeled program modules. Proceedings of the Sixth International Conference on Information Technology: New Generations, pp. 199–204 (2009)
5. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. Society for Industrial and Applied Mathematics, Philadelphia (2007)
6. Khoshgoftaar, T.M., Seliya, N., Sundaresh, N.: An empirical study of predicting software faults with case-based reasoning. Softw. Qual. J. **14**(2):85–111 (2006)
7. Menzies, T., Greenwald, J., Frank, A.: Data mining static code attributes to learn defect predictors. IEEE Trans. Softw. Eng. **32**(1):2–13 (2007)
8. Pelleg, D., Moore, A.: X-means: Extending K-means with efficient estimation of the number of clusters. Proceedings of the 17th International Conference on Machine Learning, pp. 727–734 (2000)
9. Seliya, N., Khoshgoftaar, T.M.: Software quality analysis of unlabeled program modules with semi-supervised clustering. IEEE Trans. Syst. Man Cyb A: Syst. Humans **37**(2):201–211 (2007)
10. Shatnawi, R., Li, W., Swain, J., Newman, T.: Finding software metrics threshold values using ROC curves. J. Softw. Maint. Evol.: Res. Pract. (14 Apr 2009, Published Online)
11. Tian, J.: Software Quality Engineering: Testing, Quality Assurance, and Quantifiable Improvement. Wiley, New York (2005)
12. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Networ. **16**(3): 645–678 (2005)
13. Zhong, S., Khoshgoftaar, T.M., Seliya, N.: Unsupervised learning for expert-based software quality estimation. Proceedings of the 8th International Symposium on High Assurance Systems Engineering, Tampa, FL, pp. 149–155 (2004)

Chapter 18
Models of Computation for Heterogeneous Embedded Systems

Adnan Shaout, Ali H. El-Mousa, and Khalid Mattar

Abstract The use of heterogeneous architectures in embedded systems together with the increasing complexities of hardware and software, the increased pressure to deliver full-featured products with reduced time-to-market, and the fact that more embedded systems are using dedicated hardware components (ASIC) and software running on processors is more and more increasing the complexity of designing embedded systems. This ongoing increase in complexities can be overcome with the proper usage of high-level system design techniques such as System Level Design tools and methodologies. In System Level Design, specification languages are used to build high level models of the entire system, to allow fast design space exploration. Models of Computations (MoC) are used as the underlying formal representation of a system. This article specifically investigates the specification and modeling of the computation process used in the co-design approach and its activities. Popular models of computations are presented and compared. Various specification languages for designing embedded are described and compared.

Keywords System level design · hardware/software co-design · heterogeneous embedded systems · models of computation · design languages

A. Shaout (✉) and K. Mattar
The Electrical and Computer Engineering Department,
University of Michigan-Dearborn, Dearborn, USA
e-mail: shaout@umich.edu; kmattar@ford.com

A.H. El-Mousa
Computer Engineering Department,
Faculty of Engineering & Technology,
University of Jordan, Amman, Jordan
e-mail: elmousa@ju.edu.jo

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_18,

18.1 Introduction

18.1.1 Embedded Systems

Embedded systems are special-purpose systems which are typically embedded within larger units providing a dedicated service to that unit [1]. In most embedded systems, the product manufacturer provides a function-specific software application, and end-users have limited access to altering the application running on the system. Examples of embedded systems include consumer electronics products (i.e. cell phones, PDAs, microwaves, etc.), transport control systems, plant control systems and defense systems.

Vahid et al. [2] describe the characteristics of embedded systems that differentiate them from other digital systems:

- *Single-functioned.* Embedded systems repeatedly perform a specific function.
- *Reactive and real time.* Many embedded systems, especially in the control domain, are reactive systems and must continually react to changes in the environment and meet timings constraints without delay.
- *Tightly constrained.* Embedded systems have tight constraints on design metrics. For example, embedded systems must have minimum design costs, must have small form factors and consume minimum power, especially for portable systems, must meet real time requirements, must be safe and reliable, and must have short time-to-market cycle.

A typical heterogeneous embedded system consists of: dedicated hardware parts (ASIC), programmable processors such as microprocessor and ASIP[1] components (i.e. DSP and microcontrollers), memory for data and code, peripherals such A/D, D/A and I/O units, and buses connecting the above components [3].

Traditionally, hardware synthesis tools (logic synthesis and behavior synthesis) have been used to increase productivity. However, hardware synthesis is not sufficient since embedded systems use more software content [4]. In addition, hardware synthesis methods focus on designing a single hardware chip, where more embedded systems are using heterogeneous architectures.

The complexities in designing embedded systems motivate the need for using more efficient tools and design methodologies. *System Level Design* is a methodology to help address these complexities, and enable SoC designs.

18.2 System Level Design

System Level Design is concerned with addressing the challenges encountered in designing heterogeneous embedded systems. In System Level Design, complexities are managed by (1) starting the design process at the highest level of abstraction

[1] Application Specific Instruction-Set Processor.

(System Level), (2) utilizing automated design methodologies to enable step-wise refinements during the design process (3) reusing Intellectual Property (IP) components when feasible [5,6].

The goal of System Level Design is to implement System Level specification on target architecture by refining the specification into a set of target-specific specifications.

Designing at a higher level of abstraction reduces the number of components with which the designer has to deal with, and thus increasing design productivity. This paradigm shift in design requires methodologies and automated tools to support design at higher levels abstractions.

18.2.1 System Level Design Approaches

There are three main system level design approaches: hardware/software co-design, platform-based design and component-based design [7].

- *Hardware/Software co-design* (also referred to *system synthesis*) is a top-down approach. Starts with system behavior, and generates the architecture from the behavior. It is performed by gradually adding implementation details to the design.
- *Platform-based design.* Platform-based design maps the system behavior to predefined system architecture. An example of platform-based design is shown in [8].
- *Component-based design* is a bottom-up approach. It assembles existing heterogeneous components by inserting wrappers between these components. An example of component-based design is described in [9].

18.3 Hardware/Software Co-design

Hardware/Software co-design can be defined as the cooperative design of hardware and software in order to achieve system-level objectives (functionality & constraints) by exploiting the synergism of hardware and software [6, 7]. While hardware implementation provides higher performance, software implementation is more cost effective and flexible since software. The choice of hardware versus software in co-design is a trade-off among various design metrics like performance, cost, flexibility and time-to-market. Figure 18.1 shows the flow of a typical Hardware/Software co-design system.

Generally, Hardware/Software co-design consists of the following activities: *specification and modeling*, *design* and *validation* [6].

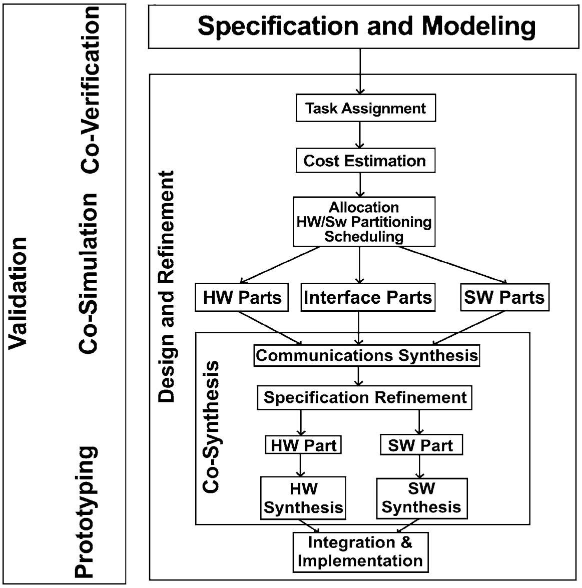


Fig. 18.1 Flow of a typical co-design system

18.3.1 Specification and Modeling

This is the first step in the co-design process. The system behavior at the system level is captured during the specification step [3]. Section 18.4 provides details about specification and modeling, including Models of Computation.

18.3.2 Design and Refinement

The design process follows a step-wise refinement approach using several steps to transform a specification into an implementation. Niemann [3] and O'Nils [6] define the following design steps:

Tasks assignment, Cost estimation, Allocation, Hardware/Software partitioning, Scheduling, and Co-synthesis. Niemann [3] classifies several design steps as part of co-synthesis: Communication synthesis, Specification refinement, Hardware synthesis and Software synthesis.

18.3.3 Validation

Informally, validation is defined as the process of determining that the design, at different levels of abstractions, is correct. The validation of hardware/software systems is referred to as *co-validation*. Methods for co-validations are [9, 10]: *Formal verification* and*Simulation*. A comparison of co-simulation methods is presented in [11].

18.4 Specification and Modeling

Specification is the starting point of the co-design process, where the designer specifies the system's specification without specifying the implementations. Languages are used to capture system specifications. Modeling is the process of conceptualizing and refining the specifications. A model is different from the language used to specify the system. A model is a conceptual notation that describes the desired system behavior, while a language captured that concept in a concrete format. A model can be captured in a variety of languages, while a language can capture a variety of models [2].

Two approaches are used for system specification, *homogeneous modeling* where one specification language is used for specifying both hardware and software components of a heterogeneous system and *heterogeneous modeling* which uses specific languages for hardware (e.g. VDHL), and software (e.g. C) [6, 12].

18.4.1 Models of Computation

A computational model is a conceptual formal notation that describes the system behavior [2]. Ideally, a Model of Computation (MOC) should comprehend *concurrency*, *sequential behavior* and *communication methods* [10]. Co-design systems use computational models as the underlying formal representation of a system. A variety of Models of Computation have been developed to represent heterogeneous systems.

The following is an overview of common MOCs based on the work in Cortes et al. [10] and Bosman [13].

18.4.1.1 Finite State Machines (FSM)

The FSM model consists of a set of states, a set of inputs, a set of outputs, an output function, and a next-state function [14]. A system is described as set of states and input values can trigger a transition from one state to another. FSMs are commonly used for modeling control-flow dominated systems. The main disadvantage of FSMs is the exponential growth of the number of the states as the system complexity rises due the lack of hierarchy and concurrency. To address the limitations of the classical FSM, researches have proposed several derivates of the classical FSM. Some of these extensions are described below.

- *SOLAR* [15] is based on the Extended FSM model (EFSM), which can support hierarchy and concurrency. In addition, SOLAR supports high level communication concepts including channels and global variables. It is used to represent high-level concepts in control-flow dominated systems, and it is mainly suited for synthesis purposes. The model provides an intermediate format that allows hardware/software designs at the system-level to be synthesized.

- *Hierarchical Concurrent FSM (HCFSM)* [3] solve the drawbacks of FSMs by decomposing states into a set of sub-states. These sub-states can be concurrent sub-states communicating via global variables. Therefore, HCFSMs supports hierarchy and concurrency. Statecharts is a graphical state machine language designed to capture the HCFSM MOC [2]. The communication mechanism in statecharts is instantaneous broadcast, where the receiver proceeds immediately in response to the sender message. The HCFSM model is suitable for control oriented/real time systems.
- *Codesign Finite State Machine (CFSM)* [16, 17] adds concurrency and hierarchy to the classical FSM, and can be used to model both hardware and software. It is commonly used for modeling control-flow dominated systems. The communication primitive between CFSMs is called an event, and the behavior of the system is defined as sequences of events. CFSMs are widely used as intermediate forms in co-design systems to map high-level languages, used to capture specifications, into CFSMs.

18.4.1.2 Discrete-Event Systems

In a Discrete Event system, the occurrence of discrete asynchronous events triggers the transitioning from one state to another. An event is defined as an instantaneous action, and has a time stamp representing when the event took place. Events are sorted globally according to their time of arrival. A signal is defined as set of events, and it is the main method of communication between processes [10]. Discrete Event modeling is often used for hardware simulation. For example, both Verilog and VHDL use Discrete Event modeling as the underlying Model of Computation [11]. Discrete Event modeling is expensive since it requires sorting all events according to their time stamp.

18.4.1.3 Petri Nets

Petri Nets are widely used for modeling systems. Petri Nets consist of places, tokens and transitions, where tokens are stored in places. Firing a transition causes tokens to be produces and consumed. Petri Nets supports concurrency and is asynchronous; however, they lack the ability to model hierarchy. Therefore, it can be difficult to use Petri Nets to model complex systems due to its lack of hierarchy. Variations of Petri Nets have been devised to address the lack of hierarchy. For example, the Hierarchical Petri Nets (HPNs) proposed by Dittrich [18].

- *Hierarchical Petri Nets* (*HPNs*) supports hierarchy in addition to maintaining the major Petri Nets features such as concurrency and asynchronously. HPNs use Bipartite[2] directed graphs as the underlying model. HPNs are suitable for modeling complex systems since they support both concurrency and hierarchy.

18.4.1.4 Data Flow Graphs

In Data Flow Graph (DFG), systems are specified using a directed graph where nodes (actors) represent inputs, outputs and operations, and edges represent data paths between nodes [3]. The main usage of Data Flow is for modeling data flow dominated systems. Computations are executed only where the operands are available. Communications between processes is done via unbounded FIFO buffering scheme [10]. Data Flow models support hierarchy since the nodes can represent complex functions or another Data Flow [6, 10].

Several variations of Data Flow Graphs have been proposed in the literature such as Synchronous Data Flow (SDF) and Asynchronous Data Flow (ADF) [18]. In SDF, a fixed number of tokens are consumed, where in ADF the number of tokens consumed is variable. Lee [19] provides an overview of Data flow models and its variations.

18.4.1.5 Synchronous/Reactive Models

Synchronous modeling is based on the synchrony hypothesis, which states that outputs are produced instantly in reaction to inputs and there is no observable delay in the outputs [12]. Synchronous models are used for modeling reactive real time systems. Cortes in [11] mentions two styles for modeling reactive real time systems: multiple clocked recurrent systems (MCRS) which is suitable for data dominates real time systems and state base formalisms which is suitable for control dominated real time systems. Synchronous languages such as Esterel [15] is used for capturing Synchronous/Reactive Model of Computation [11].

18.4.1.6 Heterogeneous Models

Heterogeneous Models combine features of different models of computation. Two examples of heterogeneous models are presented.

- *Programming languages* [20] provide a heterogonous model that can support data, activity and control modeling. Two types of programming languages: *imperative* such as C, and *declarative* languages such as LISP and PROLOG. In imperative languages, statements are executed in the same order specified in the specification. On the other hand, execution order is not explicitly specified in

[2] A graph where the set of vertices can be divided into two disjoint sets. U and V such that no edge has both end-points in the same set.

declarative languages since the sequence of execution is based on a set of logic rules or functions. The main disadvantage of using programming languages for modeling is that most languages do not have special constructs to specify a system's state [3]

Program State Machine (PSM) is a merger between HCFSM and programming languages. A PSM model uses a programming language to capture a state's actions [20]. A PSM model supports hierarchy and concurrency inherited from HCFSM. The *Spec Charts* language, which was designed as an extension to VHDL, is capable of capturing the PSM model. The *Spec C* is another language capable of capturing the PSM model. *Spec C* was designed as an extension to C [2].

18.4.2 Comparison of Models of Computation

A comparison of various Models of Computation is presented by Bosman [13], and Cortes et al. [10]. Each author compares MOCs according to certain criteria. Table 18.1 compares MOCs based on the work done in [10, 13].

18.4.3 Specification Languages

The goal of a specification language is to describe the intended functionality of a system non-ambiguously. A large number of specification languages are currently being used in embedded system design since there is no language that is the best for all applications [3]. Below is a brief overview of the widely used specification languages [2, 6]:

18.4.3.1 Formal Description Languages

Examples of formal languages are *LOTOS* and *SDL.*

- *LOTOS* is based on process algebra, and used for the specification of concurrent and distributed system.
- *SDL* used for specifying distributed real time systems, and based on extended FSM.

18.4.3.2 Real Time Languages

Esterel & StateCharts are examples of real time languages.

- *Esterel* is a synchronous programming language based on the synchrony hypothesis. It is used for specifying real time reactive systems. Esterel is based on FSM, with constructs to support hierarchy and concurrency.

Table 18.1 Comparison of models of computation [16, 20]

MOC	Origin MOC	Main application	Clock mechanism	Orientation	Time	Communication method	Hierarchy
SOLAR	FSM	Control oriented	Synch	State	No explicit timings	Remote procedure Call	Yes
HCSFM/State Charts	FSM	Control oriented/reactive real time	Synch	State	Min/max time spent in state	Instant broadcast	Yes
CFSM	FSM	Control oriented	Async	State	Events w/t time stamp	Event broadcast	Yes
Discreet Event	N/A	Real time	Synch	Timed	Globally sorted events w/t time stamp	Wired signals	No
HPN	Petri Net	Distributed	Async	Activity	No explicit timings	N/A	Yes
SDF	DFG	Signal processing	Synch	Activity	No explicit timings	Unbounded FIFO	Yes
ADF	DFG	Data oriented	Async	Activity	No explicit timings	Bounded FIFO	Yes

- *StateCharts* is graphical specification language used for specifying reactive system. StateCharts extend FSM by supporting hierarchy, concurrency and synchronization.

18.4.3.3 Hardware Description Languages (HDL)

Commonly used HDL are *VHDL*, *Verilog* and HardwareC.

- *VHDL* is IEEE standardized HW description language.
- *Verilog* is another hardware description language, which has been standardized by IEEE.
- *HardwareC* is a C based language designed for hardware synthesis. It extends C by supporting structural hierarchy, concurrency, communication and synchronization.

18.4.3.4 System Level Design Languages (SLDL)

System Level Design Languages (SLDL) are used to capture specification and model embedded system at the system abstraction level. With the increased time-to-market pressure, and to enable SoC designs, SLDS need to be able to specify and model all aspects of the system at higher abstraction level (at the System Level). This will allow early design space exploration to evaluate various design alternatives early in the design process. Most current SLDLs lack built-in support for specifying and modeling ALL aspects of a heterogeneous embedded system at the System Level. Some of these deficiencies are lack of support for:

- RTOS modeling at the System Level. This is important for modeling real time embedded system, and determining if the scheduling policy will meet time constraints and deadline at the System Level before committing to a specific RTOS implementation.
- Composing Heterogeneous models with multiple MoCs.
- Estimating power consumption at the System Level.

Examples of SLDL are SpecC and SystemC.

- *SpecC* [14] is system level design language based on ANSI C. It was developed at the University of California, Irvine to improve traditional HDL languages such as VHDL. The SpecC language models systems as a hierarchal network of behaviors and channels [3]. SpecC supports behavior and structural hierarchy, concurrency, state transition, exception handling, timing aspects and synchronization. Built on the SpecC language is the SpecC design methodology.
- *SystemC* [21] is a C++ library based language designed by the OpenSystemC Initiative (OCSI) group to improve traditional HDL languages.

Table 18.2 Comparison of specification languages

	Formal languages		Real-time languages		HDL		SLDL		
	LOTOS	SDL	Esterel	State charts	VHDL	Verilog	HardwareC	SpecC	SystemC
Structural hierarchy	F	F	F	N	F	F	F	F	F
Behavior hierarchy	F	P	F	F	P	P	P	F	F
Concurrency	F	F	F	F	F	F	F	F	F
Synchron.	F	F	F	F	F	F	F	F	F
Exception handling	F	P	F	F	N	F	N	F	F
Timing	N	P	N	P	F	F	P	F	F
State transition	F	F	N	F	N	N	N	F	F
Formal verification	F	F	F	N	N	N	N	N	N
Model executability	N	F	F	F	F	F	F	F	F
Full support for RTOS	N	N	N	N	N	N	N	N	N

F: fully supported, P: partially supported, N: not supported

18.4.4 Requirements for Specification Languages

Gajski in [14] and Niemann in [3] describe the requirements for specification languages:

- *Hierarchy* is an important feature of a specification language. Two types of hierarchy: (1) *behavior hierarchy* which allows a behavior to be decomposed of sub-behaviors, (2) *structural behavior* which allows a system to be specified as a set of interconnected components, where these components can be specified as sub-components as well.
- *State transition* is important for modeling control and reactive embedded systems.
- *Concurrency* a large number of embedded systems consist of tasks that are working concurrently.
- *Synchronization* needed when concurrent parts of exchange data.
- *Exception handling* exceptions such as reset and interrupts often occur in embedded systems. When an interrupt occurs, the system has to transition to a new state to handle the interrupt. Once the interrupt is serviced, the system has to go back to point prior to interrupt. Specification languages should be able to model exceptions.
- *Timing* is an important aspect of specifying real time embedded systems. Two timing aspects have to be specified when dealing with embedded systems: Functional timing which represents the time consumed for executing a behavior, and timing constraints which represent a range of time for executing a behavior.
- *Formal verification* desirable for specification languages since it provides a mechanism to verify the use of formal mathematical methods.
- *Support for RTOS modeling* is important for the specification of real time systems that will use a RTOS to implement dynamic scheduling.

Table 18.2 shows a comparison of different specification languages.

References

1. De Michell, G., Gupta, R.K.: Hardware/software co-design. Proc. IEEE **85**(3), 349–365 (1997)
2. Vahid, F., Givargis, T.: Embedded System Design: A Unified Hardware/Software Introduction. Wiley, Hoboken, NJ, (2002)
3. Niemann, R.: Hardware/Software Co-Design for Data Flow Dominated Embedded Systems. Kluwer, Boston, MA (1998)
4. Domer, R.: System-level modeling and design with the SpecC language. Ph. D. Dissertation, Department of Computer Science, University of Dortmund, Dortmund, Germany (2000)
5. Dömer, R., Gajski, D., Zhu, J.: Specification and design of embedded systems. it+ ti Magazine (3). Oldenbourg Verlag, Munich, Germany (June 1998)
6. O'Nils, M.: Specification, synthesis and validation of hardware/software interfaces. Doctoral thesis, Department of Electronics, Royal Institute of technology, Stockholm (1999)

7. Cai, L.: Estimation and exploration automation of system level design. Ph.D. dissertation, Department of Information and Computer Science, University of California, Irvine, CA (2004)
8. Keutzer, K., Malik, S., Newton, A.R., Rabaey, J.M., Sangiovanni-Vincentelli, A.: System-level design: orthogonalization of concerns and platform-based design. IEEE Trans. Comput-Aid. Design Integ. Circ. Syst. **19**(12), 1523–1543 (2000)
9. Cesario, W., Baghdadi, A., Gauthier, L., Lyonnard, D., Nicolescu, G., Paviot, Y., Yoo, S., Jerraya, A.A., Diaz-Nava, M.: Component-based design approach for multicore SoCs. Proceedings of 39th Design Automation Conference (DAC02), New Orleans, LA, pp. 789–794 (2002)
10. Cortes, L.A., Eles, P., Peng, Z.: A survey on hardware/software codesign representation models. SAVE Project Report, Department of Computer and Information Science, Linköping University, Linköping, Sweden (June 1999)
11. Edwards, S., Lavagno, L., Lee, E.A., Sangiovanni-Vincentelli, A.: Design of embedded systems: formal models, validation, and synthesis. Proceedings of IEEE **85**(3), 366–390 (1997)
12. Boussinot, F., de Simone, R., Ensmp-Cma, V.: The ESTEREL language. Proceedings of IEEE 79(9),1293–1304 (1991)
13. Bosman, G., Bos, I.A.M., Eussen, P.G.C., Lammel, I.R.: A survey of co-design ideas and methodologies. Master's Thesis at Vrije Universiteit Amsterdam (2003)
14. Gajski, D.D., Zhu, J., Dömer, R., Gerstlauer, A., Zhao, S.: SpecC, Specification Language and [design] Methodology. Kluwer, Boston, MA (2000)
15. Jerraya, A.A., O'Brien, K.: SOLAR: An intermediate format for system-level modeling and synthesis. In: Buchenrieder, K., Rozenblit, J. (eds.) Computer Aided Software/Hardware Engineering. IEEE Press, Piscataway, NJ (1995)
16. POLIS Group. POLIS, A framework for hardware-software co-design of embedded systems. http://embedded.eecs.berkeley.edu/research/hsc/. Accessed 5 April 2009
17. Jerraya, A.A., O'Brien, K.: SOLAR: An intermediate format for system-level modeling and synthesis. In: Buchenrieder, K., Rozenblit, J. (eds.) Computer Aided Software/Hardware Engineering. IEEE Press, Piscataway, NJ (1995)
18. Agrawal, A.: Hardware modeling and simulation of embedded applications. M.S. thesis, Department of Electrical Engineering, Vanderbilt University, Nashville, TN (2002)
19. Lee, E.A., Parks, T.M.: Dataflow process networks. Proceedings of IEEE **83**(5), 773–801 (1995)
20. Gajski, D.D., Zhu, J., Dömer, R.: Essential issues in codesign. In: Staunstrup, J., Wolf, W. (eds.) Hardware/Software Co-Design: Principles and Practice. Kluwer, Boston, MA (1997)
21. Open SystemC Initiative. SystemC. http://www.systemc.org/. Accessed on April 5 200921

Chapter 19
A Quotient-Graph for the Analysis of Reflective Petri Nets

Lorenzo Capra

Abstract The design of dynamic, adaptable discrete-event systems calls for adequate modeling formalisms and tools in order to manage possible changes occurring during system's lifecycle. A common approach is to pollute the design with details not concerning the current system behavior, rather its evolution. That hampers analysis, reuse and maintenance in general. A Petri net-based reflective model (based on classical Petri nets) was recently proposed to support dynamic discrete-event system's design, and was applied to dynamic workflow's management. Behind there is the idea that keeping functional aspects separated from evolutionary ones, and applying evolution to the (current) system only when necessary, results in a clean formal model for dynamic systems. This model preserves the ability of verifying properties typical of classical Petri nets. As a first step toward the implementation (in the short time) of a discrete-event simulator, Reflective Petri nets are provided in this paper with a semantics defined in terms of labeled state-transitions.

Keywords Petri nets · dynamic systems · evolution · quotient-graph

19.1 Introduction

Most existing discrete-event systems are subject to evolution during their lifecycle. Think e.g. of mobile ad-hoc networks, adaptable software, business processes, and so on. Designing dynamic/adaptable discrete-event systems calls for adequate modeling formalisms and tools. Unfortunately, the known well-established formalisms for discrete-event systems, such as classical Petri nets [13], lack features for naturally expressing possible run-time changes to system's structure. An approach commonly followed consists of polluting system's functional aspects with details concerning evolution. That practice hampers system analysis, reuse and maintenance.

L. Capra (✉)
Department of Informatics and Communication (D.I.Co),
Università degli Studi di Milano, MI Italy 20139
e-mail: capra@dico.unimi.it

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_19,

Reflective Petri nets [5] have been recently proposed as design framework for dynamic discrete-event systems, and successfully applied to dynamic workflows [4]. They rely on a reflective layout formed by two logical levels. The achieved clean separation between functional and evolutionary concerns results in a simple formal model for systems exhibiting a high dynamism, which should preserve the analysis capabilities of classical Petri nets. With respect to other dynamic extensions of Petri nets appeared in last decade, which set up new (hybrid) paradigms [2, 3, 9, 12], the Reflective Petri nets approach tries to achieve a satisfactory compromise between expressive power and analysis capability, through a rigorous application of reflection concepts in a consolidated Petri net framework.

On the perspective of implementing in the short time an automatic solver and a discrete-event simulation engine, Reflective Petri nets are provided in this paper with a (labeled) state-transition semantics. Any analysis/simulation techniques based on state-space inspection has to face a crucial question, that is how to recognize possible equivalent states during base-level's evolution. That major topic is managed by exploiting the symbolic state definition the particular Colored Petri net flavor [11] used for the meta-level is provided with, and represents the paper's original contribution.

The balance is as follows: background information on Reflective Petri nets and the employed Petri net formalisms are given in Sections 19.2 and 19.3. The focus is put there on those elements directly connected to the paper's main contribution, the definition of a state-transition semantics for Reflective Petri nets (Section 19.4). An application of the semantics to a dynamic system taken from literature is summarized in Section 19.5. Related works are mentioned and discussed in Section 19.6. Finally Section 19.7 is about work-in-progress.

19.2 WN's Basic Notions

The formalisms employed for the two levels (meta- and base-) of the reflective layout are Well-formed Nets (WN) [6], a flavor of Colored Petri nets (CPN) [11], and their unfolded counterpart, an extension of ordinary Place/Transition nets [13], respectively. This choice has revealed convenient for two main reasons: first, the behavior of Reflective Petri nets can be formally stated in terms of classical Petri nets state-transition; secondly, the symbolic state notion peculiar of WN makes it possible to efficiently recognize equivalent base-level's evolutions.

While retaining CPN's expressive power, WNs are characterized by a structured syntax, exploited by efficient analysis algorithms.This section does not present all the features of WNs, for which the reader can refer to [6], just introduces them informally, focusing on the symbolic marking definition. Unlike CPNs, WNs include priority levels for transition and inhibitor arcs. These features enhance the formalism expressiveness and are helpful to represent the transactional execution of evolutionary strategies.

As in CPN, places as well as transitions are associated to *color domains*, i.e., tokens in places have an identifier (color), similarly transitions are parameterized, so that there exist different *color instances* of a given transition. A *marking* $\mathbf{M}$ maps every place p to a multiset on the respective color domain $\mathcal{C}(p)$. The projection of $\mathbf{M}$ to a subset P' is denoted $\mathbf{M}[P']$. Any arc connecting p to a transition t is labeled by a function mapping every element of $\mathcal{C}(t)$ to a multiset on $\mathcal{C}(p)$.

SWN color domains are Cartesian products of *basic color classes* C_{i}. A class C_{i} may be in turn partitioned into *static subclasses* $C_{\mathrm{i,k}}$. The idea is that objects in a subclass are indistinguishable from one another.

19.2.1 The Symbolic Marking

The Symbolic Marking [7] provides syntactical equivalence relation on ordinary WN markings: two markings belong to the same SM if and only if they can be obtained from one another by means of a permutation of colors that preserve static subclasses. A SM (denoted $\widehat{\mathbf{M}}$) is formally expressed in terms of *dynamic subclasses*, and specifies the distribution of symbolic colors (tuples built of dynamic subclasses) over the WN places.

Dynamic subclasses define a parametric partition of color classes preserving the static partition: let Z_{i} and s_{i} denote the set of dynamic subclass of C_{i} (in $\widehat{\mathbf{M}}$), and the number of static subclasses of C_{i} ($s_{\mathrm{i}} \geq 1$). The j-th dynamic subclass of C_{i}, $z_{\mathrm{j}}^{\mathrm{i}} \in Z_{\mathrm{i}}$, refers to a static subclass, denoted $d(z_{\mathrm{j}}^{\mathrm{i}})$, $1 \leq d(z_{\mathrm{j}}^{\mathrm{i}}) \leq s_{\mathrm{i}}$, and has a cardinality $|z_{\mathrm{j}}^{\mathrm{i}}|$, i.e., it represents a parametric set of colors. It must hold, $\forall k : 1 \ldots s_{\mathrm{i}}$ $\sum_{\mathrm{j}:d(z_{\mathrm{j}}^{\mathrm{i}})=k} \left|z_{\mathrm{j}}^{\mathrm{i}}\right| = \mid C_{\mathrm{i,k}} \mid$.

The SM canonical form [7], based on a lexicographic ordering and a minimization of dynamic subclass distribution over the places, provides a way to uniquely represent an SM.

19.3 Reflective Petri Nets Layout

The *Reflective Petri nets* [5] approach relies on a logical layout divided in two levels. The first one, called *base-level*, is an *ordinary Petri net* (a P/T net with priorities and inhibitor arcs) representing the system prone to evolve (*base-level PN*); while the second level, called *meta-level*, consists of a *high-level Petri net* (a colored Petri net) representing the evolutionary strategies (the meta-program, following the reflection parlance) that drive the evolution of the base-level upon occurrence of certain conditions/events.

The meta-level acts on a representative of the base-level, called *reification*, which is formalized by a colored marking. The reification is used by the meta-program to observe (*introspection*) and manipulate (*intercession*) the base-level PN. Changes to the reification are reflected down to the base-level at the end of a meta-computation (*shift-down*).

The meta-level is implicitly activated (*shift-up*) at any base-level change of state. Then a strategy is selected depending on whether (a) the base-level has entered a given condition, (b) and/or any external events (simulated at meta-level) have occurred. The ability of specifying arbitrary selection conditions enhances the flexibility of the reflective layout.

The *reflective framework* is another high-level Petri net component, somehow similar to a transparent meta-layer, which is in charge of implementing base-level's introspection and intercession. The framework has a fixed layout, formed by higher-priority transitions. Intercession is performed in terms of a minimal, complete set of low-level operations (the *evolutionary interface*): addition/removal of nodes and arcs, change of transition priorities (structural changes), free moving of tokens overall the base-level PN places (state changes). If one such operation fails, the meta-program as a whole is restarted and any changes caused in the meanwhile to the reification are discarded. Trying to delete a yet not existing node is an example of failure In other words, the evolutionary strategies have a transactional semantics. After a strategy's succeeding run, changes are reflected down to the base-level Petri net.

A designer is provided with a tiny ad-hoc language, originally inspired to Hoare's CSP, to specify his/her own strategy in a simple way, without any skills in high-level Petri net being required. An automatic translation to a corresponding high-level Petri net is done.

Several strategies could be candidate for execution at a given instant: different policies might be adopted to select one, varying from a non deterministic choice, to a static assignment of priorities. According to the reflective paradigm, the base-level is unaware of the meta-program. The system designer may freely decide, using priority, to block the base-level while the meta-program is active, or to leave it running. It may also define local influence areas for some strategies, by (temporarily) locking corresponding portions of the base-level.

Let us only outline here some essential points about the interaction between base- and meta-levels:

1. The reflective framework and the meta-program share two sets of boundary colored places, denoted *reification* set and *evolutionary interface* in the sequel. Their composition, through a simple superposition of shared places, gives rise to the meta-model, called *meta-level PN*.
2. The reification is a *well-defined* marking of the reification set `Reif` formed by $\{\mathtt{reifP}, \mathtt{reifT}, \mathtt{reifA}, \mathtt{reif\Pi}, \mathtt{reifM}\}$. Such places encode structure (nodes, connections, and priorities) and current state of the base-level PN, respectively. Their color domains are built of basic classes *Place*, *Tran*, which are (logically) unbounded repositories holding all *potential* base-level nodes (they must contain the nodes of the initial base-level Petri net). We have: $\mathcal{C}(\mathtt{reifP})$, $\mathcal{C}(\mathtt{reifM})$: *Place*; $\mathcal{C}(\mathtt{reifT})$, $\mathcal{C}(\mathtt{reif\Pi})$: *Tran*; $\mathcal{C}(\mathtt{reifA})$: $\mathit{Place} \times \mathit{Tran} \times \{i, o, h\}$.

3. The isomorphism between base-level nets and reification is formalized by a bijection φ. For example a net formed by places $\{p_1, p_2, \ldots\}$, transitions $\{t_1, t_2, \ldots\}$ having priority levels $0, 1, \ldots$ respectively, by an input arc (p_1, t_1) of weight 2, an output arc (t_1, p_2) of weight 1 .., whose current marking is $\mathbf{m}(p_1) = 2$, $\mathbf{m}(p_2) = 1$, is reified as: $\mathbf{M}(\texttt{reifP}) = p_1 + p_2 + \ldots$, $\mathbf{M}(\texttt{reifT}) = t_1 + t_2 + \ldots$, $\mathbf{M}(\texttt{reif}\Pi) = t_2 + \ldots$, $\mathbf{M}(\texttt{reifA}) = 2 \cdot \langle p_1, t_1, i\rangle + \langle p_1, t_2, o\rangle + \ldots$, $\mathbf{M}(\texttt{reifM}) = 2 \cdot p_1 + p_2$.
4. A back-up copy $\texttt{Reif}_{back}$ of set `Reif` is kept. Evolutionary strategies work on `Reif`: if an operation fails, then the contents of $\texttt{Reif}_{back}$ are copied back to `Reif`, and the control passes to the base-level.
5. The shift-up is implemented in transparent way at net-level, by connecting every base-level transition to the place(s) `reifM` ($\texttt{reifM}_{back}$) by means of colored arcs. The resulting model is denoted *base-meta PN*. Consider transition t_1 of the above example: its firing makes two colors p_1 and one color p_2 be withdrawn/ added from/to `reifM` ($\texttt{reifM}_{back}$), respectively. Base-level changes of state are thus instantaneously mirrored on the reification, maintaining base-level's unawareness of the meta-level.
6. The shift-down is emulated by a homonym highest-priority meta-transition of the meta-level PN.

19.4 State-Transition Semantics for Reflective Nets

The causal connection between base- and meta-level makes it possible to formalize the behavior of Reflective Petri nets in terms of WN state-transitions:

Definition 19.1 (Reflective Petri net state). A state of a Reflective Petri net is a marking $\mathbf{M}_\mathrm{i}$ of the base-meta PN.

Let PN_0 be the (marked) base-level Petri net which models the initial system. Assume it has been connected to the meta-level. The initial state of the corresponding Reflective Petri net is obtained setting: $\mathbf{M}_0[\texttt{Reif}] = \mathbf{M}_0[\texttt{Reif}_{back}] = \varphi(PN_0)$.

Let t_c be any transition (color instance) of the base-meta PN, other than `shiftdown`. If t_c is enabled in $\mathbf{M}_\mathrm{i}$, according to the ordinary enabling rule, and $\mathbf{M}_\mathrm{j}$ is the marking reached upon the firing, we have the state-transition:

$$\mathbf{M}_\mathrm{i} \xrightarrow{t_\mathrm{c}} \mathbf{M}_\mathrm{j}$$

Only one case must be treated apart. Let *shift-down* be enabled in $\mathbf{M}_\mathrm{i}$, according to the ordinary firing rule. Then:

$$\mathbf{M}_\mathrm{i} \overset{shift-down}{\longrightarrow} \mathbf{M}'_0,$$

$\mathbf{M}'_0$ being the marking of the base-meta PN obtained by firing *shift-down* in the ordinary way, making the contents of $\texttt{Reif}_{back}$ be updated to $\texttt{Reif}$, finally (side-effect), replacing the current base-level PN with $PN'= \varphi^{-1}(\mathbf{M}_\mathrm{i}[\texttt{Reif}])$ and connecting PN' to the meta-level.

19.4.1 Handling Equivalent Evolutions

The just introduced state-transition semantics defines precisely the untimed behavior of a reflective Petri net, but suffers from two evident drawbacks affecting efficiency and effectiveness. First, the notion of state is exceedingly redundant, comprising a part, the meta-level, which outs the functional specification of a system. Secondly, there is no way of recognizing whether the system dynamics/evolution leads to equivalent conditions. The latter question is critical: the ability of deciding about finiteness and strongly-connectedness (strictly related to the ability of recognizing equivalences) is in fact mandatory for any techniques based on state-space inspection.

Recognizing equivalences in an evolving system is tricky. It may happen that after a series of transformations the base-level comes back to the original condition (state). Even more likely, the internal dynamics of the evolving system might lead to equivalent conditions. The problem is tackled by resorting to the symbolic marking notion, peculiar of WN, and the base-level reification at the meta-level.

The modeler, on his/her needs may define a *logical* partition of classes *Place*, *Tran*, possibly different from the completely split partition (implicitly) adopted when setting up the base-meta PN:

$$Place = P_1 \cup P_2 \cup \ldots P_\mathrm{k} \quad Tran = T_1 \cup T_2 \cup \ldots T_\mathrm{n}$$

The idea is simple: elements belonging to a subclasses P_i (T_j) denote indistinguishable base-level nodes, which might be freely permuted, without altering the model's semantics. Those nodes that, for any reasons, must preserve their identity during evolution, will correspond to cardinality one subclasses. The default logical partition is that in which all places/transitions can be permuted. Of course the evolutionary strategies refer to the logical partition of base-level nodes.

The causal connection between base- and meta- levels establishes an exact correspondence (at any instant) between the current base-level PN and the contents of $\texttt{Reif}_{back}$. On the light of that, we state the following state-equivalence notion, in which we refer to the logical partition of *Place* and *Tran*.

Definition 19.2 (state equivalence). Let $\widehat{\mathbf{M}}_\mathrm{i}$ be the symbolic marking obtained from $\mathbf{M}_\mathrm{i}[\texttt{Reif}_{back}]$ replacing every $p_\mathrm{i} \in P_\mathrm{k}$ ($t_\mathrm{j} \in T_\mathrm{l}$) with a corresponding dynamic subclass z^1_i (z^2_j), $d(z^1_\mathrm{i}) = k$ ($d(z^2_\mathrm{j}) = l$), $|z^1_\mathrm{i}|$ ($|z^2_\mathrm{j}|$) $= 1$. Then $\mathbf{M}_\mathrm{i} \equiv \mathbf{M}_\mathrm{j}$ if and only if $\widehat{\mathbf{M}}_\mathrm{i} \equiv \widehat{\mathbf{M}}_\mathrm{j}$.

$\widehat{\mathbf{M}}_{\mathrm{i}}$ represents an equivalence class of states (Def.19.1), so we shall use the notation $\mathbf{M} \in \widehat{\mathbf{M}}_{\mathrm{i}}$. The state-transition notion is redefined accordingly.

Definition 19.3 (visible state-transition). Let σ be a finite sequence of meta-level transition color instances other than `shiftdown` (σ possibly empty). Then $\mathbf{M}_{\mathrm{i}} \xrightarrow{t} \mathbf{M}_{\mathrm{j}}$, if and only if t is either `shiftdown` or a base-level transition, and there exist $\sigma, \mathbf{M}'_{\mathrm{i}}$ s.t. $\mathbf{M}_{\mathrm{i}} \xrightarrow{\sigma} \mathbf{M}'_{\mathrm{i}} \xrightarrow{t} \mathbf{M}_{\mathrm{j}}$ (according to the above definition).

$\mathbf{M}'_{\mathrm{i}}$, as well as any intermediate marking crossed by σ, are equivalent to $\mathbf{M}_{\mathrm{i}}$. Visible state-transitions are caused by the occurrence of either *shiftdown*, or any base-level transition. Meta-level transition sequences (σ) are not visible to an external observer.

We call reachable a state $\mathbf{M}_{\mathrm{i}}$ such that $\mathbf{M}_0 \xrightarrow{t_1} \mathbf{M}_1 \xrightarrow{t_2} \ldots \mathbf{M}_{\mathrm{i}}$. We say $\widehat{\mathbf{M}}_{\mathrm{i}}$ reachable if and only if any $\mathbf{M} \in \widehat{\mathbf{M}}_{\mathrm{i}}$ is.

Lemma 19.4. *Let* $t \in T_{\mathrm{k}}$, $\mathbf{M}_{\mathrm{i}} \xrightarrow{t} \mathbf{M}_{\mathrm{j}}$. *Then:*
$\forall \mathbf{M} \in \widehat{\mathbf{M}}_{\mathrm{i}} \ \exists t' \in T_{\mathrm{k}}, \ \mathbf{M}' \in \widehat{\mathbf{M}}_{\mathrm{j}} \ \mathbf{M} \xrightarrow{t'} \mathbf{M}'$
$\forall \mathbf{M}' \in \widehat{\mathbf{M}}_{\mathrm{j}} \ \exists t' \in T_{\mathrm{k}}, \ \mathbf{M} \in \widehat{\mathbf{M}}_{\mathrm{i}} \ \mathbf{M} \xrightarrow{t'} \mathbf{M}'$

Thanks to the above lemma we can build a *quotient-graph* in which nodes are the reachable $\{\widehat{\mathbf{M}}_{\mathrm{i}}\}$, and there is a labeled arc $\widehat{\mathbf{M}}_{\mathrm{i}} \xrightarrow{T_{\mathrm{k}}} \widehat{\mathbf{M}}_{\mathrm{j}}$ if and only if there exist $t \in T_{\mathrm{k}}, \mathbf{M} \in \widehat{\mathbf{M}}_{\mathrm{i}}, \mathbf{M}' \in \widehat{\mathbf{M}}_{\mathrm{j}}$, s.t. $\mathbf{M} \xrightarrow{t} \mathbf{M}'$.

If the meta-level PN never enters a deadlock or a livelock, then the *liveness* and *reachability* properties of the original state-transition graph are preserved.

19.5 The Dynamic Philosophers Example

The (symbolic) state-transition semantics of Reflective Petri nets has been tested on a variant of the well known dining philosophers problem which introduces a high dynamism [14]. The version here considered meets the following requirements:

- Two philosophers initially sit on the table.
- A philosopher can eat only when he/she simultaneously picks up the pair of adjacent forks, one of which is owned by the philosopher.
- A philosopher sitting on the table has two additional faculties, both requiring that the owned fork is currently available.
  - He/she can invite a colleague which is outside to join the table, sharing with him/her the owned fork.
  - He/she can leave the table, if there are at least three philosophers sited on it.
- Each philosopher is going around with his/her own fork.

The base-level Petri net representing the starting condition is depicted in Fig. 19.1. We observe that the functional aspects are described in detail, while

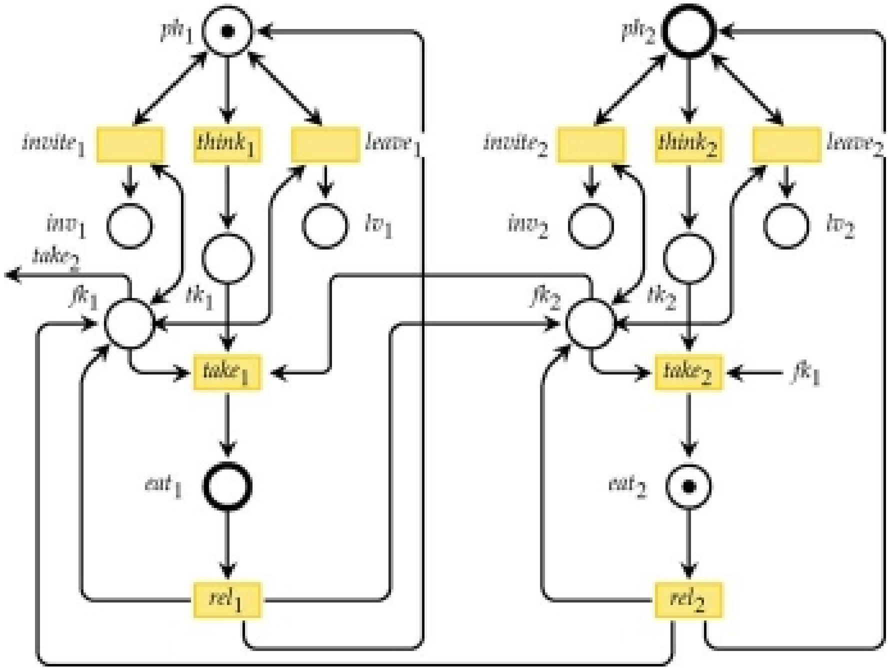


Fig. 19.1 Dynamic philosopher's base-level Petri net

the dynamic features are only sketched (transitions *invite*ᵢ, *leave*ᵢ), thus keeping the model as simple as possible. Any invitation/leaving intents activate the meta-program, which consequently implements two different strategies.

The logical partition of base-level nodes groups places/transition playing a similar role:

$$Place = Ph \cup Fork \cup Lv \cup Inv \cup \ldots \quad Tran = Invite \cup Think \cup Leave \cup \ldots$$

where $Ph = \{ph_{\mathrm{i}}\}$, $Fork = \{fk_{\mathrm{i}}\}$, $Lv = \{lv_{\mathrm{i}}\}$,..$Think = \{th_{\mathrm{i}}\}$, etc.

According to Def. 19.2 the depicted base-level net, and that having the same structure, but places eat_1, ph_2 marked instead of eat_2, ph_1, are equivalent (reachable) states of the corresponding Reflective Petri net. They can be obtained from one another by the permutation:

$$\{ph_1 \leftrightarrow ph_2, eat_1 \leftrightarrow eat_2\}$$

The *leaving* strategy is informally described in Table 19.1. The strategy is divided into an *introspection* step, which consists of checking a logical condition on the base-level reification, followed, if the check is positive, by an *intercession* phase.

Table 19.1 Leaving strategy description

introspection
There exist a marked place $l1 : Lv$ and at least 3 places of type Ph

intercession
The philosopher $ph1 : Ph$ who issued the request is identified:
Let $f1$ be the owned fork, $f2$ be the other fork used by $ph1$ ($f1, f2 : Fork$);
The philosopher $ph3$ sharing fork $f1$ with $ph1$ is identified as well:
Let $tk3 : Take$, $rel3 : Rel$ be the transitions modeling the pick-up and the release of forks by $ph3$, respectively;
$ph3$ is connected to $f2$ through a new input arc $(f2, tk3)$, and a new output arc $(rel3, f2)$;
The philosopher $ph1$ (meant as a the whole subnet) is removed, together with $f1$;
Place $l1$ is emptied;

Table 19.2 Symbolic vs. ordinary state-space size

#Philosophers	Symbolic states	Ordinary states
3	42	256
4	284	37,489
5	2, 356	176,583
6	14, 712	5,905,034
7	96, 035	*not av.*
8	476, 785	*not av.*
9	1, 207, 086	*not av.*
10	2, 978, 896	*not av.*

Symbols used in the description correspond to typed variables of the strategy specification language, which are bound from time to time to color instances.

An evidence of the effectiveness of the quotient graph based on the equivalent states notion (Def. 19.2), which comes to be live, versus the ordinary state-transition graph, is given in Table 19.2. Only visible changes of state involving the base-level are numbered. The experiment was conducted using the GreatSPN tool, with a script emulating the shift-down effect. The first column reports the problem size, i.e., the table capacity. We can appreciate a sensible reduction of the number of reached states also for small sizes, due to the high symmetry exhibited by the system during evolution. Some data about time/memory saving, not reported for the lack of space, confirm the effectiveness of the approach.

19.6 Related Works

Many efforts have been devoted in trying to extend Petri nets with dynamical features. In [15], the author is proposing his pioneering work, *self-modifying* nets, in which the flow relation between a place and a transition is a linear function of the place marking. Another major contribution of Valk is the so-called *nets-within-nets* paradigm [16], where tokens flowing through a net are in turn nets. In his work, Valk takes an object as a token in a unary elementary Petri net system, whereas the object itself is an elementary net system. Even if in the original Valk's proposal no dynamic changes are possible, and mobility is weakly supported, most extensions introduced afterward rely upon his idea.

Badouel and Oliver [2] defined a class of high level Petri nets, called *reconfigurable nets*, which can dynamically modify their own structure by rewriting some of their components. Reconfigurable nets can be unfolded to a subclass of self-modifying Petri nets for which boundedness can be decided. Mobile and dynamic Petri nets [1] integrate Petri nets with RCHAM (Reflective Chemical Abstract Machine) based process algebra.

Tokens in self-modifying, mobile/dynamic and reconfigurable nets, are passive. To bridge the gap between tokens and active objects (agents) some variations on the theme of nets-within-nets have been proposed. In [9] objects are studied as high-level net tokens having an individual dynamical behavior. Object nets behave like tokens, i.e., they are lying in places and are moved by transitions. However, they may also change their state. Reference nets [12] are a flavor of high level Petri nets which provides dynamic creation of net instances, references to other nets/tokens, and communication via synchronous channels (net-inscriptions are in `Java`).

More recent proposals have some similarity with the work we are presenting. In [3], a dynamic architecture modeling is presented which allows active elements to be nested in arbitrary and dynamically changeable hierarchies, enabling the design of systems at different levels of abstractions, by using refinements of net models. In [10], the paradigm of *nets and rules as tokens* is introduced, which permit the structure and behavior of P/T systems to be changed. The new concept is implemented using algebraic nets and graph transformations.

Most dynamic extensions of Petri nets set up new (hybrid) paradigms. While the expressive power has increased, the cognitive simplicity of Petri nets has decreased as well. As argued in [2], the intricacy of these proposals leaves little hope to obtain significant mathematical results and/or automated verification tools in a close future. The Reflective Petri nets approach is different, because it tries to achieve a satisfactory compromise between expressive power and analysis capability, through a rigorous application of reflection concepts in a consolidated high-level Petri Net framework.

19.7 Conclusions and Future Work

We have semi-formally introduced a state-transition semantics for reflective Petri nets, a formal layout based on classical Petri nets (Well formed Nets, and their unfolded counterpart) well suited to model adaptable/reconfigurable discrete-event systems. In particular, we have addressed a major topic related to recognizing equivalent system's evolutions, through the WN's symbolic state notion. We are planning to integrate the GreatSPN tool [8], that natively supports WN and their stochastic extension, SWN, with new modules for the graphical editing and the analysis/simulation of reflective Petri net models. For that purpose we are defining a stochastic process for Reflective Petri nets, in large part inspired to the SWN (GSPN) timed semantics.

References

1. Asperti, A., Busi, N.: Mobile Petri Nets. Technical Report UBLCS-96-10, Università degli Studi di Bologna, Bologna, Italy (1996)
2. Badouel, E., Oliver, J.: Reconfigurable Nets, a Class of High Level Petri Nets Supporting Dynamic Changes within Workflow Systems. IRISA Research Report PI-1163 IRISA (1998)
3. Cabac, L., Duvignau, M., Moldt, D., Rölke, H.: Modeling dynamic architectures using nets within nets. In: Ciardo, G., Darondeau, P. (eds.) Proceedings of the 26th International Conference on Applications and Theory of Petri Nets (ICATPN 2005), LNCS 3536, pp. 148–167. Miami, FL, Springer (2005)
4. Capra, L., Cazzola, W.: A reflective PN-based approach to dynamic workflow change. In: Proceedings of the 9th International Symposium in Symbolic and Numeric Algorithms for Scientific Computing (SYNASC'07), pp. 533–540. Timişoara, Romania, IEEE (2007a)
5. Capra, L., Cazzola, W.: Self-evolving Petri nets. J. Univ. Comp. Scie. **13**(13), 2002–2034 (2007b)
6. Chiola, G., Dutheillet, C., Franceschinis, G., Haddad, S.: On well-formed coloured nets and their symbolic reachability graph. In: Proceedings of the 11th International Conference on Application and Theory of Petri Nets, pp. 387–410. Paris, France (1990)
7. Chiola, G., Dutheillet, C., Franceschinis, G., Haddad, S.: A symbolic reachability graph for coloured Petri nets. Theor. Comput. Sci. B (Logic, Semantics and Theory of Programming) **176**(1& 2), 39–65 (1997)
8. Chiola, G., Franceschinis, G., Gaeta, R., Ribaudo, M.: GreatSPN 1.7: graphical editor and analyzer for timed and stochastic Petri nets. Perform. Evaluation **24**(1–2), 47–68 (1995)
9. Farwer, B. and Moldt, D. (eds.) Object Petri nets, process, and object calculi. Hamburg, Germany, Universität Hamburg, Fachbereich Informatik (2005)
10. Hoffmann, K., Ehrig, H., Mossakowski, T.: High-Level Nets with Nets and Rules as Tokens. In: Ciardo, G., Darondeau, P. (eds.) Proceedings of the 26th International Conference on Applications and Theory of Petri Nets (ICATPN 2005), LNCS 3536, pp. 268–288. Springer, Miami, FL (2005)
11. Jensen, K., Rozenberg, G. (eds.): High-Level Petri Nets: Theory and Applications. Springer, Berlin (1991)
12. Kummer, O.: Simulating synchronous channels and net instances. In: Desel, J., Kemper, P., Kindler, E., Oberweis, A. (eds.) Proceedings of the Workshop Algorithmen und Werkzeuge für Petrinetze, vol. 694 of *Forschungsberichte*, pp. 73–78. Universität Dortmund, Fachbereich Informatik (1998)
13. Reisig, W.: Petri Nets: An Introduction, vol. 4 of EATCS Monographs on Theoretical Computer Science. Springer, Berlin (1985)
14. Sibertin Blanc, C.: The hurried philosophers. In: Agha, G., De Cindio, F., Rozenberg, G. (eds.) Concurrent Object-Oriented Programming and Petri Nets, Advances in Petri Nets, LNCS 2001, pp. 536–538. Springer, Berlin (2001)
15. Valk, R.: Self-modifying nets, a natural extension of Petri nets. In: Ausiello, G., Böhm, C. (eds.) Proceedings of the Fifth Colloquium on Automata, Languages and Programming (ICALP'78), LNCS 62, pp. 464–476. Springer, Udine, Italy (1978)
16. Valk, R.: Petri nets as token objects: an introduction to elementary object nets. In: Desel, J., Silva, M. (eds.) Proceedings of the 19th International Conference on Applications and Theory of Petri Nets (ICATPN 1998), LNCS 1420, pp. 1–25. Springer, Lisbon, Portugal (1998)

Chapter 20
Using Xilinx System Generator for Real Time Hardware Co-simulation of Video Processing System

Taoufik Saidani, Mohamed Atri, Dhaha Dia, and Rached Tourki

Abstract The use of rapid prototyping tools such as MATLAB-Simulink and Xilinx System Generator becomes increasingly important because of time-to-market constraints. This paper presents a methodology for implementing real-time DSP applications on a reconfigurable logic platform using Xilinx System Generator (XSG) for Matlab. The methodology aims to improve the design verfication efficiency for such complex system. It presents architecture for Color Space Conversion (CSC) RGBTOYCbCr for video processing using Xilinx System Generator. The design was implemented targeting a Spartan3 device (3S200PQ208) then a Virtex II Pro (xc2vp7–6ff672). Obtained results are discussed and compared with an other architecture. The conversion method has been verified successfully with no visually perceptual errors in the transformed images.

Keywords Video processing · codesign environment · rapid prototyping · xilinx system generator · FPGA board

20.1 Introduction

Video processing and computer vision methods become increasingly important not only in the industrial applications but also in our daily life [1]. Video processing generally exploits tasks with very high computational demands. Such tasks can be handled by the standard processors and computers or by computers connected to the computational networks [1]. However, such approach is not always suitable that's why specialized hardware solutions based on digital signal processors (DSP) or a field programmable gate arrays (FPGA) are usually used in embedded systems [2, 3]. Xilinx System Generator allows the design of hardware system

T. Saidani (✉), M. Atri, D. Dia, and R. Tourki
EμE Laboratory, FSMonatir, 5000, Tunisia
e-mail: Saidani_taoufik@yahoo.fr; Mohamed.atri@fsm.rnu.tn;
Dhaha.Dia@issatso.rnu.tn; Rached.tourki@fsm.rnu.tn

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_20,

starting from a graphical high level Simulink environment [3,4]. System Generator extends the traditional Hardware Description Language (HDL) design providing graphical modules, and thus does not require a detailed knowledge of this complex language. The Simulink graphical language allows an abstraction of the design through the use of available System Generator blocks and subsystems [3]. This reduces the time necessary between the control design derivations and hardware implementation. In addition, the software provides for the hardware simulation and hardware-in-the-loop verification, referred to as hardware co-simulation [2,4], from within this environment. This methodology provides easier hardware verification and implementation compared to HDL based approach. The Simulink simulation and hardware-in-the loop approach presents a far more cost efficient solution than other methodologies. The ability to quickly and directly realize a control system design as a real-time embedded system greatly facilitates the design process.

The remainder of this paper is divided into six sections. After introducing, a description of Design methodology for implementation on FPGA with Xilinx System Generator is presented; Section 20.3 presents a study case which is Color Space Conversion Application. In Section 20.4, experimental results and software performances are detailed. Section 20.5 shows some discussion and comparison. This paper is concluded in Section 20.6.

20.2 Design Methodology for Implementation on FPGA with Xilinx System Generator

Efficient rapid prototyping system requires a development environment targeting the hardware design platform. The used tools are MATLAB R2007a with Simulink from MathWorks [4,5], System Generator 10.1 for DSP and ISE 10.1 from Xilinx present such capabilities (Fig. 20.1). Although the Xilinx ISE 10.1 [2,5] foundation software is not directly utilized, it is required due to the fact that it is running in the background when the System Generator blocks are implemented. The System Generator [2] environment allows for the Xilinx line of FPGAs to be interfaced directly with Simulink. In addition there are several cost effective development boards available on the market that can be utilized for the software design development phase.

MATLAB is an interactive software for numerical computations that simplifies the implementation of linear algebra routines. Powerful operations can be performed by using the provided MATLAB commands. Simulink [2,3] is an additional MATLAB toolbox that provides for modeling, simulating and analyzing dynamic systems within a graphical environment. The software allows for both modular and hierarchical models to be developed providing the advantage of developing a complex system design that is conceptually simplified.

Xilinx System Generator is a MATLAB-Simulink based design tool for Xilinx's line of FPGAs. Complex digital circuits have been developed using multiple

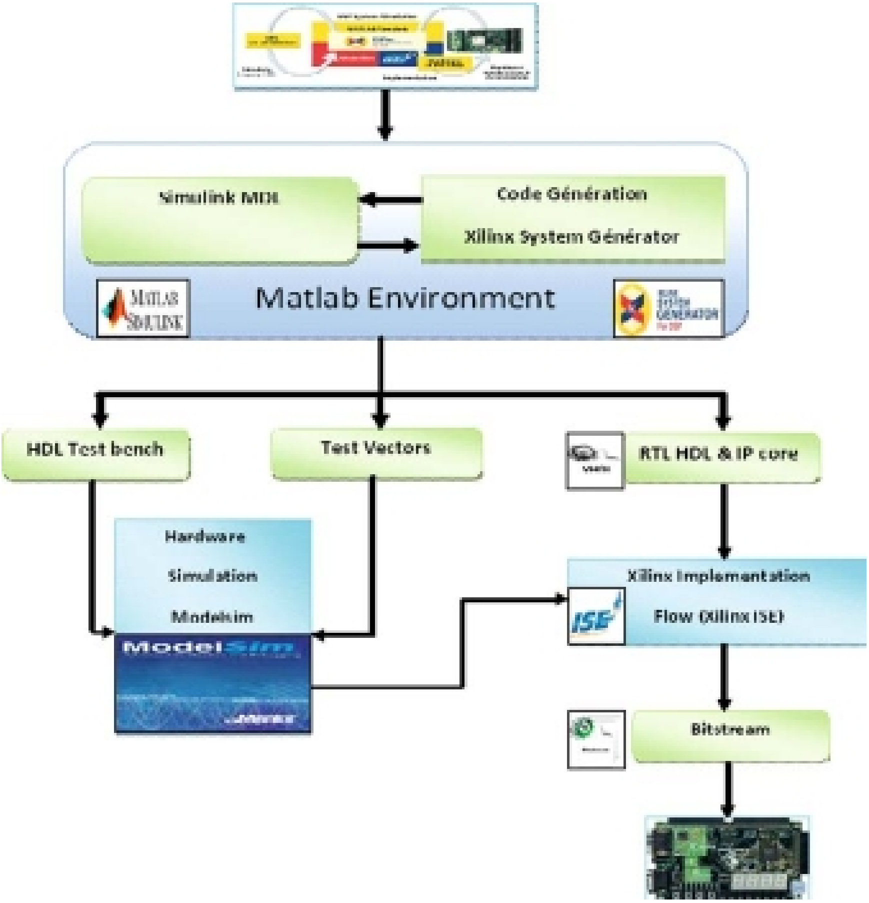


Fig. 20.1 Design methodology with Xilinx System Generator

Hardware Description Language (HDL) modules. Because of the abstraction level is very low within the HDL environment, the difficulty increases as the design becomes more complex [5].

The Xilinx Integrated Software Environment (ISE) is a powerful design environment that is working in the background when implementing System Generator blocks. The ISE environment consists of a set of program modules, written in HDL, that are utilized to create, capture, simulate and implement digital designs in a FPGA or CPLD target device [1, 2]. The synthesis of these modules creates netlist files which serve as the input to the implementation module. After generating these files, the logic design is converted into a physical file that can be downloaded on the target device.

20.3 Study Case: Color Space Conversion RGB to YCbCr

20.3.1 Overwiew

Color Space Conversion (CSC) [6,7] is an important application in image and video processing systems. CSC has been implemented in software and various kinds of hardware. Hardware implementations can achieve a higher performance compared to software-only solutions. Application specific integrated circuits (ASICs) are efficient and have good performance. However, they lack the programmability of devices such as field programmable gate arrays (FPGAs) [8,9].

Many video applications require converting video and image content from one color space to another [10–12]. Images and motion images (video) have utilized a wide variety of color spaces including: RGB, YCrCb, HSI, and other formats to represent the colors within the image [13]. Each of these color space representations has its own set of advantages and disadvantages. For example, RGB is often used for the most demanding applications where ultimate color fidelity must be maintained. Any given color that the human eye can see may be represented by a combination of the primary colors (Red – R, Blue – B, and Green – G). The human eye doesn't actually see equally well in the different color bands with our human-vision [12,14] system optimized for the red, green bands but not quite as sensitive to changes in blues. Scientist and engineers looking for was to reduce the bandwidth and/or bit rate of a video system have created other color spaces (and sampling spaces) that reduce the amount of blue information in a system while maintaining a subjectively high picture quality. Furthermore, human vision is more highly tuned to changes in brightness (black and white or gray-scale changes) than it is to changes in hue (changes from one color or another with the same brightness). Therefore, many video systems sub-sample the color information [12] (chrominance) while transmitting the black and white (luminance) in full resolutions. This sub-sampling is often applied to luminance-chrominance color space systems such as YCrCb where Y represents the luminance information and Cr and Cb are color difference signals that represent the chrominance information. In these systems all of the Y samples are used but every other color sample is dropped. These systems are referred to as 4:2:2 sampling. The 4:2:2 nomenclatures signify that for every 4 Y samples only 2 Cr and 2 Cb samples are saved. Owing to the bandwidth saving benefits of these different image formats different video equipment will adopt different color space encodings. Interoperability between such equipment often requires a device to convert the output of one video device in a given color space to the color space needed as input for the down stream device. Some examples of color space conversion are the converting of the RGB video output from a computer VGA card to YCrCb input on a TV monitor [6,13]. The opposite conversion path is also common where a video device such as a DVD player outputs YCrCb and the video needs to be converted to RGB to drive a monitor [13].

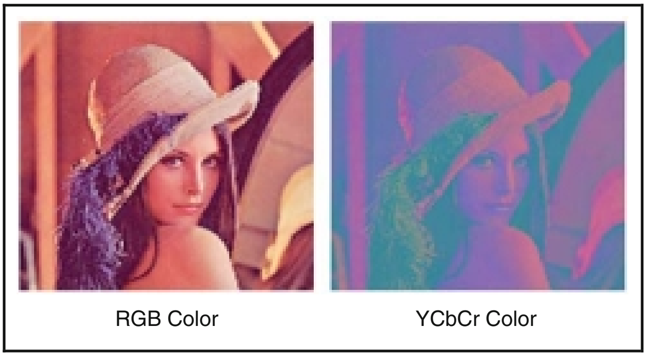


Fig. 20.2 Matlab implementation for rgb2ycbcr

20.3.2 YCbCr Color Model

YCbCr color model also belongs to the family of television transmission color models. In this color model, the luminance component is separated from the color components. Component (Y) represents luminance, and chrominance information is stored as two color-difference components. Color component Cb represent the difference between the blue component and a reference value and the color component Cr represents the difference between the red component and a reference value. The following conversion is used to segment the RGB image into Y, Cb and Cr components: The conversion matrix can be expressed as in Eq. (20.1) [13].

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \tag{20.1}$$

Among all the color models found, YCbCr seems to be better for skin detection since the Colors in YCbCr are specified in terms of luminance (Y channel) and chrominance (Cb and Cr channels). The main advantage of converting the image from RGB color model to the YCbCr color model is the influence of luminance can be removed during our video processing. Figure 20.2 shows the conversion of a RGB color model in to a YCbCr color model implemented with the function rgb2ycbcr from Matlab.

20.4 Implementation Results, Simulation and Comparisons

20.4.1 Hardware Co-simulation

Figure 20.3 shows the model that uses the top level HDL module and its Xilinx blokset for RGB to Y component. This model can be used for co-simulation.

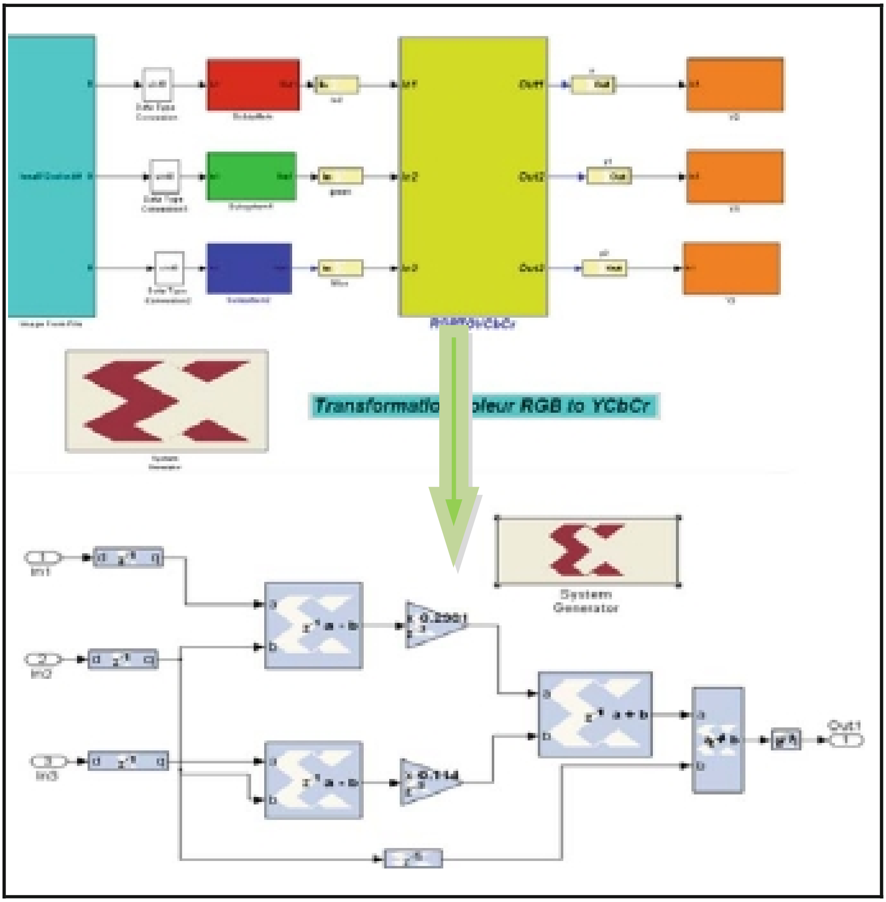


Fig. 20.3 System Generator project for simulation

Once the design is verified, a hardware co-simulation block can be generated. and then will be used to program the FPGA for the CSC design implementation. Figure 20.4 shows the model with the hardware co-simulation block. The bitstream download step is performed using a JTAG cable.

20.4.2 Simulation

After the co-simulation step the VHDL codes were automatically generated from the System Generator block sets. Behavioral and post simulation are supported by Mentor Graphics ModelSim tool (Fig. 20.5).

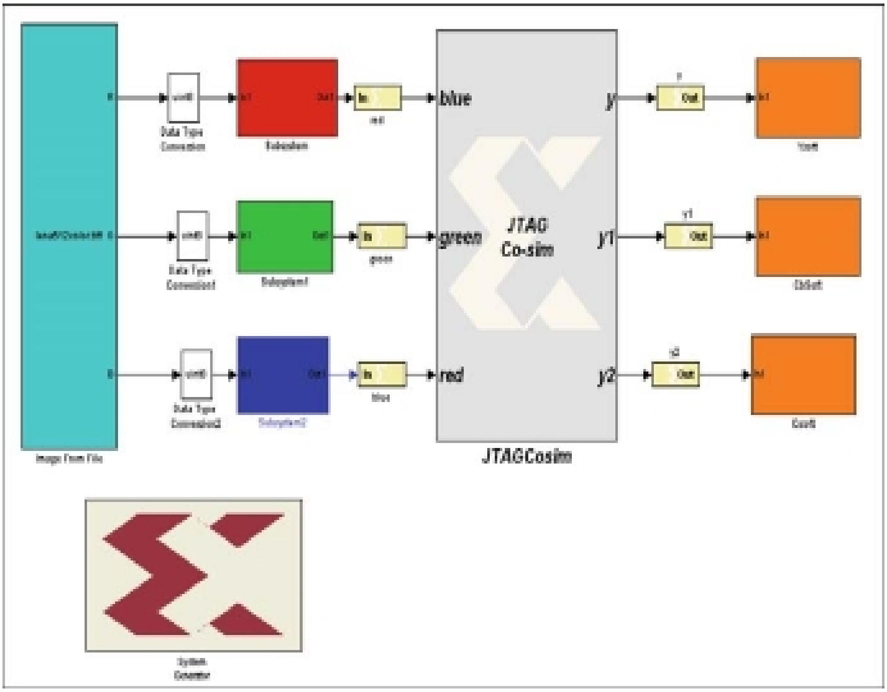


Fig. 20.4 System Generator project for hardware-in-the-loop testing

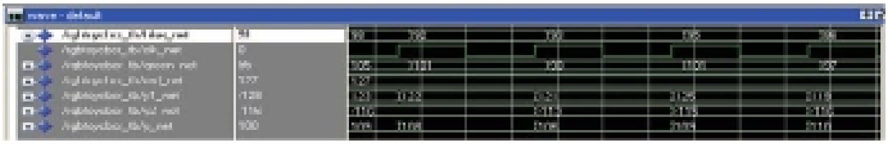


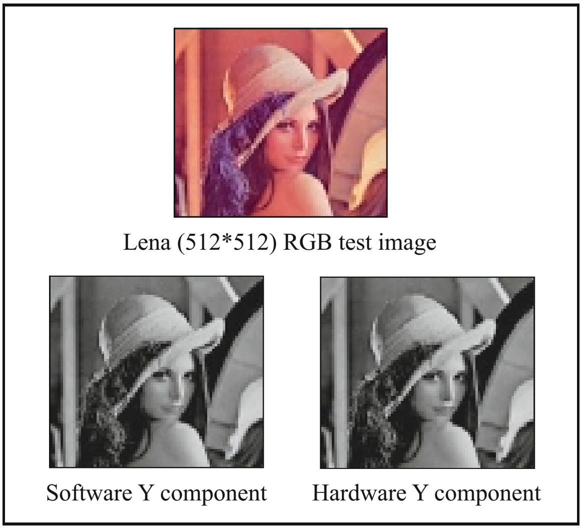
Fig. 20.5 Simulation results of the VHDL RGB to YCbCR conversion

The VHDL codes were then synthesized using Xilinx ISE 10.1i and targeted for Xilinx Spartan3 and Virtex II Pro family [2]. The optimization setting is for maximum clock speed. Table 20.1 details the resource requirements of the design. Note that in practice, additional blocks are needed for input/output interfaces, and synchronization.

The HDL-based circuit design flow is completed with the Xilinx ISE tool to perform synthesis, implementation, place & route and device programming for the whole design. For the arithmetic units, unsigned pipeline integer divider with both quotient and remainder output are parameterized and generated by Xilinx Core Generator tool [5, 9]. Multiplication uses the embedded multiplier in the hardware. The target FPGA chip is Xilinx Virtex II Pro xc2vp7–6ff672 and Spartan 3 xc3s200–5 ft256. During the Simulink-to-FPGA design flow, circuit modeling is built up

Table 20.1 FPGA resources used in the implementation for the CSC

	Spartan 3		Virtex 2 Pro	
	xc3s200–5ft256		xc2vp7–6ff672	
	Available	Used	Available	Used
Number of slices	1,920	15%	4,928	6%
Number of slice flip flop	3,840	11%	9,856	4%
Number of 4 input LUTs	3,840	14%	9,856	5%
Number of bonded IOBs	173	43%	396	18%
Number of GCLKS	8	12%	16	6%
Maximum frequency	129.721 MHz		155.063 MHz	



Lena (512*512) RGB test image

Software Y component

Hardware Y component

Fig. 20.6 Outputs from different implementations

with Simulink basic blocks and Xilinx specified blocks. Input and output data are combined with Matlab workspace, which is convenient to convert number format and debug. Figure 20.6 shows the software and hardware simulation for the CSC design for the input image.

20.5 Discussion

To provide a proper performance evaluation, the implemented CSC architecture using low cost available Spartan-II development system with Xilinx chip 2S200PQ208. The properties of other designs along with ours are listed in Table 20.2. As seen from this table, the design of the CSC proposed by [8] requires 380 CLB on the basis clock rate of 55.159 MHz.

Table 20.2 Performance comparison

	Our design		Design [8]	
	Available	Used	Available	Used
Number of slices	2,352	13%	2,352	16%
Number of slice flip flop	4,704	9%	4,704	7%
Number of bonded IOBs	140	53%	144	35%
Number of GCLKS	4	25%	4	25%
Maximum frequency	83.271 MHz		55.159 MHz	

On the other hand, our resulting architecture spent about 323 CLB with a working frequency up to 83.271 MHz. Obviously, our proposed architecture has lower complexity and improved efficiency in area, thus providing a good choice in terms of low-cost hardware.

From the development of FPGA technology, the methodology challenges the update of various EDA tools [11]. Based on the standard development flow, initial efforts have been transferred to high-level design and synthesis. There are many conversion tools such as C-to-FPGA, Stateflow diagram to VHDL Matlab-to-FPGA. The features of Simulink/Xilinx System Generator-to-FPGA [2, 4] flow can be discussed as follows.

References

1. Zemcik, P.: Hardware acceleration of graphics and imaging algorithms using FPGAs. SCCG'02: Proceedings of the 18th Spring Conference On Computer Graphics, pp. 25–32. ACM, New York (2002)
2. Xilinx System Generator User's Guide, www. Xilinx.com
3. Moctezuma, J.C., Sanchez, S., Alvarez, R., Sánchez, A.: "Architecture for filtering images using Xilinx system generator" World scientific advanced series in Electrical and Computer Engineering. Proceedings of the 2nd WSEAS International Conference on Computer Engineering and Applications, pp. 284–289 (2008)
4. The MathWorks Inc. Embedded Matlab Language User Guide (2007)
5. Ownby, M., Mahmoud, W.H.: A design methodology for implementing DSP with Xilin system generator for matlab. IEEE International Symposium on System Theory, pp. 404–408 (2003)
6. Han, D.: A cost effective color gamut mapping architecture for digital tv color reproduction enhancement. IEEE Trans. Consum. Electron. **51**(1), 168–174 (2005)
7. Bilal, M., Masud, S.: Efficient color space conversion using custom instruction in a risc processor. IEEE International Symposium on Circuits and Systems, pp. 1109, 1112 (2007)
8. Sapkal, A.M., Munot, M., Joshi, M.A.: R′G′B′ to Y′CbCr color space conversion using FPGA. Wireless, mobile and multimedia networks, 2008. IET International Conference on Digital Object Identifier, pp. 255–258, 11–12 (Jan 2008)
9. Agostini, L.V., Silva, I.S., Bampi, S.: Parallel color space converters for JPEG image compression. Microelectron. Reliability **44**(4), 697–703 (April 2004)
10. Sima, M., Vassiliadis, S., Cotofana, S., van Eijndhoven, J.T.J.: Color space conversion for MPEG decoding on FPGA-augmented trimedia processor. Proceedings. IEEE International Conference on Application-Specific Systems, Architectures, and Processors, pp. 250–259 (June 2003)
11. Han, D.: Real-time color gamut mapping method for digital tv display quality enhancement. IEEE Trans. Cons. Electron. **50**(2):691–698 (2004)

12. A. Albiol, L. Torres, and E. J. Delp.: An unsupervised color image segmentation algorithm for face detection applications. In: Proceedings. 2001 International Conference on Image Processing, vol. **2**, pp. 681–684 (2001)
13. Bensaali, F., Amira, A., Bouridane, A.: Accelerating matrix product on reconfigurable hardware for image processing applications. IEE Proceedings on Circuits Devices System **152**(3) (June 2005)
14. Kuchi, P., Gabbur, P., Bhat, S., David, S.: Human face detection and tracking using skin color modelling and connected component operators. IETE J. Res. (Special issue on Visual Media Processing) (May 2002)

Chapter 21
Perception-Based Road Traffic Congestion Classification Using Neural Networks and Decision Tree

Pitiphoom Posawang, Satidchoke Phosaard, Weerapong Polnigongit, and Wasan Pattara-Atikom

Abstract In this study, we investigated an alternative technique to automatically classify road traffic congestion with travelers' opinions. The method utilized an intelligent traffic camera system orchestrated with an interactive web survey system to collect the traffic conditions and travelers' opinions. A large numbers of human perceptions were used to train the artificial neural network (ANN) model and the decision tree (J48) model that classify velocity and traffic flow into three congestion levels: light, heavy, and jam. The both model was then compared to the Occupancy Ratio (OR) technique, currently in service in the Bangkok Metropolitan Administration (BMA). The accuracy of ANN was more than accuracy of the J48. The evaluation indicated that our ANN model could determine the traffic congestion levels 12.15% more accurately than the existing system. The methodology, though conceived for use in Bangkok, is a general Intelligent Transportation System (ITS) practice that can be applied to any part of the world.

Keywords Traffic congestion level determination · intelligent transportation system (ITS) · human judgment · artificial neural network (ANN) · decision tree (J48) · occupancy ratio (OR)

P. Posawang (✉), S. Phosaard, and W. Polnigongit
School of Information Technology, Suranaree University of Technology, 111 University Ave., Muang Nakhon Ratchasima, Nakhon Ratchasima 30000, Thailand
e-mail: infotech@sut.ac.th; s@sut.ac.th; weerap@sut.ac.th

W. Pattara-Atikom
National Electronics and Computer Technology Center (NECTEC), under the National Science and Technology Development Agency (NSTDA), 112 Thailand Science Park, Phahon Yothin Rd., Klong 1, Klong Luang, Pathumthani 12120, Thailand
e-mail: wasan@nectec.or.th

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_21,

21.1 Introduction

Accurate traffic reports are essential for congested and overcrowded cities such as Bangkok. Without traffic information, commuters might not be able to choose a proper route and might get stuck in traffic for hours. Intelligent Transportation System (ITS) with automated congestion estimation algorithm can help produce such reports. Several initiatives from both private and government entities have been proposed and implemented to gather traffic data to feed the ITS. According to our survey, most efforts focus on limited installation of fixed sensors such as intelligent traffic cameras with image processing capability. Although the investment for large-scale deployment of such camera system is capital intensive, it is desirable to have it installed due to the ability to provide visualization of real traffic conditions. It is also able to provide basic information that is essential to generate traffic reports, such as the average vehicle velocity and the volume of vehicles. Traffic congestion levels can also be derived from such parameters calculated by the traffic camera with image processing capability. The current system used to classify the congestion level is by an image processing system utilizing the Occupancy Ratio (OR) technique. A study by Charusakwong et al. [1] showed that the results from existing techniques may not be consistent with the travelers' perception. In contrast, our results show a significant improvement on consistency and suggest that a well-trained neural network using velocity and traffic flow is a promising candidate to provide an automated congestion classification.

In this study, we proposed a method to automatically classify the traffic congestion level that was consistent with motorists' perceptions by imitating their visual judgments using an artificial intelligence technique and decision tree. To accomplish this, we captured a large amount of traffic conditions, and then let the motorists indicate the congestion level for each captured image selected from the image pool. The relation patterns of these ratings along with each of their corresponding image processing information were used to train an artificial neural network (ANN) model and decision tree (J48) model. The trained ANN model and J48 model were later used to automatically classify the traffic congestion level. This approach would lift the confidence that the congestion level on the traffic reports would be consistent with motorists' perceptions. This study focuses on the accuracy optimization of the ANN model, J48 model and its accuracy when compared to the working system in Bangkok. The congestion levels that we studied were limited to three levels: light, heavy and jam, which was appropriate [2].

This chapter is organized as follows: In Section 21.2, we provide an overview of related works concerning traffic congestion reports. The methodology is explained in Section 21.3. In Section 21.4, we analyze and discuss the results, and Section 21.5 offers a conclusion.

21.2 Review Work of the Road Traffic Congestion Estimation

Many techniques to estimate traffic congestion levels were investigated to suit each type of collected data. Traffic data could be gathered automatically from two major types of sensors: fixed sensors and mobile sensors. The study in [3] applied the neural network technique to the collected data using mobile phones. It used Cell Dwell Time (CDT), the time that a mobile phone attaches to a based station, which provides rough journey time. Our work employed a similar technique but on the data captured from existing traffic cameras. The study in [4, 5] estimated the congestion level using data from traffic cameras by applying fuzzy logic, and hidden Markov model, respectively. Our work applied neural network techniques and we expected higher accuracy. The works in [6] considered traffic density and highway speeds, but our measurements reflect traffic conditions on an expressway. Studies in [7, 8] used fuzzy logic to determine continuous and six discrete levels of congestion while our method focuses on three discrete levels of congestion.

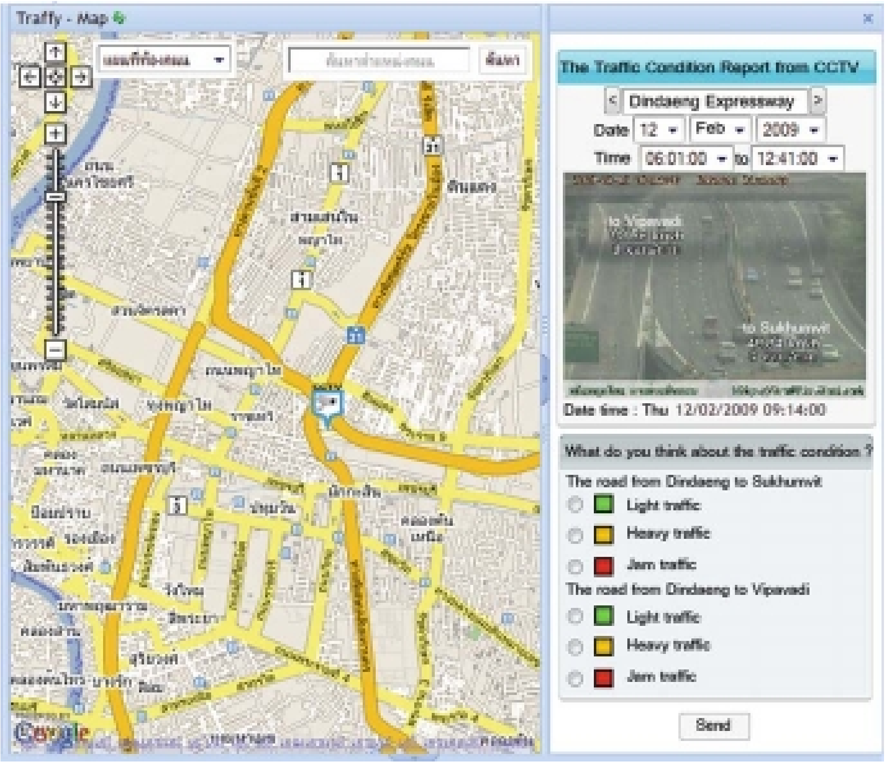
In some countries, for example, as in the study of [9, 10] found out that the main parameters used to define the traffic congestion levels are time, speed, volume, service level, and the cycles of traffic signals that delay motorists. Our work would locally investigate whether the congestion degrees are subjective to other factors, including type of roads, the time of day, the day of week. The congestion levels that we studied were limited to three levels: light, heavy and jam, which was sufficient and appropriate according to the study of [2].

In this study, we evaluated the accuracy of our technique against that of the existing system currently operating in Bangkok. In Bangkok, the processing of the traffic camera images utilizes the Occupancy Ratio (OR) technique [1]. The working principle of the OR technique is measuring the time that vehicles use to occupy a virtual indicated frame, from entering until leaving it. If the time that vehicles used to travel through the frame is short, the traffic is light. The methodology of the work was presented in the following section.

21.3 Methodology

21.3.1 Data Collection and Tools

The traffic data was captured by traffic cameras every minute. As stream video was processed, the vehicle velocity (km/h) and traffic volume (cars/min) were calculated. Then, the captured images and related information were presented to participants who then rated the traffic congestion levels. In order to ensure that the determined congestion levels were consistent with the road users' perception, we needed a large amount of samples. A web survey was used to collect the road users' opinions.



The Traffic Condition Report from CCTV

< Dindaeng Expressway >

Date 12 Feb 2009

Time 06:01:00 to 12:41:00

Date time : Thu 12/02/2009 09:14:00

What do you think about the traffic condition ?

The road from Dindaeng to Sukhumvit

○ 🟩 Light traffic

○ 🟨 Heavy traffic

○ 🟥 Jam traffic

The road from Dindaeng to Vipavadi

○ 🟩 Light traffic

○ 🟨 Heavy traffic

○ 🟥 Jam traffic

Send

Fig. 21.1 The web survey screen consists of a traffic image and data with the options for the participants to rate the congestion level

Figure 21.1 illustrates the web survey we developed to collect congestion levels judged by users corresponding with the image shown to the user. On the left side of the web survey, the participants were allowed to select from several traffic cameras mounted around congested areas of Bangkok. In this study, we collected from several highly congested roads in Bangkok, e.g., DinDaeng expressways, PraChaNuKul expressways, PTT KlongLuang highway and TubKwang highway. The captured still images, shown one at a time, along with the matching information including date and range of time were shown on the right of the screen. The participants were asked to rate the traffic congestion level according to the provided information on both in-bound and out-bound directions. The web survey was implemented by the Google Map APIs, PHP, AJAX and the PostgresSQL database. The targeted participants were general road users. The judgment data was collected between the period of Jan 10 and Jan 30, 2009. The number of total participants was 146 and there were 11,520 records of judgments. The data was used to learn by the neural network models; explanatory details are given in the following section.

21.3.2 Data Classification

In general, the architecture of a neural network consists of three node layers: one input layer, one hidden layer, and one output layer, all fully connected as shown in Fig. 21.2. The neural network model will adjust the weight of each node to reflect the patterns of the trained data [11].

The features of input data consist of (1) the day of the week (DW): Monday through Sunday, (2) the time of the day in terms of the minute of the day (MT), (3) vehicle velocity (SP) in km/h, and (4) traffic volume (VOL) in cars/min, in the input layer. The input layer is composed of ten nodes as shown in Fig. 21.2. The day of week variable constructed seven variable nodes due to its nominal status. The output layer is the targeted congestion level (CL) judged by the participants.

Decision trees are often used in classification and prediction. It is simple yet a powerful way of knowledge representation. The models produced by decision trees are represented in the form of tree structure. A leaf node indicates the class of the examples. The instances are classified by sorting them down the tree from the root node to some leaf node as shown in Fig. 21.3. Decision trees represent a supervised approach to classification. Weka uses the J48 algorithm, which is Weka's implementation of C4.5 [12] Decision tree algorithm. J48 is actually a slight improved to and the latest version of C4.5.

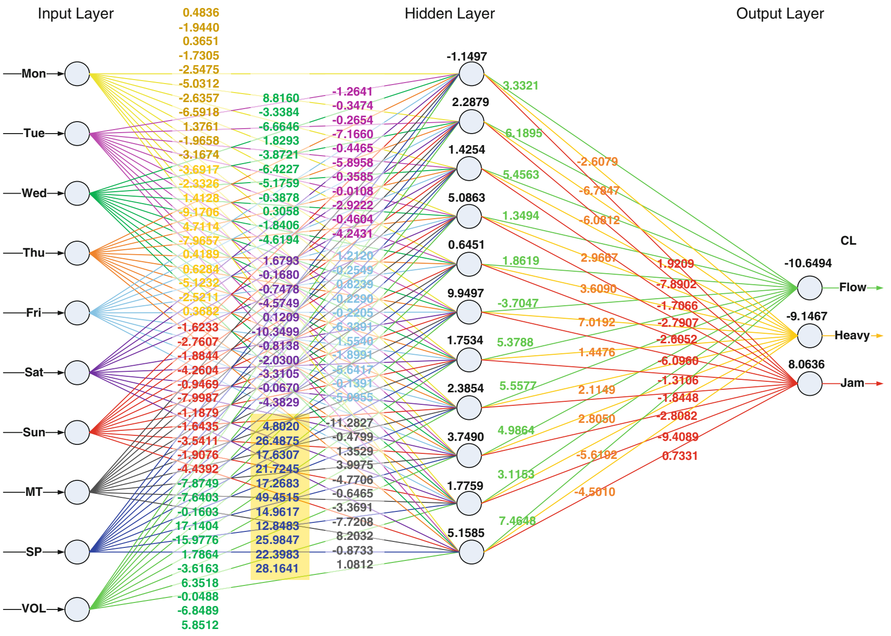


Fig. 21.2 The learned neural network model configurations

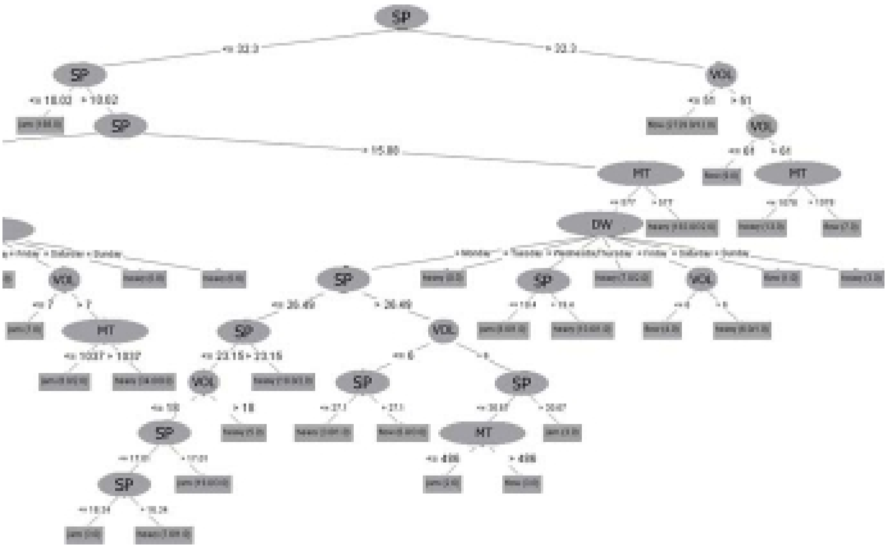


Fig. 21.3 The derived J48 decision tree

Table 21.1 Samples of training dataset

Day of the week	Time of the day	Velocity of the vehicles	Volume	Congestion level
Monday	471	12.78	5	3
Monday	472	12.88	2	3
Monday	473	11.94	4	3
Monday	474	15.37	4	3
Monday	475	25.13	3	2

We chose the Multilayer-Perceptron (Neural Network) Model and the J48 algorithm (Decision Tree) was used to train and create a learned model in the WEKA systems. Samples of training dataset are shown in Table 21.1.

21.4 Results and Evaluations

21.4.1 Performance Evaluations

Figure 21.2 represents the derived neural network model configurations labeled with the weight of each node in each layer. According to high and positive value of weights, the vehicle velocity (SP) has the highest influence to classify the congestion followed by the traffic volume, the time of the day and the day of the week.

The vehicle velocity associated with congestion levels can be illustrated as shown in Fig. 21.4 on the left (ANN model). The distribution of vehicle velocity of each congestion level is subject to further investigation.

The structure of the optimized neural network was 10-11-3 for the number of input nodes, hidden nodes, and output node, respectively; the learning rate was 0.3; and the momentum was 0.2. This DinDaeng expressways model achieved an overall highest accuracy of 94.99%, a root mean square error of 0.1583, and a precision ranging from 0.724 to 0.996. The result shows a true positive rate (TP Rate) ranging from 0.782 to 0.986, which is very high, and a false positive rate (FP Rate) ranging from 0.016 to 0.030, which is very low as shown in Table 21.2.

Figure 21.3 represents the derived decision tree model. The size of our decision tree is 97 nodes, 59 of which are leave nodes. The root node is SP attribute. This means that the vehicle velocity is the most important factor to determine the level of road traffic congestion, which is attribute the same as neural network model. The vehicle velocity associated with congestion levels can be illustrated as shown in Fig. 21.4 on the right (J48 model).

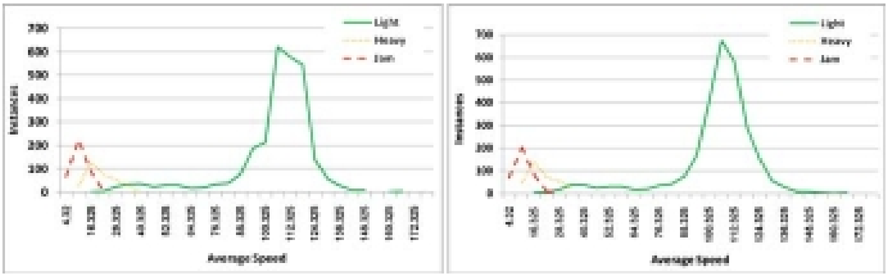


Fig. 21.4 The chart shows the number of road users' rated instances (y-axis) according to the speed of the vehicles (x-axis) along with classified congestion levels

Table 21.2 The neural network classifier's performance

				Correctly classified		
Camera	Accuracy	RMSE	Class	TP rate	FP rate	Precision
DinDaeng	94.99	0.1583	Light	0.986	0.016	0.996
			Heavy	0.782	0.030	0.724
			Jam	0.829	0.022	0.823
PraChaNuKul	87.89	0.2421	Light	0.952	0.052	0.973
			Heavy	0.875	0.112	0.695
			Jam	0.459	0.019	0.755
KlongLuang	85.54	0.2586	Light	0.942	0.104	0.946
			Heavy	0.779	0.109	0.662
			Jam	0.582	0.028	0.733
TubKwang	83.06	0.2737	Light	0.917	0.128	0.932
			Heavy	0.723	0.128	0.622
			Jam	0.554	0.030	0.711

Table 21.3 The decision tree classifier's performance

				Correctly classified		
Camera	Accuracy	RMSE	Class	TP rate	FP rate	Precision
DinDaeng	95.80%	0.1515	Light	0.988	0.034	0.991
			Heavy	0.867	0.026	0.770
			Jam	0.814	0.013	0.888
PraChaNuKul	88.30%	0.2453	Light	0.951	0.049	0.974
			Heavy	0.862	0.101	0.713
			Jam	0.525	0.025	0.729
KlongLuang	85.21%	0.2694	Light	0.941	0.116	0.940
			Heavy	0.700	0.103	0.663
			Jam	0.639	0.032	0.720
TubKwang	86.67%	0.2579	Light	0.927	0.115	0.939
			Heavy	0.762	0.092	0.708
			Jam	0.728	0.026	0.790

The structure of the optimization decision tree was the DinDaeng expressways model. This model achieved an overall highest accuracy of 95.80%, a root mean square error of 0.1515, and a precision ranging from 0.770 to 0.991. The result shows a true positive rate (TP Rate) ranging from 0.814 to 0.988, which is very high, and a false positive rate (FP Rate) ranging from 0.013 to 0.034, which is very low as shown in Table 21.3.

21.4.2 Evaluations with the Existing System

We evaluated our model against the existing system using occupancy ratio (OR) by comparing the congestion levels classified at the exact same points of time. As per Table 21.4, 84.71% of traffic incidents were classified into the same congestion level by both models of ANN and 83.53% of traffic incidents were classified into the same congestion level by both models of J48. The confusion matrix showing the congestion level classification of both models can be found in Table 21.2 of DinDaeng expressways model of ANN and Table 21.3 of DinDaeng expressways model of J48. The columns of the confusion matrix show the congestion levels classified by ANN and J48 while the rows of the confusion matrix show the congestion levels classified by OR.

For example of Table 21.4, its accuracy when compared to the working system in Bangkok. The accuracy of ANN was more than accuracy of the J48. Therefore, we chose the neural network model in the analysis data. The first column on the left was classified as light traffic by the ANN. 83.66% was classified as light traffic the same as the OR technique. However, 2.43% and 12.04% were classified as heavy and jam using the OR technique. The values in shaded cells show the percentage of mutual classification by both models.

Table 21.4 The confusion matrix between the classified congestion levels using ANN model and J48 with the occupancy ratio (or) technique from the BMA system

		ANN classification (%)			J48 classification (%)		
	Class	Light	Heavy	Jam	Light	Heavy	Jam
Occupancy ratio (or) classification (%)	Light	83.66	0.30	0.11	82.76	0.38	0.50
	Heavy	2.43	0.00	0.00	1.29	0.00	0.07
	Jam	12.04	0.41	1.05	13.59	0.62	0.77

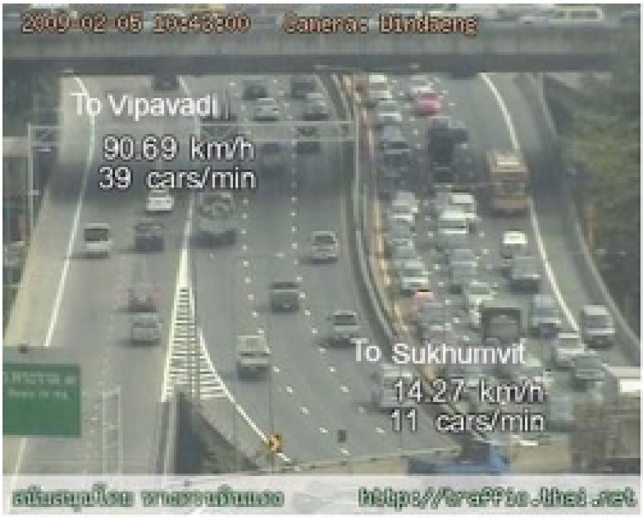


Fig. 21.5 Image classified as jam traffic by ANN and as light traffic by OR technique

We investigated the causes of different classification using the recorded traffic images, processed data and the road users' ratings. Single step differences in classification, e.g., light to heavy and heavy to jam, are due to the different points of view on congestion of motorists, in which definitive conclusions cannot be drawn. Additionally, the percentage of these differences is relatively low. We will focus our attention on the significant differences from light to jam or vice versa.

In the first case, i.e., light to jam (0.11%), the analysis revealed that the speed is between 6.04 and 14.27 km/h. All of recorded images, as one example shown in Fig. 21.7 (to Sukhumvit), clearly confirm that the traffic was in jam condition.

In the second case, i.e., jam to light (12.04%), the analysis revealed that the speed is between 70.19 and 129.41 km/h. All of recorded images, clearly confirm that the traffic was in light condition, in which our model classified accurately.

We further investigated the relationship between speed and congestion level of the existing system from the BMA system. In Figs. 21.5 and 21.6, the speed range of each congestion level clearly separates one from each other, especially for light and heavy traffic. However, the speed ranges of three congestion levels from the BMA system largely overlap. This makes it difficult and confusing to distinguish

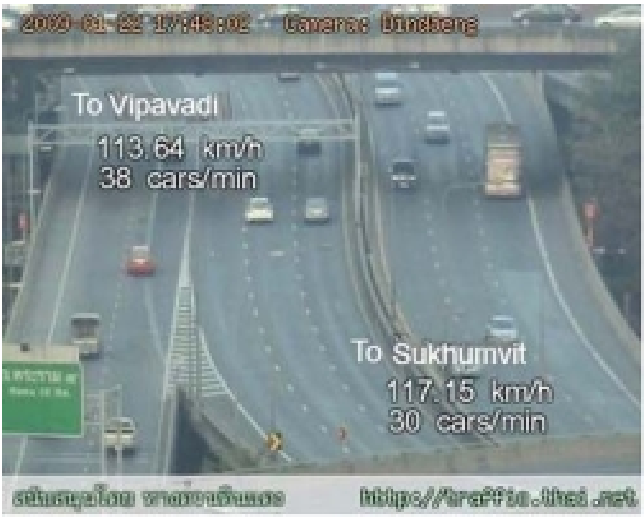


Fig. 21.6 Image classified as light traffic by ANN and as jam traffic by OR technique

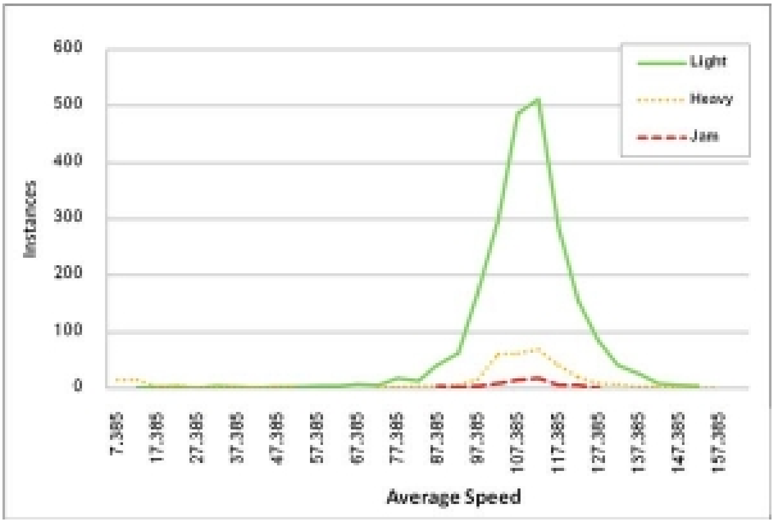


Fig. 21.7 The congestion classification derived from the BMA system

the congestion level using only speed attributes. Additionally, the figure shows a counter-intuitive relationship between speed and congestion level. For example, the speed range of jam is between 87 and 127 km/h. The preliminary results suggested that the opinion-based ANN exhibits a more concrete classification pattern.

Thus, it safe to claim that this neural network model could achieve higher accuracy in the determination of the traffic congestion level than the current system operating in Bangkok using OR technique; consistencies were greater by as much as 12.15% with the overall accuracy of the model itself reaching 94.99%.

21.5 Conclusion

The study employed the artificial neural network technique and decision tree to automatically determine the traffic congestion levels achieving an accuracy of ANN of 94.99% and the root mean square of 0.1583, based on motorists' perceptions, with the 10-11-3 node configurations. The weighting priorities of the input were vehicle velocity (km/h), traffic volume (car/min), the time of day, and the day of week. Accuracy of J48 of 95.80% and the root mean square of 0.1515, with the size of our decision tree is 97 nodes, 59 of which are leave nodes. The root node is vehicle velocity attribute. This means that its the most important factor to determine the level of road traffic congestion, which is attribute the same as neural network model.

Accuracy of the both model, when compared to the working system in Bangkok. The accuracy of ANN was more than accuracy of the J48. Therefore, we chose the neural network model in the analysis data. The ANN model was 12.15% more consistent with the motorists' perceptions than the Occupancy Ratio method used by the existing system in the Bangkok Metropolitan Administration.

References

1. Charusakwong, N., Tangittinunt, K., Choocharukul, K.: Inconsistencies between motorist's perceptions of traffic conditions and color indicators on intelligent traffic signs in Bangkok. Proceedings of the 13th National Convention on Civil Engineering, pp. TRP196–TRP202 (2008)
2. Choocharukul, K.: Congestion measures in Thailand: state of the practice. Proceedings of the10th National Convention on Civil Engineering, pp. TRP111–TRP118 (2005)
3. Pattara-atikom, W, Peachavanish, R.: Estimating road traffic congestion from cell dwell time using neural network. The 7th International Conference on ITS Telecommunications (ITST 2007), Sophia Antipolis, France (2007)
4. Pongpaibool, P., Tangamchit, P., Noodwong, K.: Evaluation of road traffic congestion using fuzzy techniques. Proceedings of IEEE TENCON 2007, Taipei, Taiwan (2007)
5. Porikli, F., Li, X.: Traffic congestion estimation using hmm models without vehicle tracking. IEEE Intelligent Vehicles Symposium, pp. 188–193 (2004)
6. Lu, J., Cao, L.: Congestion evaluation from traffic flow information based on fuzzy logic. IEEE Intell. Transport. Syst. **1**, 50–33 (2003)
7. Krause, B., Altrock, C.V.: Intelligent highway by fuzzy logic:Congestion detection and traffic control on multi-lane roads with variable road signs. 5th International Conference on Fuzzy Systems **3**, 1832–1837 (1996)
8. Alessandri, R.B.A., Repetto, M.: Estimating of freeway traffic variables using information from mobile phones. IEEE American Control Conference **5**, 4089–4094 (2003)

9. Lomax, J.T.S., Tuner, M., Shunk, G., Levinson, H.S., Pratt, R.H., Bay, P.N., Douglas, B.B.: Quantifying Congestion: Final Report. National Cooperative Highway Research Program Report 398, TRB, Washington, DC (1997)
10. Bertini, R.L.: Congestion and its extent. Access to destinations: Rethinking the transportation future of our region. Minneapolis, MN (2004)
11. Dai, H.C., Mcbeth, C.: Effects of learning parameters on learning procedure and performance of a BPNN. Neural Network **10**(8), 1505–1521 (1997)
12. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman, San Mateo, CA (1993)

Chapter 22
RDFa Ontology-Based Architecture for String-Based Web Attacks: Testing and Evaluation

Shadi Aljawarneh and Faisal Alkhateeb

Abstract String input is an issue for web application security. The problem is that developers often trust string input without checking for validity. Typically, a little attention is paid to it in a web development project, because overenthusiastic validation can tend to break the security upon web applications. In this chapter, security vulnerabilities such as SQL injection has been described and then the merits of a number of common data validation techniques have been discussed. From this analysis, a new data validation service (NDVS) which is based upon semantic web technologies, has been implemented to prevent the web security vulnerabilities and then to secure a web system even if the validation modules are bypassed. Such semantic architecture comprises of the following components: RDFa annotation for elements of web pages, interceptor, RDF extractor, RDF parser, and data validator. We carried out two experiments to address the security and the performance objectives. The results have shown that the developed service can provide a high coverage of detection and recovery and a low level of overhead times.

Keywords String · web application security · data validation · security vulnerabilities · SQL injection · semantic web technologies

22.1 Introduction

According to the recently released X-Force 2008 Trend and Risk Report, web application vulnerabilities estimated for above 54% of all vulnerabilities disclosed at end of 2007, and 74% of these vulnerabilities still had no vendor-supplied patch as of the end of 2008 [9].

S. Aljawarneh (✉)
Department of Software Engineering, Faculty of Science and IT, Al-Isra Private University,
P.O. Box 22, Al-Isra University Postoffice Amman, Jordan 11622
e-mail: shadi.jawarneh@ipu.edu.jo

F. Alkhateeb
IT Faculty, Yarmouk University, Irbid, Jordan
e-mail: alkhateebf@yu.edu.jo

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_22,

Web applications need to interact with a database in some form to persistently store data. This interaction is performed by constructing queries that manipulate or retrieve data stored in the database [6, 15].

Criminals could be able to manipulate the SQL query and tamper with data, retrieve database information or control of the database server by embedding characters that have special meaning or special commands to database. Thus, the key issue in this chapter that is that user input might be invalidated on a server that processes the SQL query. Many web developers rely on client-side validation to protect against escape characters being entered by users [11, 15].

This chapter is organized as follows: the rest of this section discusses the data validation process. The recent existing validation approaches and tools are made in Section 22.2. The proposed architecture is described, and a case study is presented in Section 22.3. The prototype implementation and system evaluation are discussed in Section 22.4. Finally, summary is offered in Section 22.5.

22.1.1 Data Validation Process

In a web application security, data validation is the process of ensuring that a web application operates on clean, correct and useful data [12]. It uses validation rules that check for correctness, meaningfulness, and security of data that are input to the web-based system. The rules may be implemented through the automated facilities of a data dictionary, or by the inclusion of explicit program validation logic [5].

There are two main principles of data validation including [5], (1) data should be validated in the data model, where the validation rules have maximum scope for interpreting the context; and (2) escaping of harmful meta-characters should be performed just before the data is processed, typically in the data access components. Web developers have adopted a number of standard validation techniques that could help to protect data integrity such as server-side validation, client-side validation, double-checking validation, and Powerforms [4]. Further details about these approaches are described in our earlier work [1, 2].

22.2 Related Work

A number of researchers have developed solutions to address this problem. Scott and Sharp [13] proposed a gateway model which is an application-level firewall on a server for checking invalid user inputs and detecting malicious script (e.g. SQL injection attack and cross-site scripting attack). This approach offers protection through the enforcement of a number of defined policies, but fails to assess the code itself or to identify the actual weaknesses.

Hassinen and Mussalo [7] have proposed a client-side encryption system to protect confidentiality, data integrity, and user trust. They encrypt data inputs using a client encryption key before submitting the content of an (X)HTML Form. However, the Java Applets can access the client's local file system and hence, a criminal can replace the original signed Applet with a faked Applet to access the client's web content. In addition, the Applet and JavaScript methods can be bypassed. If this happens, the submitted values will be in plain text.

Huang et al. [8] have used behavior monitoring to detect malicious content before it reaches users. They develop WAVES (Web application security assessment system) that performs behavior stimulation to induce malicious behavior in the monitored components. However, the testing processes cannot guarantee the identification of all bugs, and they cannot support immediate or direct security for web applications.

Jovanovic et al. [10] have developed Pixy, which is the first open source tool for statically detecting XSS vulnerabilities in PHP 4 code by means of data follow analysis which is based on a static analysis technique. Although the Pixy prototype is aimed at the detection of XSS vulnerabilities, it can be equally applied to other taint-style vulnerabilities such as SQL injection or command injection. However, Pixy does not support object-oriented features of PHP.

Balzarotti et al. [3] have presented approach to the analysis of the sanitization. This means that they combined static and dynamic analysis techniques to identify faulty sanitization procedures that can be bypassed by the criminal. Therefore, they implemented this approach in a tool, called Saner, and they applied it to a number of real-world web applications. They introduced a dynamic analysis technique that is able to reconstruct the code that is responsible for the sanitization of application inputs, and then execute this code on malicious inputs to identify faulty sanitization procedures.

It should be noted the client-side solutions (such as Client-side Encryption Approach) that target for mitigating threats to web application at the client-side could offer a protection in case of suspected XSS attacks that attempt to steal credentials of users. However, the server-side solution (such as Pixy, WAVES, and Saner) have the advantage of being able to discover a larger range of vulnerabilities, and the benefit of a security flaw by the service provider is propagated to all its clients. The server-side techniques can be further classified into dynamic and static approaches. Dynamic tools try to detect attacks while executing the audited program, whereas static analyzers scan the source code of web application for security vulnerabilities. Therefore, our point of view is to design a new approach takes of advantages the client and server sides solutions. This new approach could be able to detect the known and unknown attacks.

22.3 NDVS Design

We present a new data validation service which is based upon semantic web technologies to prevent the security vulnerabilities at the application level and to secure the web system even if the input validation modules are bypassed. As illustration in

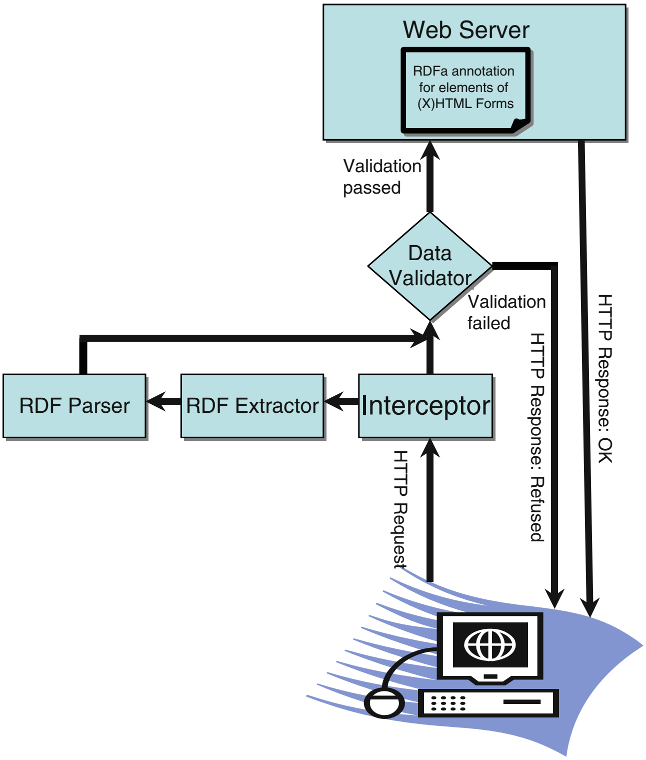


Fig. 22.1 Schematic view of NDVS architecture

Fig. 22.1, the data validation service architecture consists of the following components: RDFa annotation for elements of web pages, interceptor, RDF extractor, RDF parser, and data validator. The next section will describe the functional overview of the proposed solution.

It should be noted that the components of the proposed architecture framework do not need to run on a dedicated machine, they can be run as separate processes on the server.

22.3.1 Functional Overview

The following steps are performed:

1. Use an ontology to describe all data elements in a web application using RDFa annotation[1] and then end user requests (X)HTML form.

[1] http://wwww.w3.org/TR/xhtml-rdfa-primer/

2. Interceptor component intercepts each HTTP request at the server-side before the request arrives to web server application for processing.
3. Extracting the RDFa annotations from RDFa ontology vocabulary using the online RDFa extractor.[2]
4. Invoking the validator component to validate all user inputs.
5. If the validation is correct then the request sends to web server application for processing, otherwise, the request is refused.

22.3.2 Overview of the Framework Architecture

An illustration of RDFa ontology-based architecture is presented in Fig. 22.1. This framework consists of five components:

1. RDFa[3] annotation for elements of web pages: An URI (uniform resource identifier [14]) generalizes URL (uniform resource locater) for identifying not only web ages but any resource (human, book, an author property).

   Note that there exist tools that can be used to convert RDF/XML (in the future other formats such as N3) content into RDFa snippets that can be easily pasted into existing XHTML Web pages. Therefore, such tools will simply the process of annotating existing XHTML Web pages. One of these tools is RDF2RDFa Converter.[4]
2. HTTP Interceptor: mediates between the server and client machines by managing the HTTP requests. It intercepts HTTP request, checks the availability of HTTP request on the designated directories of web server, and invokes the RDF extractor.
3. RDF extractor: The online RDFa distiller is used to extract the RDFa annotation from the (X)HTML web page and construct the RDF ontology given in the following tables. In the RDFa Distiller, a mismatched tag or missing quotation mark causes unexplained failure. So that, the user-friendly W3 Validator service is used, and hence, at validator.w3.org, which reports some missing tags and also to save code as .xhtml. typically, RDFa distiller was designed as a check of the syntax specification, not as a user tool. It should be noted that the Distiller has some caching issues.

[2] http://www.w3.org/topic/RDFa

[3] is a specification for attributes to be used with XHTML or SVG Tiny to express structured data. The rendered, hypertext data of XHTML is reused by the RDFa markup, so that publishers do not need to repeat significant data in the document content.

[4] http://www.ebusiness-unibw.org/tools/rdf2rdfa/#convert-by-input

We use pyRdfa[5] distiller to generate the RDF triples from an (X)HTML with RDFa or SVG Tiny 1.2 file in various RDF serialization formats such as RDF/XML,[6] Turtle, N triples. pyRdfa[7] is implemented as a Python module and it is available for download.[8]

4. RDF parser: parses the form inputs and their attributes for validation process. This parser is designed to run in the Interceptor interface, allowing to process RDF on the HTTP response which is sent to a client. Typically, RDF-Parser reads RDF files into an RDF data structure and it is generally tested and expected to work with Internet Explorer with different versions. Note that, RDF datatypes and languages are supported for the RDF parser.

   We use our own parser to parse the form inputs and their attributes for validation process. The parser is based on the syntax of Turtle language,[9] which Turtle provides levels of compatibility with the existing N-Triples[10] and Notation 3[11] formats. The parser has a good performance based on tests against randomly generated ontologies and existing large ontologies.
5. Data validator: when the description is extracted using RDFa extractor, the validator takes the user inputs for validation process. The validation process checks to see if the value of user input is satisfied the conditions of its attributes (such as length, data type, minimum length, and if the value contains code or special characters) the since it was used. Any mismatching causes the content integrity check to fail. Based on whether the test passes or fails, the data validator enforces the policy that makes the decision about the next step in the process. If the integrity check passes, the web content is sent to the running process straight away. If it fails, it is refused the user request.
6. RDF parser: parses the form inputs and their attributes for validation process.
7. Data validator: when the description is extracted using RDFa extractor, the validator takes the user inputs for validation process. The validation process checks to see if the value of user input is satisfied the conditions of its attributes (such as length, data type, minimum length, and if the value contains code or special characters) the since it was used. Any mismatching causes the content integrity check to fail. Based on whether the test passes or fails, the data validator enforces the policy that makes the decision about the next step in the process. If the integrity check passes, the web content is sent to the running process straight away. If it fails, it is refused the user request.

[5] http://www.w3.org/2007/08/pyRdfa/

[6] http://www.w3.org/TR/rdf-syntax-grammar/

[7] Other implementations are available at http://rdfa.info/rdfa-implementations/

[8] http://dev.w3.org/2004/PythonLib-IH/dist/pyRdfa.tar.gz

[9] http://www.w3.org/TeamSubmission/turtle/

[10] http://www.w3.org/TR/rdf-testcases/#ntriples

[11] http://www.w3.org/DesignIssues/Notation3

22.3.3 Case Study

To illustrate our methodology we consider using our system to secure a simple employee system. Consider the following scenario: As final step in a registration transaction, employees are sent an (X)HTML form requesting their name, address, department, and qualification (Table 22.1).

Figure 22.2 illustrates the modified (X)HTML form as well as the ontology description. The shaded rows denotes to the ontology which describes each field in the (X)HTML form.

The ontology itself extracted using RDFa extractor is shown in Fig. 22.2. This ontology means that there exists someone whose first name "foaf:firstName" is the "fnm" (Note: this is the name of the label.), last name "foaf:lastname" is "lnm", employee key "vcard:KEY" is center "EMPNO" (Table 22.2). This person is a member "foaf:member" of "WORKDEPT". The employee ontology[12] is stored in the employeeontology.ttl.

Table 22.1 Snapshot of an employee (X)HTML form

```
<FORM NAME=EmployeeForm ACTION=emp_add.jsp METHOD=post>
<h2>Add Employee Record</h2>
<B><I>Employee Number: <br>(1 to 6 characters)</I></B>
<INPUT TYPE=text NAME=EMPNO>
<BR><B><I>First Name:</I></B>
<INPUT TYPE=text NAME=FNM VALUE=First Name>
<BR><B><I>Last Name: </I></B>
<INPUT TYPE=text NAME=LNME VALUE=Last Name>
<INPUT TYPE=submit NAME=Submit VALUE=Add>
</FORM>
```

```
# this is a comment
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix vcard: <http://www.w3.org/2001/vcard-rdf/3.0#>
_:someEmployee rdf:type foaf:Person.
_:someEmployee vcard:KEY 'EMPNO' .
_:someEmployee foaf:firstName 'fnm' .
_:someEmployee foaf:surname 'lnm' .
```

Fig. 22.2 Snapshot of the employee ontology is stored in the employeeontology.ttl file

[12] Note that this RDF ontology is written in the Turtle format, http://www.dajobe.org/2004/01/turtle/

Table 22.2 Snapshot of the modified (X)HTML form with the ontology description

```
<FORM NAME=EmployeeForm ACTION=emp_add.jsp METHOD=post>
<h2>Add Employee Record</h2>
<B><I>Employee Number: <br>(1 to 6 characters)</I></B>
<span property=vcard:KEY content=EMPNO/>
<INPUT TYPE=text NAME=EMPNO>
<BR><B><I>First Name:</I></B>
<span property=foaf:firstName content=fnm/>
<INPUT TYPE=text NAME=FNM VALUE=First Name>
<BR><B><I>Last Name: </I></B>
<span property=foaf:surname content=lnm/>
<INPUT TYPE=text NAME=LNME VALUE=Last Name>
<INPUT TYPE=submit NAME=Submit VALUE=Add>
</FORM>
```

22.4 Implementation and Evaluation of NDVS

The prototype implementation of this service consists of three major components: HTTP Interceptor, RDF parser, and Data validator. First, the HTTP Interceptor takes advantage of the fact that browser requests are directed at both a specific host and a specific port. In this program, the Tomcat server listens on port 8081. The utility listens for browser requests on a default port 80 and redirects to Tomcat. Responses coming to this mechanism are both sent to the client on port 80. The HTTP Interceptor is a multi-threaded java application for handling concurrent connections (requests in parallel) using multiple threads that increase the power and flexibility of a web server and client programs significantly.

Second, the RDF parser is used to parse the form inputs and their attributes. Each form input is parsed, the id of input is sent to the Data validator mechanism. It should be noted the attributes of each input also is sent to the Data validator mechanism. The third component is the Data validator. When the description is extracted using RDFa extractor, the Data validator takes the user inputs for validation process as shown in the previous section. If the integrity check passes, the data is sent to the running process straight away. If it fails, it is refused the user request. To test our approach, we implement the NDVS to verify integrity of inputs of real-world web applications. We conducted a number of experimental studies to test the reliability and performance of NDVS.

22.4.1 Case Study: Security Objective

To assess how successful the NDVS is able to detect and recover from the application vulnerabilities, we started to use an ontology to describe all data elements in a web application using RDFa annotation. This annotation takes for every (X)HTML form stored in the designated directories of the suggested web sites/applications hosted on Tomcat. Thus, over 35 attacks were performed against the target web application. These attacks exploited different types of vulnerabilities that allowed for the modification of files in the designated directories of a web Server.

22.4.2 End-to-End Performance Evaluation

The testing environment for experiments (1 and 2) is composed of Apache 1.3.29 with Tomcat container 5.01 on MS Windows Server 2003. The Tomcat web server contains a copy of target web site and shopping cart application.

A load test can be used to test an application's robustness and performance, as well as its hardware and bandwidth capacities. In these experiments, we used the Neoload application which is a stress and load testing tool to (1) test a web site's vulnerability to crashing under load and (2) check response times under the predicted load.

We have measured the end-to-end performance with the (1) implemented NDVS, (2) and the traditional validation routines. The duration of the test was dedicated by the requirements of the Neoload testing. The run-time policy was ramped-up from two users adding two users every 2 min. The virtual users were connecting at 100 Mbps through a local network. Note that, due to the lack of existing approach'es codes, this comparison only includes the NDVS and the traditional validations routines which we implemented by ourself.

All these measurements were performed from the client point of view. The average response time is the meantime necessary to process a request by each web server when the proxy, browser, and the NDVS are active. It should be noted that the communications are parts of the measured times.

In experiential study one, 5,421 hits were created, 5,397 web pages were served and 1.55 MB (total throughput) were received by client. the number of virtual users launched was between 198 and 204. Whereas, In study two, 16,396 hits were created, 14,136 web pages were served, and 29.29 MB were received by the client. Number of virtual users launched was between 150 and 474.

The results indicate that the average response time when using the traditional validation routines is better than when using the NDVS. Table 22.3 illustrates that the average response time (request) through NDVS was 3.95 s on Tomcat. when using traditional validation routines, the average response time (request) was 3.14 s on Tomcat. Therefore, the NDVS satisfies the performance objective for validation of data inputs.

Table 22.3 also illustrates that the average response time (request) through NVDS was 0.437 s on Tomcat. In case of without using any validation system, the average response time (request) was 0.527 s on Tomcat. The results indicate that the average response time when using the NVDS is greater than when not using any validation system – it is suggested that the reason for this is The HTTP connection problem. For example, at the first 3 min during the test, the wrong URL was

Table 22.3 Experimental study one and two: comparison between the response times through NVDS and the traditional validation routines, in seconds, of all requests during the test on a Tomcat web server

Experiment	System	Minimum response time (request)	Average response time (request)	Maximum response time (request)	Average page response time
1	Traditional validation routines	0.029	3.14	4.86	3.24
1	NVDS	0.219	3.95	6.01	3.97
2	Traditional validation routines	0.032	0.527	4.08	0.634
2	NVDS	0.011	0.437	5.29	0.505

requested, indicating the communication between the client and the server was disconnected. Consequently this led to an increase in the average response time. Note, the port number (8081) of connection in the URL request was not added.

22.5 Summary

In this chapter, we have surveyed the existing data validation approaches and considered their strengths, weakness, and limitations. In addition, we have compared among the existing approaches, systems and tools. In an attempt to overcome lack and bypassing of data validation problems, we have proposed a new validation service. This architecture includes a real-time framework consisting of five components: RDFa annotation for elements of web pages, interceptor, RDF extractor, RDF parser, and data validator. The results from a series of experimental tests studies appear to suggest that the NDVS can satisfy the performance objective and it might be suggested that the developed NDVS could provide a detection, and prevention of some web application attacks.

References

1. Aljawarneh, S., Alkhateeb, F.: Design and implementation of new data validation service (NDVS) using semantic web technologies in web applications. In: Proceedings of the World Congress on Engineering 2009: WCE'09, vol. I, pp. 179–184. London, UK, International Association of Engineering (2009)
2. Aljawarneh, S., Alkhateeb, F.: A semantic web technology-based architecture for new server-side data validation in web applications. In: Aldawood, A. (ed.) *ICIT'9*, Amman, Jordan Alzaytoona Univeristy (2009)
3. Balzarotti, D., Cova, M., Felmetsger, V., Jovanovic, N., Kirda, E., Kruegel, C., Vigna, G.: Saner: Composing static and dynamic analysis to validate sanitization in web applications.

In: SP '08: Proceedings of the 2008 IEEE Symposium on Security and Privacy, pp. 387–401. Washington, DC, IEEE Computer Society (2008)

4. Cardone, R., Soroker, D., Tiwari, A.: Using XForms to simplify web programming. In: WWW '05: Proceedings of the 14th international conference on World Wide Web, pp. 215–224. New York, NY, ACM (2005)
5. Corsaire: A modular approach to data validation in web applications. White paper (2006)
6. Glisson, W.B., Welland, R.: Web development evolution: The assimilation of web engineering security. In: LA-WEB '05: Proceedings of the Third Latin American Web Congress, p. 49. Washington, DC, IEEE Computer Society (2005)
7. Hassinen, M., Mussalo, P.: Client controlled security for web applications. In: Wener, B. (ed.) The IEEE Conference on Local Computer Networks 30th Anniversary, pp. 810–816. Australia, IEEE Computer Society Press (2005)
8. Huang, Y.-W., Huang, S.-K., Lin, T.-P., Tsai, C.-H.: Web application security assessment by fault injection and behavior monitoring. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, pp. 148–159. New York, NY, ACM (2003)
9. IBM: IBM Internet Security Systems X-Force Threat Insight Quarterly. http://www-935.ibm.com/services/us/iss/pdf/xftiq_09q1.pdf. Accessed 8 Oct 2009
10. Jovanovic, N., Kruegel, C., Kirda, E.: Pixy: a static analysis tool for detecting web application vulnerabilities (short paper). In: SP '06: Proceedings of the 2006 IEEE Symposium on Security and Privacy, pp. 258–263. Washington, DC, IEEE Computer Society (2006)
11. Kienle, H.M., Müller, H.A.: Leveraging program analysis for web site reverse engineering. In: WSE '01: Proceedings of the 3rd International Workshop on Web Site Evolution (WSE'01), p. 117. Washington, DC, IEEE Computer Society (2001)
12. MOCEAN, L.: Internet data validation, economy informatics. Revistaie.ase.ro/content/EN7/Mocean.pdf (2007)
13. Scott, D., Sharp, R.: Specifying and enforcing application-level Web security policies. IEEE. Knowl. Data Eng. **15**(4), 771–783 (2003)
14. Berners-Lee, T., Fielding, R., Masinter, L.: Uniform resource identifiers (URI): Generic syntax. RFC 2396, IETF. http://www.ietf.org/rfc/rfc2396.txt (1998)
15. Thompson, H.H., Whittaker, J.A.: String-based attacks demystified. In: Doctor DOBBS J. 29, 61–63 (2004). ISSN: CMP MEDIA LLC Country of publication USA. 1044-789X.

Chapter 23
Classification of Road Traffic Congestion Levels from Vehicle's Moving Patterns: A Comparison Between Artificial Neural Network and Decision Tree Algorithm

Thammasak Thianniwet, Satidchoke Phosaard, and Wasan Pattara-Atikom

Abstract We proposed a technique to identify road traffic congestion levels from velocity of mobile sensors with high accuracy and consistent with motorists' judgments. The data collection utilized a GPS device, a webcam, and an opinion survey. Human perceptions were used to rate the traffic congestion levels into three levels: light, heavy, and jam. We successfully extracted vehicle's moving patterns using a sliding windows technique. Then the moving patterns were fed into ANN and J48 algorithms. The comparison between two learning algorithms yielded that the J48 model shown the best result which achieved accuracy as high as 91.29%. By implementing the model on the existing traffic report systems, the reports will cover on comprehensive areas. The proposed method can be applied to any parts of the world.

Keywords Intelligent transportation system (ITS) · traffic congestion level · human judgment · artificial neural network (ANN) · decision tree (J48) · GPS · sliding windows · moving pattern

23.1 Introduction

Traffic reports in real-time are essential for congested and overcrowded cities such as Bangkok or even in sparse and remote areas during a long holiday period. Without these, commuters might not choose the proper routes and could get stuck in

T. Thianniwet (✉) and S. Phosaard
School of Information Technology, Institute of Social Technology,
Suranaree University of Technology, 111 University Ave., Muang Nakhon Ratchasima,
Nakhon Ratchasima 30000, Thailand
e-mail: thammasak@sut.ac.th; s@sut.ac.th

W. Pattara-Atikom
National Electronics and Computer Technology Center (NECTEC), Under the National
Science and Technology Development Agency (NSTDA), Ministry of Science
and Technology, 112 Thailand Science Park, Phahon Yothin Road, Klong 1,
Klong Luang, Pathumthani 12120, Thailand
e-mail: wasan@nectec.or.th

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_23,

traffic for hours. Intelligent Transportation System (ITS) with automated congestion estimation algorithms can help produce such reports. Several initiatives from both private and government entities have been proposed and implemented to gather traffic data to feed the ITS. According to our survey, most efforts focus on limited installation of fixed sensors such as loop-coils and intelligent video cameras with image processing capability. However, the costs of such implementations are very high due to the high cost of the devices, installation, and maintenance. Moreover, these fixed sensors are vulnerable to extreme weather typical in certain areas. Additionally, the installation of fixed sensors to cover all roads in major cities is neither practical nor economically feasible. An alternative way to collect traffic data at a lower cost with wider coverage is therefore needed.

Recently, mobile sensors or probe vehicles appeared as a complementary solution to fixed sensors for increasing coverage areas and accuracy without requiring expensive infrastructure investment. Two popular types of mobile sensors are GPS-based and cellular-based. GPS-based sensors are sensors with GPS capability and cellular-based sensors are sensors that use information from cellular networks as traffic sensors.

Cellular-based sensors are low in cost due to the large number of mobile phones and their associated infrastructures already in service. According to recent statistics, the mobile phone penetration rate in Thailand is expected to grow to 90% in 2009 [1]. However, GPS-based sensors are far more efficient to pinpoint vehicle locations, thus they can provide highly accurate vehicle movement information. Moreover, recent mobile phones have integrated GPS capability, such as Apple iPhone and several other smart phones.

In this paper, we explored a model that can automatically classified traffic congestion levels for traffic reports. The model can be further implemented in the system that combines advantages of GPS-based sensors, in that they are highly accurate, and of cellular-based sensors, in that they are highly available. This model, combine with mobile sensors, can generate traffic reports that virtually cover all of the areas that vehicles and mobile networks can reach.

This paper is organized as follows: In Section 23.2, we describe related works concerning traffic congestion reports. The methodology of the research is presented in Section 23.3. Section 23.4 provides results and evaluations, and Section 23.5 offers a conclusion and the possibilities of future work.

23.2 Related Works

Congestion level estimation techniques for various types of the collected data are our most related field. Traffic data could be gathered automatically from two major types of sensors: fixed sensor and mobile sensor. The study in [2] applied the neural network technique to the collected data using mobile phones. It used Cell Dwell Time/CDT, the time that a mobile phone attaches to a mobile phone service antenna, which provides rough journey speed. Our work employed another machine learning technique that might better suit with the characteristics of the data, and compared the neural network model with the selected technique. The GPS data would provide more precise traffic information than that roughly provided by the CDT. The studies in [3, 4] estimated the congestion level using data from traffic camera by applying fuzzy logic and hidden Markov model, respectively. Our work applied artificial neural network (ANN) and decision tree (J48) technique on mobile sensors. Using data collected from mobile sensors would cover far greater traffic ranges. The algorithm would learn over movement patterns of a vehicle. Sliding window technique with fixed window size was also used. The works of [5–7] also investigated various alternative techniques related to our work.

In some countries, for example, as in the study of [8, 9] found out that the main parameters used to define the traffic congestion levels are time, speed, volume, service level, and the cycles of traffic signal that the motorists have to wait. Our work would focus only on interpretation of vehicle velocity since our work needs to determine the congestion levels with minimal parameters. Vehicle velocity could be collected by almost all types of sensors. This made it easier, broader and more versatile for the model to be used. The congestion levels that we studied were limited to three levels: free-flow, heavy and jam, which was enough and appropriate according to the study of [10]. After we successfully derived the congestion classification model, the GPS data were planned to be collected through mobile phones attached by GPS device. The data would be sent through the data network, such as GPRS, EDGE, and so on. The next section described the methodology of the research.

23.3 Methodology

23.3.1 Collection of Empirical Data

The traffic data were collected from several highly congested roads in Bangkok, e.g., Sukhumvit, Silom, and Sathorn. A notebook attached with a USB GPS device is used to collected date, time, latitude, longitude, and vehicle velocity from GPS's GPRMC sentences. We captured images of road traffic condition by a video camera mounted on a test vehicle's dash board. Our vehicle passed through overcrowded urban areas approximately 30 km within 3 h.

In our experiment, we gathered the congestion levels from 11 subjects with driving experience up to 10 years. They watched a 3-h video clip of road survey and rated the congestion levels into three levels, light, heavy, and jam. Then, the concluded congestion levels from 11 subjects were calculated using majority vote. The judged congestion levels were then synchronized with velocity collected by the GPS device. We observed that the data were wildly fluctuated and also non-uniform, as shown in Fig. 23.1. To alleviate this oscillation, the traffic data was treated before feeding into a learning algorithm, i.e., decision tree model, which will be explained in detail in the next section.

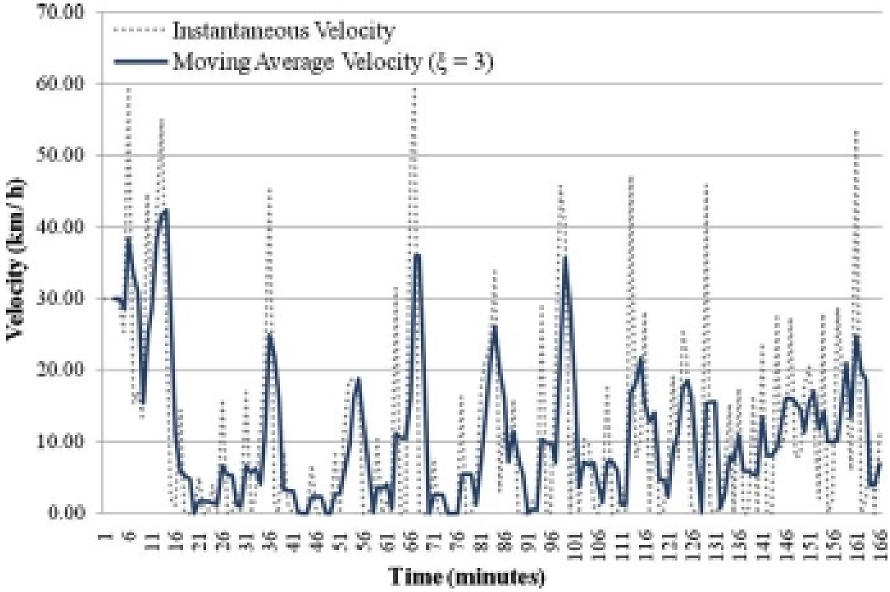


Fig. 23.1 Instantaneous velocity versus moving average velocity ($\xi = 3$)

23.3.2 Data Preparation

We minimized a set of attributes by concentrating only on the vehicle velocity and the moving pattern of a vehicle, which can infer different levels of congestion. Then, we applied three steps to prepare the data: (1) smoothening out instantaneous velocity, (2) extracting moving pattern of a vehicle using sliding windows technique, and (3) balancing the distribution of sampling data on each congestion level. Next, we will explain each procedure in details.

23.3.2.1 Smoothening Out Instantaneous Velocity

Instantaneous vehicle velocity from the GPS data usually fluctuated widely, as shown in Fig. 23.1 as the dotted line. This fluctuation made the learning algorithm difficult to determine the pattern and classify the congestion level, as in [11]. Therefore, we needed to smoothen out the fluctuation of instantaneous velocity. We applied moving average algorithm by averaging the previous ξ samples as shown in Eq. 23.1. MV_t represents the moving average velocity at time t. In our experiment, ξ was set to 3. The results of the average velocity are shown as the thick line in Fig. 23.1.

$$MV_t = \frac{\sum_{i=t}^{t-\xi} V_i}{\xi} \tag{23.1}$$

23.3.2.2 Extracting Vehicle's Moving Patterns

When the instantaneous velocity was less fluctuated by the smoothening algorithm, it was easier to investigate vehicle's moving patterns. We successfully extracted moving patterns that were practical to be efficiently learned by the learning algorithm, which can be explained as follows. Our previous work suggested we can use velocity to estimate congestion levels. Figure 23.2 illustrates the vehicle moving patterns corresponding to the congestion levels. To make the graph readable, we scale the congestion scores (1, 2 and 3) by 10, i.e., 10 = jam, 20 = heavy and 30 = light.

From Fig. 23.2, when a vehicle is moving with high velocity for a while, it means that the road traffic is light, e.g., the velocity between the time of 10 and 14. If the velocity decreases to a moderate range, it means that the road traffic is heavy, e.g., the velocity at the time of 15. And if the velocity decreases to low velocity, it means that the road traffic is jam.

Although the value of vehicle's velocity can be used to determine the congestion level, we cannot say that only a value of a vehicle's velocity at a moment can be used to accurately determine the congestion level. In a real driving situation, an instantaneous velocity can be reported at any congestion levels. For example, a vehicle needs to slow down for turning or stopping for a traffic light. In this condition, the traffic might be light but the velocity is relatively low. Figure 23.3 visualizes the chart between congestion levels and instantaneous velocity. For example, reported congestion levels with a velocity near 0 km/h were mutually reported as either light, heavy or jam.

After carefully investigating the data, we successfully mimicked humans' judgments on congestion levels based on moving patterns of a vehicle which was derived from the historical data. Sliding windows, a technique that could satisfy such moving pattern extraction, was employed. We applied fixed sliding windows of size δ to

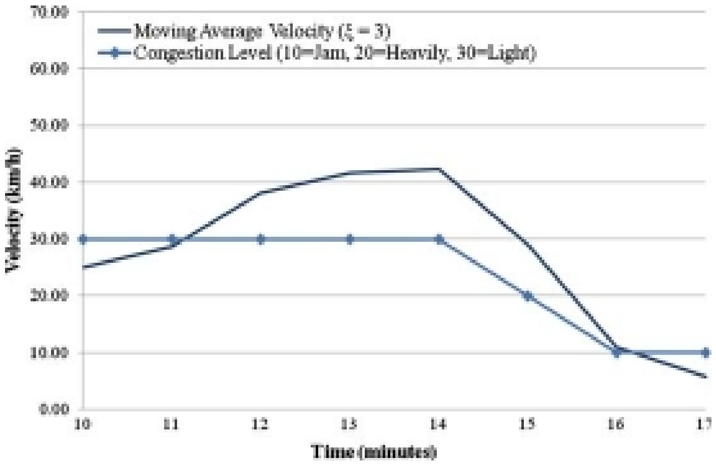


Fig. 23.2 Vehicle moving pattern and deduced congestion level

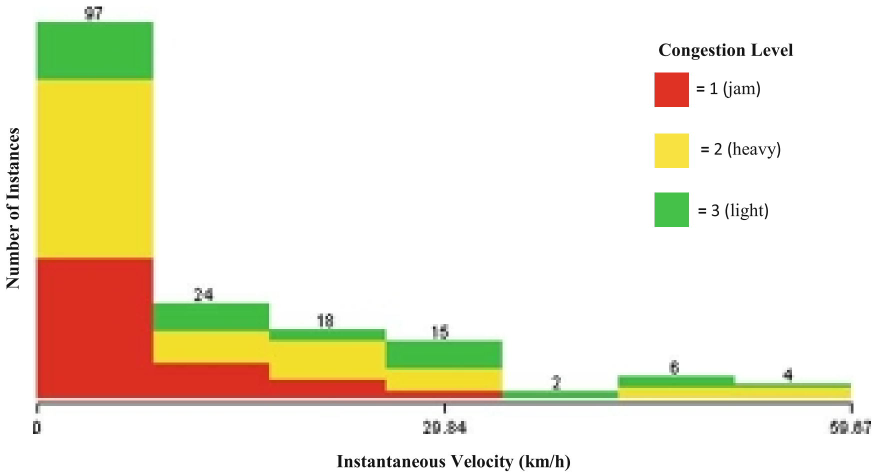


Fig. 23.3 Congestion levels versus instantaneous velocity

Table 23.1 An example of instantaneous velocity and derived attributes

Time	INS_t	MV_{t-2}	MV_{t-1}	MV_t	AMV_t	Level
13:10	44.66	–	–	–	–	3
13:11	27.79	–	–	–	–	3
13:12	42.06	–	–	38.17	–	3
13:13	55.09	–	38.17	41.65	–	3
13:14	29.83	38.17	41.65	42.33	39.89	3
13:15	2.04	41.65	42.33	28.99	31.36	2
13:16	1.11	42.33	28.99	10.99	26.03	1

capture moving patterns of a vehicle from the vehicle velocity. In our experiment, δ was set to 3, which means we captured the moving patterns by a set of three consecutive moving average velocities. The moving pattern at time t with δ equals to 3 includes three consecutive samples of moving average velocity at time t (MV_t), and two priori moving average velocities at time $t-1$ (MV_{t-1}), and $t-2$ (MV_{t-2}). We also introduced a new attribute to represent the average velocity of each sliding window (each moving pattern), called AMV_t. For the moving pattern at time t with δ and ξ set to 3, the value of AMV_t can be computed by the value of $MV_{t(\xi=5)}$. Table 23.1 demonstrates how to calculate the moving average at time t from instantaneous velocity (INS_t), and how to extract moving patterns.

The steps of how to calculate the values in Table 23.1 can be explained as follows. The moving average velocity by the time of 13:12 can be calculated by averaging the current instantaneous velocity, 42.06, with two priori velocity, 27.79 and 44.66. AMV_t is the average velocity which covers the moving pattern at time t. The value of $\mathrm{AMV_t}$ at 13:14 can be calculated by averaging instantaneous velocity beginning from 13:10, also a starting point of $\mathrm{MV_{t-2}}$, to 13:14, also the end point of $\mathrm{MV_t}$. Thus, the calculation of an $\mathrm{AMV_t}$ with δ and ξ set to 3 equals to the calculation of an $\mathrm{MV_t}$ with ξ set to 5. The last column, Level, indicates congestion levels rated by human. The values of 1, 2, and 3 represent jam, heavy, and light traffic respectively.

23.3.2.3 Balancing Class Distributions

In our experiment, we captured vehicle's moving patterns every minute from 13:00 to 15:45. Since the calculations of $\mathrm{MV_t}$ and $\mathrm{AMV_t}$ depend on previous cascading calculations, the first four instances were omitted. Therefore, there were 162 instances: 52 instances were in the class of jam traffic, 74 instances were in the class of heavy traffic, and there were only 36 instances were in the class of light traffic. Class imbalance may cause inferior accuracy in data mining learners, as [12]. Generally, classification models tend to predict the majority class if class imbalance exists. In this case, the class of heavy traffic was the majority class while the minority classes, the classes of light and jam traffic, were also highly important. Therefore, we needed to balance the class distributions to avoid the problem.

By this step, we applied a simple technique to alleviate the problem of class imbalance by applying a technique that was similar to the technique of finding a least common multiple number. The result of class balancing yielded 448 instances with 156 instances on class jam, 148 instances on class heavy, and 144 instances on class light. Then, this data set was used to train the classification model, for which we explain the details in the next section.

23.3.3 Data Classifications

The preprocessed data set was used to train and evaluate the classification model. Our data set consisted of five attributes. The first three attributes were $\mathrm{MV_{3t-2}}$, $\mathrm{MV_{3t-1}}$, and $\mathrm{MV_{3t}}$, which were three consecutive moving average velocities that represented the moving pattern. The fourth attribute was $\mathrm{AMV_{3t}}$, which was the average velocity of the corresponding moving pattern. The last attribute was Level, which was the congestion level judged by human ratings. Besides the ANN algorithm that we employed to make connections with the previous work [2], we chose the J48 algorithm, a well-known decision tree algorithm in the WEKA system, to generate a decision tree model to classify the Level. The Multilayer-Perceptron was selected for ANN in WEKA. WEKA is a machine learning software developed by the University of Waikato. It is a collection of machine learning algorithms for data mining tasks. The goal attribute of the model was set to Level. The test option was set to tenfold cross-validation.

23.4 Results and Evaluations

23.4.1 Classification Model

After successfully training the classification model, the derived neural network and the decision tree are shown in Figs. 23.4 and 23.5 respectively.

The model configuration of ANN is 4-3-3. We used the default value for the number of the hidden nodes, which is calculated by the summation of the number of the attributes and the number of the class, all divided by two. The momentum and the learning rate are set to the default 0.2 and 0.3 respectively. The model achieves an accuracy of only 55.13% with a root mean square error of 0.43. The AMV_{3t} has the highest weight in the model.

The size of our J48 decision tree is 125 nodes, 63 of which are leave nodes. The time taken to build the model is about 0.08 s. The root node is AMV_{3t} attribute. From both model, it means that the average of the moving average velocity is the most important factor to determine the level of road traffic congestion.

23.4.2 Performance Evaluations

It is obvious that the J48 algorithm achieved a much higher accuracy. The reasons behind the low accuracy of the ANN model might base on the explanations that the data set of this fluctuating vehicle moving patterns, the number of instances or

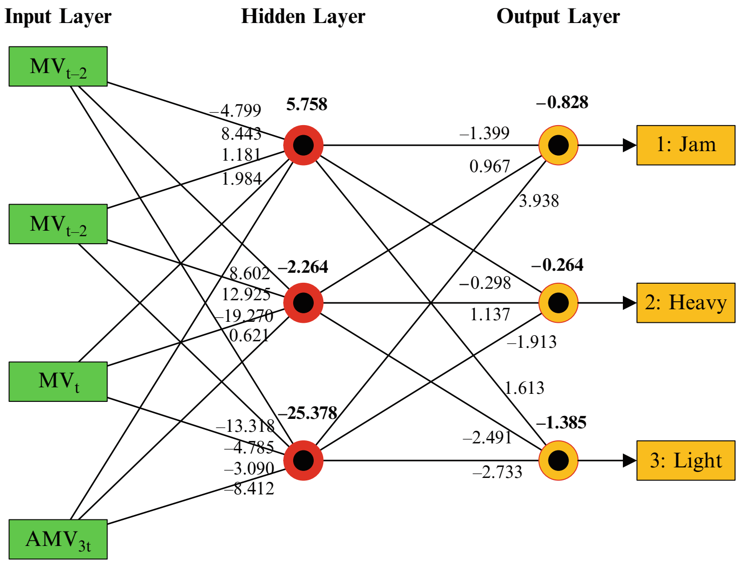


Fig. 23.4 The derived ANN model

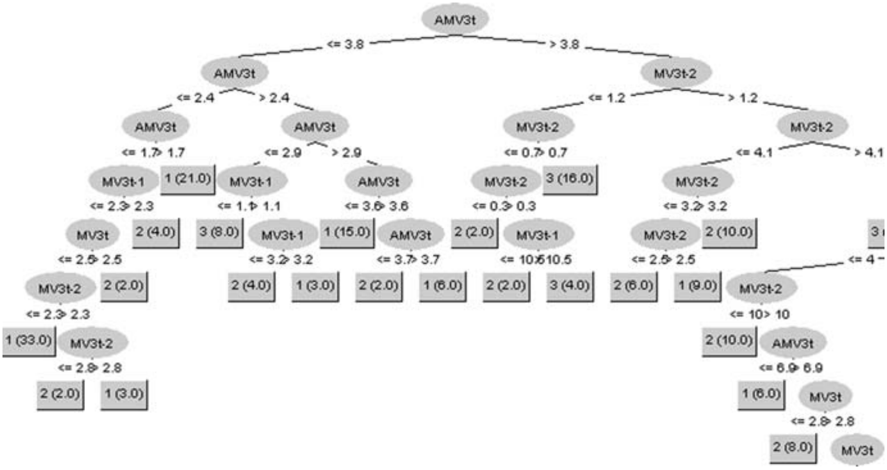


Fig. 23.5 The derived J48 decision tree

Table 23.2 The confusion matrix of the model from J48 and neural network

		J48			Neural network		
Predicted congestion level		Jam	Heavy	Light	Jam	Heavy	Light
Instances congestion level	Jam	150	4	2	88	26	42
	Heavy	20	115	13	28	52	68
	Light	0	0	144	20	17	107

Table 23.3 The classifier's performance of the model from J48

Class	TP rate	FP rate	Precision
Jam (1)	0.962	0.068	0.882
Heavy (2)	0.777	0.013	0.966
Light (3)	1.000	0.049	0.906
Average	0.913	0.044	0.918

the sliding window technique might not fit with the sigmoid function of the ANN. A further investigation can be a future research. Thus, only the J48 results will be elucidated.

The result shows a promising technique of determining congestion by the J48 algorithm with an overall accuracy of 91.29%, a root mean square error of 0.2171, and a precision ranging from 0.882 to 0.966. The result shows a true positive rate (TP Rate or sensitivity) ranging from 0.777 to 1.000, which is very high, and a false positive rate (FP Rate) ranging from 0.013 to 0.068, which is very low. Table 23.2 shows the classifier's performance for each class in details. Table 23.3 shows the result of the model evaluation by the confusion matrix.

From Table 23.3, the highest TP Rate is 1.000 on the Light class. This means that when the road traffic congestion level is light, our classifier will 100% correctly classify the traffic. The lowest TP Rate is 0.777 on the Heavy class. It can be interpreted that when the road traffic congestion level is heavy, our classifier will 77.7% correctly classify the traffic. In general outlier human perceptions could occur. Because the heavily congested level is at the middle between the light and jam level, some people may judge the traffic in the light class or in the jam class as being in the heavy class. When these judgments were fed into the classification algorithm, they were treated as noise and would be ignored. The number 20 and 13 in the confusion matrix, as per Table 23.3, is the result of misclassification on the heavy traffic class. The number 20 represents the instances of heavy class which the model misclassified as jam traffic, and the number 13 represents the instances in heavy class which the model misclassified as light traffic.

Although the classification of heavy traffic is the worst by its TP Rate, it yields the lowest value of FP Rate with 0.013, which is the best value. Moreover, it also yields the best precision score of 0.966. This means that when the classifier classifies traffic congestion patterns as a heavy traffic pattern, it will 96.6% correctly classify.

23.5 Conclusion

In this study, we investigated an alternative technique to automatically classify the road traffic congestion levels that was highly consistent with road users' judgments. The technique minimally required data from GPS devices, which can be collected from participants through mobile data networks. Vehicle velocity can be used to determine the congestion level but the instantaneous velocity fluctuated widely. We smoothened out the oscillated instantaneous velocity by averaging it with historical velocities, which was called moving average velocity. We applied a sliding windows technique to capture the consecutive moving average velocities, which was called a moving pattern. We derived a new attribute, AMV_{3t} which represents the average velocity of the corresponding moving pattern. Parameters δ and ξ were set to 3. The moving patterns were captured every minute. Then road users' judgments and related information were learned utilizing an artificial neural network (ANN) and a decision tree model (J48). The results suggested that the decision tree was better fit with the characteristics of the data. The evaluations revealed that the decision tree model achieved an overall accuracy as high as 91.29% with a precision as high as 96.6%. The root mean square error was only 0.2171.

In future study, we can optimize the δ, ξ, and time interval between two consecutive velocities, which might improve the accuracy of our model. Moreover, we plan to integrate such a model into the existing ITS system in Bangkok. The technique will also be extended to apply to cover the whole country if possible.

References

1. Phoosuphanusorn, S.: New mobile-phone users up 30%. Bangkok Post, Bangkok (May 2007)
2. Pattara-atikom, W., Peachavanish, R.: Estimating road traffic congestion from cell dwell time using neural network. The 7th International Conference on ITS Telecommunications (ITST 2007), Sophia Antipolis, France, June 2007
3. Pongpaibool, P., Tangamchit, P., Noodwong, K.: Evaluation of road traffic congestion using fuzzy techniques. Proceedings of IEEE TENCON 2007, Taipei, Taiwan, October 2007
4. Porikli, F., Li, X.: Traffic congestion estimation using hmm models without vehicle tracking. IEEE Intelligent Vehicles Symposium, pp. 188–193, June 2004
5. Lu, J., Cao, L.: Congestion evaluation from traffic flow information based on fuzzy logic. IEEE Intell. Transport. Syst. 1:50–33 (2003)
6. Krause, B., von Altrock, C.: Intelligent highway by fuzzy logic: Congestion detection and traffic control on multi-lane roads with variable road signs. 5th International Conference on Fuzzy Systems 3:1832–1837 (September 1996)
7. Alessandri, R.B.A., Repetto, M.: Estimating of freeway traffic variables using information from mobile phones. IEEE American Control Conference (2003)
8. Lomax, J.T., Tuner, S.M., Shunk, G., Levinson, H.S., Pratt, R.H., Bay, P.N., Douglas, B.B.: Quantifying Congestion: Final Report. National Cooperative Highway Research Program Report 398, TRB, Washington, DC (1997)
9. Bertini, R.L.: Congestion and Its Extent. Access to Destinations: Rethinking the Transportation Future of our Region, Minnesota, USA (2004)
10. Choocharukul, K.: Congestion measures in Thailand: State of the practice. Proceedings of the 10th National Convention on Civil Engineering, pp. TRP111-TRP118, May 2005
11. Pattara-atikom, W., Pongpaibool, P., Thajchayapong, S.: Estimating road traffic congestion using vehicle velocity. Proceedings of the 6th International Conference on ITS Telecommunications, pp. 1001–1004, Chengdu, China, June 2006
12. Drown, D.J., Khoshgoftaar, T.M., Narayanan, R.: Using evolutionary sampling to mine imbalanced data. Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA 2007), pp. 363–368, Ohio, USA, December 2007

Chapter 24
Fuzzy Parameters and Cutting Forces Optimization via Genetic Algorithm Approach

Stefania Gallova

Abstract The classification of solved signal features for manufacturing process condition monitoring has been carried out using fuzzy parameters optimization processing. In cases where assumptions in respect of nonlinear behavior cannot be made, the need to describe mathematically, ever increasing complexity become difficult and perhaps infeasible. The optimization possibilities of the fuzzy system parameters using genetic algorithms are studied. An analytical function determines the positions of the output fuzzy sets in each mapping process, that substitute the fuzzy rule base used in conventional approach. We realize case adaptation by adjusting the fuzzy sets parameters. Fuzzy parameters within optimization procedure could be multiobjective. We solve also the system for cutting process simulation, which contains the experimental model and the simulation model based on genetic algorithms. There is developed a genetic algorithm based simulation procedure for the prediction of the cutting forces. These genetic algorithms methodologies are suitable for fuzzy implementation control and for solution of large-scale problems.

Keywords Fuzzy parameter optimization · cutting forces optimization · genetic algorithm · fitness

24.1 Introduction

The genetic algorithm based simulation procedure is proposed to predict cutting forces. The procedure evaluates the cutting conditions subjected to constraints such as cutting speed, cutting width, feeding and cutting depth. We realize the transition in computer applications from *data processing* to *information processing* and then to *knowledge processing.*

S. Gallova (✉)
Pavol Jozef Safarik University in Kosice, Srobarova 2, SK-041 80 Kosice, Slovak Republic
e-mail: stefania.gallova@zoznam.sk; stefania.gallova@upjs.sk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_24,

We evaluate cutting conditions subjected to various machinery and manufacturing constraints to ensure the process quality and process efficiency. The operation of the cutting forces simulation model can be confirmed by experimental results. The solved problem applies also fuzzy set theory mathematization approach. Stable cutting experiments are defined by straight cutting using parameters of technology. Unstable cutting experiments consist of different causes of instabilities. The realized simulation process is based on genetic algorithm and on the analytical formulation of the cutting process components. We use some artificial intelligence tools, such as expert system, genetic algorithm methodology and fuzzy set theory principles. These approaches cooperate with each other. We use also progressive internet technology tools. The optimization and the simulation of machining parameters and cutting forces become easier. The configuration of combination of genetic algorithm methodology with expert system approach and web technology is illustrated in Fig. 24.1.

The used symptoms S_1 to S_n correspond to n different on-line information sources, which could be on-line measurements and controller outputs. Each symptom in the condition part of rule is coded by a 2-bit binary. The use of strong and weak functional dependencies between sets of domain fields to factor the product into a family of projections in "*normal form*" allows entry and update requests to be treated as single changes in projections that together generate a multiple change in the database state that is legal and preserves integrity. Expert system environment is used as an intelligent adviser for users. It has a large domain knowledge base that is usable for genetic algorithm procedures [1,2].

24.2 Genetic Algorithm Optimization Methodology Approach

In the absence of an a priori known best or worst fitness of a string, the solved methodology of fitness calculation is based on the idea of measuring the increase of fitness created by the secondary genetic algorithm approach. We have a parameter F_n^m, which is the best fitness of a string in the secondary population for test problem example m after generation n. A parameter p_t is the number of test problems. The *fitness* of the secondary genetic algorithm GA_{Fs} after generation n is given by [3,4]:

$$Fitness[GA_{FS}] = \frac{1}{p_t} \sum_{m=1}^{p_t} \frac{\max\left(0, F_n^m - F_0^m\right)}{1 - F_0^m} \tag{24.1}$$

By this way, we will avoid distorted fitness values arising from possibly different degrees of complexity of test problem examples. The fitness of an individual structure is a measure indicating how fitted the structure is [5].

There is realized genetic *fuzzy rule learning*. The goal of this algorithm approach is also to find effective fuzzy *If-then-else* rules to predict the class of input patterns

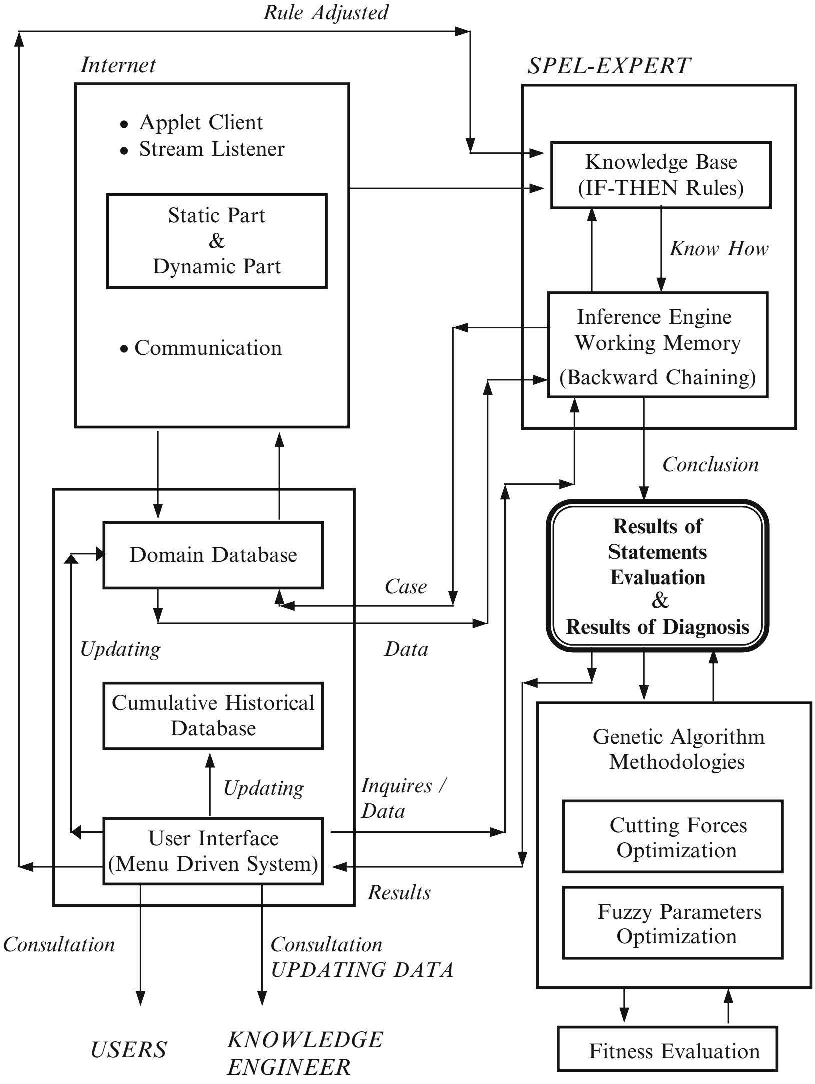


Fig. 24.1 The architecture of problem solving

correctly. The proposed learning approach is performed on intrusion detection that is a complex classification problem [1, 2, 6–8].

The random is extracted according to the patterns of the training dataset, which their consequent class is the same as the class that the algorithm works on. There is determined the most compatible combination of antecedent fuzzy sets using the six linguistic values (see Fig. 24.2).

There is calculated the compatibility of antecedent fuzzy rules with the random pattern. The final classification system is capable of detecting known intrusive

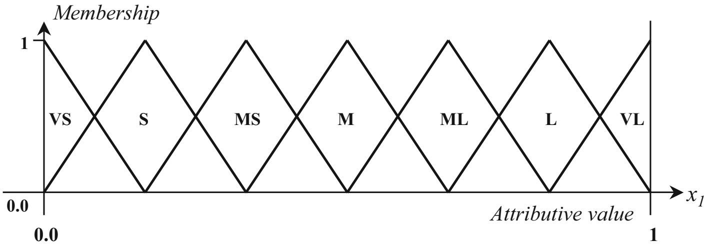


Fig. 24.2 Membership function of seven linguistic variables: VS = Very small; S = Small; MS = Medium small; M = Medium; ML = Medium large; L = Large

behavior in a computer network. Our experiments show that it has an acceptable performance.

To assign fitness to the evaluated fuzzy rule, the rule is first combined with the best rule in the other population using the following template:

if Rule1 $\geq$ 0.6
then high-permeability
else if Rule2 $\geq$ 0.6
then low-permeability
else medium-permeability,

where *Rule1* is a rule from the first population and *Rule2* is a rule from second population. The crowding technology has been introduced to induce niche like behavior in genetic algorithm search in order to maintain diversity in the population (diversity necessary for continuous evolution). Experimental results from runs where population 1 evolves rules to identify high-permeability data.

Both populations co-evolve together with comparable average fitness. This is a healthy co-evolutionary dynamics that has produced combined if-then-else rules that give more accurate permeability interpretations than that by other experiment. In both sets of experimental runs, the two populations improved very quickly at the first 190 generations. After that, the improvement is not very visible. This pattern also appears in the fitness improvement of the best combined overall permeability interpreter although to a lesser extent. One possible mean is that the best solution used to combine with individuals in the other population. This mechanism with the strength and specificity rules management can be effectively assimilated to a genetic operator. So it may be interesting to compare this solution with above mentioned genetic algorithm approach [6–8]. The learnt rules have been tested on the real process. All simulated faults were successfully diagnosed by the corresponding rules and no incorrect diagnosis occurred.

24.3 Synthesizing Process of Analytic Fuzzy Logic Controller

The process of *synthesizing analytic fuzzy logic* controller can be described as follows:

1. We define a new adaptive shapes and distribution of the input fuzzy sets by fuzzyfication interface [9–12]. See relations (24.2a, b).
2. We calculate the analytic calculation function f_1, for the activation of corresponding j-th output fuzzy set with degree f_j. (Eq. 24.3).
3. We initiate an analytic function that maps the values of input variables into positions of the centers y_{cj} of the corresponding output fuzzy sets (inference engine) (Eq. 24.4).
4. We use analytic expression for the determination of the output value from the analytic fuzzy logic controller, as function of the shapes of the output fuzzy sets and their corresponding positions of the centers and activation functions (deffuzyfication interface) (Eq. 24.5).

$$M_i^j(x_j) = \frac{1}{2^{\left(B_j . e_i^j \, |x_j|\right)}} \left\{ \frac{1 - \cos\left[2\,\pi\, e_i^j \left(x_j - x_{ci}^j + \frac{T_i^j}{2}\right)\right]}{\left(e_i^j - 1\right)\, T_i^j} \right\} \tag{24.2a}$$

with conditions:

$$-1 \le x_j \le x_{ai}^j \quad x_{ai}^j \le x_j \le x_{bi}^j \quad M_i^j(x_j) = 1 \quad x_{bi}^j \le x_j \le 1$$

$$M_i^j(x_j) = \frac{1}{2^{\left(B_j . e_i^j \, |x_j|\right)}} \left\{ \frac{1 - \cos\Bigl[2\,\pi\, e_i^j \Bigl(x_{ci}^j + \frac{T_i^j}{2} - x_j \Bigr]}{\left(e_i^j - 1\right)\, T_i^j} \right\} \tag{24.2b}$$

$$f_j = \sum_{i=1}^{n_j} M_i^j(x_j) \qquad j = 1, 2, \ldots, m \tag{24.3}$$

$$y_{cj}(t) = V_m\, F_j\, \left(1 + \left|x_j(t)\right|\right)^{A_j} \left[1 - \frac{f_1(t)}{n_j(t)} \operatorname{sgn}(x_j(t))\right] \tag{24.4}$$

$$v(t+1) = \frac{\sum\limits_{j=1}^{m} \left(f_1(t)\, y_{cj}(t) T_j \, \frac{e_{oj} + 1}{2\, e_{0j}}\right)}{\sum\limits_{j=1}^{m} \left(f_j(t)\, T_j \, \frac{e_{0j} + 1}{2\, e_{0j}}\right)} \tag{24.5}$$

where $M_i^j(x_j)$ is the ith membership function for the jth input variable, x_j *is normalized* jth *input variable,* e_i^j(I = 1, n) and $\mathrm{B_j}$ (j = 1, m) adaptation and distribution parameters, respectively for the j-th input fuzzy set, $\mathrm{V_m}$ is the maximal value of control variable v, both $\mathrm{F_j}$ and $\mathrm{a_j}$ are free parameters, $\mathrm{y_{cj}(t)}$, $\mathrm{T_j}$ and $\mathrm{e_{oj}}$ is the centre, base and adaptation parameter of the j-th output fuzzy set, respectively. The aim is to detect changes of the current process behavior and to generate analytical symptoms. The diagnosis task is accomplished by fuzzy evidential approximate reasoning scheme to handle different kinds of uncertainty that are inherently present in many real world processes, and to make decision under conflicting data or knowledge. The solved diagnostic system serves several purposes. It identifies the optimal decision boundaries between the different faulty states with as many details as possible or needed even in the presence of noise and uncertainties.

If an appropriate structure is identified, the learning task can be accomplished by any suitable training algorithm such as the classical backpropagation algorithm [2,4,9].

24.4 Parameter Optimization Procedure

Procedure of parameter optimization is realized by following way. The identification procedure of obtained parameters leads to optimization and tuning procedures. We solve the forward procedure. The functional models FM_{ij}, $I = 1, \ldots, m$; $j = 1, \ldots, l$ are identified by solving a least square problem. The backward procedures fix the functional models. The parameters of the membership functions $p_{ik}, q_{ik}; i = 1, \ldots, m$; $k = 1, \ldots, n$ are updated by an effective non-linear gradient descent optimization technique. It requires the computation of derivates of the objective function to be minimized with respect to the parameters p_{ik}, q_{ik}. We apply the optimization algorithm with variable learning rates process using.

We have a set $S = (x^p,\ s^p)^N{}_{p=1}$, such that $x^p \in X \subset R^r$; $s^p \in Y \subset R^l$, the objective is to find subsystem $y_j(x^p)$ in the form:

$$y_j(x) = \frac{\sum\limits_{i=1}^{m} FM_{ij} \prod\limits_{k=1}^{n} e^{-\frac{(x_k - p_{ik})^2}{q_{ik}^2}}}{\sum\limits_{i=1}^{m} \prod\limits_{k=1}^{n} e^{-\frac{(x_k - p_{ik})^2}{q_{ik}^2}}} \tag{24.6}$$

We minimize the function of the mean squared error:

$$E_r = \frac{1}{2} \sum_{j=1}^{l} \left(y_j - s_j^p \right)^2 \tag{24.7}$$

where $x^p \in S$.

We solve the mean p_{ik}, variance q_{ik} (i.e. ellipsoidal functions) and the adjustment of the FM_{ij}. We also assume that $p_{ik} \in X_i; q_{ik} > 0$; $FM_{ij} \in Y_j$. We solve a complex non-linear multi-input and multi-output relationship with $x = (x_1, x_2, \ldots, x_n)^T \in X \subset R^n$. Parameter x is the vector of input variables. We have also $y \in Y \subset R^l$. Parameter y is the vector of output variables. Output y and E_r depend on p_{ik}, q_{ik} only through Eq. (24.6). We have the following equations with substitution U:

$$y_j(x) = \sum_{i=1}^{m} FM_{ij} \frac{\prod_{k=1}^{n} e^{-\frac{(x_k - p_{ik})^2}{q_{ik}^2}}}{\sum_{i=1}^{m} \prod_{k=1}^{n} e^{-\frac{(x_k - p_{ik})^2}{q_{ik}^2}}}; \qquad U = \prod_{k=1}^{n} e^{-\frac{(x_k - p_{ik})^2}{q_{ik}^2}} \tag{24.8}$$

$$\begin{aligned}
\frac{\partial E_r}{\partial p_{ik}} &= \frac{\partial E_r}{\partial U} \cdot \frac{\partial U}{\partial p_{ik}} = \sum_{j=1}^{l} \left(\frac{\partial E_r}{\partial y_j} \cdot \frac{\partial y_j}{\partial U} \right) \cdot \frac{\partial U}{\partial p_{ik}} \\
&= \left[\sum_{j=1}^{l} \frac{(y_j - s_j) \cdot (FM_{ij} - y_j)}{\sum_{i=1}^{m} U} \right] \cdot \left[2 \cdot U \cdot \frac{(x_k - p_{ik})}{q_{ik}^2} \right] \\
\frac{\partial E_r}{\partial q_{ik}} &= \frac{\partial E_r}{\partial U} \cdot \frac{\partial U}{\partial q_{ik}} = \sum_{j=1}^{l} \left(\frac{\partial E_r}{\partial y_j} \cdot \frac{\partial y_j}{\partial U} \right) \cdot \frac{\partial U}{\partial q_{ik}} \\
&= \left[\sum_{j=1}^{l} \frac{(y_j - s_j) \cdot (FM_{ij} - y_j)}{\sum_{i=1}^{m} U} \right] \cdot \left[2 \cdot U \cdot \frac{(x_k - p_{ik})^2}{q_{ik}^3} \right]
\end{aligned} \tag{24.9}$$

Above-mentioned optimization procedure of parameters that is obtained by the identification procedure uses an effective training methodology. The used learning process is performed in two stages. A clustering algorithm, first of all, finds a course model that roughly approximates the underlying input-output relationship. Then the procedure of parameter optimization is performed for a better tuning of the initial structure. If an appropriate structure is identified, the learning task can be accomplished by any suitable training algorithm such as the classical backpropagation algorithm [4, 5, 7, 8].

24.5 Experiments and Results

The results of the simulation are illustrated on Figs. 24.3 and 24.4 that display position tracking error. It can be seen that after the initial oscillation that is the result of system dynamics (inertial forces) swing of the local decays quickly as approaches

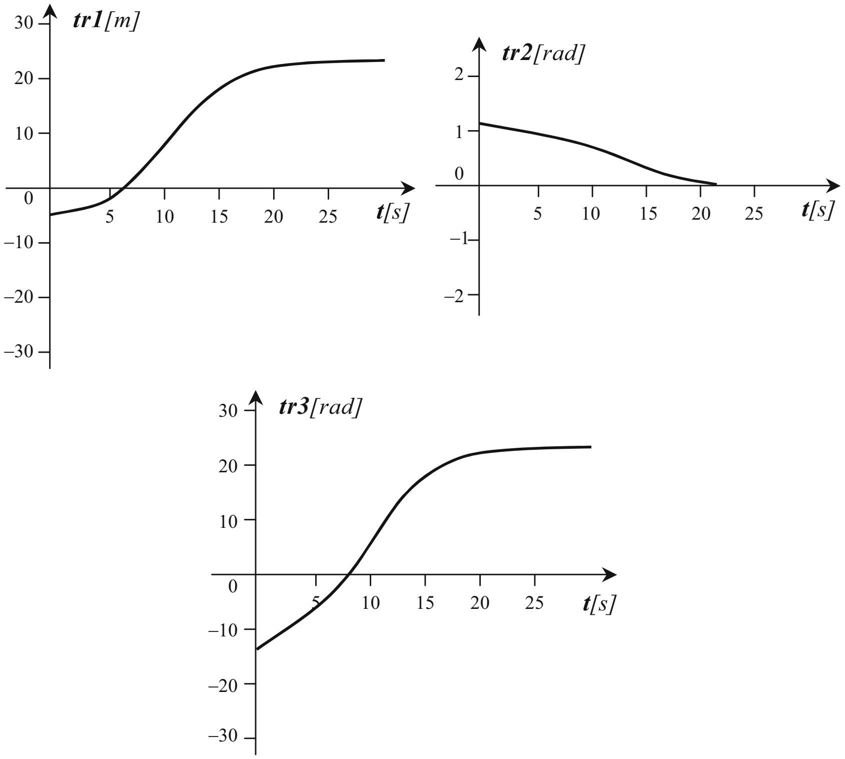


Fig. 24.3 Examples of nominal trajectories tr1, tr2, tr3 (t[s] = discrete time)

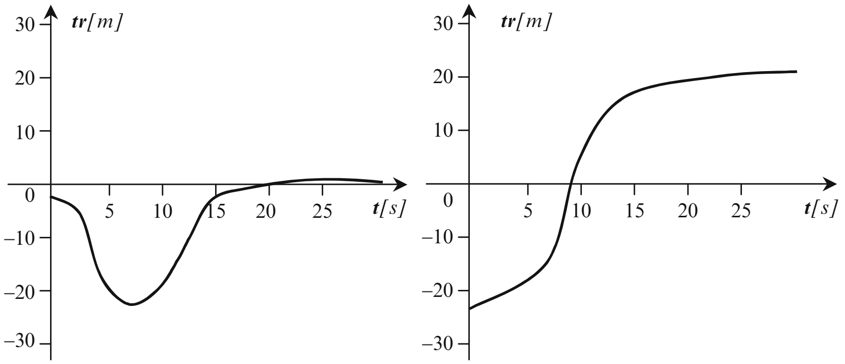


Fig. 24.4 Examples of achieved trajectories tr (t[s] = discrete time)

to the final state, resulting in almost zero oscillations at the final time of the transfer. Taken into consideration that the control algorithm is not too demanding in terms of processing time needed, proposed control scheme could be implemented on a real equipment.

24.6 Cutting Forces Optimization via Genetic Algorithm

Stable and unstable cutting experiments have been designed. As the magnitude and direction of the cutting force significantly influence the machining accuracy therefore their precise knowledge is required in precision finishing. There are process requirements concerning to the accuracy and quality that are continuously increasing [4, 10–12]. A genetic algorithm approach was applied to the simulation model to determine the parameter values of process that would result the simulated cutting forces in ball-milling.

We realize influence search of cutting speed on total average cutting force and its components at machining the various steel materials and then we realize the process optimization within genetic algorithm environment. The aim is to determine and optimize first of all a relation for all three components of the cutting force, which describes the reality well and can be applied relatively easily in the practice. Total average cutting force is calculated by following way:

$$F_c = \sqrt{F_x^2 + F_y^2 + F_z^2} \tag{24.10}$$

The system for cutting process simulation contains the experimental model and the simulation model based on genetic algorithms.

Experiments were realized on the different materials of workpieces and cutting speeds. There is an example of the simulation model to determine the process parameter values via genetic algorithm approach. The simulation model would result the simulated cutting forces in ball-milling process. Results of realized experiments and the simulation processes are illustrated in Fig. 24.5 and Table 24.1. We compare

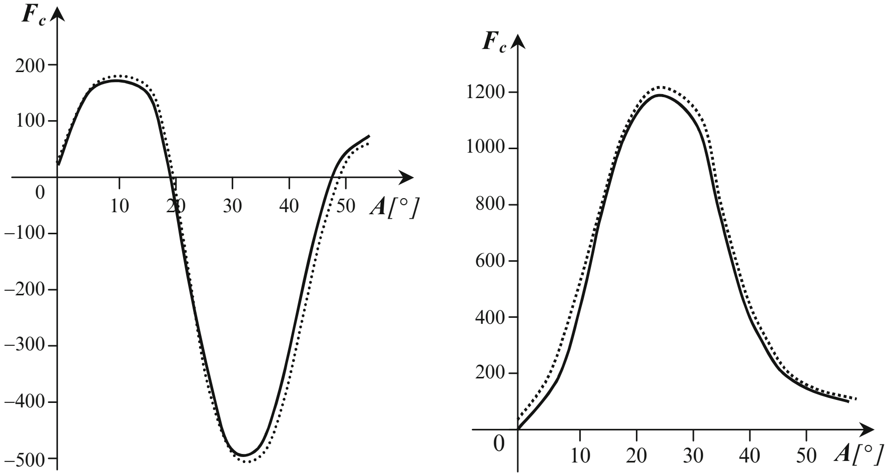


Fig. 24.5 Simulation and measurement of cutting forces

Table 24.1 Relevant parameters

Number of generations	186
Probability of the reproduction	0.79
Probability of the selection	0.67
Probability of the mutation	0.0012
Elitism function	19
Period of regeneration	10

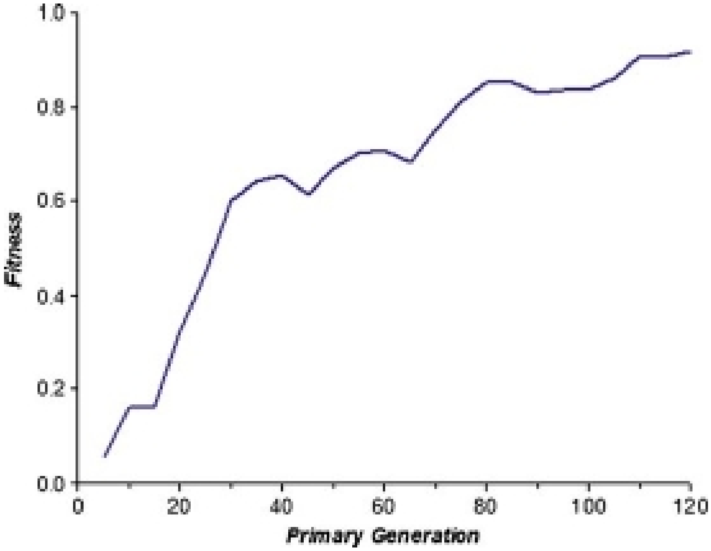


Fig. 24.6 Fitness evaluation course for used methodology approach

the simulation and experimental results. Parameter F_c represents cutting force, parameter A represents angle of the cutter rotation. The dashed line represents the simulated cutting forces. The continuous line represents the experimental cutting forces. The model of cutting forces simulation is confirmed by experimental results.

Fitness evaluation course for used methodology for solved ball-milling process is shown in Fig. 24.6.

24.7 Conclusion

The system for cutting process simulation contains the experimental model and the simulation model based on genetic algorithms. The operation of the cutting forces simulation model can be confirmed by experimental results (see Fig. 24.5).

Cutting forces are important parameters to predict machining performances of any machining operation. The force components can be obtained and displayed simultaneously on the screen for analyzing force changes via expert system tool.

In the A/D board, the analogue signal will be transformed into a digital signal form. A connecting plan blocks and channel A/D interface board are realized within interface hardware module.

The predictive modeling of machining operations requires detailed prediction of the boundary conditions for stable, safety and reliable machining. Cutting forces are relevant factors to predict machining performances of any machining operation. The classification of solved signal features for process condition monitoring has been carried out using fuzzy parameters optimization processing via genetic algorithm optimization methodology. We realize case adaptation by adjusting the fuzzy sets parameters. Fuzzy parameters within optimization procedure could be multiobjective. This genetic algorithm approach is suitable for fuzzy implementation control and for solution of large-scale problems.

Future research will be focused first of all on improving the runtime performance of solved implementation, including other genetic operators in the architecture and investigating the results of further test problems in more detail. There are will be investigated self-learning approaches of diagnostic rules through more advanced genetic and evolutionary algorithms and modified chaos theory principles. Nowadays, some achieved results seem to be very interesting. Future research will be also concerned on two populations approach only that occasionally communicate each other and on asynchronous version of co-evolution model that allows each population to have a slower and more stable evolution pace but also is suited for a parallel implementation in which each population is evolved on a separate processor. Such parallel implementation is important for the efficient processing of a large number of well logs simultaneously.

The research reported on this paper is supported by following scientific project of Slovak Grant Agency VEGA (Ministry of Education):

Intelligent Approach to Automated Diagnostic Problem Solving of Machinery Equipment (Grant Number: VEGA 1/4133/07).

References

1. Herrera-Viedma, E.: Modelling the retrieval process of an information retrieval system using an ordinal linguistic approach. Am. Soci. Inf. Sc. **6**, 460–475 (2001)
2. Gallova, S.: Fault diagnosis of manufacturing processes via genetic algorithm approach. IAENG Eng. Lett. **15**(2), 349–355 (2007)
3. Rochio, I.J.: Relevance Feedback of Information Retrieval, The Smart System Experiments in Automatic Document of Processing, pp. 313–323. Prentice-Hall, New York (1971)
4. Gallova, S.: A maximum entropy inference within uncertain information reasoning. Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 1803–1810, Les Cordeliers, Paris (2006)
5. Ballé, P.: Fuzzy model-based parity equations for fault isolation. Con. Eng. Prac. **7**(2), 261–270 (1999)
6. Brini, A.: Introduction de la Gradulaite dans le Jugement Utilisateur, Dea Report, Toulouse, France (2002)
7. Pasi, G.: A logical formulation of the boolean model and weighted boolean models. Lumis'99, University College London, England (1999)

8. Zadeh, L.: The Concept of Linguistic Variable and It's Application to Approximate Decision Making. Moscow, Mir (1976)
9. Goldberg, D.E.: Genetic Algorithms in Search, Optimization And Machine Learning. Addison-Wesley, Reading, MA (1989)
10. Lu, M., Dong, F., Fotouhi, F.: The semantic web, opportunities and challenges for next generation web applications. Inform. Res. **7**(4) (2002)
11. Kruschwitz, U.: An adaptable search system for collections of partially structured documents. IEEE Intell. Syst. **18**:44–52 (2003)
12. Novakovic, B.: Fuzzy logic control synthesis without any rule base. IEEE Trans. Syst. Man Cyber **29**(3):459–466 (1999)

Chapter 25
Encoding Data to Use with a Sparse Distributed Memory

Mateus Mendes, Manuel M. Crisóstomo, and A. Paulo Coimbra

Abstract A Sparse Distributed Memory (SDM) is a kind of associative memory suitable to work with high-dimensional vectors of random data. This memory model exhibits the characteristics of a large boolean space, which are to a great extent those of the human long-term memory. Hence, this model is attractive for Robotics and Artificial Intelligence, since it can possibly grant artificial machines those same characteristics. However, the original SDM model is appropriate to work with random data. Sensorial data is not always random: most of the times it is based on the Natural Binary Code and tends to cluster around some specific points. This means that the SDM performs poorer than expected. As part of an ongoing project, in which the goal is to navigate a robot using a SDM to store and retrieve sequences of images and associated path information, different methods of encoding the data were tested. Some methods perform better than others, and one method is presented that can offer the best performance and still maintain the characteristics of the original model.

Keywords Associative memory · sparse distributed memory · SDM · data encoding

25.1 Introduction

Kanerva [5] proposes the concept of the Sparse Distributed Memory (SDM). Kanerva figured out that such a memory model, based on the use of high dimensional binary vectors (2^n, for n of the order of 1,000), can exhibit some properties

M. Mendes (✉)
ESTGOH, Polytechnic Institute of Coimbra, R. General Santos Costa,
3400-124 Oliveira do Hospital, Portugal
and
Institute of Systems and Robotics, Pólo II, University of Coimbra, 3000 Coimbra, Portugal
e-mail: mmendes@estgoh.ipc.pt

M.M. Crisóstomo and A.P. Coimbra
Institute of Systems and Robotics, Pólo II, University of Coimbra, 3000 Coimbra, Portugal
e-mail: mcris@isr.uc.pt; acoimbra@deec.uc.pt

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_25,

similar to those of the human cerebellum. Phenomena such as “knowing that one knows”, dealing with incomplete or corrupt data and storing events in sequence and reliving them later, can be mimiced in a natural way.

In the present work, a SDM is used as the basis to navigate a robot based on sequences of views. During a learning stage the robot stores selected images it grabs from the paths, and during autonomous runs it manages to follow the same paths by correcting view matching errors which may occur, as described by [6, 8]. Kanerva proposes that the SDM must be ideal to store sequences of binary vectors, and [1] has extensively described this possibility.

Kanerva demonstrates that the characteristics of the model hold for random binary vectors. However, in many circumstances data are not random, but biased towards given points. Rao and Fuentes [10] already mention this problem, although not a solution. In the case of images, for instance, completely black or white images are not common. Authors such as [3, 11, 12] minimise this problem by adjusting the memory’s structure, so that it has more memory locations around points where they are needed, or use different addressing methods. But this only solves part of the problem, and in some cases may even fade some properties of the model.

In this work, different methods of encoding the data and computing distances between memory items were studied. The performance of the memory was assessed for each of these methods, and the results are shown. Parts of this work were also published in [9], and a cross-modal comparison of the encoding problem in robot navigation and manipulation is described in [4].

Sections 25.2 and 25.3 briefly describe the SDM and the experimental platform used. Section 25.4 explains the encoding problem in more detail and presents two widely used models of the SDM. Section 25.5 explains two novel approaches, and Section 25.6 presents and discusses the tests performed and the results obtained. Finally, in Section 25.7 some conclusions are drawn and some possible future work is pointed out.

25.2 Sparse Distributed Memories

One possible implementation of a SDM is as shown in Fig. 25.1. It comprises two main arrays: one stores the locations’ addresses (left), the other contains the data (right). In the auto-associative version of the model, as used here, the same vector can be used simultaneously as address and data, so that only one array is necessary.

Kanerva proposes that there are much less addresses than the addressable space. The actually existing locations are called “hard locations”. This is both a practical constraint and a requirement of the theory. On one hand, it’s not feasible to work with, e.g., 2^{1000} locations, using current common technology. On the other hand, the properties of the memory arise from its sparseness.

In a traditional computer memory one input address will only activate one memory location, which can then be read or written. In the SDM, addressing is done based on the use of an access circle: all the hard locations within a given radius

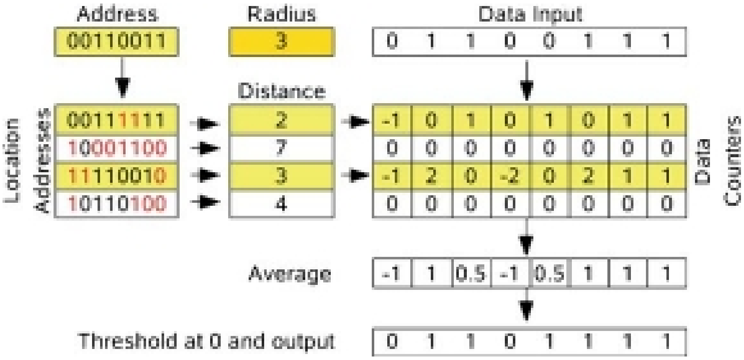


Fig. 25.1 Example of one SDM model

are activated for reading or for writing. Kanerva proposes the Hamming distance (HD) to compute the set of active locations. The HD is actually the number of bits in which two binary numbers differ: $d_h(x, y) = \sum_i(|x_i - y_i|)$. In the example (Fig. 25.1), the radius is set to 3 bits, and the input address is `00110011`. Therefore, the first and third hard locations are activated, for they differ from the input address, respectively, 2 and 3 bits.

The "hard locations" don't actually store the data as it is input to the memory: they are composed of bit counters. During a write operation, the bit counters of the selected hard locations are incremented to store ones and decremented to store zeros. During a read operation, the active bit counters are summed columnwise and averaged. If the average of the sum for a given bit is above a set threshold, then it will be one, otherwise it will be zero.

25.3 Experimental Setup

Some experiments of using a SDM to navigate a robot were carried out. The robot was manually taught a path and expected to follow it later autonomously. The robot is a small device with tank-style treads and differential drive, as shown in Fig. 25.2 and described in [7]. It contains a camera on the front and a radio communications module that permits control from a laptop computer.

Navigation based on a view sequence is based on [6]'s proposal. During the learning stage the robot captures views of the surrounding environment, as well as some additional data, such as odometric information. During the autonomous run, the robot captures images of the environment and uses them to localise itself and follow known paths, based on its previously acquired knowledge [8].

The SDM is used to store those sequences of images and related navigation data. Input and output vectors consist of arrays of bytes, whose composition is described by the equation:

$$x_i = < im_i, seq\_id, i, timestamp, motion > \qquad (25.1)$$

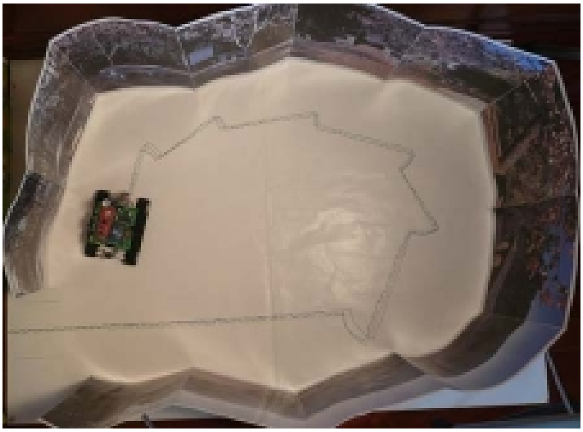


Fig. 25.2 Robot and experimental platform

im$_i$ is image i. *seq_id* is an auto-incremented 4-byte integer, unique for each sequence. It is used to identify which sequence (path) the vector belongs to. i is an auto-incremented 4-byte integer, unique for every vector in the sequence. It is used to quickly identify every image in the sequence. *timestamp* is a 4-byte integer, storing Unix timestamp. It is read from the operating system, but not being used so far for navigation purposes. *motion* is a single byte, identifying the type of movement the robot performed just before capturing *im*$_i$.

PGM images of resolution 80 × 64 are used. Since every pixel is stored as an 8-bit integer, the image alone needs $80 \times 64 = 5{,}120$ bytes $= 40{,}960$ bits. The overhead information comprises 13 additional bytes, meaning the input vector contains 41,064 bits.

25.4 Practical Problems

The original SDM model, though theoretically sound and attractive, has some faults. One fault is that of placing the hard locations in the address space. Kanerva proposes that they are placed at random when the memory is created, but many authors state that's not the most appropriate option. Rogers [12], e.g., evolves the best locations using genetic algorithms. Hely et al. [3] propose that locations must be created where there is more data to store. Ratitch and Precup [11] propose the Randomised Reallocation algorithm, which is essentially based on the same idea: start with an empty memory and allocate new hard locations when there's a new datum which cannot be stored in enough existing locations. The new locations are allocated *randomly* in the neighbourhood of the new datum address. This is the approach used here.

Another weakness of the original SDM model is that of using bit counters. This results in a low storage rate, which is about 0.1 bits per bit of traditional computer

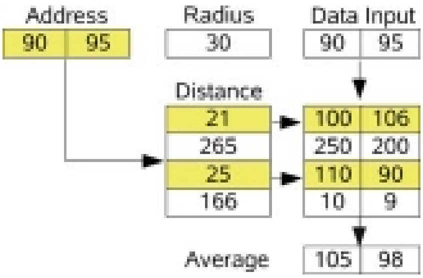


Fig. 25.3 Arithmetic SDM, which works with byte integers, instead of bit counters

memory, huge consumption of processing power and complexity of implementation. Furber et al. [2] claim their results show that the memory's performance is not significantly affected if a single bit is used to store one bit, instead of a bit counter, under normal circumstances. For real time operation of software simulations, this simplification greatly reduces the need for processing power and memory.

In this work, two models were tested: the "bitwise model" (BW), which is similar to the original model but has the bit counters replaced by single bits; and the "arithmetic model" (AR). In the AR model, the bits are grouped as integers, as shown in Fig. 25.3. Addressing is done using an arithmetic distance, instead of the Hamming distance. Learning is achieved using a kind of reinforcement learning:

$$h_t^k = h_{t-1}^k + \alpha \cdot (x^k - h_{t-1}^k), \quad \alpha \in \mathbb{R} \wedge 0 \leq \alpha \leq 1 \tag{25.2}$$

h_t^k is the k^{th} number of the hard location, at time t. x^k is the corresponding number in the input vector x and α the learning rate.

25.5 Binary Codes and Distances

In Natural Binary Code (NBC) the value of each bit depends on its position. `01` is different from `10`. This characteristic means that the HD is not proportional to the binary difference of two numbers expressed in NBC.

Table 25.1 shows the HDs between all the 3-bit binary numbers. As it shows, this distance is not proportional to the arithmetic distance (AD). The HD sometimes even decreases when the AD increases. In total, there are nine undesirable situations in the table, where the HD decreases while it should increase or, at least, maintain its previous value.

The problem with this characteristic is that the PGM images are encoded using the Natural Binary Code, which takes advantage of the position of the bits to represent different values. The performance of the SDM, therefore, might be affected because of these different criteria being used to encode the information and to process it inside the memory.

These characteristics of the NBC and the HD may be neglectable when operating with random data, but in the specific problem of storing and retrieving graylevel

Table 25.1 Hamming distances for 3-bit numbers

	000	001	010	011	100	101	110	111
000	0	1	1	2	1	2	2	3
001		0	2	1	2	1	3	2
010			0	1	2	3	1	2
011				0	3	2	2	1
100					0	1	1	2
101						0	2	1
110							0	1
111								0

Table 25.2 Example distances using different metrics

Pixel value		Distance		
im_i	im_j	Arithmetic	Hamming	Vanhala
01111111	10000000	1	8	1
11111111	00000000	127	8	1

images, they may pose serious problems. For instance, a given pixel P that has graylevel 127 (01111111) is considered 8 bits far from a very close pixel that has graylevel 128 (10000000), but only 1 bit far from distant 255 (11111111).

Vanhala et al. [13] use an approach that consists in using only the most significant bit of each byte. This still relies on the use of the NBC and is more robust to noise. However, this approach is a very rough filter, which maps the domain [0, 255] onto a smaller domain [0, 1], where only binary images can be represented. While effective reducing noise, which Vanhala reports to be the primary goal, this mapping is not the wisest solution to the original problem being discussed. To effectively store graylevel images in the SDM, using the Hamming distance, a better binary code is needed. For example, one in which the number of ones is proportional to the graylevel value of the pixel. In this aspect, Vanhala's approach should not perform well. The distance from a black pixel (00000000) to a white pixel (11111111), for instance, is the same as between two mid-range pixels which are almost the same, as in the example described above. Table 25.2 summarises some examples of distance calculations, depending on the metrics used.

25.5.1 Sorting the Bytes

An alternative approach is simply to sort the bytes in a more convenient way, so that the HD becomes proportional to the AD – or, at least, does not exhibit so many undesirable transitions.

Table 25.3 Hamming distances for 3-bit sorted numbers

	000	001	010	100	101	111	011	110
000	0	1	1	1	2	3	2	2
001		0	2	2	1	2	1	3
010			0	2	3	2	1	1
100				0	1	2	3	1
101					0	1	2	2
111						0	1	1
011							0	2
110								0

This sorting can be accomplished by trying different permutations of the numbers and computing the matrix of Hamming distances. Table 25.3 shows a different sorting, better than the NBC shown in Table 25.1. This code shows only seven undesirable transitions, while the NBC contains nine. Therefore, it should perform better with the SDM, but not outstanding. There are several sortings with similar performance (i.e., the same number of undesirable transitions). The one shown is the first that was found. If the data are uniformely distributed, then all the sortings must exhibit a similar performance. In this case, the images are equalised. Hence, the distribution of all the brightness values is such that all the values are approximately equally probable. This means that it is irrelevant which sorting is chosen, among those with the same number of undesirable transitions.

After 57,253,888 permutations the software used found one sorting with just 18,050 undesirable transitions, for all the 8-bits binary numbers. It is not clear if there is a better sorting, but up to 681,400,000 permutations no better sorting was spotted.

25.5.2 Using a Sum-Code

As written previously, using 256 graylevels it's not possible to find a suitable binary code with minimum undesirable transitions, so that one can take advantage of the representativity of the NBC and the properties of the SDM. The only way to avoid undesirable transitions at all is to reduce the number of different graylevels to the number of bits $+1$ and use a kind of sum-code. Therefore, using 4 bits it's only possible to use five different graylevels, as shown in Table 25.4. Using 8 bits, nine graylevels are available and so on. This is the only way to work with a HD proportional to the AD.

The disadvantage of this approach is, however, obvious: either the quality of the image is much poorer, or the dimension of the stored vectors has to be extended to accommodate additional bits.

Table 25.4 Sum-code to represent five graylevels

Decimal	Sum-code
0	`0000`
1	`0001`
2	`0011`
3	`0111`
4	`1111`

25.6 Tests and Results

Different tests were performed in order to assess the behaviour of the system using each of the approaches described in the previous sections. The results were obtained using a sequence of 55 images. The images were equalised, and the memory was loaded with a single copy of each image in the first test, and with five copies of each image in a second test (i.e., distribution of the data was enforced).

25.6.1 Results

Tables 25.5 and 25.6 show the average of 30 autonomous runs. The tests were performed using the arithmetic distance; using the NBC and the Hamming distance (8 bits, 256 graylevels, represented using the natural binary code); using the Hamming distance and a partially optimised sorting of the bytes, as described in Section 25.5.1; and bitwise modes in which the graylevels were reduced to the number of bits $+1$, as described in Section 25.5.2. Tests were run using from 1 to 32 bits, which means from 2 to 33 graylevels, in order to experimentally get a better insight on how the number of bits and graylevels might influence the performance of the system. However, for space constraints, the results are only shown for 1, 8, 16 and 32 bits (B02, B09, B17 and B33).

The first column of the tables specifies the operation mode. The second column specifies the number of Momentaneous Localisation Errors (MLE). A MLE occurs when one image is erroneously predicted. This happens when prediction at time t is im_i, and prediction at time $t+1$ is im_{i-n} $(n > 0)$. Since the robot is not expected to move backwards, at time $t+1$ the prediction should be, in the worst case, im_i again. Note that the occurrence of a MLE does not mean that the robot is definitely lost. It just means that one of the predictions was wrong, but the robot was always able to complete the task. Also, the number of MLEs is actually very small compared to the total number of steps necessary to complete the path. During the autonomous run, the step size is 1/16th of the step size during the learning stage. Therefore, about 880 steps are necessary to complete a sequence described by 55 images. This means that even 30 MLEs are just about 3.5% of the total length of the path.

The third column of the tables shows the distance from the input image to the closest image in the SDM (error in similarity). The fourth column is the distance to

Table 25.5 Results, when the memory contained only one copy of each image

Operation mode	M. L. errors	Dist. to closest	Dist. to 2nd closest	Inc_2 (%)	Aver. dist. to all other images	Inc_A (%)	Time (ms)
Arit.	29.0	35,240	56,511	60.36	172,236.53	388.75	7.23
NBC	31.4	7,575	8,195	8.18	9,738.01	28.56	6.90
Sort.	31.2	7,583	8,187	7.97	9,736.95	28.41	7.11
B02	29.6	161	265	65.12	1,038.70	546.22	28.20
B09	29.2	1,115	1,772	58.97	5,438.20	387.89	28.16
B17	26.4	2,234	3,530	58.01	10,802.41	383.56	28.20
B33	26.4	4,427	7,099	60.35	21,553.67	386.85	28.49

Table 25.6 Results, when the memory contained five copies of each image

Operation mode	M. L. errors	Dist. to closest	Dist. to 2nd closest	Inc_2 (%)	Aver. dist. to all other images	Inc_A (%)	Time (ms)
Arit	23.2	35,528	36,467	2.64	172,259.43	384.85	33.37
NBC	26.0	7,547	7,831	3.77	9,740.06	29.06	31.63
Sort	24.2	7,544	7,873	4.37	9,742.55	29.15	32.88
B02	27.4	165	169	2.43	1,038.38	530.34	137.86
B09	23.8	1,101	1,136	3.20	5,430.48	393.41	101.55
B17	25.2	2,212	2,263	2.28	10,802.24	388.26	139.22
B33	26.2	4,346	4,472	2.89	21,557.04	395.98	109.33

the second closest image. The fifth column shows, in percentage, the increase from the closest prediction to the second closest prediction. The sixth and seventh columns show the average distance to all the other images – this is a measure of how *successful* the memory is in separating the desired datum from the pool of information in the SDM. Finally, the last column shows the average processing time for each method. Processing time is only the memory prediction time, it does not include the image capture and transmission times. The clock is started as soon as the command is issued to the SDM and stopped as soon as the prediction result is returned.

25.6.2 Analysis of the Results

It can be confirmed that the bitwise mode using the NBC and the HD seems to be remarkably worse than the other methods, which seem to show similar results. Sorting the bytes results in a small, but not significant, improvement. Another interesting point is that the number of graylevels seems to have little impact on the selectivity of the image, for images of this size and resolution. However, higher number of bits tend to reduce the number of momentaneous localisation errors.

As for distribution of the data, the results show that distribution has a significant impact on the number of MLEs. In the arithmetic mode, for instance, distribution cut the number of MLEs by about 20%. In the other operation modes the impact is not as significant, but the MLEs in general are smaller with distribution of the data. It should be noted that Inc_2, that is, how larger the distance to the second best prediction is compared to the distance to the best prediction, is very small with distribution. This is a natural consequence of distribution of the data. Since at least five copies of each image are enforced during writing, it is expected that during prediction at least five images are found within the activation radius. The average distance to all the hard locations, however, is similar either with or without distribution.

The processing time exhibits a great variation, for the tests were run on a computer using Linux (OpenSuSE 10.2), a *best effort* operating system. Even with the number of processes running down to the minimum, there were very disparate processing times. For better precision and real time operation, a real time operating system would be recommended.

The average processing times for the arithmetic and bitwise modes (NBC and optimised code) are about 7 ms for the complete cycle to fetch the closest matching image. Using the NBC with the HD, the time is a little shorter, and using a different sorting of the bytes the time increased a little. This was expectable, since the only variation in this method was implemented using an indexed table, where each position held the sorted byte. Therefore, to compute the similarity between two pixels, two accesses had to be done to the indexed table, which considerably increases the total memory access time. A more efficient approach would be to make the conversion as soon as the images were grabbed from the camera. That is undesirable in this case, though, since other approaches are also being tested, and some of those approaches need the images without conversion.

As for the sum-code mode, using different graylevels, the processing times are all similar and about four times larger than the time of processing one image using the arithmetic mode. The reason for this is that, again, an indexed table is used to address the binary code used. And in this case there's the additional workload of processing the conversion into the desired number of gray values. In a production system, obviously, the conversion would only need to be done once, just as the images were grabbed from the camera.

25.7 Conclusions and Future Work

A Sparse Distributed Memory was tested in different operation modes. These included different methods of encoding the information, and of measuring the similarity distance between two memory items.

In the original SDM model Kanerva proposes that the Hamming distance be used to compute the similarity between two memory items. However, this method exhibits a poor performance if the data are not random. The NBC with the Hamming distance shows the worst performance in the tests that were run. By sorting some bytes the performance is slightly improved. If the bits are grouped as bytes and an arithmetic distance is used, the memory shows an excellent performance, but this can fade some characteristics of the original model, which is based on the properties of a binary space. If the number of graylevels is reduced and a sum-code is used, the performance is close to that of the arithmetic mode and the characteristics of the memory must still hold.

Although this work was performed using images as data, the results should still be valid for all non-random data, as is usually the case of robotic sensorial data.

Future work includes the study of the impact of using different encoding methods on the performance of the SDM itself, in order to infer which characteristics will still hold or fade.

References

1. Bose, J.: A scalable sparse distributed neural memory model. Master's thesis, University of Manchester, Faculty of Science and Engineering, Manchester, UK (2003)
2. Furber, S.B., Bainbridge, J., Cumpstey, J.M., Temple, S.: Sparse distributed memory using *n*-of-*m* codes. Neural Networks **17**(10), 1437–1451 (2004)
3. Hely, T.A., Willshaw, D.J., Hayes, G.M.: A new approach to kanerva's sparse distributed memories. IEEE Trans. Neural Networks (1999)
4. Jockel, S., Mendes, M., Zhang, J., Coimbra, A.P., Crisóstomo, M.M.: Robot navigation and manipulation based on a predictive associative memory. In: Proceedings of the 2009 IEEE 8th International Conference on Development and Learning (ICDL), Shanghai, China (2009)
5. Kanerva, P.: Sparse Distributed Memory. MIT Press, Cambridge (1988)
6. Matsumoto, Y., Inaba, M., Inoue, H.: View-based approach to robot navigation. In: Proceedings of 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (2000)
7. Mendes, M., Coimbra, A.P., Crisóstomo, M.: AI and memory: Studies towards equipping a robot with a sparse distributed memory. In: Proceedings of the IEEE International Conference on Robotics and Biomimetics, Sanya, China (2007)
8. Mendes, M., Crisóstomo, M., Coimbra, A.P.: Robot navigation using a sparse distributed memory. In: Proceedings of the 2008 IEEE International Conference on Robotics and Automation, Pasadena, CA (2008)
9. Mendes, M., Crisóstomo, M., Coimbra, A.P.: Assessing a sparse distributed memory using different encoding methods. In: Proceedings of the 2009 International Conference of Computational Intelligence and Intelligent Systems, pp. 37–42. London, UK (2009)
10. Rao, Rajesh P.N., Fuentes, O.: Hierarchical learning of navigational behaviors in an autonomous robot using a predictive sparse distributed memory. Mach. Learn. **31**(1–3), 87–113 (1998)
11. Ratitch, B., Precup, D.: Sparse distributed memories for on-line value-based reinforcement learning. In: ECML (2004)
12. Rogers, D.: Predicting weather using a genetic memory: a combination of kanerva's sparse distributed memory with holland's genetic algorithms. In: NIPS (1989)
13. Vanhala, J., Saarinen, J., Kaski, K.: Sparse distributed memory for multivalued patterns. In: IEEE Int. Conf. Neural Networks (1993)

Chapter 26
A Proposal for Integrating Formal Logic and Artificial Neural Systems: A Practical Exploration

Gareth Howells and Konstantinos Sirlantzis

Abstract The assimilation of formal logic into the domain of Software Engineering offers the possibility of enormous benefits in terms of software reliability and verifiability. The current paper investigates the advantages which may be gained by the software development process with the introduction of Artificial Neural Network technology into a formal software development system. The paper introduces a framework by which a programmer may define a system possessing the abstract structure of a traditional neural network but whose internal structures are taken from the formal mathematical domain of Constructive Type Theory. Two such examples are presented addressing a problem domain which has previously proved difficult to model.

Keywords Formal logic · constructive type theory · artificial neural networks · functional programming

26.1 Introduction

The application of Formal Logic in the form of Constructive Mathematics such as *Constructive Type Theory* [1, 2] to the design of major software systems can give rise to significant improvements in the reliability and dependability of the software produced. The purpose of this paper is to explore the relationship between Constructive Type Theory and Artificial Neural Networks by illustrating how two example networks may be implemented in a system derived from Type Theory. The design framework proposed offers the following advantages:

- The construction of formally derived and verifiable software system related to the domain of artificial neural networks without requiring detailed mathematical analysis by the programmer.

G. Howells (✉) and K. Sirlantzis
School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NT, UK
e-mail: W.G.J.Howells@kent.ac.uk; K.Sirlantzis@kent.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_26,

- The derived system may adapt itself to take into account differing future performance requirements within the limitations of the network architecture simulated.

The paper will introduce the structure of the framework and explore its operation by introducing two example refinement algorithms. For convenience, these will be well established algorithms taken from the field of Artificial Neural Networks. It should be emphasised however that the framework is theoretically capable of representing any arbitrary algorithm and is not restricted to these examples although a demonstration of this is outside the scope of this paper.

26.2 Constructive Type Theory

Constructive Type Theory is a formal logic [1–4] based on *Constructive Mathematics* in which proofs must be based on a demonstration of how to construct an example of the theorem or proposition being asserted. In other words, proofs by contradiction are not allowed. So, in Constructive Type Theory *proofs* can be thought of as *algorithms* to create an example of the *proposition* in hand. Furthermore, the proposition forms a *datatype definition* or at a higher level a *formal specification* for the algorithm itself.

In Constructive Type Theory, each logical proposition is accompanied by its proof object forming a pair of values termed a *Judgement*. Judgements are usually written in the form $\boldsymbol{p{:}P}$ where the proof object $\boldsymbol{p}$ bears witness to the proposition $\boldsymbol{P}$.

Each logical connective in Constructive Type Theory is associated with four rules.

- Formation (syntax)
- Introduction
- Elimination
- Computation (simplification)

A simple example is the AND ($\wedge$) rules.

26.3 The Theoretical Framework

$$\frac{A \text{ is a type } B \text{ is a type}}{(\mathrm{A} \wedge \mathrm{B}) \text{ is a type}}(FORMATION) \qquad \frac{a : A \quad b : B}{(a,b) : A \wedge B}(INTRO)$$

$$\frac{p : (A \wedge B)}{fst\ p : A}(ELIM) \qquad \frac{p : (A \wedge B)}{snd\ p : B}(ELIM)$$

$$fst(a,b) \rightarrow a(COMP) \quad snd(a,b) \rightarrow b(COMP)$$

Artificial Neural Networks [5–8] (ANN) present a powerful tool for the analysis of complex systems. However, ANN implementations are not usually amenable to formal analysis and verification of correct performance. This makes them inefficient tools for deriving algorithms meeting a given performance specification. Difficulties in their practical application centre around the derivation of suitable *weights* associated with the neural connections. On the other hand, Constructive Type Theory (CTT) offers the rigorous base of a formal logic and at the same time a formal specification tool which is capable of providing algorithmic solutions in the form of proofs of the propositions asserted.

The structure of the prototype system introduced in this paper retains the abstract structure of a neural network as a directed graph. However, unlike the conventional neural architectures in which the weights are represented, for example, as real numbers, in the prototype they are expressed as *logical implications* $A \to B$, where A and B are datatypes. If $A = \Re$ and $B = \Re$, where $\Re$ represents the set of real numbers, then our prototype may, for example, represent the conventional real valued Multi-layer Perceptron (MLP). Its training algorithm (the generalised delta rule) can be viewed as modifying only the proof portion of the proposition/proof pair.

In fact, in our prototype, both weights and activation functions of the processing elements (nodes) comprising the network are expressed as logical implications. A node with two inputs, X and Y can then be represented as an implication of the type $A \to B \to C$ where the weights associated with these two inputs can be $X \to A$ and $Y \to B$ (see Fig. 26.1). If all the types are numerical then the corresponding proof will represent a mathematical function and the prototype will describe a conventional neural network of some particular form.

Our implementation of the prototype is developed in the *Functional programming language* Haskell [9]. The advantage of using a functional language to realise our neural network prototype is threefold:

1. There exists a correspondence between Type Theoretic entities and functional programming objects simplifying the implementation
2. The expressions in a functional language preserve *referential transparency*

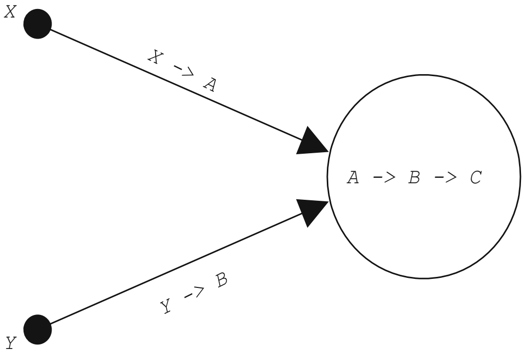


Fig. 26.1 A Schematic representation of a node in our framework, with the weights and the activation function represented as logical implications on the types of the inputs associated with it

3. Functional languages can be thought of as formal specifications and moreover, because a working prototype is produced concurrently, they can be thought of *as executable formal specifications*

In addition to these, our choice of Haskell for the implementation, which produces a compiled executable based on the C programming language, offered us a rather fast final program in comparison to the rather slowly running programs functional languages have been accused of producing.

26.4 The Prototype

The purpose is to use and implement the algorithms derived in the Constructive Type Theory framework. The derivation process is then a mechanism for modifying definitions needed by the prototype. Before presenting the definitions (in Haskell code, which in most cases is self-explanatory) a brief note on notation:

:: indicates a type signature - that is its left operand has the type of its right operand -,
[A] denotes a list of objects of type A and
-> denotes a logical implication (i.e. a function type).

Finally parentheses are used to indicate either grouping of results or clarify application of functions to a subset of their argument set (partial application of functions).

The first level of the prototype incorporates the definitions for the network itself and its components along with the functions that create them. In Haskell code these are expressed as follows:

```
data NeuralNet = NN {arch:: Net_Arch, fun::Net_Func,
                         layer_nodes:: Layers}
                  | ENN
```

where

```
type Net_Arch = Array Edge Weight
type Weight = Maybe Weight_Function
type Weight_Function = Param -> Output -> Input
type Edge = (Vertex, Vertex)
type Vertex = Int
type Layers = [[Vertex]]
type Net_Func = [Node]
```

The above code defines a neural network to be an object which has an architecture (Net_Arch), a list of nodes (Net_Func), and a list of its vertices representing their arrangement in layers (Layers) (if this is applicable to the particular network). The architecture is represented as a 2-dimensional array of weights (Weight) indexed by the edges of the network.

A Neural Net is modelled as a record type with named fields so that each one of them can be recalled by name and possibly updated. A node is defined as an object consisting of inputs (`Input`), output (`Output`), activation function (a logical implication from `Input` to `Output`), parameters associated with each one of its inputs, and error (`Delta`) to use during the training. The corresponding code reads:

```
data Node = Nd {inp:: Inputs, outp:: Output,
                 par:: Params,
                 activation_function:: Inputs -> Output,
                 error:: Delta}
type Input = [Data]
type Inputs = [Input]
type Output = [Data]
type Param = [Data]
type Params = [Param]
type Delta = [Data]
```

The second level of the abstraction implements functions which handle and update the network components. This level includes a general description of the learning process from the examples provided incorporating functions which define how to:

1. Propagate "`inputs`" into a particular "`neural_net`", given a set of associated "`parameters`", and produce the corresponding outputs.
2. Evaluate the produced outputs with respect to "`targets`", if required by the particular ANN model implemented.
3. Update the parameters of the node given a "`learning_rate`".

Training may now be defined as the composition (denoted by ".") of the three functions ("`propagate`", "`evaluate`" and "`update`"):

```
type Target = [Data]
train:: LParam -> [Input] -> [Target] -> [[Param]]
        -> NeuralNet -> NeuralNet
train learning_rate inputs targets weights neural_net
      = ((update learning_rate).
            (evaluate targets).
            (propagate inputs parameters)) neural_net
```

The definition of the update function, which is used below to define two different ANN models, may be given by:-

```
type Lparam = [Data]
update:: LParam -> NeuralNet -> NeuralNet
update lr nn = nn {fun = set_params nn ps}
```

where "`lr`" (of type "`Lparam`") is the learning rate while "`nn`" indicates the current state of the neural network. Note that by implementing the neural network as a record type with named fields, the update operation is simplified. The right hand side of the equality in the "`update`" function definition in effect means that we replace the networks element called "fun" using a function "set_params" which replaces the old parameter values stored in the network's "nn" nodes with the new ones "ps". These are calculated using the given learning parameter and the possibly the old weights' values. In the following we will see that by appropriately defining "ps"- the list of the new (updated) parameters-we are able to implement a variety of ANN models with our prototype.

The third and final level of the prototype's hierarchy comprises of the functions that form the operators on the various value types (e.g. Binary or Integer values). As soon as appropriate operators are defined at this level for any type of data, the prototype can be recompiled to produce an executable in order to process this kind of data.

Note that up to the second level of abstraction there is no requirement for any assumption about the specific type of the values that the data ("Data") can take. To retain the flexibility of easily handling the variety of data forms and network types met in areas where integration with an existing system is needed, we chose to use a *polymorphic* datatype [10] defined in Haskell code as follows:

```
data Maybe A = Just A | Nothing
```

where A is a free variable representing any datatype. This in effect means that for any particular datatype "A" (e.g. Integer), which might be appropriate for a specific application, "Data" will either carry the information represented by it, taking a value of "Just A", or carry no information at all, assuming the "Nothing" value.

26.5 A Feedforward Neural Network

The first task in the derivation of a new network implementation algorithm is to use the rules of Constructive Type Theory to derive an algorithm. In order that this paper emphasises the way such a derived prototype may be merged with the paradigm of Artificial Neural Networks, we here assume that the derivation of a traditional Neural Network Architecture has been performed. We are not seeking to say that the derivation is trivial, but that such a derivation, once performed, may be used as the basis of an automated system capable of removing errors in the algorithm by means of given examples of the problem. The derivation of the network leads to the definition of the various functions required by the prototype. The first example is a simulation of a standard Multi-layer Perceptron (MLP) [11, 12]. Firstly the node activation functions and the weight functions are derived so that they reflect the specific functional forms required by the model:

```
act_f:: NeuralNet -> Vertex -> (Inputs -> Output)
act_f nn = \i -> fnn_node_activ_func nn i
```

where "nn" is the neural network in its current state, "i" is a vertex number corresponding to the node whose activation function we define, and "x" is an input to this node. Then we have to derive "fnn_node_activ_func" to assign to each node a specific, for example if the node is in the hidden layer the appropriate function could be `data_tanh` (where `tanh` denotes the well known hyperbolic tangent function) defined as follows:

```
data_tanh Nothing = Nothing
data_tanh (Just y) = Just (tanh(y))
```

where `y` is an appropriate combination of the inputs of the node. Next we can define suitable weights (recall that the weights in our prototype are logical implications). The corresponding code is:

```
weight:: Weight_Function
weight = \par -> fnn_weight par
fnn_weight:: [Data] -> [Data] -> [Data]
fnn_weight = zipWith data_mult
```

where "`zipWith f`" is a function which when applied to two lists of objects, recursively applies "`f`" to the corresponding elements of the lists and returns the list of the results.

The final step is to set up an "`update`" function suitable for the model in hand. Recall that:

```
update lr nn = nn {fun = set_params nn ps}
```

and:

```
ps = learning_law lr nodes_to_update nn
```

where "`ps`" is the list of the new network parameters being updated according to a particular "`learning_law`", "`lr`" is the learning rate, "`nodes_to_update`" is the list of nodes whose parameters should be updated, and "`nn`" as usual is the neural network in its current (pre-updated) state. Then the last thing we need to define is the "`nodes_to_update`" list to be the list of all the non-input nodes in the network.

26.6 Forecasting the Dover Tides

A means is required to evaluate data relating to the current flows and tidal levels present within Dover harbour in Kent, UK. The main problem revolves around the complexity of the system comprising of the tides and the current flow mechanism which have proved difficult to model and forecast [13]. The task presented here is to produce predictions of the tidal levels using a series of past measurements collected via a set of sensory devices within the harbour. The data set used consisted of hourly measurements of six variables considered affecting the tide dynamics. These variables are the tide level, the air and sea temperatures, the atmospheric pressure and wind speed and direction.

The neural network used in this case had six one-dimensional input nodes (one for each one of the predictor variables), two hidden nodes and one output node.

The activation functions selected for the hidden nodes are sigmoid nonlinearities (in particular the **tanh** function which has a range of $(-1, 1)$). For the input and output nodes, the identity function was chosen.

The input variables were normalised to correspond to the effective range of the hidden nodes' activation functions, while the weights were initialised for the training phase to random values uniformly destributed in (0, 1). The training algorithm employed was the standard on-line version of Backpropagation of Errors [12]; here derived in the Constructive Type Theory framework. The network is quite minimal in its architecture. However our results show that even such a minimal network can present a sufficiently satisfying performance. The data set used for training consisted of 740 data records (each containing the six predictors mentioned above) from the sample of measurements for December 1997. The trained network was used to produce one-step-ahead predictions out-of-sample. That is a data set of the same configuration from a different month's sample, namely November 1997, was used as inputs for our predictions. Figure 26.2a presents the 740 data points from this set for the air temperature measurements. Observe that between observations 280–400 represent a failure of the sensory system during the corresponding time period, a common problem in a practical real world environment. Figure 26.2b illustrates the behaviour of the network implemented within our framework in comparison to one implemented in a conventional manner It is easy to observe, from this figure, that our implementation and the conventional network present errors with no significant differences for the time periods with no corrupted data (indices 230–280, and 400–430) and their performances deteriorate in a very similar way when data which contains corrupted measurements (indices 280–330) is presented to them.

Conventionally, there are a number of ways to address the issue of corrupted data. We chose one of the most often used in practice to compare with our proposal of employing the "domain restriction" approach to tackle the problem. This is to find a data set which presents similar characteristics, in our case a set, which contains

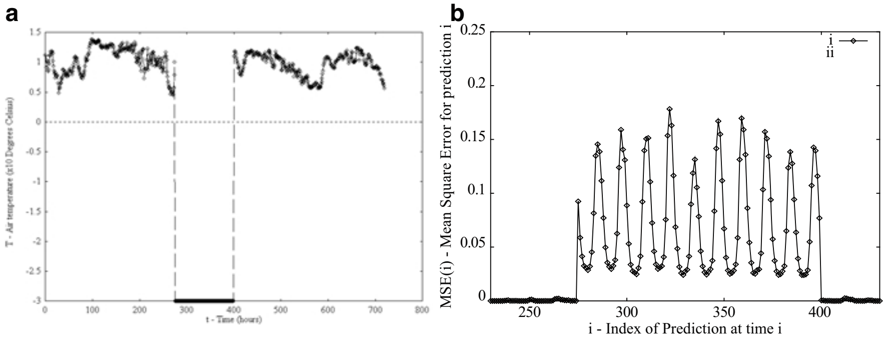


Fig. 26.2 (**a**) Air temperature measurements for November 1997, (**b**) prediction error; (i) implementation of a MLP within our framework, without domain restriction, and using a training set with no corrupted air temperature values; (ii) conventional implementation of MLP using the same training parameters and input data as in case i above

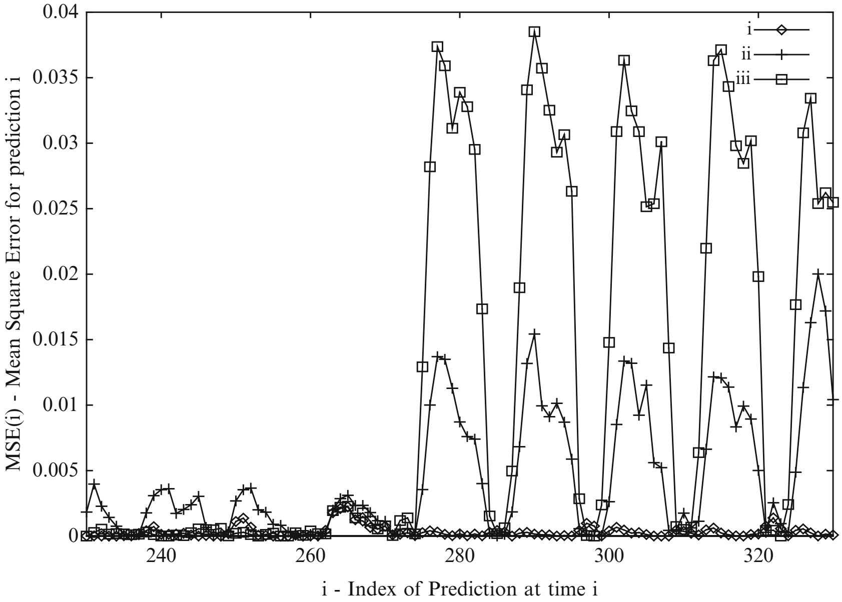


Fig. 26.3 Prediction error; (i) network using six inputs and domain restriction during the prediction phase and using a training set with no corrupted air temperature values; (ii) network using six inputs without domain restriction but using a training set which includes corrupted data; (iii) conventional implementation of MLP using the same training parameters and input data as in case (i) above

corrupted as well as correct air temperature measurements, and use it to retrain the network. Figure 26.3, shows the mean square error of the predictions produced using each of the above approaches for a period which contains both correct (indices 230–280) and corrupted data (indices 280–330). In order to have a basis for our comparisons we replot here (line with squares) the corresponding prediction error curve of the conventionally implemented MLP. The comparative advantage offered by our Prototype through the application of the "domain restriction" (line with diamonds in the figure) can be easily verified. In general, when the invalid data points are included in the training set (line with crosses) the average level of prediction error decreases which indicates some improvement.

26.7 A Self Organising Map

The second example presented is that of a Kohonen's Self-Organising Map (SOM) which is introduced to illustrate the techniques on a fundamentally different type of network to the first example. Again, we assume that the formal derivation of an appropriate architecture has already been performed. Firstly, we again derive the node activation functions and the network weights to reflect the functional form used in the SOMs. Analogously to the FNN above the corresponding code is:

```
act_f nn = \i -> som_node_activ_func nn i
weight = \par -> som_weight par
som_weight:: [Data] -> [Data] -> [Data]
som_weight = zipWith data_sbtr
```

The `som_weight` function is defined to perform a subtraction operation on the “`Data`” types of the appropriate nodes’ parameters and outputs To complete the definition of the SOM net, we need to make an appropriate modification to the “`update`” function so that it reflects the *lateral inhibition* principle used in the Kohonen type networks. Recall that:

```
update lr nn = nn {fun = set_params nn ps}
```

then:

```
ps = learning_law lr nodes_to_update nn
```

where “`ps`” is the list of the new network parameters being updated according to a particular “`learning_law`”, “`lr`” is the learning rate, “`nodes_to_update`” is the list of nodes whose parameters should be updated, and “`nn`”, as usual, is the neural network in its current (pre-updated) state. *Lateral inhibition* is expressed by defining:

```
nodes_to_update = neighbourhood nn v i k
```

where “neighbourhood” is a function with type signature:

```
neighbourhood:: NeuralNet ->Vertex ->
                    Int -> Int -> [Vertex]
```

and “`nn`” is the same as above, “`v`” is the network vertex corresponding to the output node with the minimum output and “`i`” and “`k`” are integers defining the horizontal and vertical bounds of the neighbourhood in a 2-dimensional array representing the output layer of the network.

We now present a simple example of the above SOM taken from the Dover Harbour data employing a neural network to represent the pictorial data of the current flows within the harbour represented as arrows. Due to these data being 4-dimensional vectors, we present here a simplified example version where the inputs are 2-dimensional vectors of real numbers, expressing the x-y co-ordinates of the measurement points. We trained two networks with two nodes in the input layer and four in the output layer. For each one we used a different training set consisting of 4 data points. The training set were {(0.0, 1.0), (1.0, 0.0), (2.0, 1.0), (1.0, 3.0)} in the first case and {(0.0, 1.0), (1.0, 0.0), (0.2, 2.0), (0.5, 0.6)} in the second case. Figure 26.4 shows the evolution of the corresponding network weights on the 2-dimensional weight plane during 100 training cycles. It is easy to observe convergence to the correct topological points after a small number of training steps.

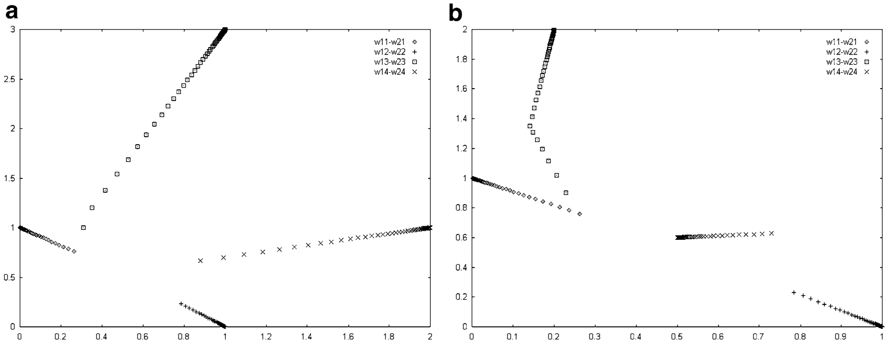


Fig. 26.4 (**a**) Kohonen network 1; evolution of the weight values on the 2-dimensional weight space during training (100 training cycles). (**b**) Kohonen network 2; evolution of the weight values on the 2-dimensional weight space during training (100 training cycles)

26.8 Conclusion

A mechanism has been introduced to illustrate the benefits of merging the areas of formal mathematics and artificial neural network. Such a technique is useful for problem domains which present difficulties in engineering an accurate solution but nevertheless possess a number of examples of valid results which may be employed as training data. This is a significant result in integrating formal logic with the field of Software Engineering.

Acknowledgement This research was supported by EPSRC grant No. GR/L78338 and by data supplied by the Dover Harbour Board Hydrographic Service.

References

1. Martin-Lof, P.: Constructive mathematics and computer programming. In: Hoare, C.A.R. (ed.) Mathematical Logic and Programming Languages. Prentice-Hall, New York (1985)
2. Thompson, S.: Type Theory and Functional Programming. Addison-Wesley, Reading, MA (1991)
3. Moran, A., Sands, D., Carlsson, M.: Erratic fudgets: a sematic theory for an embedded coordination language. Sci. Comput. Programm. **46**(1–2), 99–135 (Jan–Feb 2003)
4. Nguyen, Q.H., Kirchner, C., Kirchner, H.: External rewriting for skeptical proof assistants. J. Automat. Reason. **29**(3–4), 309–336 (2002)
5. Bishop, C.M.: Neural Networks for Pattern Recognition. Claredon, Oxford (1995)
6. Dorffner, G., Wiklicky, H., Prem, E.: Formal neural network specification and its implications for standardisation. Comput. Stand. Interface. **16**:205–219 (1994)
7. Howells, G., Fairhurst, M.C., Rahman, F.: An exploration of a new paradigm for weightless ram-based neural networks. Connect. Sci. **12**(1) (March 2000)
8. Rahman, A.F.R., Howells, G., Fairhurst, M.C.: A multi-expert framework for character recognition: a novel application of clifford networks. IEEE Trans. Neural Networ **12**(1) (January 2001)

9. Thompson, S.: Haskell, The Craft of Functional Programming. Addison-Wesley, Reading, MA (1996)
10. Kohonen, T.: Self-Organization and Associative Memory. Springer, Berlin (1984)
11. Padgett, M.L, Karplus, W.J., Deiss, S., Shelton, R.: Computational Intelligence standards: Motivation, current activities, and progress. Comput. Stand. Interface **16**:185–203 (1994)
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**:533 (1986)
13. Azoff, M.E.: Neural Network Time Series Forecasting of Financial Markets. Wiley, New York (1994)

Chapter 27
A Clustering Application in Portfolio Management

Jin Zhang and Dietmar Maringer

Abstract This chapter presents a new asset allocation model which combines a clustering technique with traditional asset allocation methods to improve portfolio Sharpe ratios and portfolio weights stability. The approach identifies optimal clustering patterns in different cluster number cases by using a population-based evolutionary method, namely Differential Evolution. Traditional asset allocations are used to compute the portfolio weights with the clustering. According to the experiment results, it is found that clustering contributes to higher Sharpe ratios and lower portfolio instability than that without clustering. Market practitioners may employ the clustering technique to improve portfolio weights stability and risk-adjusted returns, or for other optimization purposes while distributing the asset weights.

Keywords Clustering optimization · asset allocation · sharpe ratio · weights instability · differential evolution

27.1 Introduction

Asset allocation strategies can generally be classified into two major categories: the Markowitz asset allocations and the parameter-free allocations. At the Markowitz efficient frontier, an efficient portfolio yields a higher return than other portfolios given a same risk level. However, two elements in the Markowitz analysis: the assets expected returns and covariance, cannot be estimated precisely due to estimation errors and noise in the financial data. The literature suggests several approaches for reducing or avoiding error and noise impact on portfolios. For example, [6]

J. Zhang (✉)
Centre for Computational Finance and Economic Agents, University of Essex,
Wivenhoe Park, Colchester CO4 3SQ, UK
e-mail: jzhangf@essex.ac.uk

D. Maringer
Faculty of Economics and Business Administration, University of Basel,
Peter Merian-Weg 6, 4002 Basel, Switzerland
e-mail: dietmar.maringer@unibas.ch

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_27,

combined the return-based and range-based measures of volatility to improve the estimator of multivariate conditional variance–covariance matrix. On the other hand, in order to avoid estimation bias, one can simply adopt parameter-free allocations. The so-called 1/N rule, or the equally weighted investment strategy has been discussed by researchers in the literature. Windcliff and Boyle [11] proposed that the 1/N rule should be optimal in a simple market where the assets are indistinguishable and uncorrelated. However, assets may hardly be independent to each other, especially in large portfolios. Therefore, traditional asset allocations may not be enough for efficient portfolio management.

Several researchers have applied clustering techniques to portfolio management. For example, [9] employed a clustering technique to analyze mutual fund investment styles. Lisi and Corazza [7] proposed an active fund management strategy which selects stocks after clustering equities. However, the clustering techniques applied by the researchers in finance so far still follow a traditional clustering criterion: minimizing the dissimilarity between the cluster members, and maximizing the dissimilarity between clusters. This study, however, employs a different clustering criterion, namely one which is related to a portfolio performance measure: clustering assets to maximize the Sharpe ratio of portfolio returns. The proposed approach suggests an extra asset layer (i.e. the cluster portfolios), between the original assets and the terminal portfolio, in order to improve the risk-adjusted return and the weights instability. The optimal portfolios are constructed as follows. The proposed approach first groups the market assets to a series of disjoint clusters according to an optimal clustering partition; then it employs traditional asset allocations to construct a set of cluster portfolios using the cluster members; after that the method applies the same asset allocation to construct a terminal portfolio based on the cluster portfolios. According to this suggested three-layer hierarchy, the final asset weights are decided by two parts jointly: the cluster portfolio weights and the cluster member weights. A population-based evolutionary method, namely Differential Evolution (DE), is employed to tackle the clustering optimization problem. Three traditional asset allocations which cover the Markowitz allocations and the parameter-free method are employed to compute the weights of the cluster members and the cluster portfolios.

The remainder of the chapter is organized as follows. Section 27.2 introduces the optimization problem, the data for empirical experiments, and the heuristic approach to solve the clustering problem. Section 27.3 provides the experiment results and discussions. Section 27.4 concludes the chapter.

27.2 The Asset Allocation Model

27.2.1 The Optimization Problem

The optimization problem is to identify an optimal clustering partition $\mathcal{C}$, i.e. a union of optimal subsets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_G$. The optimized portfolio should yield a higher in-sample Sharpe ratio than that of the portfolios which are computed by using other clustering patterns with a same cluster number G. The optimization objective of the clustering problem can be expressed as follows:

$$\max_{\mathcal{C}} SR = \frac{\bar{r}_p - \bar{r}_f}{\sigma_p}, \tag{27.1}$$

where SR represents the Sharpe ratio; $\mathcal{C}$ denotes the optimal partition set; $\bar{r}_p$ is the average return of the portfolio; $\bar{r}_f$ refers to the risk-free return; and σ_p is a measure of the portfolio risk over the evaluation period. If one denotes G as the number of subsets in cluster set $\mathcal{C}$ and N as the number of total market assets, G should have a value range of $1 \leq G \leq N$. When G is equal to the number of either 1 or N, the optimization problem is reduced to the non-clustered case. The union of segmented assets should be equal to the original assets; and there should be no intersection between two different clusters. Let $\mathcal{C}_g$ denote the gth subset of assets, and $\mathcal{M}$ the market assets, the constraints can be expressed as:

$$G = \sharp\{\mathcal{C}_g\}, \tag{27.2}$$

$$\bigcup_{g=1}^{G} \mathcal{C}_g = \mathcal{M}, \tag{27.3}$$

$$\mathcal{C}_g \cap \mathcal{C}_j = \emptyset, \quad \forall g \neq j. \tag{27.4}$$

To avoid cases where one single cluster contains too many assets or an empty cluster exists, a cardinality constraint is imposed on the cluster size. Let $\tilde{N}^{\min}$ and $\tilde{N}^{\max}$ denote the minimum and maximum asset number of a cluster respectively, the above constraints can be described as follows:

$$\tilde{N}^{\min} \leq \sum_{s=1}^{N} I_{s \in \mathcal{C}_g} \leq \tilde{N}^{\max} \qquad \forall g \in G, \tag{27.5}$$

$$\text{where } I_{s \in \mathcal{C}_g} = \begin{cases} 1 & \text{if } s \in \mathcal{C}_g, \\ 0 & \text{otherwise,} \end{cases} \tag{27.6}$$

$$\text{with } \begin{cases} \tilde{N}^{\min} = \left\lceil \dfrac{N}{2G} \right\rceil, \\ \tilde{N}^{\max} = \left\lceil \dfrac{3N}{2G} \right\rceil. \end{cases} \tag{27.7}$$

Usually, asset allocation approaches handle the weights constraints imposed on the cluster members and the cluster portfolios, i.e., the sum of cluster member weights in a cluster, and the sum of cluster portfolio weights should be equal to 1, respectively. Apart from that, the weights can be either positive or negative, depending on whether short sales on assets are allowed. The above constraints can be written as follows:

$$\sum_{g=1}^{G} w_g = 1, \text{ and } \sum_{s \in \mathcal{C}_g} w_{g,s} = 1, \tag{27.8}$$

with either

$$w_g \geq 0, \quad w_{g,s} \begin{cases} \geq 0 & \forall s, g : s \in \mathcal{C}_g \\ = 0 & \text{otherwise} \end{cases}, \tag{27.9}$$

or

$$-\infty < w_g < +\infty, \quad w_{g,s} \begin{cases} \in (-\infty, +\infty) & \forall s, g : s \in \mathcal{C}_g \\ = 0 & \text{otherwise} \end{cases}, \tag{27.10}$$

where w_g denotes the weight of g-th cluster portfolio, and $w_{g,s}$ represents the weight of cluster member s in cluster $\mathcal{C}_g$. The terminal weight of an asset s is denoted by $\widetilde{w}_s$, i.e., the product of the cluster portfolio weight w_g and the cluster member weight $w_{g,s}$:

$$\widetilde{w}_s = w_g \cdot w_{g,s} \quad \forall g : s \in \mathcal{C}_g. \tag{27.11}$$

The above optimization problem cannot be solved by using traditional optimization methods, since the problem turns out to be NP-hard while the cluster number G increased (see [2]).

27.2.2 Asset Allocation Methods

27.2.2.1 The $1/\tilde{N}$ Allocation

In this study, the equal weights allocation is denoted as $1/\tilde{N}$, in order to distinguish it from the traditional 1/N strategy. In the $1/\tilde{N}$ allocating procedure, the cluster portfolio weights are decided by the cluster number G, while the cluster member weights are decided by the cluster size, i.e. the number of assets in the cluster. One should note that, although the cluster number G is manually assigned, the cluster size actually depends on the optimized pattern. Therefore, the cluster portfolio weights w_g are given by 1 over the cluster number G, and the cluster member weights $w_{g,s}$ in the subset $\mathcal{C}_g$ are computed by taking 1 over the number of cluster members $\sharp\{\mathcal{C}_g\}$, respectively:

$$w_g = \frac{1}{G}, \tag{27.12}$$

$$w_{g,s} = \frac{1}{\sharp\{\mathcal{C}_g\}} = \frac{1}{\sum_{s=1}^{N} I_{s \in \mathcal{C}_g}}. \tag{27.13}$$

27.2.2.2 The Markowitz MVP Allocation

It is known that the MVP is the safest portfolio yielding the minimum variance at the Markowitz efficient frontier. Usually, quadratic programming is employed to compute the cluster member weights, as well as the cluster portfolio weights:

$$\max_{\boldsymbol{w}} \lambda \boldsymbol{r}'\boldsymbol{w} - (1-\lambda)\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}, \tag{27.14}$$

where $\boldsymbol{w}$ denotes either the cluster portfolio weights vector $\boldsymbol{w}_g$ or cluster member weights vector $\boldsymbol{w}_{g,s}$; $\boldsymbol{\Sigma}$ denotes the variance-covariance matrix of the cluster portfolios or the cluster members. The λ is set at 0 for a minimum variance portfolio. Indeed the MVP is a special case located on the Markowitz efficient frontier, which may be used as a proxy to explore the cluster impact on other Markowitz efficient portfolios by changing the λ value.

27.2.2.3 The Modified Tobin Tangency Allocation

The third allocation is an extension of the Tobin's Tangency portfolio. The allocation can be expressed by using an analytical solution, while short sales is allowed. In this case, the terminal portfolio is a tangency portfolio based on the cluster tangency portfolios, which are constructed by using the cluster members as the inputs of the tangency allocation. The tangency portfolio allocation is described as follows:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{r}' \\ \boldsymbol{I}' \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \boldsymbol{r} & \boldsymbol{I} \end{bmatrix} \equiv \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \tag{27.15}$$

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \boldsymbol{r} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \dfrac{1}{b - r_f \cdot c} \\ \dfrac{-r_s}{b - r_f \cdot c} \end{bmatrix}, \tag{27.16}$$

where $\boldsymbol{I}$ is the unity vector, r_f is the risk-free rate. $\boldsymbol{r}$ represents a vector of the expected returns of either cluster portfolios or cluster members; $\boldsymbol{\Sigma}$ is a variance-covariance matrix describing the correlation of cluster portfolios or cluster members. Correspondingly, $\boldsymbol{w}$ is a vector representing the cluster portfolio weights $\boldsymbol{w}_g$ or the cluster member weights $\boldsymbol{w}_{g,s}$.

27.2.3 The Optimization Method

Heuristic methods, such as Simulated Annealing, Threshold Acceptance, and Differential Evolution (DE) provide different ways for approaching difficult combinatorial optimization problems. These methods have been applied by [4,5] and [8] to solve optimization problems in finance and econometrics. This study uses a version close to standard DE (see [10]) to tackle the discrete problem. If the row vectors $\boldsymbol{v}_p$, $p = 1, \dots, P$ denotes solutions, for each current solution p, a new solution $\boldsymbol{v}_c$ is generated as follows: randomly choosing three different members from the current population ($p_1 \neq p_2 \neq p_3 \neq p$), then linearly combining the corresponding solution vectors of the three with a probability of $(1 - \pi_1)$, or inheriting the original pth solution otherwise. With standard DE, only the population size P, the scaling factor F and the cross-over probability π_1 are considered. The extra noise is generated by adding normally distributed random number vectors with the mean value being zero to F value and the difference of two solution vectors, respectively. The noise vectors $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ have the following properties: they are random variables being zero with probabilities $\boldsymbol{\pi_2}$ and $\boldsymbol{\pi_3}$, otherwise they follows the normal distribution at $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$. Thus the modified linear combination of the cross-over procedure can be described as

$$v_c[i] := \begin{cases} v_p[i] & \text{with probabiltiy } 1 - \pi_1 \text{, or} \\ v_{p1}[i] + (F + z_1[i]) \cdot (v_{p2}[i] - v_{p3}[i] + z_2[i]), & \end{cases} \tag{27.17}$$

where π_1 is the cross-over probability. To tell the clustering pattern according to the current solution, the approach rounds up the solution to the nearest integers. After the linear combination, the algorithm updates the elitist by using the solution with higher fitness (i.e., higher Sharpe ratio value). The process is repeated until a halting criterion is met. The DE optimization procedure is described in the following pseudo-code:

Algorithm 1 Differential Evolution for Clustering Optimization

```
randomly initialize population of vectors $\boldsymbol{v}_p$, $p = 1 \dots P$
while the halting criterion is not met do
    for all current solutions $\boldsymbol{v}_p$, $p = 1 \dots P$ do
        randomly pick $p_1 \neq p_2 \neq p_3 \neq p$
        $v_c[i] \leftarrow v_{p1}[i] + (F + z_1[i])(v_{p2}[i] - v_{p3}[i] + z_2[i])$ with probability $\pi_1$,
        or $v_c[i] \leftarrow v_p[i]$ otherwise
        interpret $\boldsymbol{v}_c$ into clustering partition
        apply asset allocations to compute the weights of the terminal portfolio
        compute the fitness value, i.e. the Sharpe ratio of the portfolio
    end for
    for the current solution $\boldsymbol{v}_p$, $p = 1 \dots P$ do
        if Fitness($\boldsymbol{v}_c$) > Fitness($\boldsymbol{v}_p$)
        then $\boldsymbol{v}_p \leftarrow \boldsymbol{v}_c$
    end for
end while
```

27.2.4 Data and Performance Indicators

The study employs the adjusted daily prices of FTSE 100 stocks from January 2005 to December 2006 and the prices of DJIA 65 stocks from January 2003 to December 2004 (available from Yahoo.com). The daily returns of the assets were computed by taking the difference of the daily log price. The in-sample experiments considered the first year's data while the out-of-sample experiments used the following year's data.

Based on preliminary tests, the technical parameters of DE algorithm are reported as follows: the population size was 100; the iteration number was set at 100,000; the weighting factor F has a value of 0.7; and the cross-over probability π_1 was set at 50%. The parameters for generating the artificial noise z_1 and z_2 were: $\pi_2 = 70\%$, $\pi_3 = 30\%$, $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.03$.

Portfolio stability is important to portfolio management. The portfolio stability in this study refers to how assets weights of a portfolio are sensitive to the noises in the inputs of an allocation. A weights instability measure (see [3]) is employed to study the weights changes due to estimation errors. The measure is defined as follows:

$$\mathcal{I} = \frac{\sum_{i=1}^{N} |\widetilde{w}_{i,o} - \hat{w}_{i,e}|}{2}, \tag{27.18}$$

where $\widetilde{w}_{i,o}$ stands for the asset weights of a portfolio which are computed using the 'true' asset information; while $\hat{w}_{i,e}$ are the asset weights of the portfolio that are computed from the asset information with errors and noise. In this instability experiment, the in-sample asset returns are considered as the 'true' asset information, while the return series with 'noise' are artificially generated by adding random disturbances to certain assets returns. The instability $\mathcal{I}$ is computed by using the weights from the 'true' returns and the 'noise' returns. The simulated procedure is repeated 10,000 times, and the average instability is computed and taken as the final instability measure.

Sharpe ratio is widely used as a measure of portfolio performance by market practitioners, thus it is used as the indicator to risk-adjusted return performance of the terminal portfolios over both the in-sample and the out-of-sample period. The daily risk-free return r_f is set at zero because the rate is tiny. In addition to the above two measures, the two sample Kolmogorov–Smirnov (K–S) test is employed to judge whether the terminal portfolio return distributions over the in-sample and the out-of-sample period are statistically different after clustering.

Clustering has been widely applied in recognizing natural groups or clusters in multidimensional data based on the similarity measuring, and it will therefore be interesting to study the similarity of cluster members by taking the Sharpe ratio maximization as the clustering criterion. In the literature, a similarity measure is the Euclidean distance. The Euclidean distance is defined as

$$D_{Eu}(\vec{R}_{g,s}, \vec{R}_g) = \sqrt{\sum_{d=1}^{D} (R_{g,s,d} - R_{g,d})^2}, \text{ where } g : s \in \mathcal{C}_g, \tag{27.19}$$

where $\vec{R}_{g,s}$ represents the daily return of the cluster member s; $\vec{R}_g$ is the daily return of the cluster portfolio. D is the number of days of the observation period; $R_{g,s,d}$ and $R_{g,d}$ are the asset return and cluster return on the d-th day respectively. The Euclidean distance can be considered as a diversification measure to the cluster portfolios, i.e., the closer the Euclidean distance of cluster members, the less the diversification.

27.3 Computational Results

27.3.1 Portfolio Instability

Figures 27.1 and 27.2 report the instability $\mathcal{I}$ measure which was computed from the simulated FTSE and DJIA returns containing the noise in five cases: the G value being 1, 3, 5, 7 and 9. The two figures show the weights instability of portfolios from the MVP allocation and Tobin tangency allocation cases only, as the $1/\tilde{N}$ does not have the instability issue. As the two figures show, the instability $\mathcal{I}$ from both the MVP allocation and the Tobin allocation are reduced after clustering. This supports the conclusion that the portfolio weights from the clustered asset markets are less sensitive to changes or noise of the input parameters than those from the non-clustered markets.

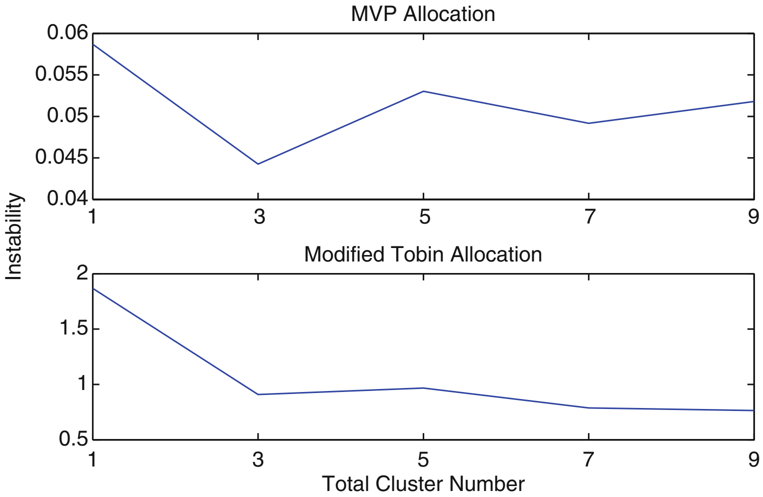


Fig. 27.1 Clustering impact on portfolio weights instability: the FTSE market

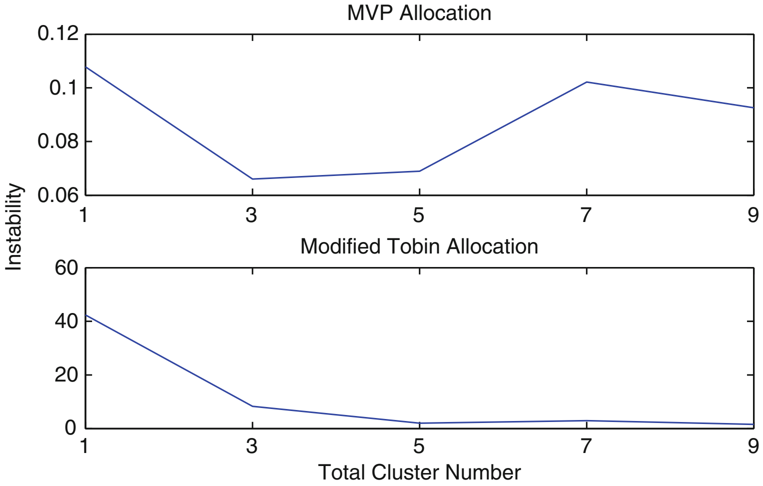


Fig. 27.2 Clustering impact on portfolio weights instability: the DJIA market

27.3.2 Sharpe Ratios and Return Distribution

Table 27.1 reports the in-sample and out-of-sample Sharpe ratios, the p-values of the K-S distribution test from using the three asset allocations. Regarding the FTSE case, it is clear that the out-of-sample Sharpe ratio is increased after the clustering. According to the p-values, the test rejects the hypothesis that the in-sample returns and the out-of-sample returns follow the same distribution at a 10% significant level. Interestingly, the non-clustered cases from the three allocations have p-values of less than 10%, which indicates that a possible structural break occurred in the FTSE market between 2005 and 2006. In order to prove that clustering helps to improve the Sharpe ratio, it is useful to perform the analysis once again upon different markets.

The right panel of Table 27.1 reports the experiment results using the DJIA market data. Again it is found that clustering the DJIA market improves the Sharpe ratios of the terminal portfolios in the out-of-sample period. Therefore, it should be confident to conclude that clustering helps to improve Sharpe ratio of terminal portfolios. Turning to the K-S test, in the first two allocations the high p-values suggest that in-sample and out-of-sample portfolio distributions are consistent, whereas the p-values from the modified Tobin allocation are close to zero, indicating that the inconsistency of the return distributions comes from the allocation method, not clustering.

Table 27.1 Sharpe ratios and p-values: the FTSE and DJIA markets

	FTSE			DJIA		
	SR(I)	SR(I)	p-values	SR(I)	SR(I)	p-values
$1/\tilde{N}$						
$G = 1$	0.159	0.077	0.095	0.069	0.087	1
$G = 3$	0.207	0.079	0.053	0.102	0.092	0.936
$G = 5$	0.213	0.078	0.041	0.104	0.092	0.967
$G = 7$	0.214	0.075	0.041	0.103	0.088	0.893
$G = 9$	0.220	0.079	0.032	0.102	0.091	0.967
MVP						
$G = 1$	0.155	0.083	0.008	0.073	0.038	0.329
$G = 3$	0.207	0.085	0.008	0.118	0.060	0.238
$G = 5$	0.230	0.082	0.006	0.129	0.068	0.167
$G = 7$	0.239	0.081	0.012	0.129	0.058	0.093
$G = 9$	0.245	0.076	0.012	0.128	0.064	0.167
Tobin						
$G = 1$	0.622	0.073	0	0.494	0.063	0
$G = 3$	0.610	0.064	0.001	0.477	0.064	0
$G = 5$	0.610	0.082	0.002	0.470	0.074	0
$G = 7$	0.611	0.088	0.001	0.467	0.081	0
$G = 9$	0.612	0.066	0.001	0.453	0.050	0

27.3.3 Cluster Properties: the Sharpe Ratio and the Euclidean Distance

This subsection studies the property of the clusters in terms of the in-sample Sharpe ratio and the Euclidean distance. The following study focuses on the FTSE market with a cluster number $G = 3$. Figure 27.3 shows the Sharpe ratios of the cluster members, and the Euclidean distance between the cluster portfolios and its corresponding members from the $1/\tilde{N}$ allocation (the cluster portfolios and cluster members are identified by using a different symbol size). The figure shows that the $1/\tilde{N}$ allocation tends to group the 'better' and 'worse' assets in terms of Sharpe ratio, while the 'better' cluster ones have heavier weights than the 'worse' ones. Figure 27.4 shows the MVP case. Similar to the case of $1/\tilde{N}$ allocation, it seems that the MVP allocation tends to group the assets with higher Sharpe ratios into a small group while putting the assets with lower Sharpe ratios into a large group. The Euclidean distances of the $1/\tilde{N}$ and MVP allocations roughly spread over a range between 0.1 and 0.3. One may not observe a clear clustering pattern from Fig. 27.5, which shows the modified Tobin case. However, the Euclidean distance is significantly different from that of the previous two allocations: the distance in this case has a range between 0.20 and 0.30. The narrowed range from the modified Tobin allocation indicates a less diversification effect than that from the first two allocations.

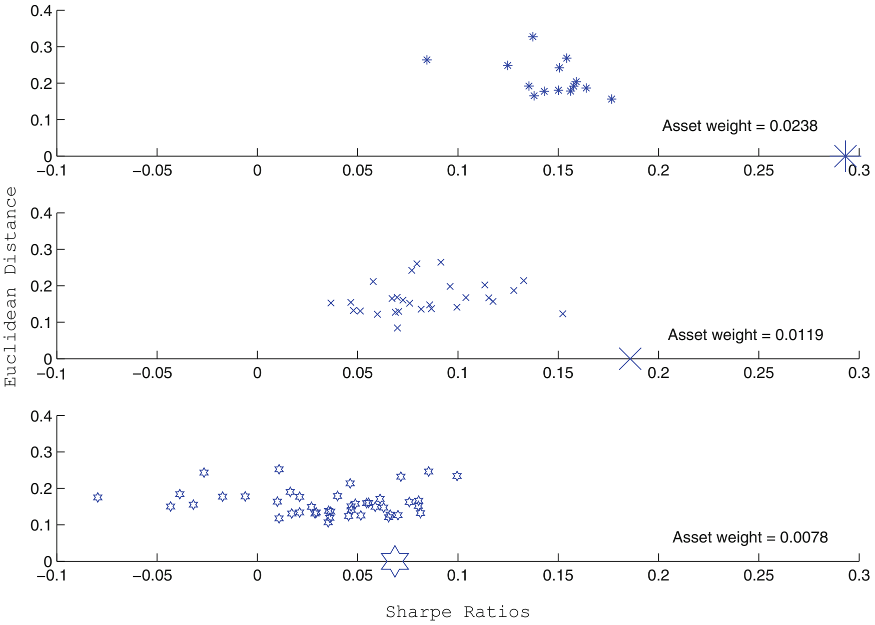


Fig. 27.3 The $1/\tilde{N}$ equal weights portfolio $G = 3$

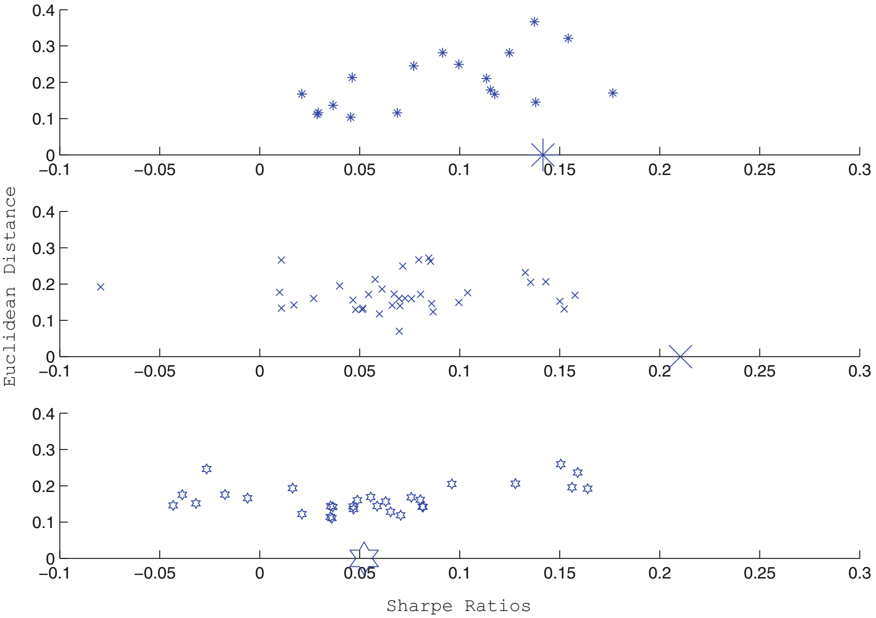


Fig. 27.4 The minimum variance portfolio $G = 3$

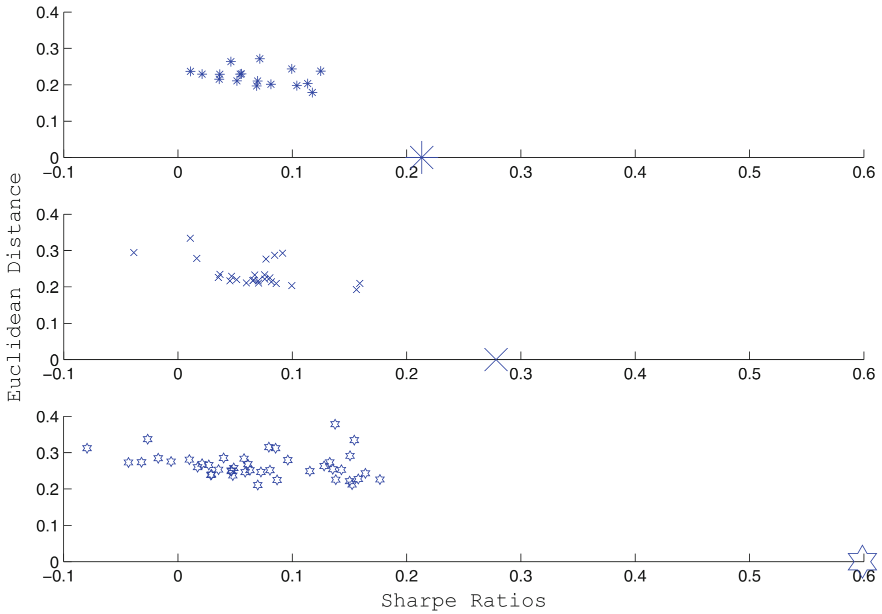


Fig. 27.5 The modified tobin $G = 3$

27.4 Conclusion

This chapter has presented an approach which combines a clustering technique with traditional asset allocation methods to improve portfolio risk-adjusted performance and weights instability. The method employed a population-based evolutionary method, namely Differential Evolution, to search optimal cluster partitions subject to maximizing the Sharpe ratio. The Optimized portfolio weights were decided by two parts: the cluster portfolios and the cluster members. The cluster member weights and cluster portfolio weights were computed by using three traditional asset allocations which included the famous mean-variance and the parameter-free allocations. Empirical studies of the clustering impact on the Sharpe ratio and the weights instability focused on two markets: the Financial Times and Stock Exchange and the Dow Jones Industrial Average markets. According to the experiment results, it was found that the portfolios incorporating clusters can have higher out-of-sample Sharpe ratios and lower weights instability than those without clustering. Portfolio managers may therefore employ the proposed clustering technique for different optimization purposes while distributing the asset weights.

References

1. Benartzi, S., Thaler, R.H.: Naive diversification strategies in defined contribution saving plans. Am. Econ. Rev. **91**, 79–98 (2001)
2. Brucker, P.: On the complexity of clustering problems. In: Optimization and Operations Research, Chapter 2, pp. 45–54. Springer, Germany (1977)
3. Farrelly, T.: Asset allocation for robust portfolios. J. Investing **15**, 53–63 (2006)
4. Gilli, M., Maringer, D., Winker, P.: Applications of Heuristics in finance. In: Handbook on Information Technology in Finance, Chapter 26, pp. 635–653. Springer, Germany (2008)
5. Gilli, M., Winker, P.: A review of Heuristic optimization methods in econometrics. Working papers, Swiss Finance Institute Research Paper Series, No. 8–12 (2008)
6. Harris, R.D.F., Yilmaz, F.: Estimation of the conditional variance – covariance matrix of returns using the intraday range. Working Papers, XFi Centre for Finance and Investment (2007)
7. Lisi, F., Corazza, M.: Clustering financial data for mutual fund management. In: Mathematical and Statistical Methods in Insurance and Finance, Chapter 20, pp. 157–164. Springer, Germany (2008)
8. Maringer, D.: Constrained index tracking under loss aversion using differential evolution. In: Natural Computing in Computational Finance, Chapter 2, pp. 7–24. Springer, Germany (2008)
9. Pattarin, F., Paterlinib, S., Minervac, T.: Clustering financial time series: an application to mutual funds style analysis. Comput. Stat. Data An. **47**, 353–372 (2004)
10. Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**, 341–359 (1997)
11. Windcliff, H., Boyle, P.: The 1/N pension investment puzzle. N. Am. Actuarial J. **8**, 32–45 (2004)

Chapter 28
Building an Expert System for HIV and Aids Information

Audrey Masizana-Katongo, Tebo Leburu-Dingalo, and Dimane Mpoeleng

Abstract We propose to develop an expert system that will provide general information on HIV and AIDS to the public of Botswana. This project is motivated by the Microsoft sponsored project conducted at the Department of Computer Science named IHISM for which the authors are members. IHISM aims to contribute to the digital divide by developing an HIV and AIDS public information portal accessible through mobile phones. The case study is based on Botswana, one of the leading countries hard hit by the HIV and AIDS pandemic where the government is working hard to facilitate the provision of education and raising awareness concerning the pandemic. Whilst IHISM is concentrating on the digital divide and mainly concerned with semiliterate people in rural settlements this alternative is for those who have access to the internet and hence can use the system wherever they are. The proposed system will act as an online 'expert' and is aiming at supporting the initiative of providing accurate information to the general public.

Keywords Expert systems · HIV and AIDS

28.1 Introduction

In January 2005 the Department of Computer Science, University of Botswana received funding from Microsoft, USA, as part of their Digital Inclusion initiative [1] which aims to bridge the gap between relatively well-developed and less-developed regions in the world. The project termed, IHISM [2] (Integrated Healthcare Information System through Mobile telephony) aims to explore the use of mobile phones as an access technology to a variety of HIV and AIDS related information required by the general public. The disease has affected every segment of Botswana society and according to UNAIDS estimates [3], one-third of Botswana's sexually-active

A. Masizana-Katongo (✉), T. Leburu-Dingalo, and D. Mpoeleng
Department of Computer Science, University of Botswana, P/Bag 0022, Gaborone, Botswana
e-mail: masizana@mopipi.ub.bw; leburutk@mopipi.ub.bw; mpoeleng@mopipi.ub.bw

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_28,

population between the ages of 15 and 49 (out of a total population of 1.5 million) are infected with the virus that causes AIDS, the highest rate in the world.

However, Botswana is also one of the leading countries at the front line of the war against AIDS and the government has demonstrated a very high level of political commitment to addressing the HIV and AIDS epidemic. The government has not only formed various partnerships to combat HIV and AIDS [4–7] but is also committed to inform the nation about the pandemic by introducing various education and communication programs [8, 9]. These programs aim at raising awareness of the diseases and offer relevant information to the general public on the dangers and management of HIV and AIDS pandemic. This response has seen Botswana exhibit high prevalence over HIV and AIDS infection as the general awareness of the dangers of the disease has been raised among the population. The government also provides public handouts which are distributed by the ministry of health as well as the hosting of public events such as HIV/AIDS fairs and pageants to try to educate and sensitize the public about the disease.

The programs described give a clear indication that the complexity and incurability of HIV/AIDS requires that the government in addition to providing clinical support, needs to also put an emphasis on the dissemination of information about the disease, prevention methods, therapeutic methods and psychological support. In this age it is evident that one of the best ways to do this is through the utilization and provision of ICT technologies. To support this initiative, the IHISM project has pledged to develop a general HIV and AIDS information portal which would be queried by the users using mobile phone technology. The information service portal would allow the general public to request for information on topics related to HIV and AIDS such as descriptions, infection, testing, counselling and support, opportunistic diseases and paediatric care etc. The portal represents this information in the form of Frequently Asked Questions (FAQ) service where the user inputs a query on any of the subjects. The project is on-going and has achieved thus far a milestone of deliveries and has influenced a number of research projects within the department.

This project has emerged as a by-product of the IHISM concept by developing a system which provides similar services in the form of an expert system to the public with internet access. The system is meant to act as an online ‘expert’ in HIV and AIDS information such that some information may have to be derived through inference as opposed to simple data retrieval. The system will accept as input, an FAQ from the user and provide the most relevant answer to the question. Dissemination of accurate information about HIV is an essential element of the national AIDS prevention strategies and thus information service provided by this system has to be accurate. The challenge is that users may ask the questions differently in pursuit of the same answer and the system should be able to systematically analyse the questions and provide an appropriate answer. The main challenge is to determine the various forms in which a typical FAQ question could be mapped to the relevant answer. Another challenge is to determine the degree to which the answer is relevant to a particular question. To address the former we derive various keywords for any input question and to address the latter we employ the use of certainty factors offered within the Corvid Exsys Expert system developer [10] used to develop the system (Fig. 28.1).

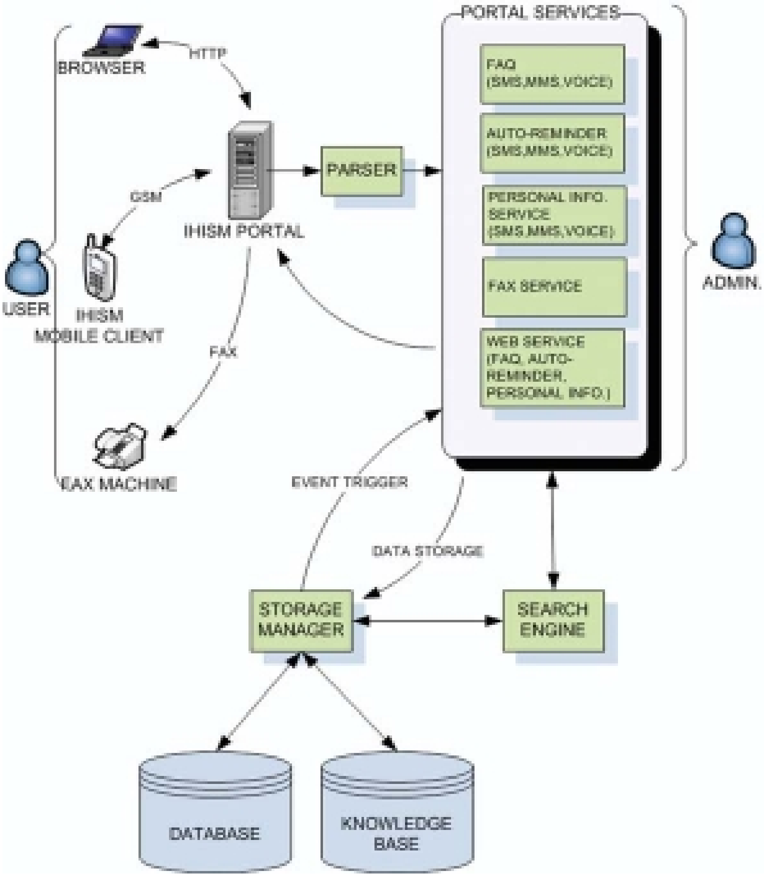


Fig. 28.1 IHISM architecture

28.2 Health Expert Systems

The application of IT research and development to support health and medicine is an emerging research area with significant potential. Major initiatives to improve the quality, accuracy and timeliness of healthcare data and information delivery are emerging all over the world. The Agency for Healthcare Research and Quality (AHRQ) [11], of the US Department of Health Services (HHS), awarded grants and contracts to promote the use of health information technology. Computerized systems including expert systems have been used to carry out efficient and effective data processing on complex problems to support various problem domains since the 1970s. Jackson [12] defines an expert system as a "…computing system capable of representing and reasoning about some knowledge–rich domain, such as internal medicine or geology, with a view to solving problems and giving advice." Since the advent of artificial intelligence in the 1970s which saw the birth of expert systems, various domains have taken advantage of this technology. The most popular application has been in the area of health and medicine. MYCIN [12], developed in 1970 at the Stanford University, is one of the most popular medical expert system used to assist to diagnose and treat blood diseases. MYCIN was the pioneer in demonstrating how a system can be used to successfully perform medical diagnosis.

Another Early expert system is the PACE(Patient Care Expert System (Pace) [13] which was conceived in 1977 with the purpose being to make "intelligent selections" from the overwhelming and ever changing information related to health in order to facilitate patient care. The system started off as an educational system for the nursing profession. Throughout the years the system evolved and went through many development generations to a point where it became an advanced clinical management system capable of supporting the entire health care field to diagnose and care for patients with pulmonary diseases.

Another expert system called, MITIS system, was developed in 2004 at the National Technical University of Athens. The MITIS system was developed to assist in the management and processing of obstetrical, gynecological, and radiological medical data [14]. The concept behind this system is to record and store information from experts in medical departments of gynecology, radiology and obstetrics to provide a centralized mechanism for managing patient information within and outside a hospital.

HIV and AIDS information is required to be accurate, timely and easy to understand. As it is essential that HIV and AIDS information distributed should be from an expert source, expert systems are a preferential way to undertake this task. The advent of HIV and AIDS has prompted the development of knowledge base systems in this problem domain. Most of these systems are still at research and developmental levels. In the recent years, researchers are developing Question and Answer systems on HIV and AIDS to assist mainly medical practitioners. Most of these systems differ in the way they were developed and take different forms of HIV and AIDS expertise delivery to the intended users including online. Atalay et al. [15] developed a system that diagnoses HIV-patients and prescribes the correct drug regimen for them. Temesgen et al. [16] developed a Questions and Answers support system named CHESS which provides answers to questions logged in by people living with HIV and AIDS infection. Sloot et al. [17] employs grid technology to develop an HIV expert system which delivers expert knowledge accessed from distributed databases of various infectious diseases.

Another system is the Customized Treatment Strategies for HIV (CTSHIV) [18], a knowledge-base system for management of HIV-Infected Patients. CTSHIV expert system was developed at the University of California Irvine as a tool that "...recommends individualized treatment strategy for HIV patients..." The concept behind the system is to analyze the strains of HIV in a patient and find out what antiretroviral agent the strain is resistant to. Based on this information the system is then able to determine which antiretroviral drugs are appropriate for that particular patient. Although CTSHIV is aimed at physicians as a diagnostic tool it does have the advisory capabilities in that it refers users to more information in respect to recommendations being generated.

This project proposes a more general information providing version of the CTSHIV expert system which would provide HIV and AIDS information to the public of Botswana based on the FAQ logged in by the public.

28.3 HIV and Aids Expert System

In describing the basic concept of an expert system, Giarratano Riley [19] describes three components being the user, who supplies facts and information to the expert system or receives expert advice from the system, the knowledge base which contains knowledge, and finally the inference engine that uses rules and the knowledge base to draw conclusions in response to a user's query.

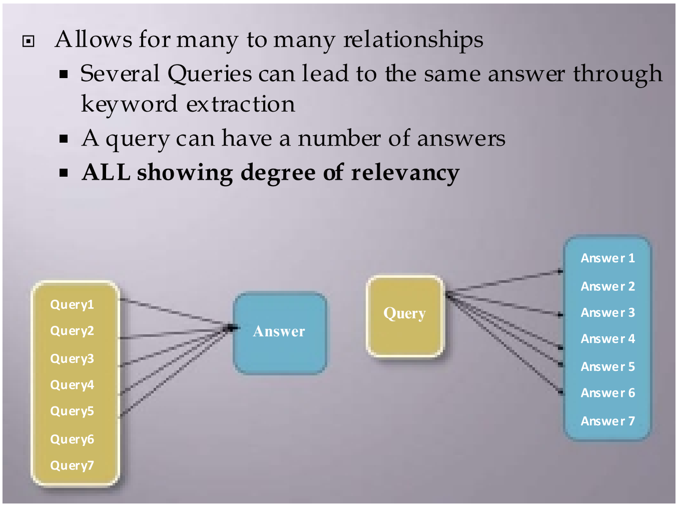
The proposed expert system consists of basically the same components. The system consists of three main components; the user interface, inference engine and knowledge base. The user interacts with the system through the interface that consists of graphical screens that allow the user to type in a question and display responses from the system. The variables that are used at the user interface are the user question and answer. The user variable is used with the question logic block to ask a question while the answer is used by the results block to publish the response to the user.

The answers evaluated for this task are stored in the knowledge base. The knowledge base is built from mainly Frequently Asked Questions (FAQ) and answers manual of a local HIV/AIDS information call center, Ipoletse [9]. Ipoletse in English means '*Ask for yourself*', meaning if you need information just pick the phone and have the information given to you. The call center was set up by Botswana government in 2002 with a mission to provide free information on HIV and AIDS to the public. The manual consists of over 200 questions and answers on HIV/AIDS related information. This knowledge base also acts as a store for rules used by the inference engine in the extraction of keywords and a meta block that holds the answers and their ranks.

In order to extract the keywords for each question and answer pair, an online survey was set up using a web survey tool called Limesurvey [20]. A sample of a 100 students at the University of Botswana were provided with the questions from the manual and asked to provide another way of asking the same questions so that a sample of different words could be mapped to a particular question and mapped onto a relevant answer. As an example in asking the question,

Q1: 'What is HIV'. A1: HIV is. . . . . .

The input from the sample users offered variations of the same questions such as '*what do you 'understand'* about HIV; '*describe*' HIV, What is '*meant'* by HIV' etc. The statistical analysis of the extracted words was performed in order to derive the most frequently used words (FUW). In the case of this question the FUW set include ***define, do, does, understand, describe, and explain*** etc. with the given percentages as in Fig. 28.1. The chart shows that for this particular question, HIV is pretty much



- Allows for many to many relationships
  - Several Queries can lead to the same answer through keyword extraction
  - A query can have a number of answers
  - **ALL showing degree of relevancy**

Fig. 28.2 Query and answer structure

the main keyword and a default word followed in some capacity by the verbs like do, does and adverbs like describe, define, understand and describe (Fig. 28.2).

By employing significant FUW's we derive phrases that are in fact variations of the possible combinations of the words that can be accepted as the question. A keyword would comprise of one or more combination of question default words, tags (HOW, WHAT) and FUW. In order to provide flexibility, each question is allowed to have several keywords to map to the correct answer (Figs. 28.3 and 28.4).

The inference engine can be considered the brain of the system. It is responsible for the "reasoning" of the system. In the proposed expert system, the inference engine extracts keywords from input by the user and processes them to check the validity of the question asked. The system is rule based and thus implements inferencing by utilizing the IF THEN rule to draw conclusions as to which answer is to be retrieved for a relevant query or question. To reach a recommendation the system uses the encoded information of keywords to determine if they qualify for any answer in the system. To make this assessment the weight process logic block utilizes confidence variables to rate the relevancy of keywords combination to any particular input or query (Fig. 28.5).

28.4 Development of the System

The system was developed using Exsys CORVID development tool [10]. This product provides an advanced environment to develop Knowledge Automation Expert Systems in various domains and has good features to allow web applications. The inference engine uses the problem-solving logic to emulate the decision

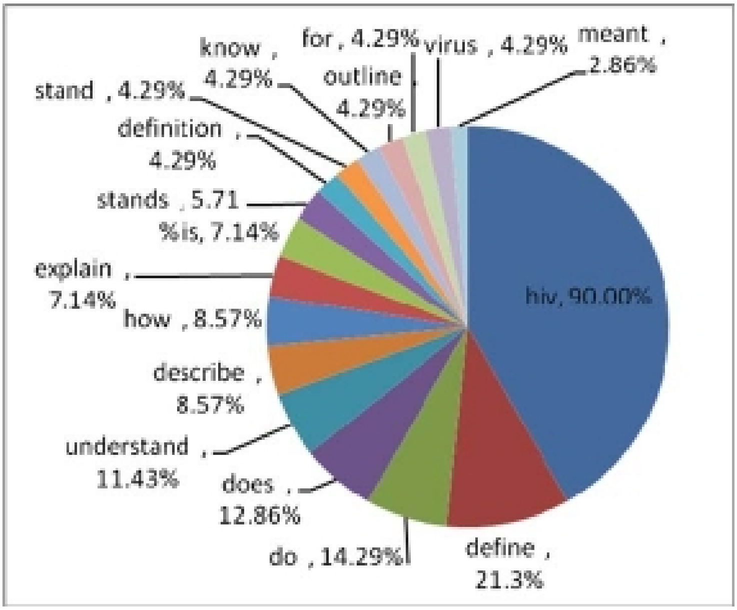


Fig. 28.3 Percentage FUW for Q1

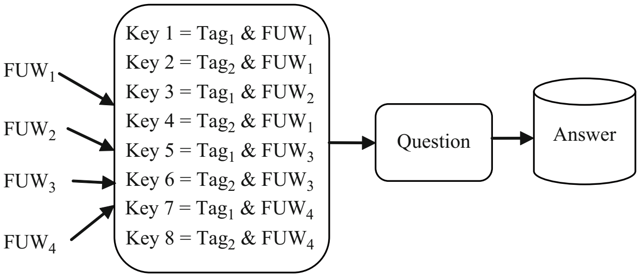

FUW_1
FUW_2
FUW_3
FUW_4
Key 1 = Tag_1 & FUW_1
Key 2 = Tag_2 & FUW_1
Key 3 = Tag_1 & FUW_2
Key 4 = Tag_2 & FUW_1
Key 5 = Tag_1 & FUW_3
Key 6 = Tag_2 & FUW_3
Key 7 = Tag_1 & FUW_4
Key 8 = Tag_2 & FUW_4
Question
Answer

Fig. 28.4 Keyword extraction

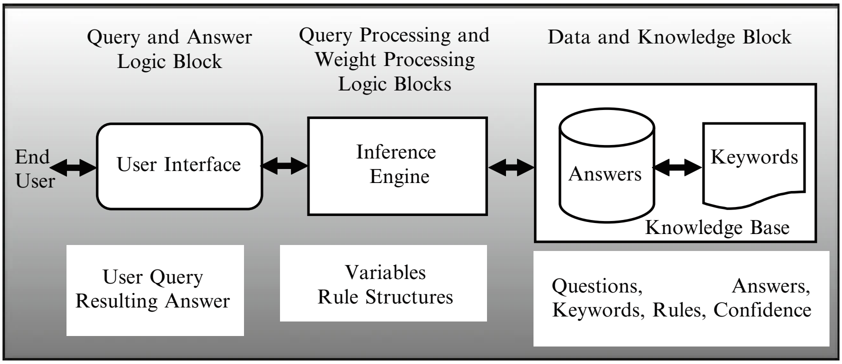


Fig. 28.5 Components of the expert system

making of domain experts. It allows the domain experts to easily describe their decision-making steps in a logical manner using tree-structured logic diagrams described as rules. People interact with the system as if they were talking to the expert and in this case an expert in HIV and AIDS knowledge. The rules of the system are defined, organized and structured using one or more logic blocks in tree-structured diagrams. The logic may be a complex branching tree that systematically covers all possible input cases, or a simple diagram that correlates with a few rules. Logic Blocks are created and maintained in a visual, intuitive development environment using the IF/THEN representation.

Figure 28.6 below demonstrates the main processes of user input, processing and system response. First the user types in a question at the system interface. The system uses rules to assess if the question's keywords match those in the system. If a match is not found the system outputs a no answer response. If a match exists in the system the keywords are assigned weights referred to as confidence variables. The weights for all identified keywords are then summed. This sum is used to measure how accurate the question is, relative to predefined real life user questions (Fig. 28.7).

Therefore if a combination of keywords closely resembles a predefined question then the sum for that combination would be high. The system then accesses

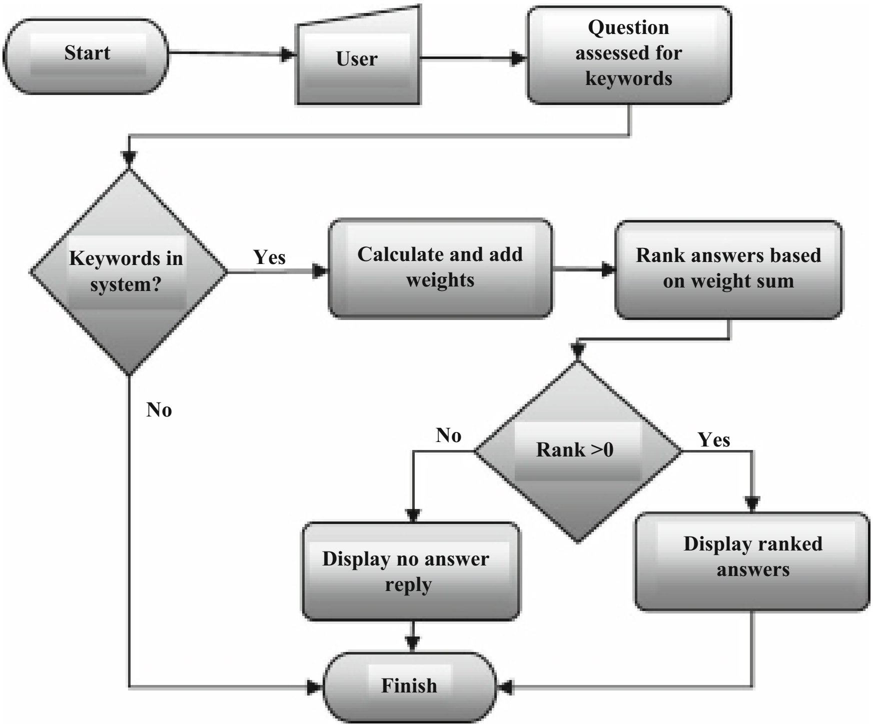
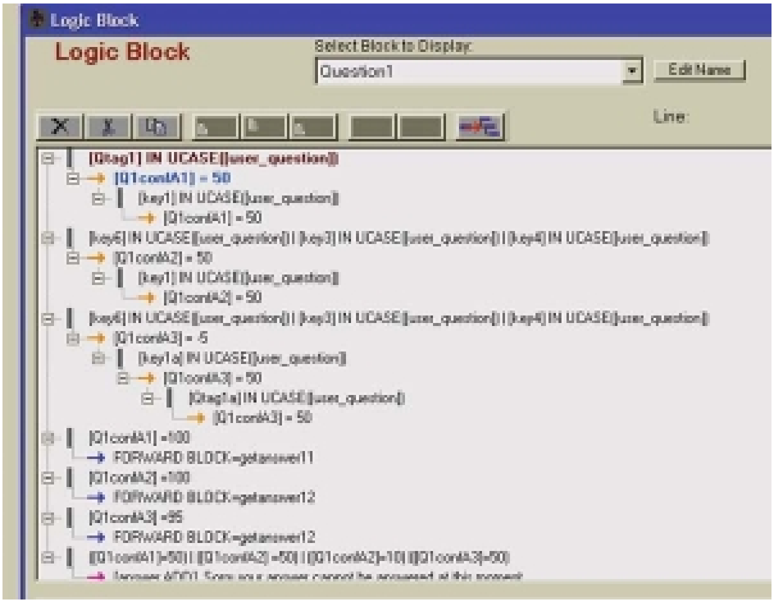


Fig. 28.6 System flowchart



```
[Qtag1] IN UCASE[[user_question]]
    [Q1conlA1] = 50
        [key1] IN UCASE[[user_question]]
            [Q1conlA1] = 50
[key6] IN UCASE[[user_question]] | [key3] IN UCASE[[user_question]] | [key4] IN UCASE[[user_question]]
    [Q1conlA2] = 50
        [key1] IN UCASE[[user_question]]
            [Q1conlA2] = 50
[key6] IN UCASE[[user_question]] | [key3] IN UCASE[[user_question]] | [key4] IN UCASE[[user_question]]
    [Q1conlA3] = -5
        [key1a] IN UCASE[[user_question]]
            [Q1conlA3] = 50
                [Qtag1a] IN UCASE[[user_question]]
                    [Q1conlA3] = 50
[Q1conlA1] =100
    FORWARD BLOCK=getanswer11
[Q1conlA2] =100
    FORWARD BLOCK=getanswer12
[Q1conlA3] =95
    FORWARD BLOCK=getanswer12
[[Q1conlA1]=50] | [[Q1conlA2] =50] | [[Q1conlA2]=10] [[Q1conlA3]=50]
    [answer ADD] Sorry your answer cannot be answered at this moment
```

Fig. 28.7 A typical logic block showing question 1 variations

answers that are related to the predefined question and ranks the answers based on the weighted sum. If none of the answers get a ranking of over 0 then the system outputs a no response reply. Otherwise the system outputs a list of answers ranked according to their relevancy to the question.

28.5 System Interactivity

A sample of twenty students was requested to interact with the system to verify its efficiency and accuracy. The diagram on Fig. 28.8 shows the user interface that will be presented when the user launches the system. User interaction screens are designed using HTML templates that allow the full power of HTML to be utilized to create highly interactive but easy to use interfaces. The diagram shows a sample of the interface where the user enters the query and clicks on the submit button. The user is allowed to enter the long question as they understand it and the system is then supposed to extract the keywords from it.

As shown in Fig. 28.8, the retrieved answer is sometimes the selection of a single answer or a list of possible answers attached with a degree of relevancy to the question asked arranged in their order of certainty. The program on request, can explain how it arrives at its conclusion. The input question is 'Define HIV' and on analysing it, using Fig. 28.4, we realise that these are amongst the FUW for ***A1***. Hence the system can convert these two words to become a keyword formed from the two FUW's,

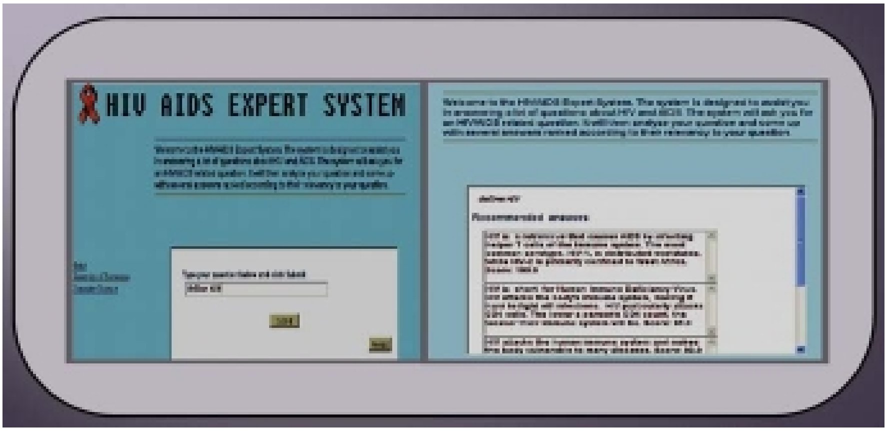


Fig. 28.8 User input and output interfaces

Define (21%) & HIV (90%), perhaps the most relevant keyword for this question. Therefore the first answer $A1$ is a perfectly relevant one with 100% certainty as shown in Fig. 28.8. The next retrieved answer has a measure of relevancy to the input question but less than ***A1*** and so on. More than 90% of the participants agreed that the expert system was not only easy to use but was also able to almost always provide a relevant answer to the input question. They also felt comfortable asking questions and felt the system could be very useful in their own HIV and AIDS health care support.

The results not only give more information to the input question but also show the relevancy of the answer to the question and hence asking a question can lead to other relevant important question that you did not ask for. Most importantly the system indicates that an HIV and AIDS question can be derived using keywords which the question represented accurately in most of the cases.

28.6 Conclusion

An Expert System has been developed and a sample of participants used to build the data and test the effectiveness of the system. The results show that the system has been developed successfully and found to be a good system to support the public with timely information on HIV and AIDS. However, we cannot rule out the possibility of some input data being outside the collected data, thus rendering the knowledge base incomplete and more likely not to provide the expected feedback to users in some cases. As a result, there is a need to simulate other variations of the FUW's not provided by the sample. We therefore recognize the need to develop a procedure to derive further guesses which could result in a great deal more guesses for expanding the knowledgebase. The system is also intended to be exhaustive and thus the knowledge base is going to expand phenomenally over time as various other sources of HIV information are used. We therefore recognize the potential research areas which include designing a more formal or algorithmic keyword extraction method for the questions.

References

1. Microsoft Digital Inclusion Program. http://research.microsoft.com/ur/us/fundingopps/rfps/DigitalInclusion_2005_RFP.aspx. Accessed 17 February 2009
2. IHISM Project website www.ub.bw/ihism. Accessed 18 February 2009
3. UNAIDS/WHO. Report on the Global AIDS Epidemic (2006)
4. Botswana National Strategic Framework for HIV AND AIDS, NACA (2003)
5. The NACA Project. http://www.naca.gov.bw/about.htm. Accessed 18 February 2009
6. African Comprehensive HIV AND AIDS Partnerships (ACHAP). http://www.achap.org/profile.html. Accessed 21 February 2009
7. Botswana Harvard Partnership. http://www.hsph.harvard.edu/bhp/. Accessed 21 February 2009
8. MASA – The National Antiretroviral Therapy Programme of Botswana, Botswana Ministry of Health. http://www.moh.gov.bw/index.php?id=192. Accessed 21 February 2009
9. IPOLETSE Call Centre for HIV AND AIDS, Botswana Ministry of Health. http://www.mri.co.bw/html/call-ipoletse.htm. Accessed 21 February 2009
10. EXYS CORVID Development Environment. http://www.technology-reports.com/company.asp?id=163. Accessed 21 February 2009
11. Agency for Healthcare Research and Quality (AHRQ). http://www.ahrq.gov/RESEARCH/hitfact.htm. Accessed 20 January 2009
12. Jackson, P.: MYCIN: Medical diagnosis using production rules. In: Introduction to Expert Systems. Addison-Wesley, Reading, MA, pp. 102–114 (1986)
13. Evans, S.: PACE-Then and Now. In: The Pace System: An Expert Consulting System For Nursing, pp. 26–32. Springer, Berlin (1996)
14. Matsopoulus, G.K., et al.: MITIS: A WWW-based medical system for managing and processing gynecological-obstetrical-radiological data. Comput. Methods Program. Biomed. **76**(1), 53–71 (October 2004)
15. Atalay, B., Potter, W.D., Haburchak, D.: HIVPCES: A WWW-based HIV patient care expert system. In: Computer-Based Medical Systems, 1999. 12th IEEE n, pp. 214–219 (1999)
16. Temesgen, Z., et al.: A comprehensive health enhancement support system (CHESS) for people with HIV infection. Springer, AIDS and Behavior **10**(1), 35–40 (January 2006)
17. Sloot, A.V., Boukhanovsky, K., Boucher, C.A.: A grid-based HIV expert system. IEEE Comput. Soc. **1**, 471–486 (2005)
18. Pazzani, M., et al.: CTSHIV: A Knowledge-Based System for the Management of HIV-Infected Patients, University of California (Irvine),(1999). http://ieeexplore.ieee.org/iel3/5136/13957/00645169.pdf?tp=\&arnumber=645169\&isnumber=13957. Accessed 20 January 2009
19. Giarratano, J.C., Riley, G.D.: Introduction To Expert Systems. Expert Systems, 4th edn. Thomson Course Technology, p. 19, 29, 32, 34 (2005)
20. Limesurvey, a survey application. http://www.limesurvey.org/. Accessed 17 January 2009

Chapter 29
A Multi-Objective Approach to Generate an Optimal Management Plan in an IMS-QSE

Ahmed Badreddine, Taieb Ben Romdhane, and Nahla Ben Amor

Abstract Recently, we have proposed a new process-based approach to implement an integrated management system grouping Quality, Security and Environment (QSE) avoiding weaknesses of parallel implementations. Our approach is based on three phases i.e. *Plan*, *Do* and *Check and Act*. This paper proposes an implementation of the most important part of the plan phase, consisting in the definition of an appropriate QSE management plan. The principle of this implementation is the transformation of already existing bow ties into a multi-objective influence diagram (MID) which is one of the most commonly used graphical decision models for reasoning under uncertainty. More precisely, we propose to map existing *bow ties* which are a very popular and diffused risks analysis tool into a MID, then to evaluate it in order to generate an optimal management plan.

Keywords Integrated management system · risk management · multi-objective influence diagram · bow tie

A. Badreddine (✉) and N.B. Amor
LARODEC, Institut Supérieur de Gestion de Tunis, 41 Avenue de la liberté,
2000 Le Bardo, Tunisie
e-mail: Badreddine.ahmed@hotmail.com; Nahla.benamor@gmx.fr

T.B. Romdhane
Institut National des Sciences Appliquées et de la Technologie, Centre Urbain Nord
BP 676 - 1080, Tunisie
e-mail: Benromdhane.t@planet.tn

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_29,

29.1 Introduction

The evolution of the current industrial context and the increasing of the competition pressure, led the companies to adopt new concepts of management. In this context, the implementation of international norms relative to different management systems became a real need and target for many organizations. In particular, the implementation of the three standards ISO 9001 [4], OHSAS 18001 [12] and ISO 14001 [5] relative to quality, security and environment, respectively, can be considered as a widespread phenomenon around the world. Nevertheless, the major difficulty of such an implementation is that these three management systems were proposed separately and thus their combination is not an obvious task since they have common and confused procedures. Thus, if they are adopted without any care about their interactivities, several weaknesses relative to duplicate management tasks suggested by the three standards (e.g. written procedures, checking, control forms, etc.) can be observed. Hence, proposing an integrated management system including quality, security and environment management systems also known as QSE management system have drawn the attention of both academics and practitioners. These researches studied the integration of the three systems from various viewpoints relative, essentially, to the definition of its success criteria [7–9, 16]. However, a few studies have developed effective methodologies and approaches.

Recently, we have proposed a new process-based approach to implement an integrated management system using three integration factors namely the *process approach*, the *risk management* and a *global monitoring system* [1]. This approach covers the whole PDCA scheme (i.e. Plan, Do, Check, Act) by gathering its steps into three phases such that the first one concerns the *Plan* phase, the second, the *Do* phase and the third the *Check* and the *Act* phases as illustrated by Fig. 29.1. This paper proposes an implementation of the most important part of the first phase, consisting in the definition of an appropriate QSE management plan. Our idea is to handle all quality, security and environment objectives issued from the requirements and the expectations of stakeholders (i.e. customers, employees, population, environment, etc.) through a multi-objective influence diagram (MIDs) [11] which is one of the most commonly used graphical decision models for reasoning under uncertainty with multiple objectives. More precisely, we propose to map existing *bow ties* which are a very popular and diffused risks analysis tool into a MID, then to evaluate it in order to generate an optimal management plan.

The remainder of this paper is organized as follows: Section 29.2 presents a brief recall on the new process based approach for implementing an IMS. Section 29.3 proposes a multi-objective approach to define an appropriate QSE management plan. Indeed, a transformation algorithm from existing *bow ties* into a multi-objective influence diagrams will be proposed. Finally Section 29.4 presents an illustrative example in the petroleum field involving a decision problem faced during the definition of a QSE management plan for the *TOTAL TUNISIA company*.

29.2 A Brief Recall on the New Process-Based Approach

This section presents a brief recall on the new process based approach for implementing an integrated Quality, Security and Environment management system. This approach is based on three integrated factors [1]: *Risk management* to increases the compatibility and the correspondence between the three systems, *Process-based approach* to deal with coordination and the interactions between the activities of a company, *Monitoring System* to ensure the monitoring of the global system and the integration as a continuous improvement of the performance.

The proposed approach is illustrated by Fig. 29.1, where the different steps cover the whole PDCA (Plan, Do, Check, Act) scheme. The idea here is to gather these steps into three phases such that the first one concerns the *Plan* phase, the second, the *Do* phase and the third the *Check* and the *Act* phases. These three phases can be detailed as follows [1]:

- **Plan Phase** This phase is composed of six steps: the first consists in setting up all quality, security and environment objectives issued from the requirements and the expectations of stakeholders (i.e. customers, employees, population,

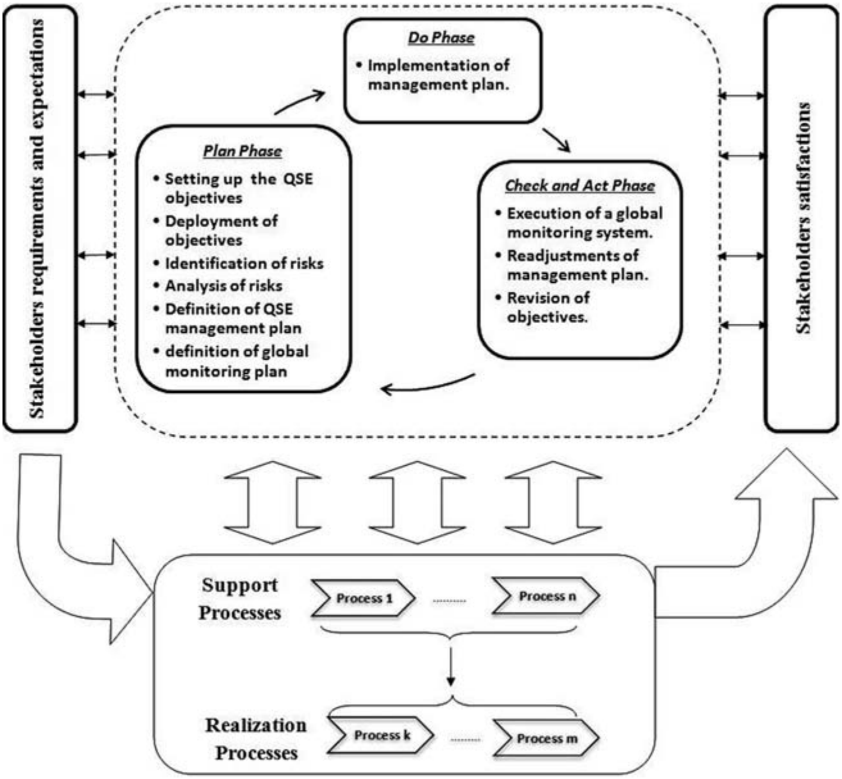

Stakeholders requirements and expectations
Do Phase
• Implementation of management plan.
Plan Phase
• Setting up the QSE objectives
• Deployment of objectives
• Identification of risks
• Analysis of risks
• Definition of QSE management plan
• definition of global monitoring plan
Check and Act Phase
• Execution of a global monitoring system.
• Readjustments of management plan.
• Revision of objectives.
Stakeholders satisfactions
Support Processes
Process 1
Process n
Realization Processes
Process k
Process m

Fig. 29.1 Proposed process-based approach for IMS [1]

environment, etc.). In the second, we will deploy all these objectives in each process. The third step consists in the analysis of each process with respect to the pre-set objectives defined in the second one in order to identify the sources of hazard and possible targets leading to a possible failure to reach up the objectives. In the fourth step, each identified risk has to be analyzed in term of potential consequences in each management area. In the fifth step we have to define a QSE management plan to implement selected treatments as preventive and corrective actions, in order to reduce levels of risks already identified and to improve the efficiency of the IMS. To this end, we have to consider the interaction between the different management areas, indeed some decisions can be beneficial for some management areas and harmful for others. Finally, the sixth step is devoted to the definition of an appropriate monitoring plan, in order to ensure the well implementation of the QSE management plan.

- **Do Phase** This phase has as input the QSE management plan and the corresponding global monitoring plan generated from the *plan* phase and will implement the selected treatments. Note that we have to define the appropriate Scheduling to optimize the resources in order reach up the objectives more efficiently.
- **Check and Act Phase** Once the do phase achieved, this phase will finalize the process of integration by the measure of the effectiveness of different decisions and their readjustments via three steps. In the first one, we have to measure all the indicators already defined in order to evaluate the effectiveness of selected treatments and to estimate the degree of achievement of objectives. For this reason, we have to aggregate the indicators of each objective. In the second step, a readjustment of the management plan will be done in order to satisfy unreached objectives. Although, some objectives may not be reached, that is why we should revise some of the initial assigned objectives in order to make their satisfaction possible, in this context we propose the third step (i.e. revision of objectives) in order to contribute to sustainable development.

29.3 A Multi-Objective Approach for a QSE Management Plan

In this section, we propose an implementation of the most important part of the *Plan* phase consisting in the definition of an appropriate QSE management plan . In fact, our idea is to use the risk management as integrating factor and to consider the different interactions between policies, objectives and resources of the quality, security and environment standards. Several approaches for risk evaluation exist, within the most famous ones we can mention, *preliminary risk analysis* (APR), *hazard operability* (HAZOP), *failure mode and effects analysis* (FMEA), and tree-based technique such that *fault tree analysis*, *event tree analysis* and *bow tie analysis*. Unfortunately, these methods are not appropriate to deal with many management areas simultaneously and they are usually limited to technical level. Moreover, since 2003 it is necessary to respect the law 2003-699 [6] relative to the introduction of probability concepts in any risk analysis which is not the case of all these tools. In the literature some researches has been carried out to take into account this law. Most of these researches are based on tree-based techniques which offer a flexible structure to be used with probability concepts. Moreover, several approaches concerning the introduction of probabilistic concepts with risk analysis are particularly focalized on *Bayesian networks* which are a popular tool for representing uncertainty in artificial intelligence [14]. These approaches can be divided into three classes:

- In the first class the principle is to *transform* a risk analysis tool into a Bayesian network. This idea was first introduced by Bibbio et al. [2] which propose a mapping from fault tree analysis into Bayesian networks. In the same context, léger et al. [10] propose to extend the technical bow tie analysis to a global system, including human beings and organizations.
- The principle of the second class is the *fusion* of a risk analysis tool and a Bayesian network. We can mention in particular the work of Trucco et al. [15] where Bayesian networks are used as an extension of the fault tree in order to introduce the social activity in the evaluation of the latter.
- The third class does not require any risk analysis tools. In fact each identified risk will be directly modeled by a Bayesian network as proposed by Palaniappan [13].

The first problem with these methods in that they deal with a unique management area, so they cannot be applied in the context of a fully integrated management system. Moreover, the fact that these methods are based on Bayesian networks presents a real weakness since this graphical model is not really appropriate to generate optimal decisions. In fact, the powerful of Bayesian networks consists in their ability in reasoning under uncertainty and not in decision making area. For this reason, several extensions where proposed in order to extend them to the decisional aspect. Thus, our objective is to model a more efficient risk management tool by using an appropriate graphical decisional model. More precisely, we propose to use *influence diagrams* which are an extensions of Bayesian networks able to provide optimal solutions while maximizing decision makers utilities. Moreover, given the multi-objective aspect of our problem, we will use *multi-objective influence diagrams* (MIDs) which are a new variant of influence diagrams dedicated to such a problems. Thus our idea is to map existing *bow ties* which are a very popular and diffused risks analysis tool into a MID, then to evaluate it in order to generate an appropriate QSE management plan. Before detailing our approach we propose a brief recall on bow tie analysis and multi-objective influence diagrams.

29.3.1 Bow Tie Method

The bow tie method is a very popular and diffused probabilistic technique developed by shell for dependability modeling and evaluation of large safety–critical systems. The principle of this technique is to built for each identified risk R_i (also called *top event* (TE)) a bow tie representing its whole scenario on the basis of two parts, as

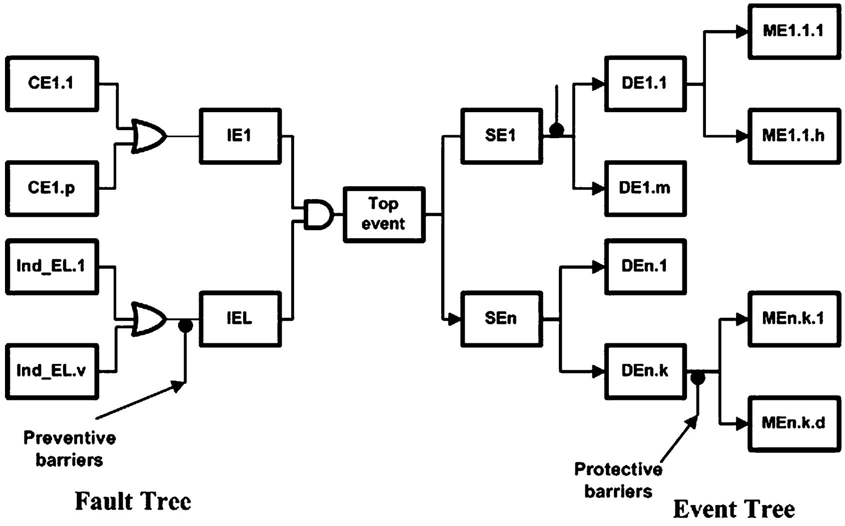


Fig. 29.2 A bow tie analysis model

shown in Fig. 29.2: The first part corresponds to the left part of the scheme which represents a *fault tree* defining all possible causes leading to the (TE). These causes can be classified into two kinds: the first are the initiator events (IE) which are the principal causes of the TE, and the second are the undesired or critical events (IndE and CE) which are the causes of the IE. The construction of the left part proceeds in top down manner (from TE to IndE and CE). The relationships between events and causes are represented by means of logical AND and OR gates. The second part corresponds to the right part of the scheme which represents an *event tree* to reach all possible consequences of the TE. These consequences can be classified into three kinds: second events (SE) which are the principal consequences of the TE, dangerous effects (DE) which are the dangerous consequences of the SE and finally majors events (ME) of each DE. The construction of the event tree proceeds as the fault tree i.e. in top down manner. The bow tie also allows to define in the same scheme the *preventive barriers* to limit the occurrence of the TE and the *protective barriers* to reduce the severity of its consequences. In spite its widely use in many organizations, this method remains limited by its technical level and by the graphical presentation of different scenarios without any suggestion about optimal decisions regarding the objectives expected.

29.3.2 Multi-Objective Influence Diagrams

Influence diagrams (IDs), initially proposed by Howard and Matheson [3], are within most commonly used graphical decision models for reasoning under uncertainty. Their success is due their clarity and their simplicity since their topology (chance node, value node and decision node) is easily comprehensible by decision makers. Moreover their evaluation provides the optimal solutions while maximizing the decision makers utilities. Formally, an influence diagram has two components:

1. **Graphical Component** (or qualitative component) is a directed acyclic graph (DAG) denoted by $G = (N, A)$ where A is the set of arcs in the graph and N its node set. The node set N is partitioned into subsets C, D and V such that:

   - $C = \{C_1, \ldots, C_n\}$ is a set of chance nodes which represent relevant uncertain factors for decision problem. Chance nodes are represented by circles.
   - $D = \{D_1, \ldots, D_m\}$ is a set of decision nodes which depict decision options. These nodes should respect a temporal order. Decision nodes are represented by rectangles.
   - $V = \{V_1, \ldots, V_k\}$ is a set of value nodes which represent utilities to be maximized, they are represented by lozenges.

   Arcs in A have different meanings according to their targets. We can distinguish *Conditional arcs* (into chance and value nodes), those that have as target chance nodes represent probabilistic dependencies and *Informational arcs* (into decision nodes) which imply time precedence.

   Influence diagrams are required to satisfy some constraints to be *regular*, in particular value nodes cannot have children and there is a directed path that contains all of the decision nodes. As a result of this last constraint, influence diagrams will satisfy the *no-forgetting* property in the sense that a decision node and its parents should be parents to all subsequent decision nodes.
2. **Numerical Component** (or quantitative component) consists in evaluating different links in the graph. Namely, each conditional arc which has as target a chance node C_i is quantified by a conditional probability distribution of C_i in the context of its parents. Such conditional probabilities should respect the probabilistic normalization constraints. Chance nodes represent uncertain variables characterizing decision problem. Each decision alternative may have several consequences according to uncertain variables. The set of consequences is characterized by a utility function. In IDs, consequences are represented by different combinations of value node's parents. Hence, each value node is quantified by a utility function, denoted by U, in the context of its parents. The definition of the numerical component is in general done by experts and decision makers.

Once the ID constructed it can be used to identify the optimal policy, this can be ensured via evaluation algorithms which allow to generate the best strategy yielding to the highest expected utility. Standard IDs are usually limited to single objectives or a combined one. Recently, they have been extended to deal with multiple objectives decision problems (MIDs) [11] by gathering different objectives in a unique value node. To evaluate such diagrams Micheal et al. [11] have proposed a direct evaluation algorithm based on arc reversal and node deletion. This algorithm can be outlined as follows:

Algorithm 1: Direct evaluation of MID

```
Data: MID
Result: Optimal decision regarding the objectives
begin
    1. Check the regular property of MID.
    2. Remove barren nodes (i.e. nodes without successors).
    3. If a chance node exists with the value node as its sole successor then remove it and
    update the utility function of the value node.
    4. If any node remains in the diagram then return to step 3 otherwise terminate the
    algorithm.
    5. If there is a decision node which is a direct predecessor of the value node such that the
    remaining predecessors of the value node are informational predecessors of the decision
    node, then:
    - remove it,
    - update the utility function of the value node,
    - remove any barren node.
    If any node remains in the diagram then return to step 3 otherwise terminate the algorithm.
    6. Find a chance node i which is a direct predecessor to the value node such that it has no
    decision node as successor.
    7. Find a chance node j which is a direct successor of i such that there is no other directed
    path between i and j and reverse the arc between i and j. If i has any other successors repeat
    step 6.
    8. Remove the chance node i with the arc reversal transformation (probability table trans-
    formation).
    9. If any node remains in the diagram return to step 3 otherwise terminate the algorithm.
end
```

29.3.3 Transformation of Bow Ties into a MID

In order to generate the optimal QSE management plan satisfying all the objectives, we propose a mapping from existing bow ties to a multi-objective influence diagram. In fact, our idea is to gather all the QSE required objectives in the same value node, then each identified risk and its respective scenario occurrence from initiators to final consequences will represent a chance node, and finally the barriers (i.e. preventive and corrective) will be mapped as decision nodes in order to define the appropriate QSE management plan. Once this building phase achieved, we should quantify the resulted multi-objective influence diagram as explained previously. To this end, we propose a transformation procedure (i.e. Algorithm 2) ensuring an automatic transformation from existing bow ties to an alternative model (MID) that allows the generation of optimal strategies.

Let $BT_1..BT_n$ the set of bow ties and $O_1..O_k$ the set of objectives. Let R_i be top event of BT_i and F_i be its occurrence. Let IE_i (resp. $CE_i, IndE_i, SE_i, DE_i, ME_i$) be the set of initiator (resp. critical, undesired, second, dangerous, majors) events in BT_i. Let Cq_i (resp. Cs_i, Ce_i) be the consequence on quality (resp. security, environment) in BT_i. Let X_i and Y_i be any set of events in BT_i, then $Ar(X_i, Y_i)$ is a function which returns the set of arcs relative to all links between X_i and Y_i in BT_i. For instance $Ar(IE_i, CE_i)$ is the set of arcs relative to all links between IE_i and

CE_i in BT_i. Let $ArCq_i$ (resp. $ArCs_i$, $ArCe_i$) the set of major events which have a possible links to Cq_i (resp. Cs_i, Ce_i) in BT_i. Let $PreB_i$ (resp. $ProB_i$) be the set of *preventive* barriers (resp. *protective*) barriers in BT_i. Let $PE(.)$ (resp. $SE(.)$) be a function which returns the set of *precedent* (res. *successive*) events of any barrier in BT_i. Let D the set of all barriers. Let $ArpB$ the set of additional arcs relative to the links between each element of D to each event. Let *ord* be the order relative to different decision nodes relative to existing barriers in $BT_1..BT_n$, this order can be defined by experts. Let nb(.) be a function returning the number of elements of a given set. Algorithm 2 outlines the major steps of our approach.

Algorithm 2: Transformation of bow ties into a *regular.* MID

```
Data: $BT_1..BT_n$; $O_1..O_k$; $ArCq_1..ArCq_n$; $ArCs_1..ArCs_n$; $ArCe_1..ArCe_n$;$ArpB$; $ord$
Result: MID
begin
    Building phase: - $C \leftarrow \emptyset, D \leftarrow \emptyset, V \leftarrow \emptyset, A \leftarrow \emptyset$
    - Gather all the QSE objectives $O_i$ (i=1..k) in the same value node $V_{QSE}$
    - $V \leftarrow V_{QSE}$
    for $i \leftarrow 1..n$ do
        % Create $R_i$ and $F_i$ and connect them
        $C \leftarrow C \cup R_i \cup F_i$
        $A \leftarrow A \cup (R_i \rightarrow V_{QSE}) \cup (F_i \rightarrow R_i)$
        % Create all the events and connect them
        $C \leftarrow C \cup IE_i \cup CE_i \cup IndE_i \cup SE_i \cup DE_i \cup ME_i$
        $\forall IE_{ij} \in IE_i, A \leftarrow A \cup (IE_{ij} \rightarrow F_i)$
        $\forall SE_{ij} \in SE_i, A \leftarrow A \cup (F_i \rightarrow SE_{ij})$
        $A \leftarrow A \cup Ar(IE_i, CE_i) \cup Ar(IE_i, IndE_i) \cup Ar(SE_i, DE_i) \cup Ar(DE_i, ME_i)$
        % Create $Cq_i$, $Cs_i$, $Ce_i$ and connect them
        $C \leftarrow C \cup Cq_i \cup Cs_i \cup Ce_i$,
        $A \leftarrow A \cup (Cq_i \rightarrow R_i) \cup (Cs_i \rightarrow R_i) \cup (Ce_i \rightarrow R_i)$
        $\forall ArCq_{ij} \in ArCq_i, A \leftarrow A \cup (ArCq_{ij} \rightarrow Cq_i)$
        $\forall ArCs_{ij} \in ArCs_i, A \leftarrow A \cup (ArCs_{ij} \rightarrow Cs_i)$
        $\forall ArCe_{ij} \in ArCe_i, A \leftarrow A \cup (ArCe_{ij} \rightarrow Ce_i)$
        % Handel barriers
        $D \leftarrow D \cup PreB_i \cup ProB_i$
        $\forall PreB_{ij} \in PreB_i, \forall ProB_{ij} \in ProB_i,\ A \leftarrow A \cup (PreB_{ij} \rightarrow PE(PreB_{ij})) \cup (ProB_{ij} \rightarrow SE(ProB_{ij}))$
    % Additional links $A \leftarrow A \cup ArpB$
    % Connect decision nodes while respecting the precedence order. $n_1 \leftarrow nb(D)$
    for $k \leftarrow 1..(n_1 - 1)$ do
        for $l \leftarrow (k+1)..n_1$ do
            $A \leftarrow A \cup (D_{ord(k)} \rightarrow D_{ord(l)})$
    Quantification phase: Assign the numerical values for each node in the MID.
end
```

It is important to note that this algorithm provides a regular influence diagram satisfying the no-forgetting property.

29.4 Case Study

The case study presented in this section, has been released in the petroleum field. This application involves a decision problem faced during the definition of a QSE management plan for *TOTAL TUNISIA company* which is certified in quality, security and environment management systems. Due to the lack of space we will only consider three objectives (O_1: Gain market share by providing superior all-round service to the customer, O_2: Minimize the environmental waste and O_3: Increase safety staff) and two risks (R_1: A major fire and explosion on tanker truck carrying hydrocarbon and R_2: A fire in container).

The first step is to proceed to the bow tie analysis of the two identified risks (i.e. R_1 and R_2) as shown in Figs. 29.3 and 29.4. Note that in BT_1 we have five preventive barriers (i.e. *Periodic preventive maintenance tank valve* (PMV), *Periodic preventive maintenance to minimize exhaust failure* (PME), *Education and Training Task* (ETT), *Prohibition to park the trucks close the site after loading* (PPT), and *Fire simulation* (FS)) and four protective barriers (i.e. *A fix or tractable canal to prevent incident along the site* (FTC), *Blast protection window film* (BPW),

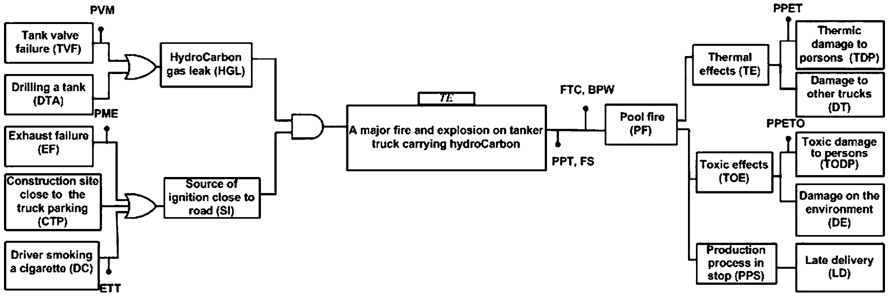

PVM
Tank valve failure (TVF)
Drilling a tank (DTA)
PME
Exhaust failure (EF)
Construction site close to the truck parking (CTP)
Driver smoking a cigarette (DC)
ETT
HydroCarbon gas leak (HGL)
Source of ignition close to road (SI)
TE
A major fire and explosion on tanker truck carrying hydroCarbon
FTC, BPW
PPT, FS
Pool fire (PF)
PPET
Thermal effects (TE)
Thermic damage to persons (TDP)
Damage to other trucks (DT)
PPETO
Toxic effects (TOE)
Toxic damage to persons (TODP)
Damage on the environment (DE)
Production process in stop (PPS)
Late delivery (LD)

Fig. 29.3 A bow tie analysis of R_1

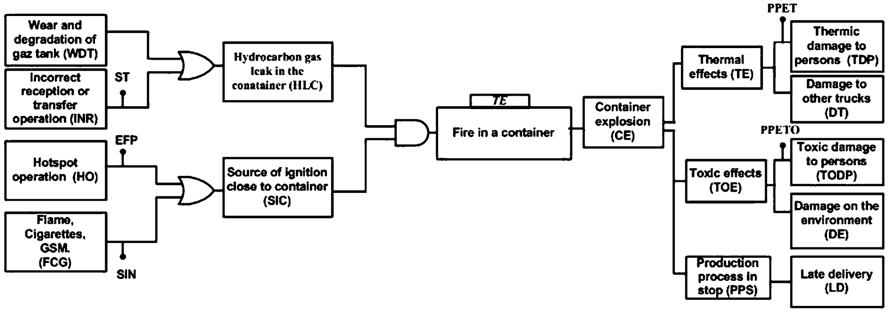

Wear and degradation of gaz tank (WDT)
Incorrect reception or transfer operation (INR)
ST
EFP
Hotspot operation (HO)
Flame, Cigarettes, GSM. (FCG)
SIN
Hydrocarbon gas leak in the conatainer (HLC)
Source of ignition close to container (SIC)
TE
Fire in a container
Container explosion (CE)
PPET
Thermal effects (TE)
Thermic damage to persons (TDP)
Damage to other trucks (DT)
PPETO
Toxic effects (TOE)
Toxic damage to persons (TODP)
Damage on the environment (DE)
Production process in stop (PPS)
Late delivery (LD)

Fig. 29.4 A bow tie analysis of R_2

Personal Protective equipment to limit thermal effects (PPET) and *Personal Protective equipment to limit toxic effects* (PPETO)). In the same way, in BT_2 we have three preventive barriers (i.e. *Establish fire permit* (EFP), *Setting instructions* (SIN) and *Successive training* (ST)) and two protective barriers (i.e. *Personal Protective equipment to limit thermal effects* (PPET) and *Personal Protective equipment to limit toxic effects* (PPETO)), Personal Protective equipment to limit toxic effects.

Once the bow tie analysis achieved, we will apply our transformation procedure (i.e. Algorithm 2) with the following input data:

- BT_1, BT_2, O_1, O_2, O_3
- $ArCq_1 = ArCq_2 = \{LD, DT\}$ since *Late delivery* (LD) and *Damage to trucks* (DT) have consequences on quality
- $ArCs_1 = ArCs_2 = \{TDP, TODP\}$ since *Toxic damage to persons* (TDP) and *Thermic damage to persons* (TODP) have consequences on security
- $ArCe_1 = ArCe_2 = \{DE, DT\}$ since *Damage on the environment* (DE) and *Damage to trucks* (DT) have consequences on the environment
- The additional arcs defined in $ArpB$ are (FS, Ce_1), (ST, TVF) and (ST, Cs_2) since *Fire simulation* (FS) is considered as pollutant for the environment (Ce_1), *Successive trainings* (ST) can increase *Tank valve failure rates* (TVF) and successive trainings (ST) can have an impact on security (Cs_2)
- In order to respect the precedence order relative to different decision nodes relative to existing barriers in BT_1 and BT_2 (i.e. PVM, PME, ETT, PPT, FS, FTC, BPW, PPET, PPETO, ST, EFP, SIN), we will consider *ord* $= \{6, 4, 5, 3, 2, 7, 8, 1, 9, 10, 11, 12\}$.

The resulted MID is represented by Fig. 29.5. Then, we should proceed to the *quantification phase*. For the lack of space we cannot give numerical data here (for instance the table relative to V_{QSE} contains $10^2 + 2^3$ entries since we have 10 binary and 2 ternary decision nodes). Once the transformation achieved, we can apply the evaluation algorithm proposed by [11]. The final output of this algorithm is the optimal decision satisfying all the objectives while maximizing decision makers utilities. This decision corresponds to the QSE management plan. For our illustrative example the optimal decision is the following: PPET=T,FS=R, PPT=T, PME=T, ETT=T, PMV=T, FTC=T, BPW=T, PPETO=T, ST=R, EFP=T, SIN=T. It is clear that if we limit our analysis to BT_1 and BT_2, we cannot define the appropriate management plan regarding all the objectives. This is not the case with the resulted MID since its evaluation enabled us to generate the appropriate management plan satisfying all the QSE objectives while maximizing decision makers utilities.

29.5 Conclusion

This paper proposes a first implementation of a new-process based approach for integrating Quality, Security and Environment management systems that we have proposed in [1]. This implementation concerns the most important part of the plan

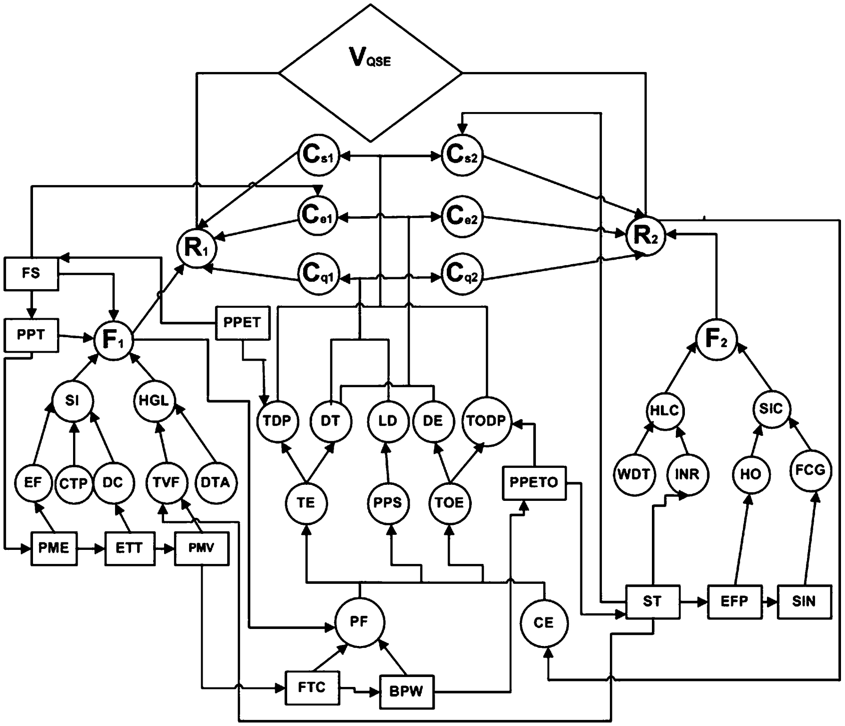


Fig. 29.5 The resulted MID

phase relative to the elaboration of an appropriate QSE management plan. This implementation is based on the transformation of existing bow ties into a multi-objective influence diagram. This choice was motivated by the fact that bow ties are very popular and diffused risk analysis tools allowing to define in the same scheme the whole scenario from initiators events to finale consequences. Moreover, it defines all the possible actions and decisions as preventive and corrective barriers to reduce the occurrence and the severity of each risk identified. Also the multi-objective influence diagram are one of the most appropriate graphical decision models for reasoning under uncertainty in addition to the fact that they allow the manipulation of different objectives which feats well with our problem since we deal with the three standard QSE. To obtain the optimal and appropriate QSE management plan, we have proposed a transformation procedure (i.e. Algorithm 2) to provide an automatic transformation from the bow ties model to an alternative model (MID) that facilitates the calculation of optimal strategies. This implementation will directly affect the remaining parts of our integration system since it will provide the QSE management plan, which should be executed in the Do phase. As a future work we propose to implement a whole decision support system relative to our process-based approach.

References

1. Badreddine, A., Ben Romdhane, T., Ben Amor, N.: A new process-based approach for implementing an integrated management system: quality, security, environment. In: Proceedings of the International Conference on Industrial Engineering, IMECS 2009, pp. 1742–1747 (2009)
2. Bobbio, A., Portinale, L., Minichino, M., Ciancamerl, E.: Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. Reliab. Eng. Syst. Safe. **71**, 249–260 (2001)
3. Howard, R.A., Matheson, J.E.: Influence diagrams. The Principles and Applications of Decision Analysis (1984)
4. ISO 9001:2008: Quality managment system. Requirements. ISO (2008)
5. ISO 14001:2004: Environmental managment system. Requirements with guidance for use. ISO (2004)
6. J0 175: Law n 2003-699 concerning the prevention of technological and natural risks, and to the repair of damage. Off. J. July 30 (2003)
7. Jorgensen, T.H., Remmen, A., Mellado, M.D.: Integrated management systems - three different levels of integration. J. Cleaner Prod. **14**, 713–722 (2006)
8. Jorgensen, T.H.: Towards more sustainable management systems: through lide cycle managment and integration. J. Cleaner Prod. **16**, 1071–1080 (2008)
9. Labodova, A.: Implementing integrated management systems using a risk analysis based approach. J. Cleaner Prod. **12**, 571–580 (2004)
10. Léger, A., Duval, C., Weber, P., Levrat, E., Farret, R.: Bayesian network modelling the risk analysis of complex socio technical systems. In: 4th Workshop on Advanced Control and Diagnosis, France (2006)
11. Micheal, D., Yacov, Y.H.: Influence diagrams with multiple objectives and tradeoff analysis. IEEE Trans. Syst. Man Cyb. **34**, 293–304 (2004)
12. OHSAS 18001:2000: Occupational health and safety management systems-specification. BSI: British Standard Institution (2000)
13. Palaniappan, R.: Bayesian networks: application in safety instrumentation and risk reduction. ISA Trans. **46**, 255–259 (2007)
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, Los Altos (1989)
15. Trucco, P., Cagno, E., Ruggeri, F., Grande, O.: A bayesian belief network modelling of organisational factors in risk analysis: a case study in maritime transportation. Reliab. Eng. Syst. Safe. **93**, 845–856 (2008)
16. Zeng, S.X., Shi, J.J., Lou, G.X.: A synergetic model for implementing an integrated management system: an empirical study in China. J. Cleaner Prod. **15**, 1760–1767 (2007)

Chapter 30
Topological Spatial Relations for Circular Spatially Extended Points: An Overview

Maribel Yasmina Santos and Adriano Moreira

Abstract Spatial relations between geographical objects include topological spatial relations, that are relations that are preserved under continuous transformations of the space, such as rotation or scaling. Research on topological spatial relations between different types of objects (points, lines, or regions) has been undertaken for many years. In this chapter, it is presented the particular case of the topological spatial relations that can exist between a Circular Spatially Extended Point and a region, a Circular Spatially Extended Point and a line, and two Circular Spatially Extended Points. For the several topological spatial relations, the geometrical representation of the relations is also presented.

Keywords Qualitative spatial reasoning · topological spatial relations · circular spatially extended point

30.1 Introduction

Human beings use qualitative identifiers extensively to simplify reality and to perform spatial reasoning more efficiently. Spatial reasoning is the process by which information about objects in space and their relationships are gathered through measurement, observation or inference and used to draw valid conclusions regarding the relationships between the objects [1]. Qualitative spatial reasoning [2] is based on the manipulation of qualitative spatial relations.

Spatial relations between geographical objects have been classified into several types [3, 4], including direction relations [5], distance relations [6] and topological relations [7]. Topological relations are those spatial relations preserved under continuous transformations of the space, such as rotation or scaling.

M.Y. Santos (✉) and A. Moreira
Information Systems Department, Algoritmi Research Centre, University of Minho,
Campus de Azurém, 4800–058 Guimarães, Portugal
e-mail: maribel@dsi.uminho.pt; adriano@dsi.uminho.pt

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_30,

Research on topological spatial relations between different types of objects (points, lines and regions) has been undertaken for many years, identifying the topological spatial relations between them, and proving the existence of such relations by demonstrating their geometric realization. Some of the works undertaken so far include the identification of the topological spatial relations between regions [8], between lines [8], between regions and lines [8,9], between regions with broad boundaries [10], between a spatially extended point and a region [11], between broad lines [12], and between lines with broad boundaries [12], only to mention a few.

This chapter presents an overview of the particular case of the topological spatial relations that include a Circular Spatially Extended Point (CSEP), namely the relations that exist between a CSEP and a region, a CSEP and a line, and between two CSEPs. A CSEP is a region-like object characterized by the inclusion of a point and a region that defines the area of influence of that point (see Fig. 30.1). A CSEP represents a complex object in the sense that the point and its region of influence are not dissociable.

The relevance of identification of such topological spatial relations is associated with the need to conceptualize the spatial relations that can exist among several objects in the geographical space. The obtained models can be used as a computational framework for spatial reasoning. Their implementation in a system, like a Geographical Information System, allows the representation and manipulation of complex objects associated with complex realities.

This chapter is organized as follows. Section 30.2 gives a brief overview of the concepts associated with qualitative spatial reasoning and a description of the several types of spatial relations: direction, distance and topological. Section 30.3 presents the topological spatial relations that exist between a CSEP and a region. Section 30.4 describes the topological spatial relations existing between a CSEP and a line, while Section 30.5 summarizes the topological spatial relations that can exist between two CSEPs. Section 30.6 concludes this chapter summarizing the presented work.

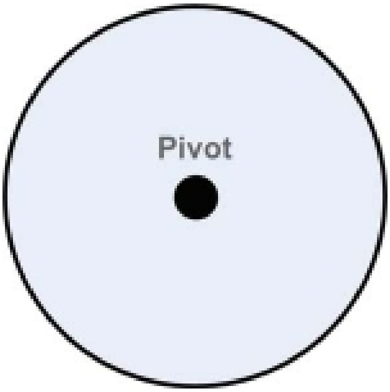


Fig. 30.1 A circular spatially extended point

30.2 Qualitative Spatial Reasoning

Qualitative spatial reasoning [2] is based on the manipulation of qualitative spatial relations, for which composition tables facilitate reasoning, thereby allowing the inference of new spatial knowledge. Geographical Information Systems allow for the storage of geographic information and enable users to request information about geographic phenomena. If the requested spatial relation is not explicitly stored in the database, it must be inferred from the available information. The inference process requires searching relations that can form an *inference path* between the two objects where the relation is requested [13]. The composition operation combines two contiguous paths in order to infer a third spatial relation. A composition table integrates a set of inference rules used to identify the result of a specific composition operation. To be possible the definition of composition tables, the spatial relations that can exist between the objects in analysis must be identified. This chapter is concerned with a particular type of spatial relation – topological spatial relations, and with a special type of object – a CSEP. However, next sections give an overview of the main concepts associated with the different types of spatial relations.

30.2.1 Direction Spatial Relations

Direction relations describe where objects are placed relative to each other. Three elements are needed to establish an orientation: two objects and a fixed point of reference (usually the North Pole) [3, 5]. Cardinal directions can be expressed using numerical values specifying degrees (0°, 45°...) or using qualitative values or symbols, such as North or South, which have an associated acceptance region. The regions of acceptance for qualitative directions can be obtained by projections (also known as half-planes) or by cone-shaped regions (see Fig. 30.2).

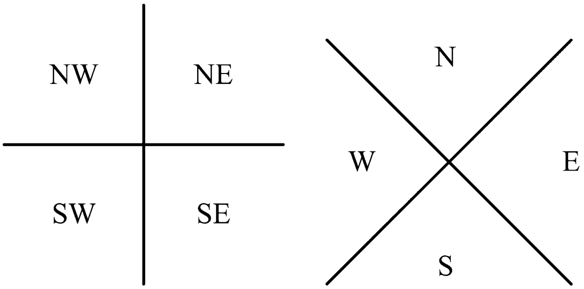


Fig. 30.2 Direction relations by projection and cone-shaped systems

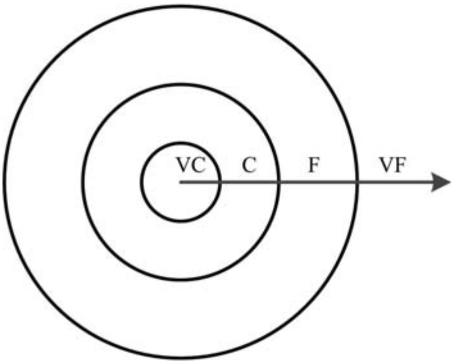


Fig. 30.3 Qualitative distances

30.2.2 Distance Spatial Relations

Distances are quantitative values determined through measurements or calculated from known coordinates of two objects in some reference system. The most frequently used definition of distance can be achieved using the Euclidean geometry and Cartesian coordinates. In a two-dimensional Cartesian system, it corresponds to the length of the shortest possible path (a straight line) between two objects, which is also known as the Euclidean distance [13]. Qualitative distances must correspond to a range of quantitative values specified by an interval. The adoption of the qualitative distances *very close* (VC), *close* (C), *far* (F) and *very far* (VF), intuitively describe distances from the nearest to the furthest [6] (see Fig. 30.3).

30.2.3 Topological Spatial Relations

Topological relations are those relationships that are invariant under continuous transformations of space such as rotation or scaling. There are eight topological relations that can exist between two planar regions without holes: *disjoint*, *covered by*, *covers*, *inside*, *contains*, *meet*, *equal* and *overlap* (see Fig. 30.4). These relations can be defined considering the intersections between the two regions, their boundaries and their complements [7].

30.3 Topological Spatial Relations Between a CSEP and a Region

The study of the topological spatial relations existing between a CSEP and a region was addressed by Lee and Flewelling [11]. In their work, a Spatially Extended Point is defined as a point-object that have a region of influence associated with

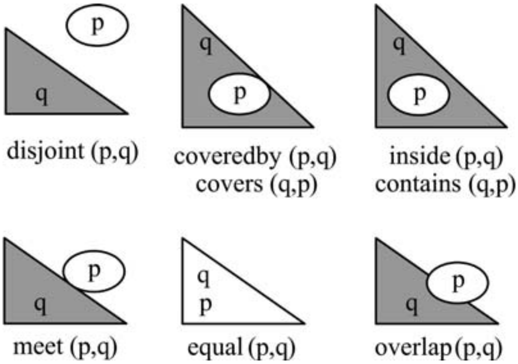


Fig. 30.4 Topological spatial relations between regions

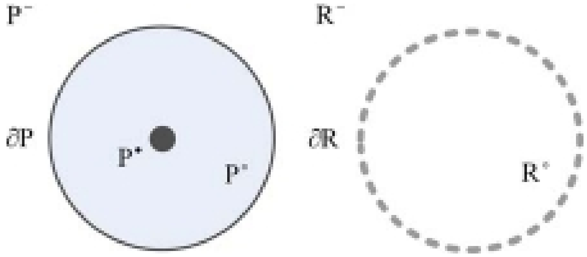


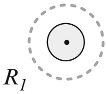
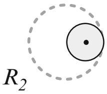
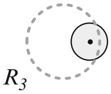
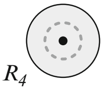
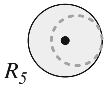
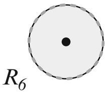
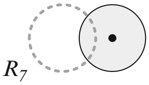
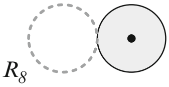
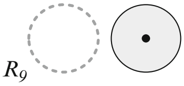
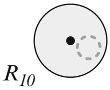
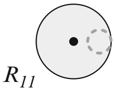
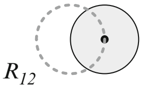
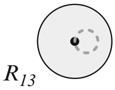
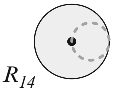
Fig. 30.5 Components of a CSEP and a Region

it and that also moves with it. The point that integrates the Spatially Extended Point and its associated region forms an object of hybrid dimensionality exhibiting more spatial relations than the union of point-region and region-region relations [11]. For the identification of the topological spatial relations that exist between a Spatially Extended Point, here assumed as a CSEP, and a region, the point-set (or 9-intersection) approach was used [7, 8].

A CSEP (P) has its own interior (P°), boundary (∂P) and exterior (P^{-}). While it shares the same concepts of interior (R°), boundary (∂R) and exterior (R^{-}) of a region (R), the CSEP is distinguished from a general region by the identification of a point within the interior called the pivot ($P^{\bullet}$). The pivot is conceptually similar to a 0-dimension object (see Fig. 30.5).

The topological spatial relations were identified using a 4 × 3 matrix [11]. The 4 × 3 matrix identifies the intersections ($\cap$) between the pivot ($P^{\bullet}$), interior (P°), boundary (∂P) and exterior (P^{-}) of a CSEP (P) and the interior (R°), boundary (∂R) and exterior (R^{-}) of a region (R). Each relation (R) between a CSEP (P) and a region (R) is characterized by 12 (4 × 3) intersections with empty ($\oslash$) or non-empty ($\neg\ \oslash$) values depending on how the geographical objects are related (Eq. 30.1). Fourteen spatial relations between a CSEP and a region were identified [11]. The analysis of the geometric realization [11] of the several topological spatial relations

Table 30.1 Topological spatial relations between a CSEP and a Region

Relation	Intersection matrix	Relation	Intersection matrix
	$\begin{bmatrix}1&0&0\\1&0&0\\1&0&0\\1&1&1\end{bmatrix}$		$\begin{bmatrix}1&0&0\\1&0&0\\1&1&0\\1&1&1\end{bmatrix}$
	$\begin{bmatrix}1&0&0\\1&1&1\\1&1&1\\1&1&1\end{bmatrix}$		$\begin{bmatrix}1&0&0\\1&1&1\\0&0&1\\0&0&1\end{bmatrix}$
	$\begin{bmatrix}1&0&0\\1&1&1\\0&1&1\\0&0&1\end{bmatrix}$		$\begin{bmatrix}1&0&0\\1&0&0\\0&1&0\\0&0&1\end{bmatrix}$
	$\begin{bmatrix}0&0&1\\1&1&1\\1&1&1\\1&1&1\end{bmatrix}$		$\begin{bmatrix}0&0&1\\0&0&1\\0&1&1\\1&1&1\end{bmatrix}$
	$\begin{bmatrix}0&0&1\\0&0&1\\0&0&1\\1&1&1\end{bmatrix}$		$\begin{bmatrix}0&0&1\\1&1&1\\0&0&1\\0&0&1\end{bmatrix}$
	$\begin{bmatrix}0&0&1\\1&1&1\\0&1&1\\0&0&1\end{bmatrix}$		$\begin{bmatrix}0&1&0\\1&1&1\\1&1&1\\1&1&1\end{bmatrix}$
	$\begin{bmatrix}0&1&0\\1&1&1\\0&0&1\\0&0&1\end{bmatrix}$		$\begin{bmatrix}0&1&0\\1&1&1\\0&1&1\\0&0&1\end{bmatrix}$

allows us to identify and present in this chapter the corresponding intersection matrices (following Eq. 30.1). Both elements, geometric realization and intersection matrices, are presented in Table 30.1.

$$R(\mathrm{P},\mathrm{R}) = \begin{bmatrix} \mathrm{P}^{\bullet} \cap \mathrm{R}^{\circ} & \mathrm{P}^{\bullet} \cap \partial\mathrm{R} & \mathrm{P}^{\bullet} \cap \mathrm{R}^{-} \\ \mathrm{P}^{\circ} \cap \mathrm{R}^{\circ} & \mathrm{P}^{\circ} \cap \partial\mathrm{R} & \mathrm{P}^{\circ} \cap \mathrm{R}^{-} \\ \partial\mathrm{P} \cap \mathrm{R}^{\circ} & \partial\mathrm{P} \cap \partial\mathrm{R} & \partial\mathrm{P} \cap \mathrm{R}^{-} \\ \mathrm{P}^{-} \cap \mathrm{R}^{\circ} & \mathrm{P}^{-} \cap \partial\mathrm{R} & \mathrm{P}^{-} \cap \mathrm{R}^{-} \end{bmatrix} \tag{30.1}$$

30.4 Topological Spatial Relations Between a CSEP and a Line

The identification of the topological spatial relations that exist between a CSEP and a line was carried out by Santos and Moreira [14, 15]. In their work, the CSEP represents a complex object in the sense that the point and its region of influence are not dissociable. The identification of the topological spatial relations was carried out as a tool for the prediction of mobile users' future positions in a context-aware

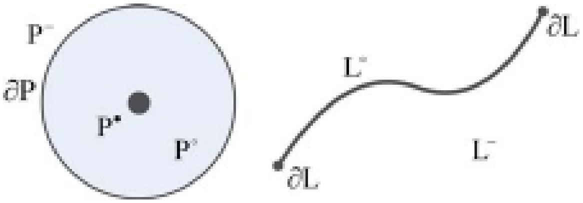


Fig. 30.6 Components of a CSEP and a Line

mobile environment. With the topological spatial relations it is possible to identify the conceptual neighborhood graphs that state the possible transitions between spatial relations and, therefore, the possible movements that a mobile user can do in a road network [16]. The selection of a CSEP is associated with the need to associate a certain degree of uncertainty to the position of a mobile user reported by an arbitrary positioning system.

The formalism used in the identification of the topological spatial relations between a CSEP and a line is based on the algebraic approach proposed by Egenhofer (the 4- and 9-intersections models) [8]. The topological spatial relations were identified, using also a 4×3 matrix (as presented in the previous section). The conditions that allowed the identification of the spatial relations were revised and their formal proofs were also undertaken [15]. The 4×3 matrix identifies the intersections ($\cap$) between the pivot ($\mathrm{P}^\bullet$), interior (P°), boundary ($\partial\mathrm{P}$) and exterior (P^-) of a CSEP (P) and the interior (L°), boundary ($\partial\mathrm{L}$) and exterior (L^-) of a line (L) (see Fig. 30.6).

Each relation (R) between a CSEP (P) and a line (L) is characterized by 12 (4×3) intersections with empty ($\oslash$) or non-empty ($\neg\oslash$) values depending on how the geographical objects are related (Eq. 30.2).

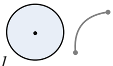
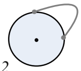
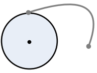
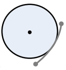
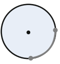
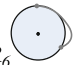
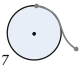
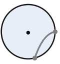
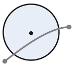
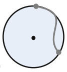
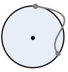
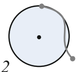
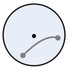
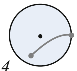
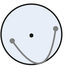
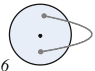
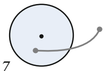
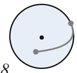
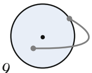
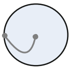
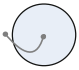
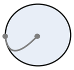
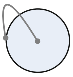
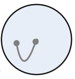
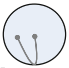
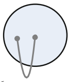
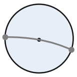
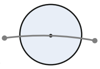
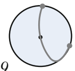
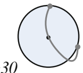
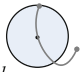
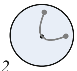
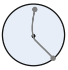
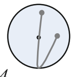
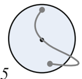
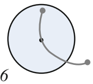
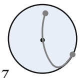
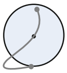
$$R(\mathrm{P},\mathrm{L}) = \begin{bmatrix} \mathrm{P}^\bullet \cap \mathrm{L}^\circ & \mathrm{P}^\bullet \cap \partial\mathrm{L} & \mathrm{P}^\bullet \cap \mathrm{L}^- \\ \mathrm{P}^\circ \cap \mathrm{L}^\circ & \mathrm{P}^\circ \cap \partial\mathrm{L} & \mathrm{P}^\circ \cap \mathrm{L}^- \\ \partial\mathrm{P} \cap \mathrm{L}^\circ & \partial\mathrm{P} \cap \partial\mathrm{L} & \partial\mathrm{P} \cap \mathrm{L}^- \\ \mathrm{P}^- \cap \mathrm{L}^\circ & \mathrm{P}^- \cap \partial\mathrm{L} & \mathrm{P}^- \cap \mathrm{L}^- \end{bmatrix} \tag{30.2}$$

After the application of 14 conditions [15], 38 topological spatial relations were identified. Table 30.2 presents the geometric realization of the 38 topological relations, as well as their corresponding intersection matrices.

30.5 Topological Spatial Relations Between Two CSEPS

The last set of topological spatial relations presented in this chapter is associated with the relations that exist between two CSEPs. This work was undertaken by Wuersch and Caduff that used those relations in a route instructions system [17, 18]. This system, designed for pedestrian navigation, uses two CSEPs: one representing

Table 30.2 Topological spatial relations between a CSEP and a Line

Relation	Matrix	Relation	Matrix	Relation	Matrix
 R_1	$\begin{bmatrix}0&0&1\\0&0&1\\0&0&1\\1&1&1\end{bmatrix}$	 R_2	$\begin{bmatrix}0&0&1\\0&0&1\\0&1&1\\1&0&1\end{bmatrix}$	 R_3	$\begin{bmatrix}0&0&1\\0&0&1\\0&1&1\\1&1&1\end{bmatrix}$
 R_4	$\begin{bmatrix}0&0&1\\0&0&1\\1&0&1\\1&1&1\end{bmatrix}$	 R_5	$\begin{bmatrix}0&0&1\\0&0&1\\1&1&1\\0&0&1\end{bmatrix}$	 R_6	$\begin{bmatrix}0&0&1\\0&0&1\\1&1&1\\1&0&1\end{bmatrix}$
 R_7	$\begin{bmatrix}0&0&1\\0&0&1\\1&1&1\\1&1&1\end{bmatrix}$	 R_8	$\begin{bmatrix}0&0&1\\1&0&1\\0&1&1\\0&0&1\end{bmatrix}$	 R_9	$\begin{bmatrix}0&0&1\\1&0&1\\1&0&1\\1&1&1\end{bmatrix}$
 R_{10}	$\begin{bmatrix}0&0&1\\1&0&1\\1&1&1\\0&0&1\end{bmatrix}$	 R_{11}	$\begin{bmatrix}0&0&1\\1&0&1\\1&1&1\\1&0&1\end{bmatrix}$	 R_{12}	$\begin{bmatrix}0&0&1\\1&0&1\\1&1&1\\1&1&1\end{bmatrix}$
 R_{13}	$\begin{bmatrix}0&0&1\\1&1&1\\0&0&1\\0&0&1\end{bmatrix}$	 R_{14}	$\begin{bmatrix}0&0&1\\1&1&1\\0&1&1\\0&0&1\end{bmatrix}$	 R_{15}	$\begin{bmatrix}0&0&1\\1&1&1\\1&0&1\\0&0&1\end{bmatrix}$
 R_{16}	$\begin{bmatrix}0&0&1\\1&1&1\\1&0&1\\1&0&1\end{bmatrix}$	 R_{17}	$\begin{bmatrix}0&0&1\\1&1&1\\1&0&1\\1&1&1\end{bmatrix}$	 R_{18}	$\begin{bmatrix}0&0&1\\1&1&1\\1&1&1\\0&0&1\end{bmatrix}$
 R_{19}	$\begin{bmatrix}0&0&1\\1&1&1\\1&1&1\\1&0&1\end{bmatrix}$	 R_{20}	$\begin{bmatrix}0&1&0\\1&0&1\\0&1&1\\0&0&1\end{bmatrix}$	 R_{21}	$\begin{bmatrix}0&1&0\\1&0&1\\1&0&1\\1&1&1\end{bmatrix}$
 R_{22}	$\begin{bmatrix}0&1&0\\1&0&1\\1&1&1\\0&0&1\end{bmatrix}$	 R_{23}	$\begin{bmatrix}0&1&0\\1&0&1\\1&1&1\\1&0&1\end{bmatrix}$	 R_{24}	$\begin{bmatrix}0&1&0\\1&1&1\\0&0&1\\0&0&1\end{bmatrix}$
 R_{25}	$\begin{bmatrix}0&1&0\\1&1&1\\1&0&1\\0&0&1\end{bmatrix}$	 R_{26}	$\begin{bmatrix}0&1&0\\1&1&1\\1&0&1\\1&0&1\end{bmatrix}$	 R_{27}	$\begin{bmatrix}1&0&0\\1&0&1\\0&1&1\\0&0&1\end{bmatrix}$
 R_{28}	$\begin{bmatrix}1&0&0\\1&0&1\\1&0&1\\1&1&1\end{bmatrix}$	 R_{29}	$\begin{bmatrix}1&0&0\\1&0&1\\1&1&1\\0&0&1\end{bmatrix}$	 R_{30}	$\begin{bmatrix}1&0&0\\1&0&1\\1&1&1\\1&0&1\end{bmatrix}$
 R_{31}	$\begin{bmatrix}1&0&0\\1&0&1\\1&1&1\\1&1&1\end{bmatrix}$	 R_{32}	$\begin{bmatrix}1&0&0\\1&1&1\\0&0&1\\0&0&1\end{bmatrix}$	 R_{33}	$\begin{bmatrix}1&0&0\\1&1&1\\0&1&1\\0&0&1\end{bmatrix}$
 R_{34}	$\begin{bmatrix}1&0&0\\1&1&1\\1&0&1\\0&0&1\end{bmatrix}$	 R_{35}	$\begin{bmatrix}1&0&0\\1&1&1\\1&0&1\\1&0&1\end{bmatrix}$	 R_{36}	$\begin{bmatrix}1&0&0\\1&1&1\\1&0&1\\1&1&1\end{bmatrix}$
 R_{37}	$\begin{bmatrix}1&0&0\\1&1&1\\1&1&1\\0&0&1\end{bmatrix}$	 R_{38}	$\begin{bmatrix}1&0&0\\1&1&1\\1&1&1\\1&0&1\end{bmatrix}$		

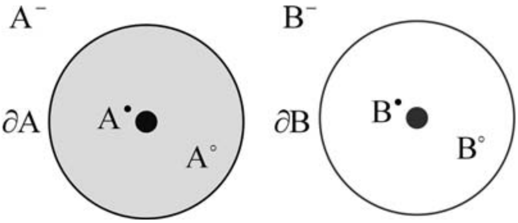


Fig. 30.7 Components of the two CSEPs

the user's location and the other representing a waypoint that is used to define paths for pedestrians in a pedestrian guiding system.

The identification of the topological spatial relations used a 4 × 4 intersection matrix, where it is possible to verify the intersections (∩) between the pivot ($A^\bullet$), interior (A°), boundary (∂A) and exterior (A^-) of a CSEP (A), and the pivot ($B^\bullet$), interior (B°), boundary (∂B) and exterior (B^-) of another CSEP (B) (see Fig. 30.7).

Each relation (R) between a CSEP A and a CSEP B is characterized by 16 (4 × 4) intersections with empty (∅) or non-empty (¬∅) values depending on how the geographical objects are related (Eq. 30.3).

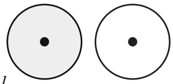
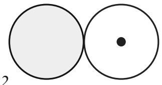
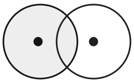
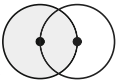
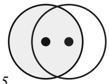
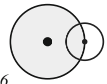
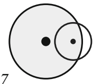
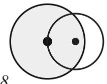
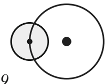
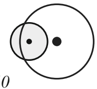
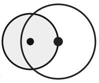
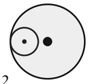
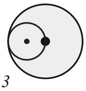
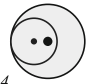
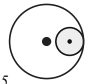
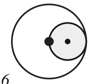
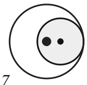
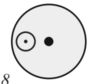
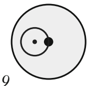
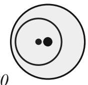
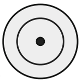
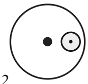
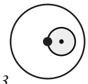
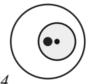
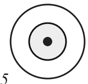
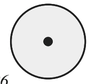
$$R(\mathrm{A},\mathrm{B}) = \begin{bmatrix} A^\bullet \cap B^\bullet & A^\bullet \cap B^\circ & A^\bullet \cap \partial B & A^\bullet \cap B^- \\ A^\circ \cap B^\bullet & A^\circ \cap B^\circ & A^\circ \cap \partial B & A^\circ \cap B^- \\ \partial A \cap B^\bullet & \partial A \cap B^\circ & \partial A \cap \partial B & \partial A \cap B^- \\ A^- \cap B^\bullet & A^- \cap B^\circ & A^- \cap \partial B & A^- \cap B^- \end{bmatrix} \tag{30.3}$$

Applying a set of constraints, for the specific case of the topological spatial relations that can exist between two CSEPs [17], 26 topological relations were identified. The analysis of the geometric realization [17] of the several topological spatial relations allows us to identify and present in this chapter the corresponding intersection matrices (following Eq. 30.3). Both elements, geometric realization and intersection matrices, are presented in Table 30.3.

30.6 Conclusion

This chapter presented the topological spatial relations that exist between a CSEP and a region, between a CSEP and a line, and between two CSEPs. The objective was to summarize the work related with CSEP, a complex object that integrates a point and a region. The several topological spatial relations presented can be used in the identification of the composition tables or the conceptual neighborhood graphs that allow spatial reasoning with these different types of objects.

Table 30.3 Topological spatial relations between two CSEPs

Relation	Matrix	Relation	Matrix
 R_1	$\begin{bmatrix}0&0&0&1\\0&0&0&1\\0&0&0&1\\1&1&1&1\end{bmatrix}$	 R_2	$\begin{bmatrix}0&0&0&1\\0&0&0&1\\0&0&1&1\\1&1&1&1\end{bmatrix}$
 R_3	$\begin{bmatrix}0&0&0&1\\0&1&1&1\\0&1&1&1\\1&1&1&1\end{bmatrix}$	 R_4	$\begin{bmatrix}0&0&1&0\\0&1&1&1\\0&1&1&1\\1&1&1&1\end{bmatrix}$
 R_5	$\begin{bmatrix}0&1&0&0\\1&1&1&1\\0&1&1&1\\0&1&1&1\end{bmatrix}$	 R_6	$\begin{bmatrix}0&0&0&1\\0&1&1&1\\1&1&1&1\\0&1&1&1\end{bmatrix}$
 R_7	$\begin{bmatrix}0&0&0&1\\1&1&1&1\\0&1&1&1\\0&1&1&1\end{bmatrix}$	 R_8	$\begin{bmatrix}0&0&1&0\\1&1&1&1\\0&1&1&1\\0&1&1&1\end{bmatrix}$
 R_9	$\begin{bmatrix}0&0&1&0\\0&1&1&1\\0&1&1&1\\1&1&1&1\end{bmatrix}$	 R_{10}	$\begin{bmatrix}0&1&0&0\\0&1&1&1\\0&1&1&1\\1&1&1&1\end{bmatrix}$
 R_{11}	$\begin{bmatrix}0&1&0&0\\0&1&1&1\\1&1&1&1\\0&1&1&1\end{bmatrix}$	 R_{12}	$\begin{bmatrix}0&0&0&1\\1&1&1&1\\0&0&1&1\\0&0&0&1\end{bmatrix}$
 R_{13}	$\begin{bmatrix}0&0&1&0\\1&1&1&1\\0&0&1&1\\0&0&0&1\end{bmatrix}$	 R_{14}	$\begin{bmatrix}0&1&0&0\\1&1&1&1\\0&0&1&1\\0&0&0&1\end{bmatrix}$
 R_{15}	$\begin{bmatrix}0&1&0&0\\0&1&0&0\\0&1&1&0\\1&1&1&1\end{bmatrix}$	 R_{16}	$\begin{bmatrix}0&1&0&0\\0&1&0&0\\1&1&1&0\\0&1&1&1\end{bmatrix}$
 R_{17}	$\begin{bmatrix}0&1&0&0\\1&1&0&0\\0&1&1&0\\0&1&1&1\end{bmatrix}$	 R_{18}	$\begin{bmatrix}0&0&0&1\\1&1&1&1\\0&0&0&1\\0&0&0&1\end{bmatrix}$
 R_{19}	$\begin{bmatrix}0&0&1&0\\1&1&1&1\\0&0&0&1\\0&0&0&1\end{bmatrix}$	 R_{20}	$\begin{bmatrix}0&1&0&0\\1&1&1&1\\0&0&0&1\\0&0&0&1\end{bmatrix}$
 R_{21}	$\begin{bmatrix}1&0&0&0\\0&1&1&1\\0&0&0&1\\0&0&0&1\end{bmatrix}$	 R_{22}	$\begin{bmatrix}0&1&0&0\\0&1&0&0\\0&1&0&0\\1&1&1&1\end{bmatrix}$
 R_{23}	$\begin{bmatrix}0&1&0&0\\0&1&0&0\\1&1&0&0\\0&1&1&1\end{bmatrix}$	 R_{24}	$\begin{bmatrix}0&1&0&0\\1&1&0&0\\0&1&0&0\\0&1&1&1\end{bmatrix}$
 R_{25}	$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&1&0&0\\0&1&1&1\end{bmatrix}$	 R_{26}	$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$

References

1. Sharma, J.: Integrated spatial reasoning in geographic information systems: combining topology and direction. Ph.D. thesis, University of Maine (1996)
2. Abdelmoty, A., El-Geresy, B.A.: A general method for spatial reasoning in spatial databases. Proceedings of the Fourth International Conference on Information and Knowledge Management, Baltimore (1995)
3. Frank, A.U.: Qualitative Spatial Reasoning: cardinal directions as an example. Int. J. Geo. Inform. Syst. **10**(3), 269–290 (1996)
4. Papadias, D., Sellis, T.: On the qualitative representation of spatial knowledge in 2D space. Very Large Database. J. **3**(4), 479–516 (1994)
5. Freksa, C.: Using orientation information for qualitative spatial reasoning. In: Frank, A.U., Campari, I., Formentini, U. (eds.) Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Springer, Berlin (1992)
6. Hernández, D., Clementini, E., Felice, P.D.: Qualitative distances. Proceedings of the International Conference COSIT '95, Semmering, Austria (1995)
7. Egenhofer, M.J.: Deriving the composition of binary topological relations. J. Visual Lang. Comput. **5**(2), 133–149 (1994)
8. Egenhofer, M.J., Herring, J.R.: Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical Report, Department of Surveying Engineering, University of Maine (1991)
9. Egenhofer, M.J., Mark, D.M.: Modeling conceptual neighborhoods of topological line-region relations. Int. J. Geo. Inform. Syst. **9**(5) 555–565 (1995)
10. Clementini, E., Felice, P.D.: Approximate topological relations. Int. J. Approx. Reason. (16) 173–204 (1997)
11. Lee, B., Flewelling, D.M.: Spatial organicism: relations between a region and a spatially extended point. In GIScience 2004, Extended Abstracts and Poster Sessions, University of Maryland (2004)
12. Reis, R., Egenhofer, M.J., Matos, J.: Topological relations using two models of uncertainty for lines. Proceedings of the 7th International Symposium on Spatial Data Accuracy Assessment in Natural Resources and Environment Sciences, Lisbon (2006)
13. Hong, J.-H.: Qualitative distance and direction reasoning in geographic space. Ph.D. thesis, University of Maine (1994)
14. Santos, M.Y., Moreira, A.: Topological spatial relations between a spatially extended point and a line for predicting movement in space. Proceedings of the 10th AGILE Conference on Geographic Information Science, Aalborg (2007)
15. Santos, M.Y., Moreira, A.: How a circular spatially extended point and a line can be topological related? Proceedings of the 2009 International Conference of Computational Intelligence and Intelligent Systems, London (2009)
16. Santos, M.Y., Moreira, A.: Conceptual neighborhood graphs for topological spatial relations. Proceedings of the 2009 International Conference of Computational Intelligence and Intelligent Systems, London (2009)
17. Wuersch, M., Caduff, D.: Refined route instructions using topological stages of closeness. Proceedings of the 5th International Workshop on Web and Wireless Geographical Information Systems, Lausanne (2005)
18. Wuersch, M., Caduff, D.: Region-based pedestrian navigation: route instructions based on topological stages of closeness. Proceedings of the 4th International Conference on Geographic Information Science – GIScience'2006, Munich (2006)

Chapter 31
Spatial Neighbors for Topological Spatial Relations: The Case of a Circular Spatially Extended Point

Maribel Yasmina Santos and Adriano Moreira

Abstract This paper presents the conceptual neighborhood graphs with the transitions that exist between several topological spatial relations. The analyzed topological spatial relations include those that exist between a circular spatially extended point and a region, between a circular spatially extended point and a line, and between two circular spatially extended points. The conceptual neighborhood graphs were identified using the snapshot model. In this model, the identification of neighborhood relations is achieved looking at the topological distance existing between pairs of spatial relations. The obtained graphs are suitable for reasoning about gradual changes in topology. These changes can be associated with the motion of objects and/or deformations over time.

Keywords Conceptual neighborhood graph · topological spatial relations · circular spatially extended point · snapshot model

31.1 Introduction

The relevance of identifying topological spatial relations is associated with the need to conceptualize the spatial relations that can exist among several objects in the geographical space. The work described in this chapter is associated with the topological spatial relations that integrate a geographical object represented by a Circular Spatially Extended Point (CSEP). A CSEP is a region-like object characterized by the inclusion of a point and a region that defines the area of influence of that point (see Fig. 31.1).

Research on topological spatial relations, between a CSEP and a region [1], between a CSEP and a line [2, 3], and between two CSEP [4, 5], was already

M.Y. Santos (✉) and A. Moreira
Information Systems Department, Algoritmi Research Centre, University of Minho,
Campus de Azurém, 4800-058 Guimarães, Portugal
e-mail: maribel@dsi.uminho.pt; adriano@dsi.uminho.pt

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_31,

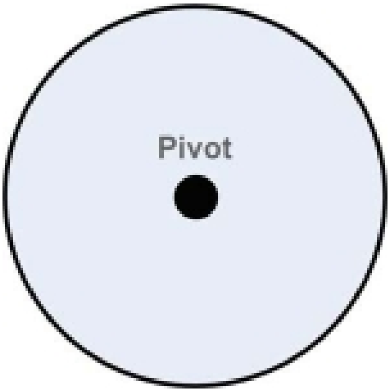


Fig. 31.1 A circular spatially extended point

carried out. These works allowed the identification of the topological spatial relations that exist between these different objects. After the verification of the topological spatial relations it is possible to identify the conceptual neighborhood graphs that state the possible transition between these relations. This chapter summarizes the work undertaken so far in this field, as well as identifies and presents the respective graphs whenever they are not available in the literature.

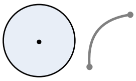
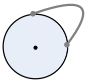
This chapter is organized as follows. Section 31.2 introduces the principles associated to the snapshot model for the identification of conceptual neighborhood graphs. Section 31.3 presents the conceptual neighborhood graph that states the possible transitions for the topological spatial relations existing between a CESP and a region. Section 31.4 describes the identification of the graph but now for the topological spatial relations existing between a CSEP and a line. Section 31.5 illustrates the conceptual neighborhood graph for the particular case of the topological spatial relations between two CSEPs. Section 31.6 concludes this chapter summarizing the presented work.

31.2 Conceptual Neighborhood Graphs: Snapshot Model

Geographic objects and phenomena may gradually change their location, orientation, shape, and size over time. A qualitative change occurs when an object deformation affects its topological relation with respect to other objects. Models for changes of topological relations are relevant to spatio-temporal reasoning in geographic space as they derive the most likely configurations and allow for predictions (based on inference) about significant changes [6].

In a conceptual neighborhood graph, nodes represent spatial relations and edges are created to link similar relations. Different definitions of similarity lead to different graphs involving the same set of relations. Usually, conceptual neighborhood graphs are built considering situations of continuous change, representing the possible transitions from one relation to other relations. Those graphs are useful for reducing the search space when looking for the next possible situations [7].

Table 31.1 Topological distance: an example

Relation R_1	Relation R_2	Topological distance
		$M_1 - M_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$
$M_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	$M_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$	$\tau_{R_1,R_2} = 2$

One of the approaches to identify a conceptual neighborhood graph is using the snapshot model. This model compares two different topological relations with no knowledge about the potential transformations that may have caused the change [8]. The comparison is made by considering the topological distance between two topological relations [6]. This distance determines the number of corresponding elements, empty and non-empty, with different values in the corresponding intersection matrices. Next sections present different intersection matrices, which depend on the geographical objects in analysis.

The definition of topological distance (τ) between two spatial relations (R_A and R_B) is the sum of the absolute values of the differences between corresponding entries of the intersections verified in the corresponding matrices (M_A and M_B) [6]. Considering a 12-intersection matrix, the topological distance is calculated by Eq. (31.1).

$$\tau_{R_A,R_B} = \sum_{i=1}^{3} \sum_{j=1}^{4} \left| \, M_A\left[i,j\right] - M_B\left[i,j\right] \, \right| \tag{31.1}$$

As an example, consider the topological spatial relations illustrated in Table 31.1. Using relation R_1 and relation R_2 [2, 3], and their corresponding matrices M_1 and M_2, the calculated topological distance between these two topological spatial relations is 2.

31.3 Conceptual Neighborhood Graph: A CSEP and a Region

The topological spatial relations existing between a CSEP and a region were identified by Lee and Flewelling [1] following the 9-intersection formalism proposed by Egenhofer and Herring [9]. This 9-intersection matrix was used in many works, one of them identifying the topological spatial relations that can exist between two regions [9]. A region (R) is a 2D geographical object that has an interior (R°), a boundary (∂R) and an exterior (R^-). A CSEP (P) is a complex object that integrates a point and a region, and that it is constituted by a pivot ($P^\bullet$), an interior (P°), a boundary (∂P) and an exterior (P^-). Figure 31.2 shows the several components to these two objects.

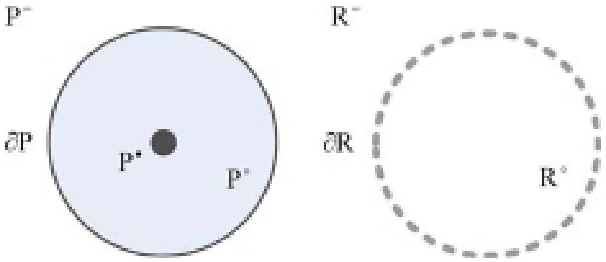


Fig. 31.2 Components of a CSEP and a region

Table 31.2 Topological distance: spatial relations between a CSEP and a Region

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}
R_1	0	1	4	6	7	4	6	7	6	8	9	6	8	9
R_2	1	0	3	7	6	3	5	6	7	9	8	5	9	8
R_3	4	3	0	4	3	6	2	5	6	6	5	2	6	5
R_4	6	7	4	0	1	4	6	7	6	2	3	6	2	3
R_5	7	6	3	1	0	3	5	6	7	3	2	5	3	2
R_6	4	3	6	4	3	0	8	7	8	6	5	8	6	5
R_7	6	5	2	6	5	8	0	3	4	4	3	2	6	5
R_8	7	6	5	7	6	7	3	0	1	5	4	5	7	6
R_9	6	7	6	6	7	8	4	1	0	4	5	6	6	7
R_{10}	8	9	6	2	3	6	4	5	4	0	1	6	2	3
R_{11}	9	8	5	3	2	5	3	4	5	1	0	5	3	2
R_{12}	6	5	2	6	5	8	2	5	6	6	5	0	4	3
R_{13}	8	9	6	2	3	6	6	7	6	2	3	4	0	1
R_{14}	9	8	5	3	2	5	5	6	7	3	2	3	1	0

As a CSEP integrates four parts, Lee and Flewelling [1] used a 12-intersection matrix to verify the intersections (∩) between the pivot ($P^{\bullet}$), interior (P°), boundary (∂P) and exterior (P^{-}) of a CSEP (P) and the interior (R°), boundary (∂R) and exterior (R^{-}) of a region (R). Each relation (R) between a CSEP (P) and a region (R) is characterized by empty (∅) or non-empty (¬∅) values depending on how the geographical objects are related (Eq. 31.2).

$$R(\mathrm{P},\mathrm{R}) = \begin{bmatrix} \mathrm{P}^{\bullet} \cap \mathrm{R}^{\circ} & \mathrm{P}^{\bullet} \cap \partial\mathrm{R} & \mathrm{P}^{\bullet} \cap \mathrm{R}^{-} \\ \mathrm{P}^{\circ} \cap \mathrm{R}^{\circ} & \mathrm{P}^{\circ} \cap \partial\mathrm{R} & \mathrm{P}^{\circ} \cap \mathrm{R}^{-} \\ \partial\mathrm{P} \cap \mathrm{R}^{\circ} & \partial\mathrm{P} \cap \partial\mathrm{R} & \partial\mathrm{P} \cap \mathrm{R}^{-} \\ \mathrm{P}^{-} \cap \mathrm{R}^{\circ} & \mathrm{P}^{-} \cap \partial\mathrm{R} & \mathrm{P}^{-} \cap \mathrm{R}^{-} \end{bmatrix} \tag{31.2}$$

The analysis of the obtained 14 topological spatial relations allowed the identification of the corresponding intersection matrices and the calculation of the topological distance among them. Table 31.2 presents the calculated topological distances. In this table, the conceptual neighbors identified by Lee and Flewelling [1] are shown with a gray shadow.

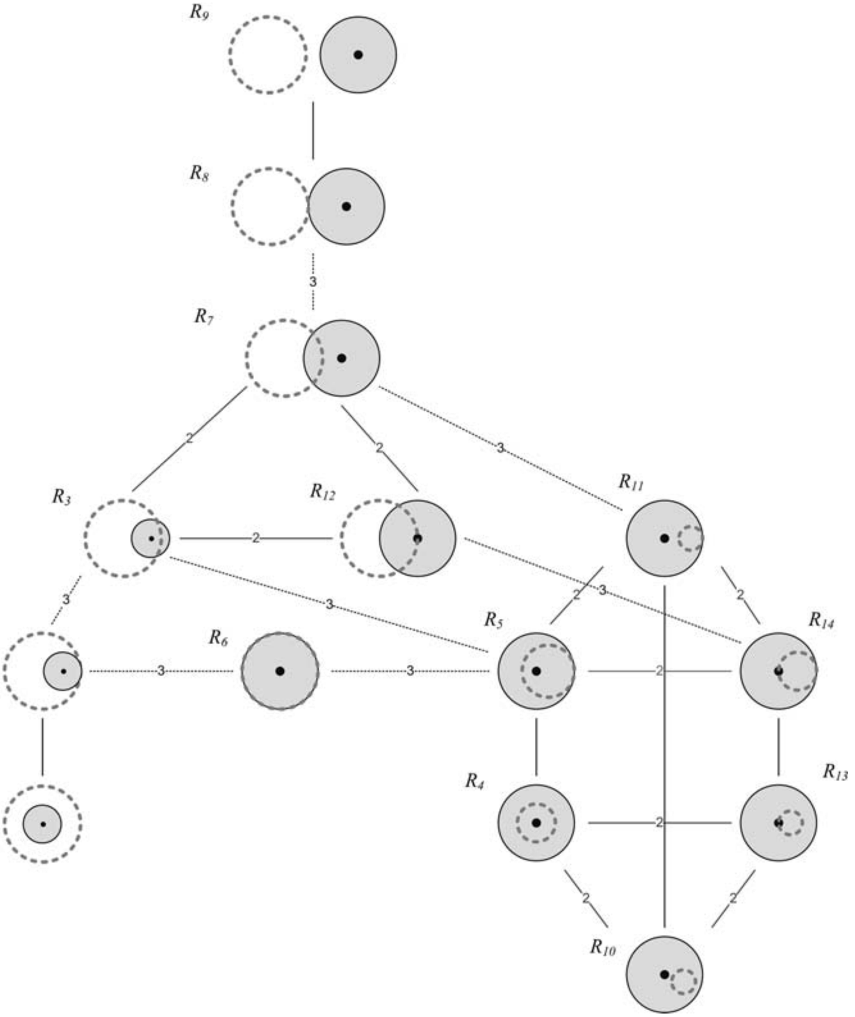


Fig. 31.3 Conceptual neighborhood graph: relations between a CSEP and a region

The several shadowed cells link spatial relations that are considered neighbors. In some situations the minimal topological distance is 1. However, and in this particular case of the topological spatial relations that exist between a CSEP and a region, there are few of these links. It was necessary to use topological distances of 2 and 3 to obtain a graph that links all the topological relations. The graph presented by Lee and Flewelling [1] is shown in Fig. 31.3. In this graph all the distances different than 1 are marked in the corresponding links.

Based on the topological distance, the conceptual neighborhood graph presented in Fig. 31.3 extended the graph available for the topological spatial relations between two regions [6]. There are eight topological relations that exist between two regions [9]. In the particular case of the topological relations between a CSEP and a region, six new spatial relations emerged as a consequence of the intersection of the CSEP's pivot with the interior (R_3, R_4 and R_5) and the boundary (R_{12}, R_{13} and R_{14}) of the region.

31.4 Conceptual Neighborhood Graph: A CSEP and a Line

Another work that used a 12-intersection matrix was carried out by Santos and Moreira [2,3], in the identification of the topological spatial relations that can exist between a CSEP and a line. For this particular case, the 12-intersection matrix showed also to be appropriate since a line (L) integrates an interior (L°), a boundary (∂L) and an exterior (L⁻). Figure 31.4 shows the several components of the considered geographical objects.

Each relation (R) between a CSEP (P) and a line (L) is characterized by 12 intersections with empty (∅) or non-empty (¬∅) values depending on how the geographical objects are related (Eq. 31.3).

$$R(\mathrm{P},\mathrm{L}) = \begin{bmatrix} \mathrm{P}^{\bullet} \cap \mathrm{L}^{\circ} & \mathrm{P}^{\bullet} \cap \partial\mathrm{L} & \mathrm{P}^{\bullet} \cap \mathrm{L}^{-} \\ \mathrm{P}^{\circ} \cap \mathrm{L}^{\circ} & \mathrm{P}^{\circ} \cap \partial\mathrm{L} & \mathrm{P}^{\circ} \cap \mathrm{L}^{-} \\ \partial\mathrm{P} \cap \mathrm{L}^{\circ} & \partial\mathrm{P} \cap \partial\mathrm{L} & \partial\mathrm{P} \cap \mathrm{L}^{-} \\ \mathrm{P}^{-} \cap \mathrm{L}^{\circ} & \mathrm{P}^{-} \cap \partial\mathrm{L} & \mathrm{P}^{-} \cap \mathrm{L}^{-} \end{bmatrix} \qquad (31.3)$$

For the obtained 38 topological spatial relations, the several intersection matrices and the topological distances (Table 31.3) between them were obtained [2,3].

In this case, the analysis of Table 31.3 shows that for the majority of the topological relations the minimum distance to its neighbors is 1. The minimum topological distance between one relation and its neighbors is 2 only in the case of relation R_{21}. Figure 31.5 shows the corresponding conceptual neighborhood graph obtained by applying the snapshot model.

The graph is virtually divided in three parts. In the upper part, the 19 topological relations do not include any intersection between the pivot of the CSEP and the line. If the pivot of the spatially extended point is ignored, making a CSEP equal to a region, these 19 topological spatial relations correspond to the 19 topological spatial

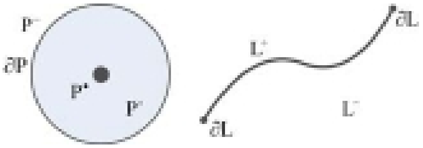
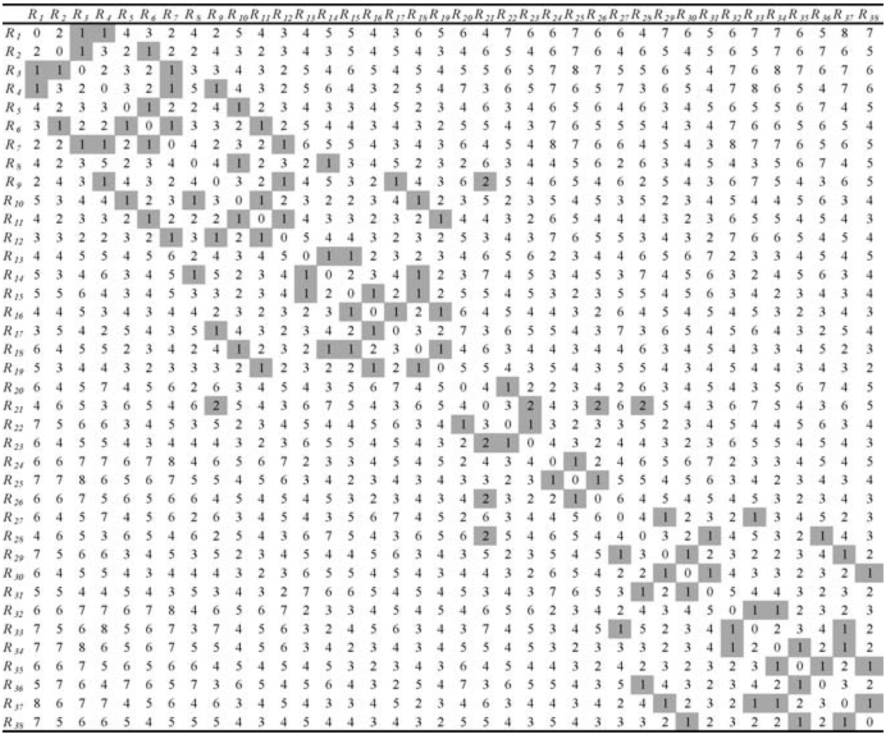


Fig. 31.4 Components of a CSEP and a line

Table 31.3 Topological distance: spatial relations between a CSEP and a line



	R_{1}	R_{2}	R_{3}	R_{4}	R_{5}	R_{6}	R_{7}	R_{8}	R_{9}	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}	R_{20}	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}	R_{27}	R_{28}	R_{29}	R_{30}	R_{31}	R_{32}	R_{33}	R_{34}	R_{35}	R_{36}	R_{37}	R_{38}
R_{1}	0	2	1	1	4	3	2	4	2	5	4	3	4	5	5	4	3	6	5	6	4	7	6	6	7	6	6	4	7	6	5	6	7	7	6	5	8	7
R_{2}	2	0	1	3	2	1	2	2	4	3	2	3	4	3	5	4	5	4	3	4	6	5	4	6	7	6	4	6	5	4	5	6	5	7	6	7	6	5
R_{3}	1	1	0	2	3	2	1	3	3	4	3	2	5	4	6	5	4	5	4	5	5	6	5	7	8	7	5	5	6	5	4	7	6	8	7	6	7	6
R_{4}	1	3	2	0	3	2	1	5	1	4	3	2	5	6	4	3	2	5	4	7	3	6	5	7	6	5	7	3	6	5	4	7	8	6	5	4	7	6
R_{5}	4	2	3	3	0	1	2	2	4	1	2	3	4	3	3	4	5	2	3	4	6	3	4	6	5	6	4	6	3	4	5	6	5	5	6	7	4	5
R_{6}	3	1	2	2	1	0	1	3	3	2	1	2	5	4	4	3	4	3	2	5	5	4	3	7	6	5	5	5	4	3	4	7	6	6	5	6	5	4
R_{7}	2	2	1	1	2	1	0	4	2	3	2	1	6	5	5	4	3	4	3	6	4	5	4	8	7	6	6	4	5	4	3	8	7	7	6	5	6	5
R_{8}	4	2	3	5	2	3	4	0	4	1	2	3	2	1	3	4	5	2	3	2	6	3	4	4	5	6	2	6	3	4	5	4	3	5	6	7	4	5
R_{9}	2	4	3	1	4	3	2	4	0	3	2	1	4	5	3	2	1	4	3	6	2	5	4	6	5	4	6	2	5	4	3	6	7	5	4	3	6	5
R_{10}	5	3	4	4	1	2	3	1	3	0	1	2	3	2	2	3	4	1	2	3	5	2	3	5	4	5	3	5	2	3	4	5	4	4	5	6	3	4
R_{11}	4	2	3	3	2	1	2	2	2	1	0	1	4	3	3	2	3	2	1	4	4	3	2	6	5	4	4	4	3	2	3	6	5	5	4	5	4	3
R_{12}	3	3	2	2	3	2	1	3	1	2	1	0	5	4	4	3	2	3	2	5	3	4	3	7	6	5	5	3	4	3	2	7	6	6	5	4	5	4
R_{13}	4	4	5	5	4	5	6	2	4	3	4	5	0	1	1	2	3	2	3	4	6	5	6	2	3	4	4	6	5	6	7	2	3	3	4	5	4	5
R_{14}	5	3	4	6	3	4	5	1	5	2	3	4	1	0	2	3	4	1	2	3	7	4	5	3	4	5	3	7	4	5	6	3	2	4	5	6	3	4
R_{15}	5	5	6	4	3	4	5	3	3	2	3	4	1	2	0	1	2	1	2	5	5	4	5	3	2	3	5	5	4	5	6	3	4	2	3	4	3	4
R_{16}	4	4	5	3	4	3	4	4	2	3	2	3	2	3	1	0	1	2	1	6	4	5	4	4	3	2	6	4	5	4	5	4	5	3	2	3	4	3
R_{17}	3	5	4	2	5	4	3	5	1	4	3	2	3	4	2	1	0	3	2	7	3	6	5	5	4	3	7	3	6	5	4	5	6	4	3	2	5	4
R_{18}	6	4	5	5	2	3	4	2	4	1	2	3	2	1	1	2	3	0	1	4	6	3	4	4	3	4	4	6	3	4	5	4	3	3	4	5	2	3
R_{19}	5	3	4	4	3	2	3	3	3	2	1	2	3	2	2	1	2	1	0	5	5	4	3	5	4	3	5	5	4	3	4	5	4	4	3	4	3	2
R_{20}	6	4	5	7	4	5	6	2	6	3	4	5	4	3	5	6	7	4	5	0	4	1	2	2	3	4	2	6	3	4	5	4	3	5	6	7	4	5
R_{21}	4	6	5	3	6	5	4	6	2	5	4	3	6	7	5	4	3	6	5	4	0	3	2	4	3	2	6	2	5	4	3	6	7	5	4	3	6	5
R_{22}	7	5	6	6	3	4	5	3	5	2	3	4	5	4	4	5	6	3	4	1	3	0	1	3	2	3	3	5	2	3	4	5	4	4	5	6	3	4
R_{23}	6	4	5	5	4	3	4	4	4	3	2	3	6	5	5	4	5	4	3	2	2	1	0	4	3	2	4	4	3	2	3	6	5	5	4	5	4	3
R_{24}	6	6	7	7	6	7	8	4	6	5	6	7	2	3	3	4	5	4	5	2	4	3	4	0	1	2	4	6	5	6	7	2	3	3	4	5	4	5
R_{25}	7	7	8	6	5	6	7	5	5	4	5	6	3	4	2	3	4	3	4	3	3	2	3	1	0	1	5	5	4	5	6	3	4	2	3	4	3	4
R_{26}	6	6	7	5	6	5	6	6	4	5	4	5	4	5	3	2	3	4	3	4	2	3	2	2	1	0	6	4	5	4	5	4	5	3	2	3	4	3
R_{27}	6	4	5	7	4	5	6	2	6	3	4	5	4	3	5	6	7	4	5	2	6	3	4	4	5	6	0	4	1	2	3	2	1	3	4	5	2	3
R_{28}	4	6	5	3	6	5	4	6	2	5	4	3	6	7	5	4	3	6	5	6	2	5	4	6	5	4	4	0	3	2	1	4	5	3	2	1	4	3
R_{29}	7	5	6	6	3	4	5	3	5	2	3	4	5	4	4	5	6	3	4	3	5	2	3	5	4	5	1	3	0	1	2	3	2	2	3	4	1	2
R_{30}	6	4	5	5	4	3	4	4	4	3	2	3	6	5	5	4	5	4	3	4	4	3	2	6	5	4	2	2	1	0	1	4	3	3	2	3	2	1
R_{31}	5	5	4	4	5	4	3	5	3	4	3	2	7	6	6	5	4	5	4	5	3	4	3	7	6	5	3	1	2	1	0	5	4	4	3	2	3	2
R_{32}	6	6	7	7	6	7	8	4	6	5	6	7	2	3	3	4	5	4	5	4	6	5	6	2	3	4	2	4	3	4	5	0	1	1	2	3	2	3
R_{33}	7	5	6	8	5	6	7	3	7	4	5	6	3	2	4	5	6	3	4	3	7	4	5	3	4	5	1	5	2	3	4	1	0	2	3	4	1	2
R_{34}	7	7	8	6	5	6	7	5	5	4	5	6	3	4	2	3	4	3	4	5	5	4	5	3	2	3	3	3	2	3	4	1	2	0	1	2	1	2
R_{35}	6	6	7	5	6	5	6	6	4	5	4	5	4	5	3	2	3	4	3	6	4	5	4	4	3	2	4	2	3	2	3	2	3	1	0	1	2	1
R_{36}	5	7	6	4	7	6	5	7	3	6	5	4	5	6	4	3	2	5	4	7	3	6	5	5	4	3	5	1	4	3	2	3	4	2	1	0	3	2
R_{37}	8	6	7	7	4	5	6	4	6	3	4	5	4	3	3	4	5	2	3	4	6	3	4	4	3	4	2	4	1	2	3	2	1	1	2	3	0	1
R_{38}	7	5	6	6	5	4	5	5	5	4	3	4	5	4	4	3	4	3	2	5	5	4	3	5	4	3	3	3	2	1	2	3	2	2	1	2	1	0

relations identified for line-region relations [8]. The middle of the graph contains the relations in which one of the boundaries of the line intersects the pivot of the CSEP. The lower part of the graph contains the topological relations in which the pivot of the CSEP is intersected by the interior of the line. These three parts are linked by relation R_{21} that presents edges to relations R_{9}, R_{23}, R_{26} and R_{28} with the minimum topological distance of 2. All the other edges, as previously mentioned, link spatial relations with topological distance equal to 1.

Another approach that can be used for the identification of a conceptual neighborhood graph is by using the smooth-transition model. The smooth-transition model states that two relations are conceptual neighbors if there is a smooth-transition from one relation to the other. A smooth-transition can be defined as an infinitesimally small deformation that changes the topological relation [8]. The conceptual neighborhood graph obtained following the principles of the smooth-transition model, for the topological spatial relations that exist between a CSEP and a line, is available at [10].

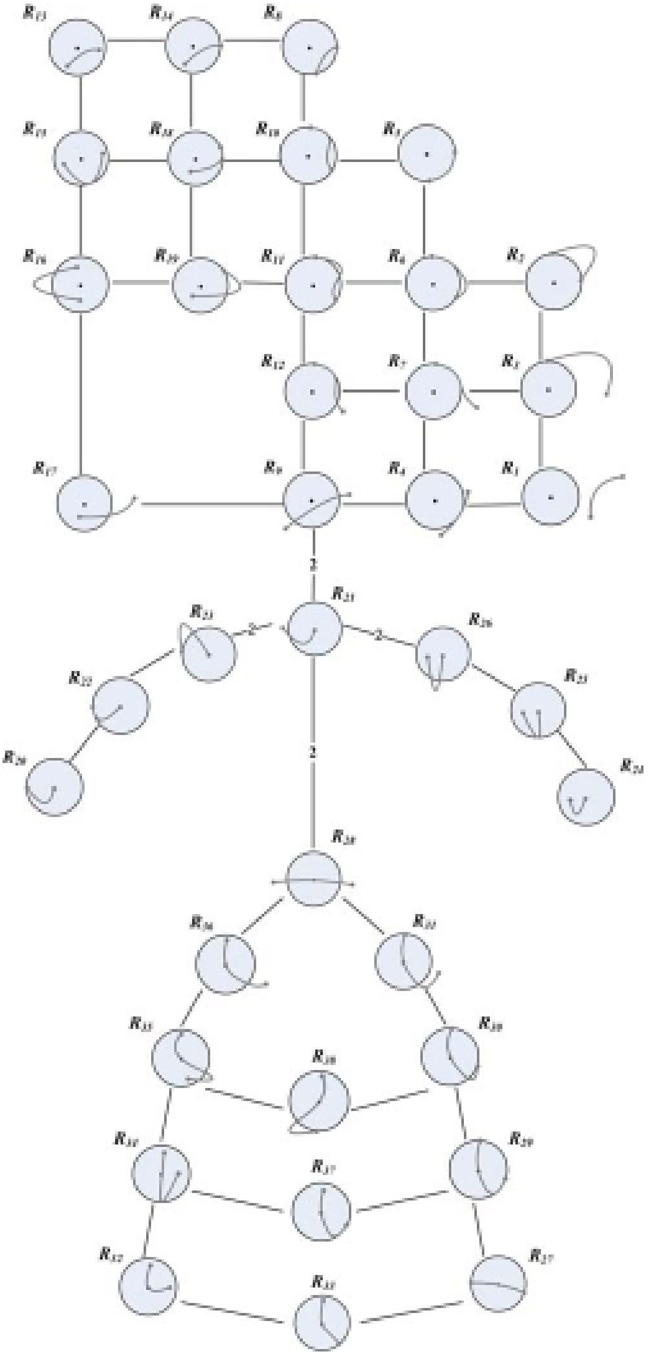


Fig. 31.5 Conceptual neighborhood graph: relations between a CSEP and a Line

31.5 Conceptual Neighborhood Graph: Two CSEPS

The topological spatial relations that include a CSEP were also identified in the case the two geographical objects are represented by CSEP [4]. In this particular case, we are dealing with two complex objects, each of them integrating a point and a region. Following the descriptions made so far for the geographic objects in analysis, in previous sections, each of these CSEP integrates four components. A CSEP (A), it integrates a pivot ($A^{\bullet}$), an interior (A°), a boundary (∂A) and an exterior (A^{-}). For a CSEP (B), its components also include a pivot ($B^{\bullet}$), an interior (B°), a boundary (∂B) and an exterior (B^{-}) (see Fig. 31.6).

Each relation (R) between a CSEP A and a CSEP B is characterized by 16 (4 × 4) intersections with empty (∅) or non-empty (¬∅) values depending on how the geographical objects are related (Eq. 31.4).

$$R(\mathrm{A},\mathrm{B}) = \begin{bmatrix} A^{\bullet}\cap B^{\bullet} & A^{\bullet}\cap B^{\circ} & A^{\bullet}\cap \partial B & A^{\bullet}\cap B^{-} \\ A^{\circ}\cap B^{\bullet} & A^{\circ}\cap B^{\circ} & A^{\circ}\cap \partial B & A^{\circ}\cap B^{-} \\ \partial A\cap B^{\bullet} & \partial A\cap B^{\circ} & \partial A\cap \partial B & \partial A\cap B^{-} \\ A^{-}\cap B^{\bullet} & A^{-}\cap B^{\circ} & A^{-}\cap \partial B & A^{-}\cap B^{-} \end{bmatrix} \qquad (31.4)$$

For the identified 26 topological spatial relations [4], the identification of the corresponding intersection matrices allowed us to calculate the topological distance between spatial relations. This process is essential for the identification of the conceptual neighborhood graph. Table 31.4 presents the resulting topological distances.

For the particular case of the topological spatial relations between two CSEPs, the topological distance (Table 31.4) between neighbors is 1 in just a few cases. There is the need to use topological distances of 2, 3 and 4 in order to obtain a graph that links all the spatial relations. To help in the visualization of these situations, Table 31.4 presents different grey grades in the shadowed cells. Linking these conceptual neighbors, the graph presented in Fig. 31.7 was obtained.

Analyzing the obtained graph (Fig. 31.7), it is possible to verify that its upper part integrates CSEP that are disjoint and that successively come closer to each other until they meet and then an overlap is verified. This overlapping occurs in several ways leading to different spatial relations. These relations are linked with a topological distance of 2. As the overlapping relation progress to a contain or contained

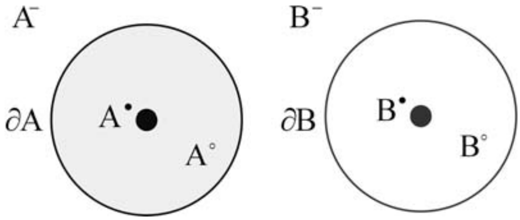
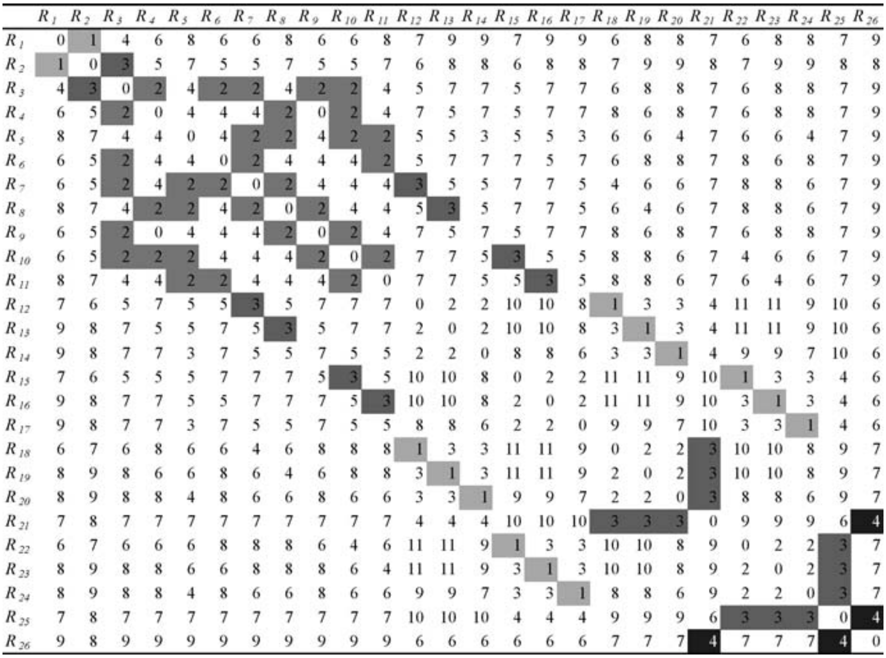


Fig. 31.6 Components of two CSEPs

Table 31.4 Topological distance: spatial relations between two CSEPs



	R_{1}	R_{2}	R_{3}	R_{4}	R_{5}	R_{6}	R_{7}	R_{8}	R_{9}	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}	R_{20}	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}
R_{1}	0	1	4	6	8	6	6	8	6	6	8	7	9	9	7	9	9	6	8	8	7	6	8	8	7	9
R_{2}	1	0	3	5	7	5	5	7	5	5	7	6	8	8	6	8	8	7	9	9	8	7	9	9	8	8
R_{3}	4	3	0	2	4	2	2	4	2	2	4	5	7	7	5	7	7	6	8	8	7	6	8	8	7	9
R_{4}	6	5	2	0	4	4	4	2	0	2	4	7	5	7	5	7	7	8	6	8	7	6	8	8	7	9
R_{5}	8	7	4	4	0	4	2	2	4	2	2	5	5	3	5	5	3	6	6	4	7	6	6	4	7	9
R_{6}	6	5	2	4	4	0	2	4	4	4	2	5	7	7	7	5	7	6	8	8	7	8	6	8	7	9
R_{7}	6	5	2	4	2	2	0	2	4	4	4	3	5	5	7	7	5	4	6	6	7	8	8	6	7	9
R_{8}	8	7	4	2	2	4	2	0	2	4	4	5	3	5	7	7	5	6	4	6	7	8	8	6	7	9
R_{9}	6	5	2	0	4	4	4	2	0	2	4	7	5	7	5	7	7	8	6	8	7	6	8	8	7	9
R_{10}	6	5	2	2	2	4	4	4	2	0	2	7	7	5	3	5	5	8	8	6	7	4	6	6	7	9
R_{11}	8	7	4	4	2	2	4	4	4	2	0	7	7	5	5	3	5	8	8	6	7	6	4	6	7	9
R_{12}	7	6	5	7	5	5	3	5	7	7	7	0	2	2	10	10	8	1	3	3	4	11	11	9	10	6
R_{13}	9	8	7	5	5	7	5	3	5	7	7	2	0	2	10	10	8	3	1	3	4	11	11	9	10	6
R_{14}	9	8	7	7	3	7	5	5	7	5	5	2	2	0	8	8	6	3	3	1	4	9	9	7	10	6
R_{15}	7	6	5	5	5	7	7	7	5	3	5	10	10	8	0	2	2	11	11	9	10	1	3	3	4	6
R_{16}	9	8	7	7	5	5	7	7	7	5	3	10	10	8	2	0	2	11	11	9	10	3	1	3	4	6
R_{17}	9	8	7	7	3	7	5	5	7	5	5	8	8	6	2	2	0	9	9	7	10	3	3	1	4	6
R_{18}	6	7	6	8	6	6	4	6	8	8	8	1	3	3	11	11	9	0	2	2	3	10	10	8	9	7
R_{19}	8	9	8	6	6	8	6	4	6	8	8	3	1	3	11	11	9	2	0	2	3	10	10	8	9	7
R_{20}	8	9	8	8	4	8	6	6	8	6	6	3	3	1	9	9	7	2	2	0	3	8	8	6	9	7
R_{21}	7	8	7	7	7	7	7	7	7	7	7	4	4	4	10	10	10	3	3	3	0	9	9	9	6	4
R_{22}	6	7	6	6	6	8	8	8	6	4	6	11	11	9	1	3	3	10	10	8	9	0	2	2	3	7
R_{23}	8	9	8	8	6	6	8	8	8	6	4	11	11	9	3	1	3	10	10	8	9	2	0	2	3	7
R_{24}	8	9	8	8	4	8	6	6	8	6	6	9	9	7	3	3	1	8	8	6	9	2	2	0	3	7
R_{25}	7	8	7	7	7	7	7	7	7	7	7	10	10	10	4	4	4	9	9	9	6	3	3	3	0	4
R_{26}	9	8	9	9	9	9	9	9	9	9	9	6	6	6	6	6	6	7	7	7	4	7	7	7	4	0

by relation between the two CSEPs, the graph is divided into two branches. The left-hand side with the region A being contained by B, and the right-hand side with the region A containing B.

31.6 Conclusion

This paper presented the conceptual neighborhood graphs that represent the transitions that can occur between the topological spatial relations that include a CSEP. Three graphs, stating the transition between the topological spatial relations, were presented: for the relations between a CSEP and a region, between a CSEP and a line, and between two CSEPs.

The graphs were analyzed in terms and their structure and possible transitions attending to the topological distance among relations.

This work constitutes a basis for dealing with spatial objects that can be represented geometrically by a CSEP, and is suitable for reasoning about gradual changes in topology. These changes can be associated with the movements of objects and/or deformations over time [11].

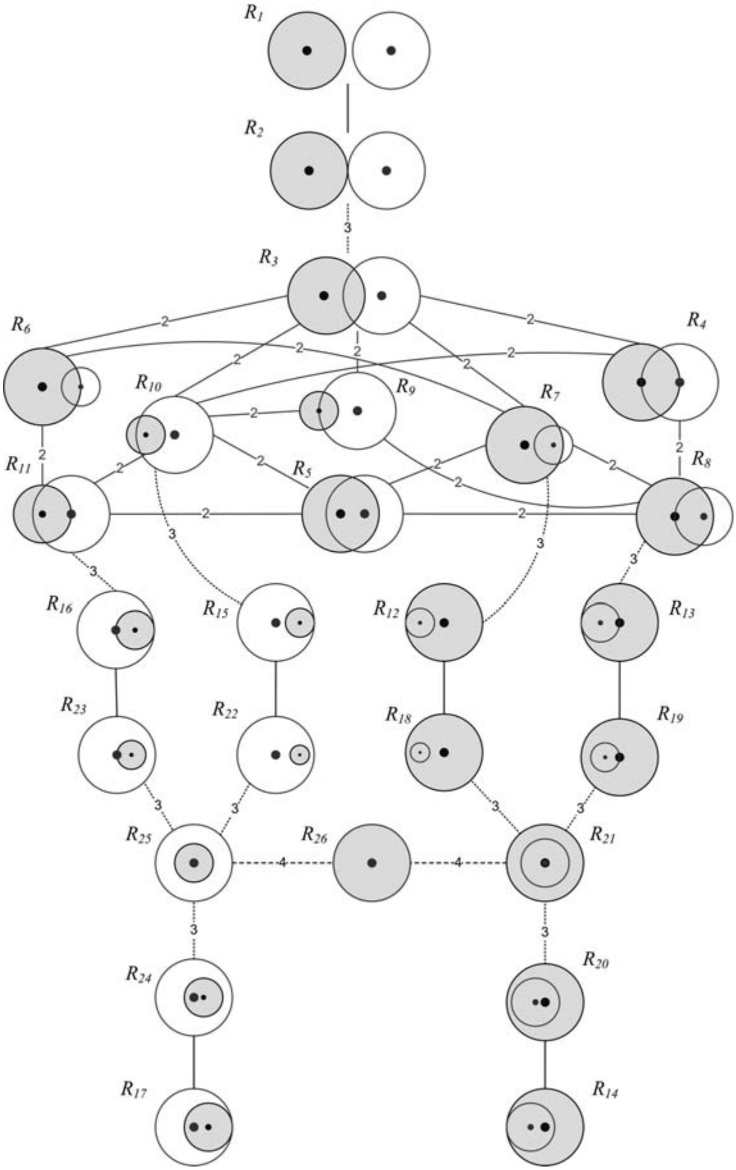


Fig. 31.7 Conceptual neighborhood graph: relations between two CSEPs

References

1. Lee, B., Flewelling, D.M.: Spatial organism: relations between a region and a spatially extended point. In: GIScience 2004, Extended Abstracts and Poster Sessions, University of Maryland, Maryland (2004)

2. Santos, M.Y., Moreira, A.: Topological spatial relations between a spatially extended point and a line for predicting movement in space. In: Proceedings of the 10th AGILE Conference on Geographic Information Science, Aalborg (2007)
3. Santos, M.Y., Moreira, A.: How a circular spatially extended point and a line can be topological related? In: Proceedings of the 2009 International Conference of Computational Intelligence and Intelligent Systems, London (2009)
4. Wuersch, M., Caduff, D.: Refined Route instructions using topological stages of closeness. In: Proceedings of the 5th International Workshop on Web and Wireless Geographical Information Systems, Lausanne (2005)
5. Wuersch, M., Caduff, D.: Region-based pedestrian navigation: route instructions based on topological stages of closeness. In: Proceedings of the 4th International Conference on Geographic Information Science – GIScience'2006, Munich (2006)
6. Egenhofer, M.J., Al-Taha, K.K.: Reasoning about gradual changes of topological relations. In: Frank, A.U., Campari, I., Formentini, U. (eds.) Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Springer, Berlin (1992)
7. Reis, R., Egenhofer, M.J., Matos, J.: Topological relations using two models of uncertainty for lines. In: Proceedings of the 7th International Symposium on Spatial Data Accuracy Assessment in Natural Resources and Environment Sciences, Lisbon (2006)
8. Egenhofer, M.J., Mark, D.M.: Modeling conceptual neighborhoods of topological line-region relations. Int. J. Geo. Inform. Syst. **9**(5), 555–565 (1995)
9. Egenhofer, M.J., Herring, J.R.: Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical Report, Department of Surveying Engineering, University of Maine (1991)
10. Santos, M.Y., Moreira, A.: Conceptual neighborhood graphs for topological spatial relations. In: Proceedings of the 2009 International Conference of Computational Intelligence and Intelligent Systems, London (2009)
11. Clementini, E., Felice, P.D.: Approximate topological relations. Int. J. Approx. Reason. 16(2), 173–204 (1997)

Chapter 32
Multi-Agent Exploration Inside Structural Collapses

Panteha Saeedi and Soren Aksel Sorensen

Abstract Autonomous navigation in unknown cluttered environments is one of the main challenges for search and rescue robots inside collapsed buildings. Being able to compare different search strategies in various search fields is crucial to attain fast victim localization. Thus we discuss an algorithmic development and proliferation of realistic afterdisaster test fields for search and rescue simulated robots. In this paper we characterized our developed search environments by their fractal dimensions. This index has shown to be a discriminative index for narrow pathways inside confined and cluttered spaces in our simulation test fields.

Keywords Multi-agent · exploration algorithms · fractal dimensions

32.1 Introduction

Imagine a scenario that, buildings are collapsed, major roads become impassable and previously established network and power have failed.... The disaster victims must be rescued within 48 hs. Human rescuers must take quick decisions, and try to get victims to safety often at their own risk. They must verify the location and status of victims. They also should ascertain the stability of the structures as quickly as possible thus medics and fire fighters can enter the disaster area and save victims. All of these tasks are performed mostly by human and trained dogs, often in a very dangerous and risky situation. In some disasters, officials find out that even trained dogs for search and rescue are unable to climb across much of the rubble. In addition to this, the dust-laden air diminished the dogs sense of smell. However robots can be sent into areas where structurally unstable or contaminated and are not safe for navigation by human or search and rescue dogs.

The 2008, earthquake in southern Sichuan (China) have shown that the number of trapped victims can be significantly high in a densely populated areas. Rescue

P. Saeedi (✉) and S.A. Sorensen
Computer Science Department, University College London, WC1E 6BT, London
e-mail: p.saeedi@cs.ucl.ac.uk; s.sorensen@cs.ucl.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_32,

specialist teams are therefore required at a number of affected sites to locate the victims in the rubble, rescue them and give initial medical treatment. The rescue teams should be equipped with adequate numbers of search and rescue robots to enhance their rescue performance. These robots are required to collect vital information both from victims and their surrounding area. The aim for these robots is that they obtain such information faster and more accurate than aid workers, whilst they can squeeze into spaces that are too small for men to enter.

Inside unstructured and partially collapsed buildings in disaster areas, there are unknown, and cluttered parts known as 'life safe voids' [9]. In Urban Search And Rescue (USAR) operations, robots are sent inside these confined spaces. A robot platform on order of 0.33 m^2 or even less is required to enter these voids via their narrow pathways. Traditional tracked robots are not always able to crawl inside these small voids and look for survivors. Additionally USAR has an extremely rugged terrain, so it demands power, while smaller robots carries smaller batteries. Unfortunately most of the energy might be spent getting to an area of interest – i.e. *life safe voids*. As a result of these limitations we assume a marsupial team [1] consists of a mother robot which transports and supports one or more daughters.

Our goal is to evaluate the performance of multi-robot search techniques inside the structural collapses, for limited sensing mini robots, by reproducing various cluttered test fields. Standardisation of the challenges, through the use of simulated search terrains, can enable a meaningful comparison of exploration algorithms. It seems reasonable to describe cluttered and confined voids as a mixture of open space voids with maze-like paths with the connectivity of adjacent places (connected cavities). Additionally a *discriminative index* is required to define the complexity of search fields and provide reference problems through performance metrics. In other words, it allows the researchers to create synthetic fields with which to compare their search strategies.

Studies have ascertained that traditional performance measures were incomplete and inadequate for analysing control and exploration tasks inside complex environments [13]. Therefore they introduced a novel analysis approach based on *fractal path tortuosity*. This index shows how well a robot can handle the difficulties in an environment, from robots movement path. For instance the movement path of the robot inside an environment with a high density of obstacles and closed passages (dead ends) has a larger fractal path tortuosity than to an open space environment with few obstacles. However , we verify that the fractal dimension of pathways (tortuosity) inside the search fields, can be also applied as a discriminative index to characterise test arenas for exploration robots.

32.2 Converting Real World to Simulated World

Development in the area of mobile robotics exploration strategies is often supported by a simulation, since it is significantly cheaper and allows much faster and more flexible experimentation than using physical robots. Furthermore, modelling the

USAR scenario, particularly pancake collapses, are not very practical. Thus, we map the real world to a simulated world. To be able to convert the real world to the simulated world we describe three major elements developed for the simulation: *World model*, *Robot model*, and *Configuration interface*.

– World Model: Based on analysis of a variety of disaster site images and fire fighters reports, we identified the simulated world. The basis for measuring performance are defined according to the published performance guide for Urban Search and Rescue robots by the US Department of Homeland Security [11].

The simulation environment is divided into several equal square cells. Defined *2D simulation environment* frame works can be defined with different frame sizes.

– Robot Model: In this activity we have selected Millibots as the most suitable robot for our application. We will model this robot as our agent and send it inside our grid cell simulation environment to perform various exploration algorithms. The robot should be able to navigate inside various dense, cluttered and confined search fields. The natural choice for motion inside such an environment is the *obstacle avoidance methods*. In this regard there are several sensor – action schemes define in this thesis. All simulated robots are equipped with a ring of eight simulated Infrared (IR) sensors spaced *45 degrees* apart of each other. An IR range reading provides information concerning empty or occupied volumes in the space subtended by the beam (45° cone) for the present sensors that covers adjacent grid cells around the agent. It has been considered that each robot has the view of eight cells in its surrounding area. Cells can be empty or occupied. Mobile agents also can move one cell only in any direction, if the chosen cell is empty. They are aware of their direction, e.g. North, with their equipped simulated gyroscope. Additionally we assume that they are able to detect trapped victims from their heat (body temperature), and CO_2 signature.

Victims are created as having parts that are detectable (e.g. uncovered head) and parts that are undetectable, they are covered. To distinguish live victims from other warm objects, the agent is also equipped with an *NDIR* (non dispersive infrared absorbance) sensor, which is able to recognise humans from their CO_2 signature. Parts of victims undetectable by both IR and NDIR are classed as obstacles. Agent can select of the adjacent cells, according to its sensor readings, to move to, if it reads empty space. We define a fuzzy system with a fuzzy number 'D' that its true values are defined as $D = [-180, +180]$. To fuzzify D we should apply grade of membership on linguistic variable that are the equivalent of the interval D. This is known as fuzzy set '*Degree*' with members of $\{\theta_E, \theta_{NE}, \theta_N, \theta_W, \theta_{NW}, \theta_S, \theta_{SE}, \theta_{SW}\}$.

A simple input–output relations is described by set of rules. A simple fuzzy system is introduced. The first step for designing the fuzzy system is to partition the input space as illustrated in Fig. 32.1a. The second step is deciding the control value of each rule area. In this regard when single input e.g. θ_E is selected by the agent, agent applies the fuzzy rules according to Table 32.1 therefore the output is for the agent to move east. The knowledge of east for the agent is defined in Table 32.1.

The last stage for designing a fuzzy system is to determine the final system output from the designated fuzzy rules. In our fuzzy rule set, each rule has a 45 membership values, we need to define what is the strength of the rules for 45 membership values.

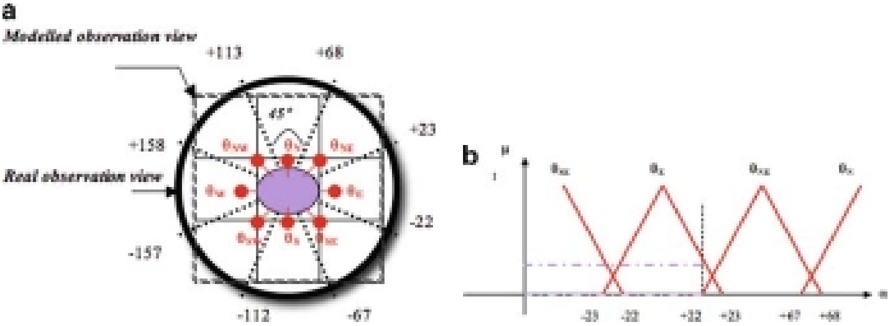


Fig. 32.1 (**a**) Eight segmented vision by infrared sensors, (**b**) triangle norm operator for the fuzzy rule members

Table 32.1 Angle grid fuzzy rule table

Input	Fuzzy rule	Output
θ_{NE}	**IF** $\alpha \in \{+23, \ldots, \mathbf{+45}, \ldots, +67\}$	**THEN** agent moves north east
θ_{N}	**IF** $\alpha \in \{+68, \ldots, \mathbf{+90}, \ldots, +112\}$	**THEN** agent moves north
θ_{NW}	**IF** $\alpha \in \{+113, \ldots, \mathbf{+135}, \ldots, +157\}$	**THEN** agent moves north west
θ_{E}	**IF** $\alpha \in \{-22, \ldots, \mathbf{0}, \ldots, +22\}$	**THEN** agent moves move east
θ_{SE}	**IF** $\alpha \in \{-23, \ldots, \mathbf{-45}, \ldots, -67\}$	**THEN** agent moves south east
θ_{S}	**IF** $\alpha \in \{-68, \ldots, \mathbf{-90}, \ldots, -112\}$	**THEN** agent moves south
θ_{SW}	**IF** $\alpha \in \{-113, \ldots, \mathbf{-135}, \ldots, -157\}$	**THEN** agent moves south west
θ_{W}	**IF** $\alpha \in \{-158, \ldots, \mathbf{180}, \ldots, +158\}$	**THEN** agent moves west

We applied the fuzzy number operator Fig. 32.1b. We define membership grade for each value. μ is the membership grade of the our fuzzy set. For instance for moving agent's EAST, $\mu = 1$ when $\alpha = 45$. θ_E and θ_{NE} intersect at grades of membership $\mu = 0.045$ and $\mu = 0$ for $\alpha = 22$. The greatest of these is 0.045 that is the accepted value of the comparison. Thus, for $\alpha = 22$ the agent moves to east.

All our agents are able to communicate with each other, robot–robot interaction. They interact with each other by sending visual information through their eight triangular *RGB* full colour LEDs. Agents should communicate visually while wireless communication is unreliable in indoor environments. Visual communication as well is restricted to one cell only in each direction.

– Configuration interface: Our configuration interface is based on the*object-oriented CLOWN* formalism that controls the agents exploration techniques in addition to search field development. CLOWN formalism is based on Petri nets approach. It defines many object-oriented concepts, with special regards to a net-based notion of inheritance. The main building block of a CLOWN specification is the elementary class. The semantics of a class is given by corresponding objects. The communication between objects is managed by the equally synchronous execution of corresponding methods.

32.3 Simulated Agent Tasks

In this section, we define our simulated agents by describing their functionality and tasks. There are two main tasks defined for our modelled robots inside the grid cell simulated environment:

(a) They are in charge of creating search fields, as a result generating *life safe voids* [9]. For this task they perform a random walk and looking for victims – i.e. ANT algorithm. Their trail across the field will be recorded and defined as the pathways while unexplored cells will be subsequently selected as obstacles.

(b) They are used to evaluate the performance of search robots modelled inside the simulation environment of various exploration algorithms.

Our simulation tool simulates robots on the sensor level. In this framework we only modelled one kind of robot, i.e. millibot. Thus, there is only one size of robot (i.e. 80 mm) and they are able to move only one cell in each time steps. All the modelled robots have a `SENSOR-SET` that includes simulated Infrared, gyroscope, and NIDR sensors. The user is able to reconfigure the `SENSOR-SET`. Moreover we are able to send different number of robots inside the simulated search filed, from one robot to a team of hundred robots.

The ANT algorithm [12] is one of the most popular random exploration techniques. In this work the ANT algorithm has been selected to generate simulation environments, testing the developed fields to introduce discriminative index and we also compare its performance to our novel algorithm. When agent ANTs are released at the entrance of the confined simulation search field, they immediately start their random walk inspired by "Brownian movement" [7]. We define the Brownian motion on our Ant agent, on a regular lattice, as follow: Suppose that the ant executes a random walk by taking a step every time step on a *2D lattice*, ant starts at time $t = 0$ from an arbitrary point of the lattice. The number of possible orientations is $\phi = 4$ for $2D$ square lattice (i.e. grid cell). Our Ant agent walks in each direction with an equal probability, the probability of choosing any given walking direction is $P = \frac{1}{\phi}$ at each time step. These Ant agents follow a straight line of random length headed to some initial random direction. When they reach the end of this segment, a new direction is selected. They can select one direction out of four directions $\{\theta_E, \theta_N, \theta_W, \theta_S\}$ Agent selects one of the sensor directions randomly if it reads empty space. We eliminate the intermediate directions (e.g. NE) to have equal steps.

There are boundaries defined to limit the environment and set the bounded field for agents to navigate within. Agents should change direction every time they meet the borders or occupied cells. This is also known as the end of their initial segment. However there is a selected time limit for all ANT agents. Every time an agent reaches its time limit it choose a random direction and resets the ω. The generation of Brownian random walks, requires random numbers that fall into a Gaussian distribution. The *Box–Muller* [2] method is used to generate a pair of Gaussian random numbers (GRN) from a pair of uniform numbers. Time segment lengths (i.e. ω) are randomly drawn every time a new direction is chosen, and it is derived as

$$\omega = \left| \frac{1}{GRN} \times \Delta t \right| seconds \tag{32.1}$$

The time delay and therefore the time step, is here assumed to be $\Delta t = 36$ seconds – corresponding to that of *Millibot*.

32.4 Simulated 2D Search Field

The main goal of generating any test bed is to facilitate the trial and evaluation of idea that have promise in the real world. As discussed before inside structural collapses there are confined voids connected through narrow pathways [3]. These life safe voids are the most demanding parts for exploration robots. Search robots should explore the entire maze of obstacles and debris in search of victims and the appropriate exploration technique must secure that the overall victim discovery time is minimised.

Our simulated agents, by performing Ant Brownian movement, can develop various $2D$ simulated environments. In this section, we describe the algorithmic approach to introduce a search field generator. index to differentiate these generated test fields. The simulation environment is divided into several equal square cells, length L. Here we have selected the grid size to be $L = 100$ mm.

In real search and rescue scenarios, for a fast observation larger robots are applied. They are able to avoid obstacles, move over rubbles, and have a larger line of sight comparing to small robots. However they are not able to squeeze inside small gaps and voids. Due to limitation in their functionality they can deploy mini robots whenever it is required to perform an exhaustive search operation. Thus in our scenario there is a mother robot [8] sitting at the start point and sending in smaller robots, one by one inside the bounded field to perform the next stage of the search operation.

There are four state cells available for our simulation grid search field:

1 – Occupied by obstacle or undetectable victim's body part, **2 –** Occupied by victim's detectable part i.e. only two adjacent cells, **3 –** Occupied by other agent, **4 –** Empty. The first three states make the cell un-traversable by the agents. We assumed that each agent occupies only one cell every time it moves. Any cell that is explored by our agent will be indicated on the simulation field. Therefore empty and unexplored cells are white, while explored cells are illustrated with grey colour. This feature demonstrates the time sliced evaluation of agents behaviour, e.g. how well the strategy is able to spread its agents inside the search field.

To generate various confined and cluttered 2D simulation environments we follow this algorithmic approach [10]:

Algorithm 1 Search Field Generation

```
Input: Frame size; No. of victims; No. of agents;
Output: Frame work;

Locate the victim(s) randomly inside the empty search field
Select a narrow restricted entrance in the corner (start point).
while Victim is not located do
    Send in the agent(s) to perform ANT algorithm
    Record explored cells as pathways & untouched cells as obstacles.
end while
```

To create a life safe void, the victim(s) is located randomly inside an empty field. We are able to change the number of victims and robots and follow the steps above to generate various simulation environments. All the paths taken by all the agents will be considered as pathways and untouched free cells will become obstacles. In this case we may have multi paths, with only one leading to the victim. Therefore, by running our ANT algorithm several times we are able to generate various search paths inside the empty field. The advantages of such a search field generator are that it creates an accessible space, with a complexity that is rigidly defined through its fractal dimension.

32.5 Discriminative Index

There are several discriminative indices that is able to differentiate landscape patterns. In this regard, *fractal dimension* is an index of complexity of shapes on landscapes. Fractal dimension provides an objective way to quantify the fractal properties of an object and distinguish it from other similar objects. For instance the rugged coastline of Great Britain has been calculated to have a fractal dimension about 1.25, while the smooth South Africa is only slightly rougher than a straight line, with the fractal dimension of 1.02.

In literature, self-affine fractal dimensions have been used in various applications to indicate fractal complexity. Every fractal has a numeric fractal dimension that can be used to indicate fractal figure complexity. There are several methods to calculate fractal dimension. The box counting method has been used in various applications such as graphic image processing [4]. Box counting superimposes a regular box of length γ on the object and counts the number of occupied cells (n_i) in every type of cells. We follow the power law relationship defined by Voss et al. [14] to calculate the fractal dimension of our simulation search field. The Power-law relationship is

$$n_i = K\gamma^{-D} \tag{32.2}$$

Where K is the total number of grid cells available on our search field of all states "$K = \sum_{i=1}^{N} n_i(\gamma)$". We have N states of cells where n_i is the number of individuals belonging to the i_{th} types of cells ($i \in I = \{1, 2, \ldots, N\}$) with the size of γ. Here we consider only two types of cells (occupied and free, $N = 2$).

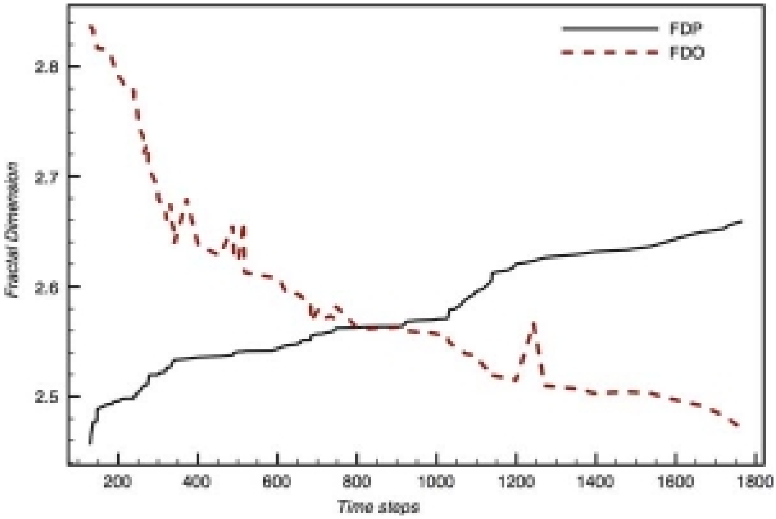


Fig. 32.2 Fractal dimension of obstacles and pathways (one time step = 36 s)

D is a metric dimension, therefore its definition depends on metric scaling properties. Figure 32.2 illustrates the correlation between victim discovery time and fractal dimension. 101 simulated search fields (generated by our search field generator) have been chosen to determine the most suitable index for search terrain complexity. These fields are differentiated by their FDO (Fractal Dimension of Obstacles) and FDP (Fractal Dimension of Pathways). Fractal dimension, considering the obstacles, has a negative correlation to time steps (with little fluctuations), in other words to search field complexity. While there is a positive correlation between fractal dimensions and exploration terrain, considering the pathways. Thus FDP is selected as our best discriminative index (with a less fluctuation) for the search field complexity. The larger the FDP, the more complex is the generated search terrain. Each time step in our graphs is the average of running the ANT algorithm 100 times.

The default values in our experiments are: a frame size of 33 × 33 squared grid cells, and as discussed above we vary the environment according to their fractal dimension. The number of the agents in this experiment was set as the best performance for ANT algorithm. Thus we selected three agents to run the experiment. In the result of this experiment we were convinced to use FD of path tortuosity as complexity index in our search field generator. This index is able to differentiate between random developed environments according to their path complexity, since Brownian movement of our agents has developed them.

FracLac [5] is an image analyse software, developed for image J tool. This tool is used for objectively analysing complexity and heterogeneity as well as some other measures of binary digital images. It has a global binary grid scan option that applies box-counting technique across an image. It calculates the fractal dimension of a complex search environment using the mathematical procedure presented above.

Therefore all users are able to produce the search fields by our algorithmic approach and save the image as a binary image (0 for obstacles and 1 for pathways) and scan it by J image to estimate the FDP of the generated search fields.

Search agents are sent inside the void from restricted entrance path to explore the unknown search environment, with its various branches in different lengths. The average time steps that single agent takes to locate a first trapped victim in each field, is recorded as *victim discovery time*. This is our primary fitness metric applied to select the discriminative index for our generated search fields.

32.6 Validation

In Fig. 32.3 we compare TREE exploration Algorithm against ANT algorithm for a range of fractal dimensions (i.e. 100 search fields). TREE algorithm, an implementation of the ANT on-line STC algorithm by Gabriely and Rimon [6]. In this algorithm the agents run a Depth-first Search (DFS) algorithm during the incremental spanning tree construction. We fixed the size of sweep covering width *D* to the length of our grid cell, i.e. 100 mm.

The default values in this experiment are defined as same as above. Three agents are sent in to perform a search algorithm to explore a single trapped victim. In contrast to fractal dimension, which is independent of goal positioning, discovery time is highly sensitive to goal state positioning for all algorithms except the ANT algorithm. The graph depicts the average VDTs of 2,000 executions in which a two cells goal was placed in 20 different positions, and located 100 times.

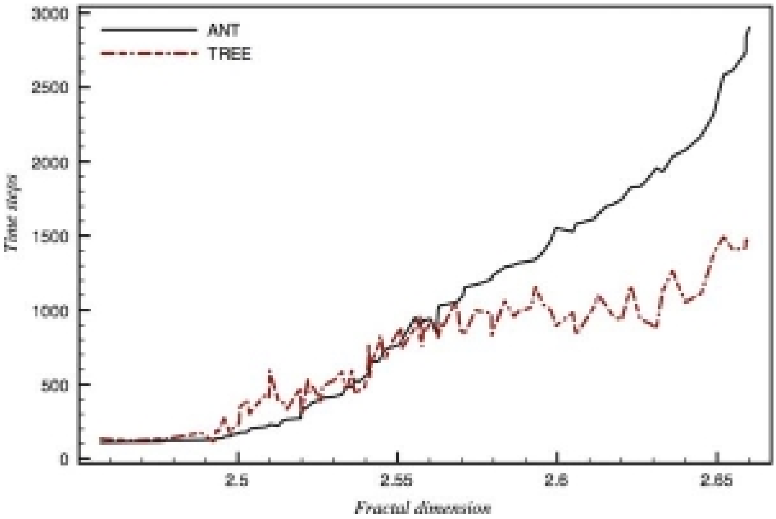


Fig. 32.3 TREE algorithm against ANT algorithm (one time step = 36 s)

From the experimental results ANT algorithm simulated search robot is repeating the same old behaviour. However, the TREE algorithm is able to reduce the redundancy of revisiting the search environment inside the narrow gaps and small voids of a collapsed building. Each time step in our graphs is the average of running the exploration 100 times.

The larger is the FDP, the more complex is the generated search terrain. Thus ANT agents, with high redundancy behaviour, spend a longer time to discover a victim.

32.7 Discussion

By learning new cooperative behaviours, multi-robot system is able to minimize the overall victims discovery time, inside the narrow gaps and small voids of a collapsed building. However there is no such a search field generator available to generate various random after-disaster search fields to allow researchers for detailed testing of their exploration techniques, prior to physical tests. We discuss how to model the dangerous, cluttered and confined parts of NIST red course in a simulation environment. By deploying fractal path tortuosity, search field generator is able to differentiate its random confined fields.

Small robots must sacrifice mobility, sensing, and power to achieve their desired scale. However to remain effective, they must adopt new techniques to overcome these limitations.

References

1. Balch, T.: Behavioural diversity in learning robot teams. Ph.D. thesis, College of Computing, Georgia Institute of Technology, USA (1998)
2. Box, G.E.P., Muller, M.E.: A note on the generation of random normal deviates. Ann. Math. Stat. **29**(2), 610–611 (1958)
3. Jacoff, A., Messina, E., Evans, J.: A reference test course for autonomous mobile robots. In: Proceeding of SPIE-AeroSense Conference, Orlando (2001)
4. Jones, A.L.: Image segmentation via fractal dimension. Technical Report, Air-Force Institute of Technology, Wright-Patterson AFB, Ohio (1987)
5. Karperien, A.: http://rsb.info.nih.gov/ij/plugins/frac-lac.html. 2008
6. Matsuo, Y., Tamura, Y.: Tree formation multi-robot system for victim search a devastated indoor space. In: Proceeding of IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), vol. 2, pp. 1071–1076. Sendal, Japan (2004)
7. Merloti, P.E., Lewis, J.: Simulation of artificial ants behaviour in a digital environment. In: Proceeding of International Conference on Artificial Intelligence (ICAI05), Las Vegas (2005)
8. Murphy, R., Assumes, M., Bugajsk, M., Johnson, T., Kelley, N., Kiefer, J., Pollock, L.: Marsupial-like mobile robot societies. In: Proceeding of Fourth International Conference on Autonomous Agents, pp. 364–365. ACM (1999)
9. Murphy, R.R.: Activities of the rescue robots at the world trade centre from 11–12 september 2001. Proc. IEEE Robot. Autom. Mag. **3**, 851–864 (2004)

10. Saeedi, P., Sorensen, S.A.: An algorithmic approach to generate after-disaster test fields for search and rescue agents. In: Proceedings of International Conference of Word Congress Engineering(IAENG-WCE), London (2009)
11. Security, Homeland: National Institute of Standards & Technology – Urban Search and Rescue Robot Performance Standards (2008)
12. Svennebring, J., Koenig, S.: Building terrain-covering ant robots: a feasibility study. Auton. Robots **16**(3), 313–332 (2004)
13. Voshell, A.W.M., Woods, D.D.: Overcoming the keyhole in human–robot coordination: simulation and evaluation. In: Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting (2005)
14. Voss, R.F.: Fractals in nature: from characterization to simulation. The Science of Fractal Images, pp. 21–70, Springer-Verlag, New York, USA (1988)

Chapter 33
Diagnostic Problem Solving by Means of Neuro-Fuzzy Learning, Genetic Algorithm and Chaos Theory Principles Applying

Stefania Gallova

Abstract The complexity and reliability demands of contemporary industrial systems and technological processes require the development of new fault diagnosis approaches. Performance results for finding the best genetic algorithm for the complex real problem of optimal machinery equipment operation and predictive maintenance are presented. A genetic algorithm is a stochastic computational model that seeks the optimal solution to an objective function. A methodology calculation is based on the idea of measuring the increase of fitness and fitness quality evaluation with chaos theory principles applying within genetic algorithm environment. Fuzzy neural networks principles are effectively applied in solved manufacturing problems mostly where multisensor integration, real-timeness, robustness and learning abilities are needed. A modified Mamdani neuro-fuzzy system improves the interpretability of used domain knowledge.

Keywords Mamdani neuro-fuzzy system · genetic algorithm · fitness · fuzzy rule · chaos theory · metric entropy

33.1 Introduction

Due to the large number of process variables and their complex interconnections in a machinery environment, the pertinent knowledge system is mostly qualitative and incomplete.

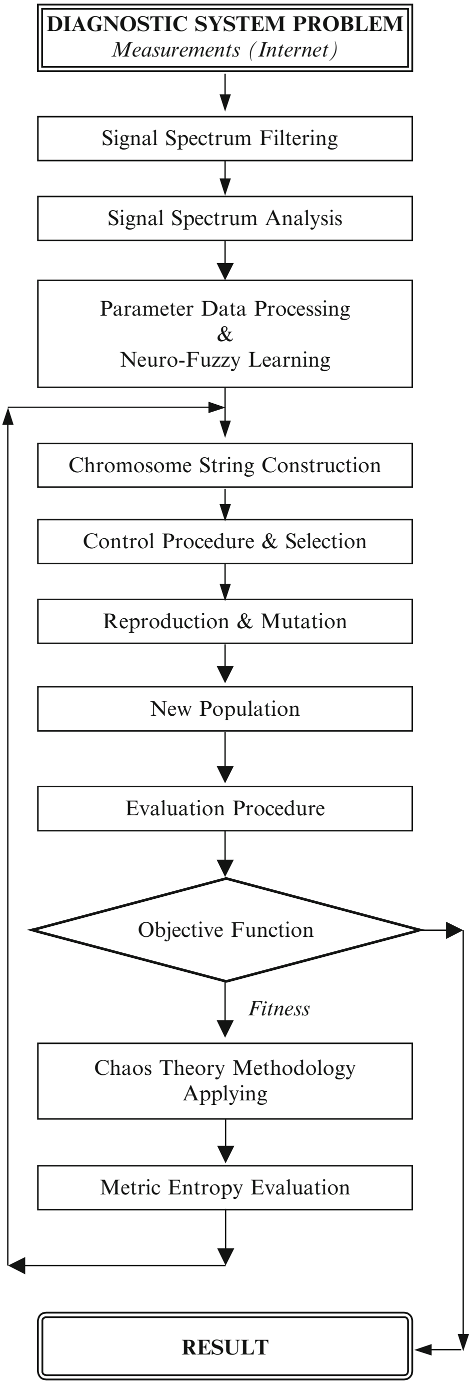
A cardinal rule of solved approach is illustrated in Fig. 33.1.

The diagnostic parameters of machine can be represented by a linguistic variable with fuzzy set definition through "fuzzyfication" in terms of a defined fuzzy membership function. Fuzzy rules describe the qualitative relations between the major

S. Gallova (✉)
Pavol Jozef Safarik University in Kosice, Srobarova 2, SK-041 80 Kosice, Slovak Republic
e-mail: stefania.gallova@zoznam.sk; stefania.gallova@upjs.sk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_33,

Fig. 33.1 Optimization problem architecture



operation conditions (i.e. various operation conditions, vibrations, tool accuracy, safe load, working load and so on). In practice, classical tools solve this task by using of sensors great number, or by special sensors creating. This approach increases costs and reduces the reliability. It is possible to compensate this unfavourable state effectively by higher quality of analysis at lower requirements for definiteness of inputs [1].

Chaos theory is a progressive field, which in this time is not sufficiently utilized at failure analysis as possible effective tool. Nowadays, used tools of analysis, or statistic parameters of random phenomena do not find out all possibilities, that is all possible information about real technical state of equipments that the given random signal could provide. Chaos theory principles provide very valuable complementary information, which is not possible to obtain by another means of analysis. Genetic algorithm is a suitable medium for effective application of chaos theory principles for real diagnostic problem solving. It is possible to work with inaccurate, vague, indefinite, non-unambiguous information that real practice in predominant extent provides.

Fuzzy neural networks principles are effectively applied in solved manufacturing problems as a first step mostly where multisensor integration, real-timeness, robustness and learning abilities are needed. A modified Mamdani neuro-fuzzy system improves the interpretability of used domain knowledge by parameter data processing.

These progressive approaches to complex diagnostic problem solving of practice provide the high effective tools for optimal machine condition and for security of machine equipment reliability.

33.2 A Mamdani Neuro-Fuzzy Network Module

The idea of the proposed methodology is based on the parameters estimation of the diagnosed system. The comparison of the estimated parameters with the a priori estimated parameters of the nominal system allows performing both the fault detection and identification.

There is removed the last layer performing the division thus the system has two outputs. The error on the first output will be computed taking into account desired output from learning data. Desired signal on the second output is constant and equals "one" (see Fig. 33.2). After learning causes structures according to the initial idea, we can build the modular system. In such system the rules can be arranged in arbitrary order (see Fig. 33.3). At the beginning of the numerical simulation, input fuzzy sets are determined by the modified fuzzy c-means clustering algorithm. Then, all parameters are tuned by the backpropagation algorithm [1–5].

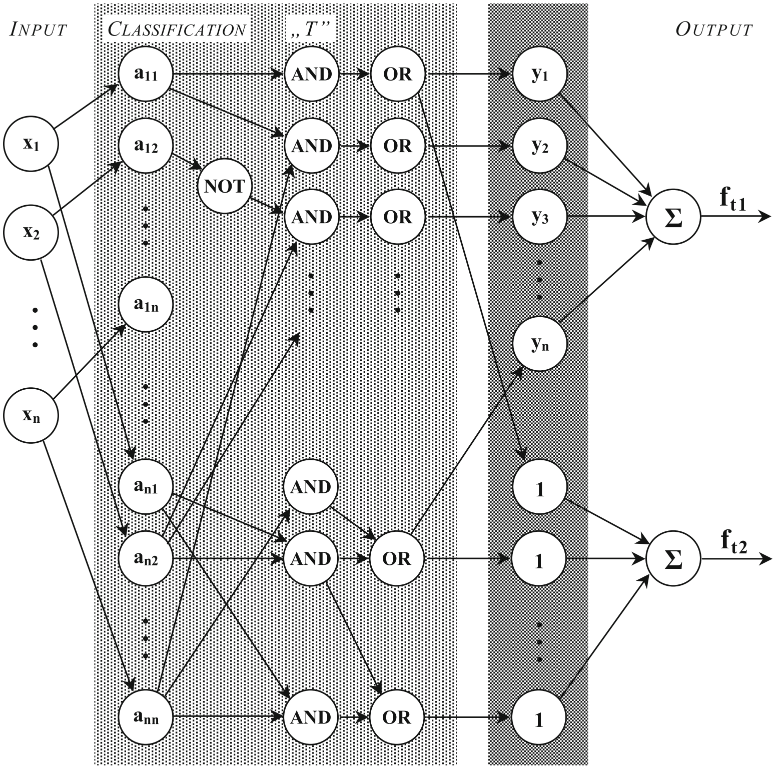


Fig. 33.2 A modified neuro-fuzzy system architecture

Solved network was trained and the related equations are presented below:

$$x(t) = [x_1(t), \ldots,\ x_n(t)] \qquad f(r_j(t)) = v_j(t)$$

$$f(e_j(t)) = y_j(t) \qquad r_j(t) = \sum_{i=1}^{n+m} w_{ij}\ x_i(t)$$

$$e_j(t) = \sum_{i=1}^{k} w_{ij}\ v_i(t) \tag{33.1}$$

where x(t) is the input vector, r_j(t) and l_j(t) are input signals provided to the hidden and output layer neurons. Parameter k stands for the size of the hidden layers. Parameters v_j(t) stands for the neurons activations in the hidden layer at time t, and y_j(t) stands for the activations of the neurons in the output layer at time t, parameters w_{ij} are neurons weights values.

The process under consideration within experimental implementation is diagnostic system for solved complex manufacturing processes control. For example,

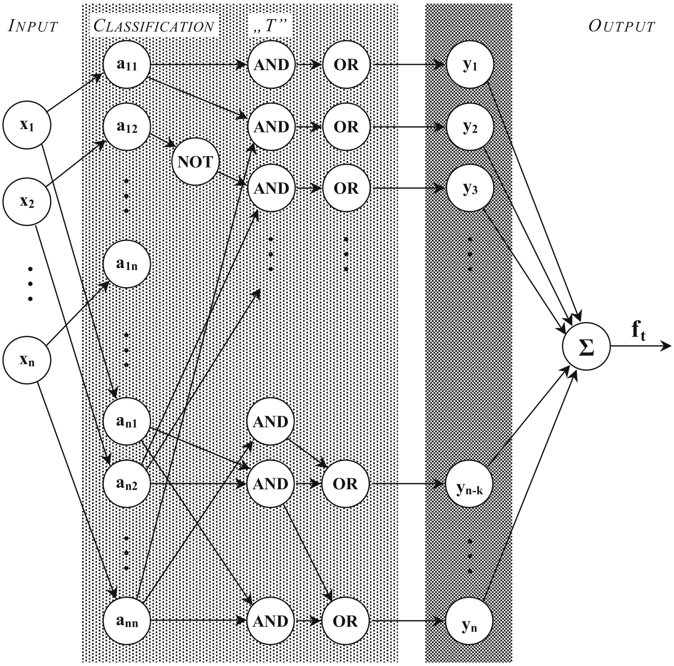


Fig. 33.3 A solved neuro-fuzzy system architecture after learning

accuracy of the top composition model is studied. The fundamental model is used as the real process within identification experiment. We obtain two sets of data, namely training and test sets.

Neuro-fuzzy systems using can provide better results than the one based on traditional indices. Classifiers can be combined to improve network accuracy. Solved subsystems are trained by different datasets. There is also realized a detection in web implementation.

33.3 Problem Solving by Means of Chaos Theory Principles

Further, we analyze appropriate indicators of chaos approach, which are first of all fractal dimensions, metric entropies and Lyapunov exponents [6–12].

A general used formula for calculation of multidimensional vector data code is given by:

$$\frac{(k^2 - k + 1)}{R} = V_1 \times V_2 \times \ \cdots \ \times V_i \quad with \quad (V_1,\ V_2, \quad \cdots \quad V_i) = 1 \quad (33.2)$$

We enumerate a set of nodes of $V_1 \times V_2 \times, \ldots, \times V_i$ (grid). Each node of the grid meets exactly R-times.

We consider fuzzy rule base of Takagi-Sugeno system form. The Takagi-Sugeno fuzzy model is described by a set of fuzzy implications, which characterize local relations of a system in the state space [1]. The main feature of this model is to express the local dynamics of each fuzzy rule, i.e. implication by a linear-state-space system model, and the overall fuzzy system is then modeled by fuzzy "blending" of these local linear systems models through some suitable membership functions. Approaches to chaotify discrete-time and continuous-time Takagi-Sugeno fuzzy systems are very different. For simplicity, we have a following illustrative example.

We consider a nonchaotic discrete-time Takagi-Sugeno fuzzy model, given by these rules:

$$\begin{aligned} &IF\ x(t)\ is\ M_{f1} \quad THEN \quad x(t+1) = H_1\ x(t) + v(t) \\ &IF\ x(t)\ is\ M_{f2} \quad THEN \quad x(t+1) = H_2\ x(t) + v(t) \end{aligned} \tag{33.3}$$

with $x(t) \in [-b, +b], b > O$ with the membership functions M_1, M_2, where:

$$H_1 = \begin{bmatrix} b & 0,288 \\ 1 & 0 \end{bmatrix} \qquad H_2 = \begin{bmatrix} -b & 0,288 \\ 1 & 0 \end{bmatrix}$$

$$M_1 = \frac{1}{2}\left(1 - \frac{x(t)}{b}\right) \qquad M_2 = \frac{1}{2}\left(1 + \frac{x(t)}{b}\right) \tag{33.4}$$

The controlled Takagi-Sugeno fuzzy system is described as follows:

$$x(t+1) = \sum_{i=1}^{n} \mu_i(t)\ H_i\ x(t)\ v(t) = \sum_{i=1}^{n} \mu_i(t)\ H_i\ x(t)\ + \delta\ \sin\left(\frac{\pi}{\delta}\ \beta\ x(t)\right) \tag{33.5}$$

The controller is taken as a sinusoidal function. In the simulation, the magnitude of the control input is experimentally chosen to be $\delta = 0.09$. Thus, $||v(t)||_\infty < \delta$, and can also be regarded as a control parameter. Without control, the Takagi-Sugeno fuzzy model is stable.

Let the solved dynamical system is generally described by the differential equation [1, 8, 9]:

$$x(t) \equiv \frac{dx}{dt} = F(x,\ \mu) \tag{33.6}$$

where x is a vector whose components are the dynamical variables of the system, $F\ (x, \mu)$ is a nonlinear function of x and μ stands for the control parameter. Let $x_1(\mu)$ and $x_2(\mu)$ be two different solutions of Eq. (33.6). A critical point is therefore defined by the equation:

$$x_1(\mu_c) = x_2(\mu_c) \tag{33.7}$$

which is an implicit equation for μ_c. The Jacobian matrix $G(x, \mu)$ associated with $F\ (x, \mu)$ is defined through:

$$F(x + \delta x, \mu) = F(x, \mu) + G(x, \mu)\delta x + O(\delta x)^2 \tag{33.8}$$

The fundamental property that can be proved by using the implicit function theorem is:

$$\det G(x, \mu_c) = 0 \tag{33.9}$$

This result implies critical slowing down. This is easily shown by a linear stability analysis of the solution $x_1(\mu)$. We have the following assumption:

$$(x, \mu) = x_1(\mu) + \varepsilon x'(\mu) \cdot \exp(\lambda t) + 0 \cdot \varepsilon \tag{33.10}$$

The characteristic equation for λ is:

$$\det\{\lambda \mathrm{I} - G(x_1, \mu)\} = 0 \tag{33.11}$$

where I is the unit matrix, the ε-exponent is interpreted as a dimension. Hence at $\mu = \mu_c$ one root (at least) of Eq. (33.11) will vanish, implying an infinite relaxation time. This result remains true if $x_1(\mu)$ and $x_2(\mu)$ are time-periodic solutions of relation Eq. (33.6). We realize the direct estimate of fractal dimension from experimental data (embedding theorem and related topics) with a particular attention to the effect of filtering on a chaotic signal.

We first define a *partition* (covering) of the phase space, as a collection of disjoint open sets with variable size ε_i. In this way, we can associate a mass p_i to each element E_i of the partition as:

$$m_i = \int_{E_i} d\mu \tag{33.12}$$

The mass m_i can be evaluated from the fraction of points belonging to E_i, when a sufficiently large number of the set points, is generated according to the measure μ. From the obvious consideration that $m \sim \varepsilon^1$ in the case of a plane, it follows that the ε-exponent is to be interpreted as a dimension. Accordingly, we can define a *local* dimension α_i in terms of size and mass of the ith element of the covering: $m_i \sim \varepsilon_i^{\alpha\, i}$ where ε_i is assumed to be sufficiently small, and we let α_i explicitly depend on the index i.

This is a crucial point that makes also a statistical approach most appropriate. Metric entropy using improves this approach.

We start with metric entropies which are defined in terms of a sequential measurement, i.e. a series of observations of a trajectory at equally spaced times ($t_n = n\Delta t$, in the case of continuous time, and $t_n = n$ for maps). To be more specific, we consider a discrete-time dynamical system. We have the mapping that is strictly deterministic. We have the initial condition x_o with infinite precision, the trajectory $x_{n=}F^n(x_o)$ is uniquely determined. Let us also assume that an observation of the system be done with limited resolution. Therefore, we introduce a partition of the phase-space in L regions E_j. When the representative point x_n is in the element E_j, the *reading instrument* displays the value j. In this way, a symbolic sequence $S_N = \{s_n,\ n = 1, \ldots, N\}$ can be associated to each trajectory $x_n,\ n = 1, \ldots, N$ (for finite N, S_N is also named word). The symbol $s_n \in [1, L]$ represents the index of the partition element visited by the trajectory at time n. The order of symbols in the word is crucial in determining the metric entropy. The conditional probability of being in subset E_{ij} at time n, given an initial condition in E_i, is:

$$p(j,n\,|\,i) = \frac{\mu(F^n\, E_i\ \cap E_j)}{\mu(E_i)} \tag{33.13}$$

where μ is the invariant measure. The mass $\mu(E_i)$ is the same as that contained in $F^n\, E_i$, and the conditional probability is normalized in such a way that

$$P(j,\ 0|\mathrm{I}) = \delta_{\mathrm{i,j}} \tag{33.14}$$

We have for a mixing system:

$$\lim_{n\to\infty} p(j,n\,|\,i) = \mu(E_j) \tag{33.15}$$

That is, the image of the initial set covers the whole attractor and no correlation survives between arrival and starting element. The action of the map F in phase-space is translated into a shift of symbols in the associated space:

$$x_n \to x_{n+1} \Rightarrow \{\ldots,\ \mathrm{s_{n-1}},\ \mathrm{s_n},\ \mathrm{s_{n+1}}, \ldots\} \to \{\ldots,\ \mathrm{s_n},\ \mathrm{s_{n+1}},\ \mathrm{s_{n+2}}, \ldots\} \tag{33.16}$$

The time origin in the (doubly-infinite) symbol sequence is moved one place to the right. When a generating partition is not known, as in the case of experimental data or in most of computer simulations, it is possible to divide the searching space of size $\varepsilon_\mathrm{i} < \varepsilon$ (they are usually taken all equal for simplicity) and evaluate the *metric entropy* as:

$$K(q) = \lim_{\varepsilon\to 0}\ \lim_{N\to\infty} \frac{1}{N}\frac{1}{(1-q)}\ \ln \sum_{S_N} p^q\ (S_N) \tag{33.17}$$

where the limit $\varepsilon \to 0$ guarantees that a generating partition is finally obtained. The performed analysis assumes implicitly the knowledge of all coordinates of each attractor's point in phase space. This is certainly the case of all numerical simulations, but it is not always possible in experiments. Sometimes, just one variable can be measured at different times.

In this way, we construct the values and parameters of genes within solved domain chromosome. We solve metric entropy for each element or for solved chromosome within genetic algorithm running. This chromosome implementation by genetic algorithm approach is near to real diagnostic situation representation. Nowadays, some experiments were realized. Achieved results are very promising.

33.4 Obtained Results

The Most Significant Decisions Indicators are illustrated in Fig. 33.4.

A remarkable property of the model is an ability to reproduce the maximum number of combinatorial varieties in the systems with a limited number of elements and bonds. Among the experimental variants provided for the diagnostic system problems solving methodologies, the proposed approach clearly dominates the other variants. It provides qualitatively higher performance level of implementation. Practically, there are no instable parameters courses.

Fitness function is near to the best solution (as expected) within solved phase of run. The solved approach requires the more complex real-valued gene coding. This mechanism with the strength and specificity rules management can be effectively assimilated to a genetic operator.

The learnt rules have been tested on the real process. All simulated faults were successfully diagnosed by the corresponding rules and no incorrect diagnosis occurred.

A genetic algorithm has to maintain a balance between the preservation of good combinations of genes, and the exploration of new combinations. We adopt a successful strategy for achieving this balance which has been to combine a highly explorative, or disruptive crossover with elitism, in which a fraction of the best individuals found so far survive into the next generation. Elitism gives better individuals more chances of mating to produce fit offspring, an advantage when their offspring will frequently be poor. We compare *Fitness evaluation* for classical (without chaos

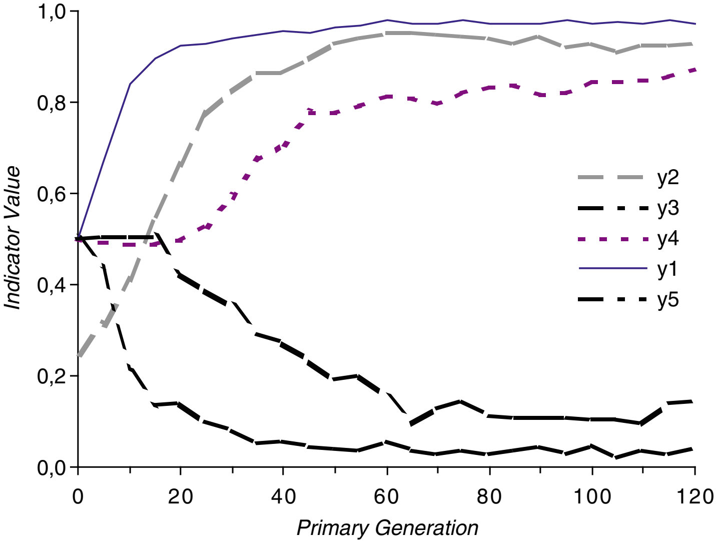


Fig. 33.4 The most significant decisions indicators: y1 = crossover units course, y2 = selection method course, y3 = elitist model course, y4 = mutation function course, y5 = value replacement course

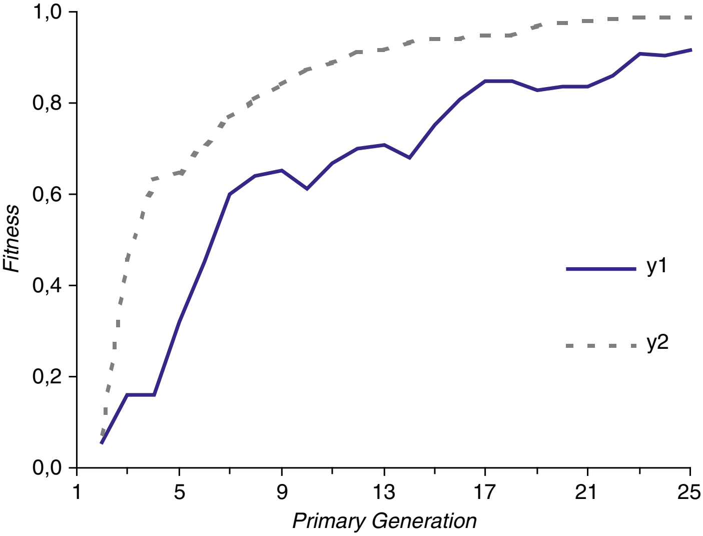


Fig. 33.5 Fitness evaluation course: y1 = Fitness curve for the classical genetic algorithm approach; y3 = fitness curve for the proposed methodology with chaos theory principles applying

theory principles applying) approach and solved methodology approach. Judged input parameters for experimental simulations have been the same for both methodological cases [1, 7]. The result is illustrated in Fig. 33.5.

33.5 Conclusion

Solved neuro-fuzzy system seems to perform both the fundamental requirements of intelligent manufacturing, i.e. real-time nature, uncertainties handling, learning ability and managing both symbolic and numerical information, and the expectation to generate sufficient rule set also for larger problems, which would be handled by usual neuro-fuzzy models only with severe difficulties [1, 13–15]. The solved system is a good illustration of how a genetic algorithm approach to a complex practical combinatorial problem can provide an extremely robust solution with several practical advantages. The main difficulty in the problem is that there typically is a multitude of local extreme, which happen to be located close to a bounding constraints, conventionally imposed at the given threshold, and that anyway has to be imposed out of safety considerations. From this point of view, the proposed methodology achieves the best results. The solved genetic algorithm approach is feasible by implementing it in a multi-transputer environment.

Future research will be focused first of all on improving the runtime performance of solved implementation, including other genetic operators in the architecture and investigating the results of further test problems in more detail. There are will be investigated self-learning approaches of diagnostic rules through more advanced genetic and evolutionary algorithms and modified chaos theory principles. Further research will be also focused on exploring the relation between fuzzy logic approach and chaos theory, and combining fuzzy and chaos control technologies for real applications. Nowadays, some achieved results seem to be very interesting.

We realize the direct estimate of fractal dimension from experimental data (embedding theorem and related topics) with a particular attention to the effect of filtering on a chaotic signal. Genetic algorithm will have been the main rule discovery algorithm. It will concern to obtain self-organisation of a kind of communication protocol among a solved population.

The research reported on this paper is supported by following scientific project of Slovak Grant Agency VEGA (Ministry of Education): *Intelligent Approach to Automated Diagnostic Problem Solving of Machinery Equipment* (Number: VEGA 1/4133/07).

References

1. Gallová, Š.: Some effective techniques applying for complex diagnostic problem solving via genetic algorithm approach. In: Computational Intelligence: Methods and Applications, pp. 183–207. EXIT, Warsaw, ISBN 978-83-60434-50-5 (2008)
2. Abraham, NB., Lugiato, L.A., Narducci, L.M.: Instabilities in active media. J. Soc. Am. B. (1999)
3. Korytkowski, M., Rutkowski, L., Scherer, R.: On combining backpropagation with boosting. In: Proceedings of International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence, Vancouver, Canada (2006)
4. Wang, L.: Adaptive Fuzzy Systems and Control. PTR Prentice-Hall, Englewood Cliffs, NJ (1994)
5. Mrugalski, M.: Neural Network Based Modelling of Non-Linear Systems. University of Zielona Gora, Poland (2004)
6. Li, Z., Park, J.B., Joo, I.H., Chen, G., Choi, I.H.: Anticontrol of chaos for discrete TS fuzzy systems. IEEE Trans. Circ. Syst. 1 **49** (2) (2002)
7. Gallová, Š.: A maximum entropy inference within uncertain information reasoning. In: Information Processing and Management of Uncertainty in Knowledge-based Systems: Proceedings, pp. 1803–1810, Paris, Les Cordeliers, E.D.K., Paris, 2–7 July 2006, ISBN sss-X (2006)
8. Mandel, P., Erneux, T.: Dynamic versus static stability. In: Hilger, A (ed.) Frontiers in Quantum Optics, Bristol, Boston, MA (1986)
9. Badii, R., Politi, A.: Strange attractors. Phys. Lett. **104A**, 303 (1984)
10. Goldman, S.A., Rivest, R.L.: A non-iterative maximum entropy algorithm. In: Koval, L.N., Lemmu, F.J. (eds.) Ucertainty in Artficial Intelligence, Vol. 2, pp. 133–148, North-Holland (1988)
11. Hamming, R.W.: Coding and Information Theory, Prentice-Hall, Englewood Cliffs, NJ (1980)
12. Ballé, P.: Fuzzy model-based parity equations for fault isolation. Contr. Eng. Prac. **7**, 261–270 (1999)
13. Zhang, J., Roberts, P.D.: On-Line Process Fault Diagnosis Using Neural Network Techniques, Institute of Measurement and Control (1992)
14. Montana, D., Davis, L.: Training feedforward neural networks using genetic algorihms. In: Proceedings of the 11th International Joint Coference on Artificial Intelligence, pp. 762–767 (1989)
15. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA (1989)

Chapter 34
The New Measure of Robust Principal Component Analysis

Dyah E. Herwindiati and Sani M. Isa

Abstract Principal Component Analysis (PCA) is a technique to transform the original set of variables into a smaller set of linear combinations that account for most of the original set variance. The data reduction based on the classical PCA is fruitless if outlier is present in the data. The decomposed classical covariance matrix is very sensitive to outlying observations. ROBPCA is an effective PCA method combining two advantages of both projection pursuit and robust covariance estimation. The estimation is computed with the idea of minimum covariance determinant (MCD) of covariance matrix. The limitation of MCD is when covariance determinant almost equal zero. This paper proposes PCA using the minimum vector variance (MVV) as new measure of robust PCA to enhance the result. MVV is defined as a minimization of sum of square length of the diagonal of a parallelotope to determine the location estimator and covariance matrix. The usefulness of MVV is not limited to small or low dimension data set and to non-singular or singular covariance matrix. The MVV algorithm, compared with FMCD algorithm, has a lower computational complexity; the complexity of VV is of order $O(p^2)$.

Keywords Determinant · generalized variance · outlier · principal component analysis · robust · vector variance

34.1 Introduction

Some practical problems arise in data mining when a large number of variable are measured. This is usually due to the fact that more than one variable may be measuring the same information. The one of variables can be written as a near linear combination of the other variables, and the number of correlated variables will

D.E. Herwindiati (✉) and S.M. Isa
Tarumanagara University, Jln Let. Jend. S Parman 1, Jakarta 1140, Indonesia
e-mail: dyah.fti.untar@gmail.com; sani.fti.untar@gmail.com

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_34,

increase when the number of variables increase. To have the good analysis it is necessary to eliminate the redundant information by creating a new set of variables that extract the essential characteristics of the information.

PCA is a technique to transform the original set of variables into a smaller set of linear combinations that account for most of the original set variance. The basic idea of PCA is to describe the dispersion of an array of n points in p – dimensional space by introducing a new set of orthogonal linear coordinates so that the sample variances of the given points are in decreasing order of dimension, Gnanadesikan [1].

A principal component analysis focused on reducing the dimensionality of a data set in order to explain as much information as possible. The first principal component is the combination of variables that explains the greatest amount of variation. The second principal component defines the next largest amount of variation and is independent to the first principal component. This step will be continued for the entire principal components corresponding to the eigenvectors of covariance matrix sample.

The data reduction based on the classical PCA becomes unreliable if outliers are present in the data. The decomposed classical covariance matrix is very sensitive to outlying observations. The first component consisting of the greatest variation is often pushed toward the anomalous observations, so that a single outlier can significantly shift the mean and inflate the covariance matrix.

Barnett and Lewis [2] define an outlier to be data which is inconsistent relative to the other group of data. The word 'not consistent' on the definition is not easy to be formulated in general situations. This reason makes people, up to now; develop better methods in identifying outliers. Identifying multivariate outliers is not trivial; there are many procedures to identify outliers. We need a good method to identify them. A good method must be specific and sensitive. Specific means that it is able to say that a 'good' data is really good, and sensitive means that it is able to say that a 'bad' data is really bad. The concept of sensitive developed more operational after Hampel et al. [3] introduced the influential functions. To satisfy the needs of a good method, Huber et al. [4] introduced a new method for the robust principal component (ROBPCA).

ROBPCA is PCA method combining two advantages of both projection pursuit and robust covariance estimation. Based on our experience in computations, ROBPCA is an effective and efficient method for dimension reduction and identifying of the anomalous observations. The ROBPCA estimator is computed by the MCD ideas of covariance matrix; which is constrained by minimum determinant of covariance matrix (CD). As multivariate dispersion CD is often used in many applications of statistical data analysis, but an insuffiency of CD as measure dispersion is CD $= 0$ not certainly implies that $\vec{X}$ is of degenerate distribution in the mean vector $\vec{\mu}$. The good properties of ROBPCA and the limitation of CD tend us to propose the new measure of robust principal component based on minimum vector variance (MVV).

MVV is a measure minimizing vector variance to obtain the robust estimator. The vector variance (VV) is multivariate dispersion that is formulated as $Tr\left(\Sigma^2\right)$, geometrically VV is a square of the length of the diagonal of a parallelotope generated by all principal components of $\vec{X}$ [5]. The usefulness of $Tr\left(\Sigma^2\right)$ is not limited to small or low dimension data set and to non-singular covariance matrix. VV can be used efficiently for very large and high dimension data sets or even for singular covariance matrix. The MVV algorithm, compared with FMCD algorithm, has a lower computational complexity; the complexity of MVV is of order $O(p^2)$. FMCD is the fast algorithm minimizing the determinant of covariance matrix.The objective of this paper is to propose the minimum vector variance (MVV) as new measure of robust PCA.

34.2 The Classical Principal Component Analysis (PCA)

The principal component analysis is primarily a data analytic technique describing the variance covariance structure through a linear transformation of the original variables. The technique is a useful device for representing a set of variables by a much smaller set of composite variables that account for much of the variance among the set of original variables.

Suppose that the random vector $\vec{X}$ of p components has the classical covariance matrix S which is a $p \times p$ symmetric and positive semi definite. Covariance matrix S can be reduced to a diagonal matrix L which is a particular orthogonal matrix U such that $U'SU = L$

The diagonal elements of $L, \lambda_1, \lambda_2, \ldots, \lambda_p$, are called the characteristic roots or eigenvalues of S, the columns of U are called the characteristic vectors or eigenvectors of S. For $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, the principal components are uncorrelated linear combinations $\vec{Y}$ whose variances are as large as possible. The first principal component (PC) is given by $\vec{Y}_1 = \vec{U}_1' X$ which has the largest proportion of total variance.

The proportion of total variance the k principal component is often explained by the ratio of the eigenvalues $\lambda_k = \sum_{i=i}^{k} \lambda_i$. The determination of k is an important role to the PCA analysis. A larger k gives a better fit in PCA, but a larger k has the larger redundancy of information. The replacement of original variable p to the k principal component must be considered as a goal in optimizing.

The decomposed classical covariance matrix S is very sensitive to outlying observations. The k principal component becomes unreliable if outliers are present in the original variable p. The k principal component consisting of the largest proportion of total variance S is often pushed toward the outliers. As an example, consider the collinear data given by generated random $n = 100$ and $p = 3$, the eigen analysis of the correlation matrix of that data is listed in Table 34.1, and the plotting of each observation appears in Fig. 34.1.

The outlier detection by using classical PCA approach is presented on Fig. 34.2a. The figure explains the outliers can not be detected though the cumulative proportion of eigen value is almost 100%. To analyze the data set, we use the second classical component having 98.4% of characteristics cumulative proportion.

Table 34.1 The Eigen analysis of the collinear data with four outliers

	The first PC	The second PC	The third PC
Eigenvalue	2.7927	0.1589	0.0484
Proportion	0.931	0.053	0.016
Cumulative	0.931	0.984	1

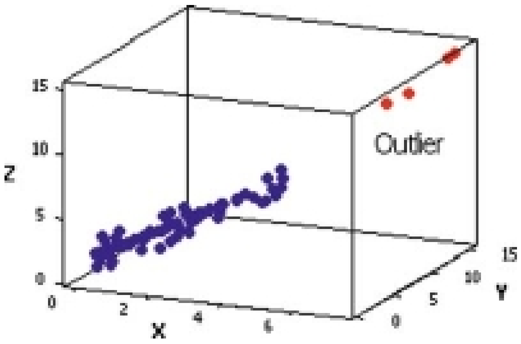


Fig. 34.1 The scatter plot of the collinear data with four outliers

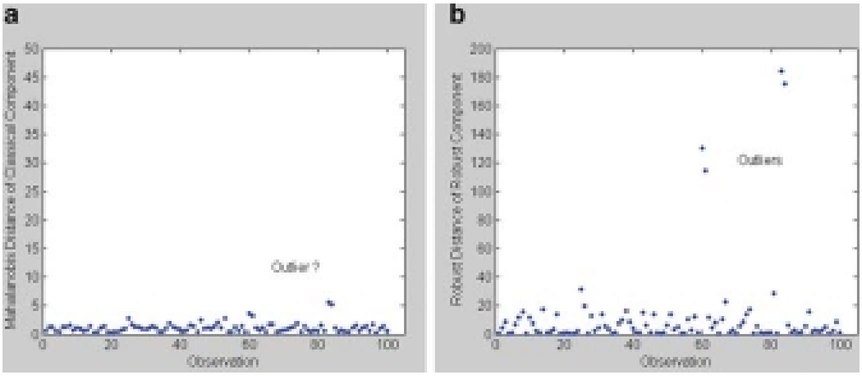


Fig. 34.2 The outlier detection by using: (**a**) Classical PCA (**b**) MVV Robust PCA

Figure 34.2b illustrates the outlier detection by robust approach with the first component. The detection is close to the real situation, four outliers can be separated clearly from the clean data. The influence of outliers can not shift the location estimator significantly and the masking effect can be handled well. The result is difference with the previous outcome. To make sense out of the robust approach, we discuss the algorithm of MVV robust PCA in the Section 34.3.

The following illustration is the second example of PCA which is classical to the process of clustering flowers. There are three categories of flowers; red color for Red Hisbiscus, Purple color for Linum Narbonense, and yellow color for Oxalis Pes-Caprae Each pixel of the image can be represented as a point in a 3D RGB color space. The visual contents of the images are extracted and described by color feature vectors. Figure 34.3 illustrates those flowers.

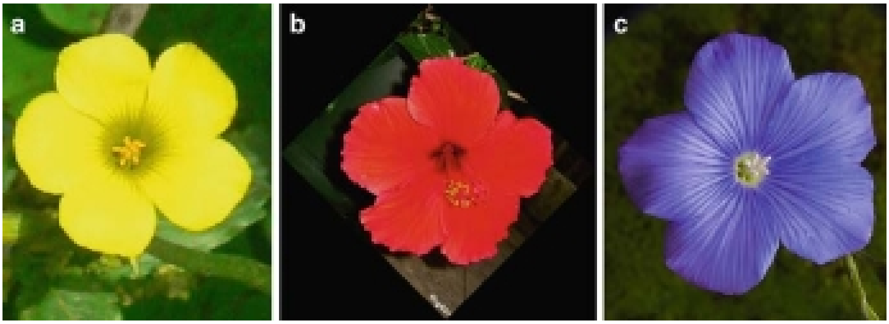


Fig. 34.3 The images of flower (**a**) Oxalis Pes-Caprae, (**b**) Red Hisbiscus and (**c**) Linum Narbonense

Table 34.2 The eigen value of flower image

Ordered eigen value λ	Cumulative proportion of λ
4.1296	0.459
3.1250	0.806
0.9316	**0.910***
0.5320	0.969
0.1638	0.987
0.0686	0.995
0.0332	0.998
0.0157	1.000
0.0005	1.000

We can easily categorize these three flowers by their colors although they have almost no different shapes. The classical PCA will be used to cluster the flowers. Table 34.2 contains the ordered λ and cumulative proportion of λ. The result of the clustering involving the three largest or biggest components with cumulative proportion of 91% total of variation turns out to show a 'bad' clustering

Figure 34.4 gives the description of categorized flowers based on their colors. The figure explains that the components having the 'best' low rank approximation to original data can not separate the three categorized flower colors. To enhance the clustering, the robust PCA will be discussed in the next section.

34.3 The Robust PCA Using Minimum Vector Variance (MVV)

A measure of dispersion is a measure which explains how far a group of data spread out. Two famous measures of multivariate dispersion are often used in the applications. They are the total variance (TV) and the generalized variance (GV). Generalized variance is often called as covariance determinant (CD). Related with the covariance matrix Σ, TV is defined as $Tr\,(\Sigma)$ and CD is defined as $|\Sigma|$. The role of TV in general can be found in the problem of reduction on data dimension, such as in the analysis of principal component, analysis of discriminant and canonical

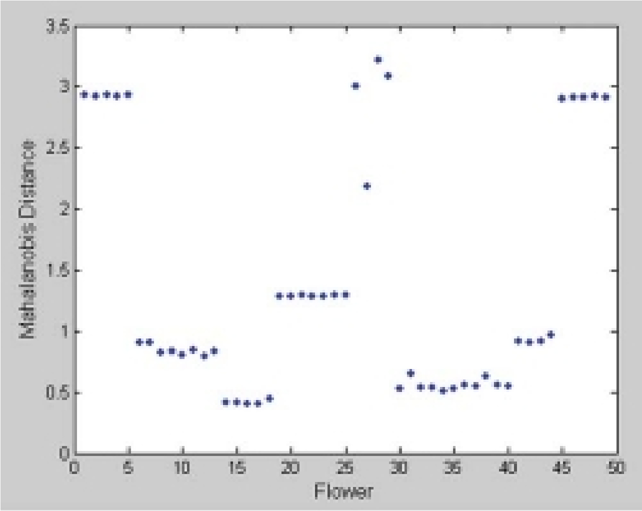


Fig. 34.4 The clustering of flower using classical PCA

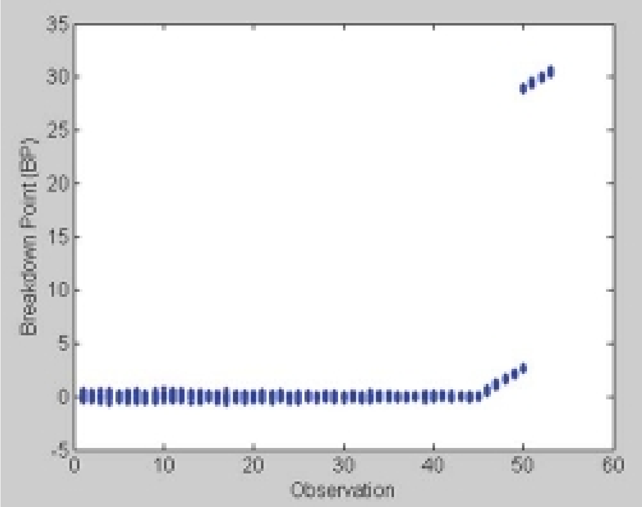


Fig. 34.5 Breakdown point of MVV robust PCA using $h = [(n + k + 1)/2]$

analysis [6]. The role of CD can be found on every literature of multivariate analysis. The limitation of TV is very natural, because TV is merely involving variances without involving the structure of covariance. Meanwhile CD involves both of them, the structure of variance and covariance. That is way CD has a wider role on application [5], including the role on various robust methods. Even though CD has wider applications than TV, but CD has a limitation too (Figs. 34.5 and 34.6).

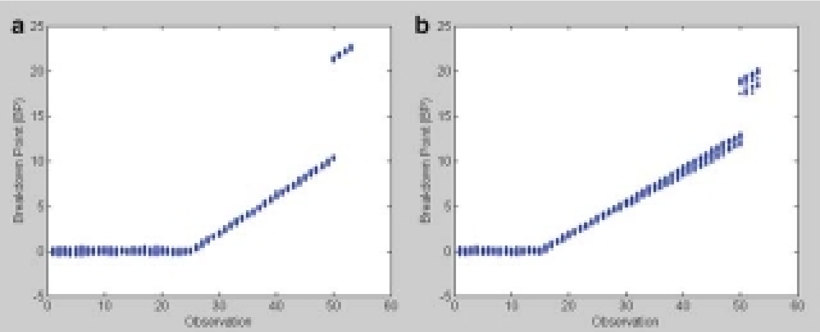


Fig. 34.6 Breakdown point of MVV robust PCA using (**a**) $h = 0.75\,n$ and (**b**) $h = 0.85\,n$

Alt and Smith [7] stated that the main limitation lies on the property that CD $= 0$ when there is a variable of zero variance or when there is a variable which is a linear combination of other variables. Due to this limitation Djauhari [5] proposed a different concept of multivariate dispersion measure, called the vector variance (VV). Geometrically VV is the square of the length of the diagonal of a parallelotope generated by all principal components of $\vec{X}$

Suppose $\vec{X}$ is a random vector of covariance matrix Σ of dimension $(p \times p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are eigen values of Σ,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

The structure of TV, GV (CD) and VV can be formulated as,

$$\text{TV} = Tr\,(\Sigma) = \lambda_1 + \ \lambda_2 \ + \cdots + \ \lambda_p \tag{34.1}$$

$$\text{CD} = |\Sigma| = \lambda_1\ \lambda_2\ \cdots\ \lambda_p \tag{34.2}$$

$$\text{VV} = Tr\left(\Sigma^2\right) = \lambda_1^2 + \ \lambda_2^2 \ + \cdots + \ \lambda_p^2 \tag{34.3}$$

Computations of VV are very efficient. The efficiency of VV is of order $O(p^2)$ compare with CD by using Cholesky decomposition which is of order $O(p^3)$.

Regarding the efficient computation of VV, Herwindiati et al. [8] proposed Minimum Vector Variance (MVV) in obtaining the robust estimator by minimizing vector variance. The algorithm of MVV has no significant difference to Rousseeuw and van Driessen's FMCD (1999) except that the criterion used here is not MCD but MVV.

In the outlier labeling process, MVV is an effective and an efficient method, but MVV still takes a few more times in the computation when the dimension p is larger than 100; that are around 110.531 s. Huber et al. [4] introduced a new method for robust principal component (ROBPCA). ROBPCA is an effective and an efficient PCA method which combines two advantages of both projection pursuit and robust covariance estimations. The ROBPCA estimator is computed by the minimum of GV or CD. The effectiveness of ROBPCA and the inssuficient characteristic of CD tend us to propose the new measure of robust principal component based on minimum vector variance (MVV). The algorithm of MVV robust PCA is composed as follows,

Stage 1. Start with a singular value decomposition of the mean centered data matrix $\vec{X}_{n,p} - 1_n \bar{\vec{X}} = U_{n,r} L_{r \times r} V'_{r,p}$, with $U'U = I_r = V'V$, $\bar{\vec{X}}$ is classical mean vector, L is an $r \times r$ diagonal matrix, and I_r is the $r \times r$ identity matrix. Chose the k-principal component consisting of the major part of total variance.

Stage 2. Estimate the location and covariance matrix using MVV robust approach.

1. Let H_{old} be an arbitrary subset containing $h = [(n+k+1)/2]$ data points. Compute the mean vector $\bar{\vec{X}}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}. Then compute, $d^2_{H_{old}}(i) = \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right)^t S^{-1}_{H_{old}} \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right)$ for all $i = 1,\ 2, \ldots,\ n$
2. Sort these distances in increasing order
$$d^2_{H_{old}}(\pi(1)) \le d^2_{H_{old}}(\pi(2)) \le \cdots \le d^2_{H_{old}}(\pi(n))$$
3. Define $H_{new} = \left\{\vec{X}_{\pi(1)}, \vec{X}_{\pi(2)}, \cdots, \vec{X}_{\pi(h)}\right\}$
4. Calculate $\bar{\vec{X}}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$
5. If $Tr\left(S^2_{H_{new}}\right) = 0$, repeat steps 1 to 5

If $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$, the process is stopped.

Otherwise, the process is continued until the k-th iteration if

$$Tr\left(S_1^2\right) \ge Tr\left(S_2^2\right) \ge Tr\left(S_3^2\right) \ge \cdots \ge Tr\left(S_k^2\right) = Tr\left(S_{k+1}^2\right)$$

Stage 3. Identify the labeled outlier by using robust MVV distance.

Let $\vec{T}_{MVV}$ and S_{MVV} be the location and covariance matrix given by that process. Robust squared Mahalanobis distance is defined as,

$$d^2_{MVV}\left(\vec{X}_i, \vec{T}_{MVV}\right) = \left(\vec{X}_i - \vec{T}_{MVV}\right)^t S^{-1}_{MVV}\left(\vec{X}_i - \vec{T}_{MVV}\right) \quad \text{for all } i = 1, 2, \ldots, n.$$

Observations which have a large distance $d^2_{MVV}\left(\vec{X}_i, \vec{T}_{MVV}\right)$ will be labeled as outliers or suspects.

Compared to FMCD algorithm, the MVV algorithm has a lower computational complexity. As VV is the sum of square of all elements of the covariance matrix, the computational complexity of VV is of order $O(p^2)$. On the other hand, based on Cholesky decomposition for large value of p, the number of operations in the computation of CD is equal to $p + p(p-1) + (p-1)\sum_{i=1}^{p-1}(p-i-1)(p-i)$ which is of the order of $O(p^3)$.

The subset h in the first step has the important role in the estimator. Hubert et al. [4] suggested taking subset $h = \max\{[\alpha n], [(n + k_{\max} + 1)/2]\}$, where α is chosen as any real value between 0.5 and 1, $k_{\max}$ as a maximal number of components that will be computed. In this paper we chose $h = [(n + k + 1)/2]$.

The choice of this subset is due to the 'reality' of breakdown points that are found in our computation experience. Compared to the other subset h, the breakdown point of $h = [(n + k + 1)/2]$ is more stable. The break down point is tool measuring how many data can be changed to be infinity before they are meaningless crushed to bits [3]. The following figures reveal the fact.

34.4 The Performance of MVV Robust PCA

34.4.1 The Clustering Flower Images

This section discusses the work of MVV through the example of flowers clustering in Section 34.2. We will categorize the flowers according to their colors; 40 images of Red Hisbiscus, 15 images of Linum Narbonense, and 19 images of Oxalis Pes-Caprae. The color moment is used in order to get the color feature of those flowers. The extraction of each pixel in the color feature is represented as a point in a 3D RGB color space, see Fig. 34.3.

MVV robust PCA is used to cluster the flowers based on their color. The excellent result of clustering can be seen in Fig. 34.7. Every flower is perfectly categorized into its group, as can be seen below.

34.4.2 The Identification of Anomalous Data in High and Large Dimension

MVV Robust PCA also works well in the process of identification of anomalous data in high and large dimension, assuming that anomalous data is suspected as outlier. For this purpose, we generate $n = 400$ random data from a mixture of p-variate normal distribution $(1-\varepsilon)\,N_p\left(\vec{\mu}_1,\ I_p\right) + \varepsilon N_p\left(\vec{\mu}_2, I_p\right)$ with $p = 300$; $\varepsilon = 0.1$ where $\vec{\mu}_1 = \vec{0}$, $\vec{\mu}_2 = 10\vec{e}$ and $\vec{e} = \vec{e} = (1\ 1\ \cdots\ 1)^t$ is of p dimension.

The identification process is done quite well by MVV robust PCA. The suspected outliers can be clearly separated and is located far away from the group of clean data. The separating process needs only less than 4 s (Fig. 34.8).

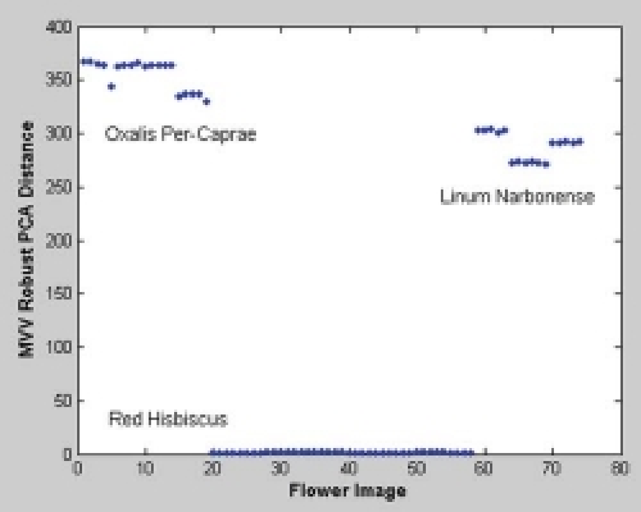


Fig. 34.7 Scatter plot of MVV robust PCA for clustering flower images

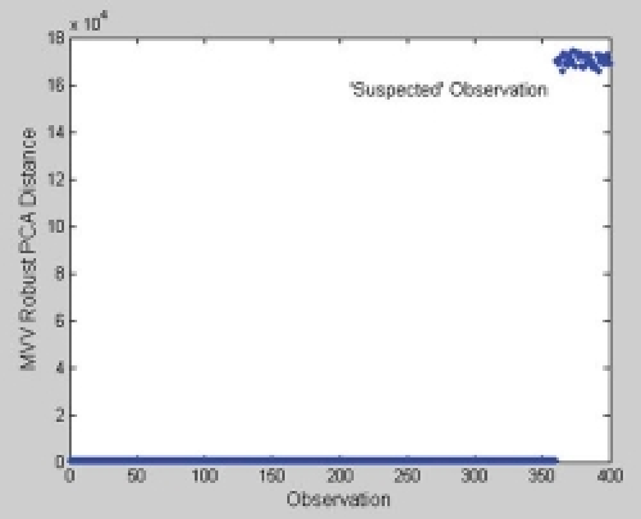


Fig. 34.8 The outlier labeling in high and large dimension data

34.4.3 The Computation Time of MVV Robust PCA

Hubert et al. [4] described that the computation of ROBPCA on Pentium IV with 2.40 GHz is 3.06 s for $n = 39$, $p = 226$ and 3.19 s for $n = 111$, $p = 11$. Compared to ROBPCA, the computation time of MVV robust PCA is not slower (Fig. 34.9). To see the time effectiveness of MVV robust PCA can be seen in the following

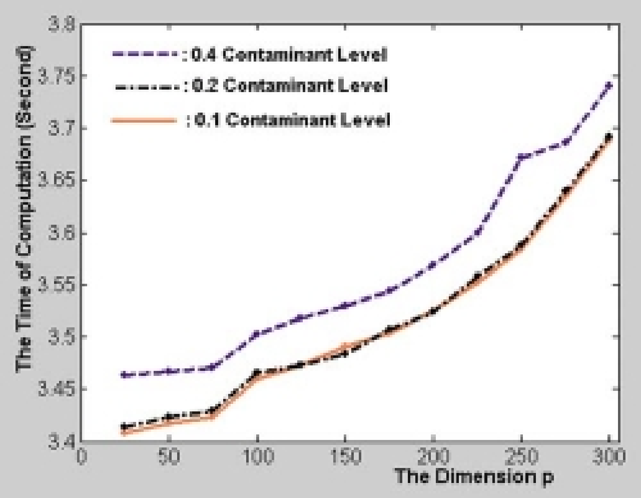


Fig. 34.9 The computation time of MVV with contaminant level $\varepsilon = 0.1$; $\varepsilon = 0.2$ and $\varepsilon = 0.4$

figures which show that the computation process between the dimensional data of $p = 25$ to $p = 300$ with $n = 100$, $\varepsilon = 0.1$, and $\varepsilon = 0.2$ from a mixture model $(1 - \varepsilon)\, N_p\left(\vec{\mu}_1, I_p\right) + \varepsilon N_p\left(\vec{\mu}_2, I_p\right)$

The figure tells us that the additional contaminant ε and also the change of p dimension does not produce a significant difference in time. Even if we compare the amount of contaminant $\varepsilon = 0.1$ and $\varepsilon = 0.4$ for the same dimension, we find no significant difference of time in second (the difference is around 0.1 s for $p = \mathit{300}$).

34.5 Conclusion

MVV robust PCA is an effective and an efficient method to identify outlier in a high and large dimension. MVV robust PCA is also an impressive method for interpreting the application of PCA, such as the clustering process. From the aspects of computation of several p-dimensions, MVV robust PCA gives the promising results.

References

1. Gnanadesikan, R.: Method for Statistical Data Analysis of Multivariate Observations. Wiley, New York (1977)
2. Barnett, V., Lewis, T.: Outliers in Statistical Data, 2nd edn. Wiley, New York (1984)
3. Hampel, F.R., Ronchetti, E.M., Rousseuw, P.J., Stahel, W.A.: Robust Statistics. Wiley, New York (1985)

4. Hubert, M., Rousseeuw, P.J., vanden Branden, K.: ROBPCA: a new approach to robust principal component analysis. J. Technomet. **47**, 64–79 (2005)
5. Djauhari, M.A.: Improved monitoring of multivariate process variability. J. Quality Technol. **37**(1), 32–39 (2005)
6. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 2nd edn. Wiley, New York (1984)
7. Alt, F.B., Smith, N.D.: Multivariate process control. Handbook Statis **7**, 333–351 (1988)
8. Herwindiati, D.E., Djauhari, M.A., Mashuri, M.: Robust multivariate outlier labeling. J. Commun. Statis. Simul. Comput. **36**(6) (2007)
9. Hawkins, D.M.: The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. J. Comput. Statis Data Anal. **17**, 197–210 (1994)
10. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis, 2nd edn. Wiley, New York (1988)
11. Long, F., Zhang, H., Feng, D.D.: Multimedia Information Retrieval and Management. Spinger, Berlin (2003)
12. Rousseeuw, P.J., van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. J. Technomet. 41, 212–223 (1999)

Chapter 35
The Affects of Demographics Differentiations on Authorship Identification

Haytham Mohtasseb and Amr Ahmed

Abstract There is lots of previous studies concern the language difference in text regarding the demographics attribute. This investigation is different by presenting a new question: is male style more consistent than female or the opposite? Furthermore, we study the style differentiation according to age. Hence, this investigation presents a novel analysis of the proposed problem by applying authorship identification across each category and comparing the identification accuracy between them. We select personal blogs or diaries, which are different from other types of text such as essays, emails, or articles based on the text properties. The investigation utilizes couple of intuitive feature sets and studies various parameters that affect the identification performance. The results and evaluation show that the utilized features are compact while their performance is highly comparable with other larger feature sets. The analysis also confirmed the usefulness of the common users' classifier, based on common demographics attributes, in improving the performance for the author identification task.

Keywords Web mining · information extraction · psycholinguistic · machine learning · authorship identification · demographics differentiation

35.1 Introduction

Blogs are one of the most popular forms of users' contribution to the web contents. There are many categorizations of blogs which are differing in the content, publishing methodology, and even in the type of readers. Personal blog, or online diary, is the most famous category in which the blogger expresses his feelings, show creativity, and communicate with other people faster than emails or any other media. In addition, there are some targeted or focused blogs which focus on a specific subject

H. Mohtasseb (✉) and A. Ahmed
School of Computer Science, University of Lincoln, Brayford Pool, Lincoln, LN1 1LH, UK
e-mail: hmohtasseb@lincoln.ac.uk; aahmed@lincoln.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_35,

such as news blogs, political blogs, and educational blogs. Our research is focused on the personal blogs category. We selected one of the famous personal blog sites, namely the "*LiveJournal*"[1]. LiveJournal is a free personal blog website forming a community on the internet that contains millions of users publishing their own ongoing personal diaries.

The misuse of online messaging is getting increased as the users could be anonymous and hidden in the cyberspace. This encourages the harmful behavior and publishing illegal contents. The need to authenticate those texts is being more important and raises the call for new techniques that can capture the identity regardless the IP address or any other information that could be not true. Authorship identification is a good solution and able to discover the author of anonymous message via modeling the stylistic attributes of previous known messages of the author.

Authorship identification in blogs has various motivations and challenges. Identifying the author of anonymous blog posts could be useful in various applications. This includes online security where it is valuable to extract the patterns of authors who may participate in different blog sites with different identities. However, the task has its associated challenges. The large number of authors is one of the key factors in authorship identification. In particular, scaling existing solutions with the huge, and increasing, number of authors is a challenge. Moreover, there are many factors that have important roles and affect the performance of identification process such as the text length, the number of posts per author, and the type of authors.

In this study, we address the above issues by applying authorship identification on online diaries corpus using a different type of linguistic features and analyze the affects of demographics attributes on the identification results. The remainder of the chapter is organized as follows. In the next section, we review the existing related work in authorship identification. The following section describes the utilized feature sets. Our main work follows in Section 35.4, with the proposed framework and experiments. Results and discussions come next. Finally, the chapter is concluded, and future work is also highlighted.

35.2 Related Work

Early work on authorship identification, on the Federalist Papers, is back to 1964 [1]. In this early work, a set of function words, which were not topic-related features, were utilized. Since then, Authorship identification has been researched in various text domains, such as emails, forums, and books as discussed below.

De Vel analyzed stylistics attributes to discover forensics in emails [2]. Although they achieved relatively good results, this may not be applicable straight-forward on the blogs due to the different nature of the text in emails and blogs. Generally, email text is shorter than diaries text and it is usually a topical dialogue between

[1] http://www.livejournal.com

two authors, while online diaries text is from the author to the public, at least the intended group. Also in books and literature, Gamon [3] utilized the part-of-speech (POS) tri-grams and other features to find out the correspondent author out of just three writers. The main differences from our work are the smaller number of authors, and the nature of book text. Text in books is normally too long compared to text in diaries. And usually, there is a specific topic in the book. Books are also expected to be well written and proof read, which results in much less grammatical and syntactical errors than the case in personal blogs.

In the domain of web forums, Abbasi et al. [4] used a collection of lexical, syntactical, structural, and content-specific features to find out the extreme patterns of writing on web forums. It may look that the text in web forums is similar to that in the personal blogs. But regularly there is a subject to be discussed in the forum, which in contrast to diaries that contains usually general ideas and thoughts on various and mixed issues.

Recently, Abbasi et al. [5] presented the "Writeprints" technique, which separately model the features of each individual author, instead of using one model for all the authors. They build a writeprint for each author using the author's key features. Authorship attribution was also manipulated also in probabilistic approaches using Markov chains of letters and words [6]. The above two methodologies are different in which they need to build an individual model for each author instead of just one model that classify all the authors. Although one model for each author will best represent the author style, this requires comparing the features from the new text against all the authors' models rather than testing through just one classification model.

From the above, it can be seen that author identification in personal blogs or diaries has received little attention. Consequently, no specific standard features are confirmed or solidly recommended due to the differentiation in the properties of text in each context. In the work presented in this paper, we address the above issues by applying authorship identification on an online diaries corpus using a different type of linguistic features and analyze those factors that affect the identification results.

35.3 Feature Set

A very important concern in text classification is the selection of features. In our investigation, we chose LIWC the Linguistic Inquiry Word Count [7], MRC Psycholinguistic database [8], and a collection of syntactic features. Those feature sets that have been selected have psychology basis, and known relate well with the author's style and/or personality. The properties of diaries text as they contains lots of feelings, personal activities, and thoughts are more captured using our selected features sets. The selected 88 LIWC features are grouped into four types:

1. Standard linguistic features (e.g., total word count, word per sentence, pronouns, punctuations, articles, time)
2. Psychological features (e.g., affect, cognition, biological processes)

3. Personal concerns features (e.g., work, sports, religion, sexuality)
4. Paralinguistic features (assents (e.g., agrees, ok), fillers (e.g., err, umm), non fluencies (e.g., I mean, you know))

In LIWC, the features are more of categories based on their intuitive meaning, including psychology and affect. These features (or categories) are evaluated by calculating scores from a number of related words that are defined in the LIWC dictionary [7]. This means that the calculated word frequency is not used directly, but rather contributes to the final scores of multiple LIWC features. For example, the word "*cried*" is contributing to the calculation of the scores of five features: *sadness*, *negative emotion*, *overall affect*, *verb*, and *past tense verb*. Moreover, the LIWC can handle the different stems of the word, which is one of the common issues in natural language processing NLP. So the stem *hungr* captures the words *hungry*, *hungrier*, *hungriest* and so on dictionary [7].

The MRC database contains psycholinguistic statistics for over 150,000 words. It includes frequencies among the Psychology lexicon such as: number of phonemes, number of syllables, imagebility rating, letters count, part-of-speech information, and familiarity rating.

35.4 Framework

In this section, we describe the design of the framework and the experiments for identifying the authors of blog posts. After grabbing the data corpus from the web, the extraction phase converts each post to a features vector containing the corresponding features values. This changes the input data from unstructured text space into features vectors space. The setup of our framework is depicted in Fig. 35.1.

First, we divided the input features vectors into groups according to three parameters: the post length, the number of authors, and the number of posts per author. Each group is manipulated individually by the classification algorithm. Support Vector Machines (SVM) has been selected as the machine learning algorithm in our framework. SVM is one of the best algorithms in this domain.

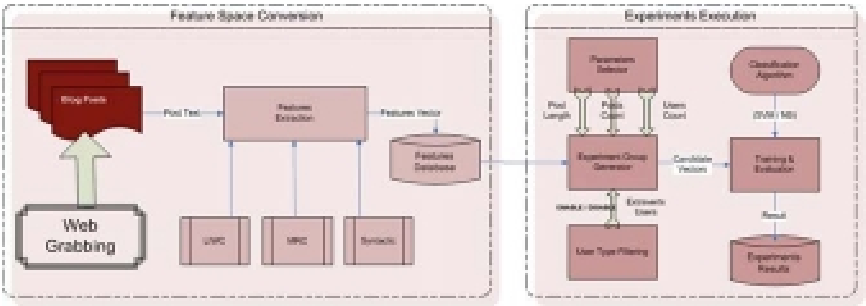


Fig. 35.1 Authorship identification framework

For each experiment's data group, SVM is trained and tested by applying tenfold cross validation. This means that there are ten cycles of validation and the identification accuracy will be calculated among the average of them. In each cycle, 90% of the dataset is used for training and the remaining 10% is used for testing. We selected the implemented SVM algorithm (SMO) in the WEKA toolbox with linear kernel [9] for machine learning algorithms in our framework.

We choose eight different numbers of authors, five different post counts per user, and 11 different post lengths. This makes 440 groups in total. The second contribution of this paper is to study the effect of pre-filtering the candidate authors that are selected in the sampling stage of the classifier. In this study, we present the feasibility of building a classifier that contains the users which have common demographics attributes such as personality, gender, and age. We try to find the type of personality either extraversion or introversion that is more correlated with authorship identification. Finally, we compare the identification accuracy between males and females, and then across three different age groups.

35.5 Results

It should be mentioned that given that we have three parameters investigated simultaneously, the result would ideally need to be represented in a four dimensional space. However this may not be easy to be visualized. So, Fig. 35.2 depicts a

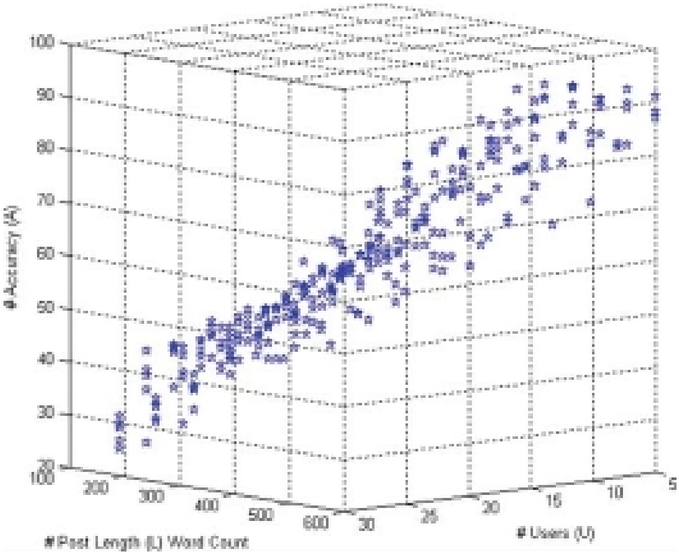


Fig. 35.2 Identification accuracy (users/post length)

selected 3D cube that represents the identification results according to the number of users and the post length. The results, as presented in Fig. 35.2 justify the effective parameters ranges in which the identification percentage is more accurate.

The trends in Fig. 35.2 indicate that the identification accuracy is enhanced when there are more words in the post (post length). Although the selected features are less effective in short posts, having more posts (posts size) improves the identification accuracy. This happens because having more posts (post size) provides more text that the same author has written with different styles and contents, which is in turn included in the learning process.

35.5.1 Demographics Differentiations

One of the big problems in authorship identification is to identify the author among large number of authors. Building different classifiers according to the type of users will decrease the number of the potential authors to be involved in each classifier. This would help in scaling the solutions with the increase in the number of authors. In this section, we discuss how the demographics attributes will affect the final identification result. We start by studying the personality factor and then move toward gender and age variation.

35.5.1.1 Personality

Writing diaries to be read publicly and describing the details of the private life to everybody on the internet is an indication that the bloggers are Extraverts [10]. Extraversion is one factor of the Big Five personality traits model [11]. The extravert person could be described for example as sociable, assertive, friendly, and playful. On the other hand, bloggers can be considered as introverts because they are writing using nicknames on the blogging site, hiding their real identity [10].

We chose to test the authorship identification for those who are extraverts in their text. Although the corpus does not contain any tagging for extraversion, we extracted the extraversion value automatically using a personality recognition software system[2] which computes estimates of personality scores along the Big Five dimensions. Figure 35.3 displays the classification accuracy average for five users in the different post lengths between the extraverts and the corresponding unfiltered users. The results indicate that those who have a high extraversion score are better classified in the authorship identification process.

[2] http://mi.eng.cam.ac.uk/∼farm2/personality/recognizer.html

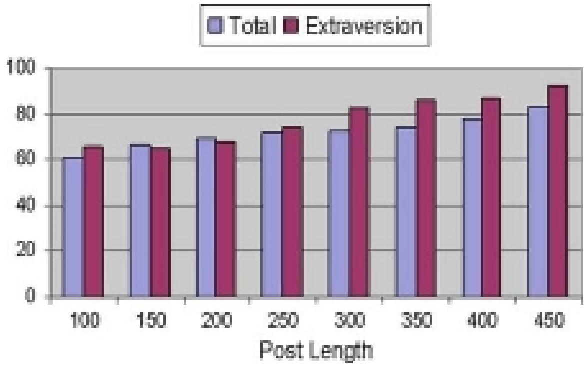


Fig. 35.3 Classification accuracy comparison between extraverts and generic users

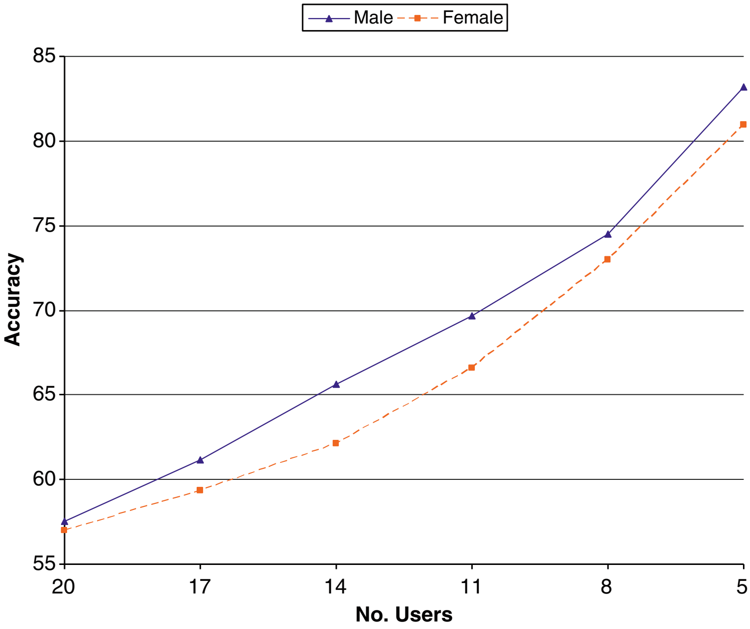


Fig. 35.4 Authorship identification comparison between males and females

35.5.1.2 Gender and Age

There are lost of previous studies concern the language difference regarding the demographics attribute. This experiment presents a novel analysis by applying authorship identification across each category and comparing the identification accuracy between them. For this experiment we use different data from another corpus [12] where the gender and age information of authors are available (Figs. 35.4 and 35.5).

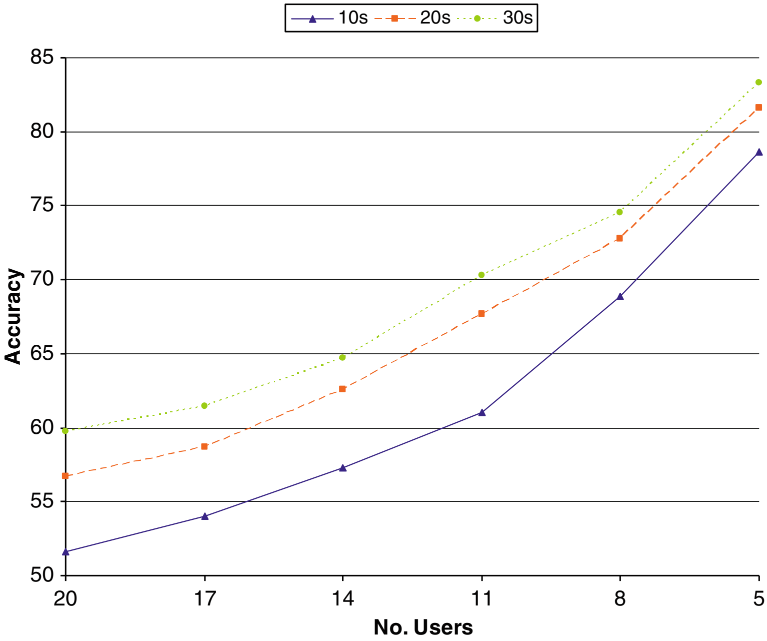


Fig. 35.5 Authorship identification comparison across three age groups

35.6 Conclusion

In this chapter, we presented our investigation of identifying the bloggers in online diaries by mining the diaries text of each blogger. The investigation contains majorly three contributions. The first one was by utilizing two psycholinguistic features, namely the LIWC and MRC feature sets together, on the personal blogs for blogger/authorship identification. The second one was the analyzing of the effect of various parameters on the identification performance. This included the number of authors in the data corpus, the post size or the word count, and the number of posts for each blogger. Finally, we studied the identification outcome for shared-attribute authors.

We found that the post length, or the number of words in the text, is highly contributing to the author style attribution. Having more words facilitates more accurate and stable identification performance as the author style can be more appropriately captured. The results provided the preferred ranges of those parameters, which can be used as recommendation for further studies in authorship identification in personal blogs.

The results of testing authorship identification across common-attribute authors confirm that the identification accuracy is highly related with the type of users. This motivated us in future to build multi-layer classification system that includes several classifiers based on selected demographics attributes.

References

1. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist. Addison-Wesley, Reading, MA (1964)
2. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. ACM SIGMOD Record **30**, 55–64 (2001)
3. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)
4. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intell. Syst. **20**, 67–75 (2005)
5. Abbasi, A., Chen, H.: Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inform. Syst. **26**, 29 (2008)
6. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking. In: Proceeding of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 482–491 (2006)
7. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum Associates, Mahway (2001)
8. Wilson, M.: MRC Psycholinguistic Database: Machine Usable Dictionary, Information Division Science and Engineering Research Council (1987)
9. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, CA (2005)
10. Nowson, S., Oberlander, J.: The identity of bloggers: openness and gender in personal weblogs. In: Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs (2006)
11. NORMAN, W.T.: Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. J. Abnorm. Soc. Psychol. **66**, 574–583 (1963)
12. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (2006)

Chapter 36
Anonymous ID Assignment and Opt-Out

Samuel S. Shepard, Renren Dong, Ray Kresman, and Larry Dunning

Abstract The networked society places great demand on the dissemination and sharing of private data. As privacy concerns grow, anonymity of communications becomes important. This paper addresses the issue of anonymous ID assignment to nodes in a distributed network and how it can be integrated with secure mining algorithms to allow nodes, that have privacy concerns, a capability to opt out of the mining computation.

We propose two algorithms for ID assignment and evaluate their performance. We use them in the design of a protocol that allows a node to opt out of data mining, and investigate the collusion resistance capability of the resulting protocol.

Keywords Anonymity · data mining · secure sum · Hamiltonian cycles · privacy · opt-out

36.1 Introduction

Anonymity of communications [17] is important to a number of application realms. These include health care, business, e-vote, education, web search, mobile communication, and others.

Suppose that a customer has defaulted on loans. Banks as a group would want to know this information. Yet, a bank that has made loans to a customer who is not creditworthy would want the customer name to be shared with other banks,

S.S. Shepard (✉)
Department of Medicine, University of Toledo, Health Science Campus,
Toledo, OH 43614, USA
e-mail: sammysheep@gmail.com

R. Dong, R. Kresman, and L. Dunning
Department of Computer Science, Bowling Green State University
Bowling Green, OH 43403, USA
e-mail: drenren@gmail.com; kresman@cs.bgsu.edu; duning@cs.bgsu.edu

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_36,

but would not want to embarrass itself by divulging its (the bank's) name to other banks [2]. Use and transmission of medical data has strict privacy requirements. Medical facilities [1] do not wish to provide any information about an individual's DNA sequence or medical history, but still wish to help one another to discover the relationship between drugs and DNA sequence.

As e-voting becomes popular privacy concerns need to be addressed [12]. The e-voting system must guarantee a number of properties such as anonymity, completeness, correctness and uniqueness. Similarly, nodes in mobile communication need dynamic IDs so that their static IDs remain private [9]. As another example, unconditionally secure anonymity was introduced by Chaum [5] as the dining cryptographers (DC) problem. It allows any diner to pick up the meal tab for the party in an anonymous manner.

One way to provide anonymity is to assign IDs to nodes in a network such that no node is aware of what ID is assigned to a specific node. We describe and evaluate such protocols in this paper. We then describe the application of these protocols to distributed data mining (DDM).

DDM handles privacy concerns through privacy preserving data mining (PPDM). Consider the popular data mining operator, secure sum (SS) [7]; each node contributes a value to the mining process and at the end of the computation every node figures out the sum of the values, all in a distributed manner.

SS is vulnerable to collusion. Clifton et al. [6, 7] lay the foundation for a cycle-partitioned secure sum (CPSS) algorithm by noting how any node in a secure sum can divide its value into random shares such that each share is passed along a different secure sum circuit or cycle. Collusion protection is then achieved when no node has the same neighboring nodes twice for each cycle. Even if nodes collude to disover someone's value, the value is not exposed as long as the number colluding is less than a threshold (see also [16]).The algorithm of [15] relies on SS and on the need for a fully-connected network but sends data to nodes in a tree-like pattern for each round of the algorithm. Encryption has also been a staple method in PPDM to ensure privacy, see for example [3, 10].

However, there are instances where a node fears that its value may be discovered by others. A node that has proprietary or strategic knowledge may refuse to participate in the mining computation because it thinks that this information may be compromised. Or, a node may feel that participation violates or oversteps the boundaries set forth by legal statues. Finally, if a node's value is an outlier, then others can discover who contributed this value, and perhaps estimate the value too. Whenever a small group of nodes have too great a contribution to the sum, or rather, if too many nodes contribute too little, the possibility to disclose an estimate of private information exists even if fewer nodes collude than would be required for randomly distributed data.

In sum, we want node(s) to have an ability to cancel the mining process whenever they want, without anyone knowing who cancelled the mining process. This feature is also known as anonymous opt-out. This paper uses anonymous ID as a basis, to propose distributed mining algorithms that help address such privacy concerns.

36.2 Anonymous ID Assignment Algorithms

Suppose the set P denotes the number of stations in the network with $|P| = M$. We want to assign unique, anonymous, IDs to these M nodes. We first develop an algorithm for anonymous ID assignment (an "AIDA" algorithm) that is K collusion resistant using a network topology of edge-disjoint Hamiltonian cycles (EDHC) that we name "AIDA-H". Afterwards we present an improved algorithm "AIDA-D" that uses a decentralized topology with greater collusion resistance.

36.2.1 AIDA-H – Using Edge-Distjoint Hamiltonian Cycles

AIDA-H works by distributing a slot array, S, to each node, where $\mathit{length}(S) = N, N \geq M$ nodes. Each node randomly selects exactly one slot from the array and adds 1 to that slot. When the slot array returns to the initiating node, the initiating node broadcasts to each other node the element indices for all 1 values found in ascending order. See Algorithm 1 for the pseudocode of this approach. All pseudocode is presented in the chapter Appendix 36.A.

If a collision occurs ($\mathit{value} > 1$) or a 0 is found in the S array, that element index will not be broadcast. The node that chose an index in the broadcast will then set its ID number to the position of their index relative to the ordering of the broadcast (this relative order spans each round of the algorithm). For example, if $\mathit{indices} = \{23, 57, 89, 110\}$ have been broadcast, then the node choosing index 89 in S will be ID #3 – a unique, anonymous ID. In subsequent rounds the remaining nodes will be assigned using the same method. This logic can be seen in lines 23–26 of Algorithm 1. Suppose also that there are C edge-disjoint Hamiltonian cycles. Similar to [7], we may transmit a random partition of the S array into partial sums through each cycle using CPSS. In lines 7–9 and 13–15 the partitions of the slot array S are transmitted, whereas on lines 19–21 each array element j of S is recovered from the random partitions of each cycle i by the initiating node. This process may be repeated until the number of indices broadcast matches the number of nodes in the data mining operation.

K-collusion resistance for the anonymity in AIDA-H is conferred by sending S via cycle partitioned secure sum (CPSS) [14], that is, the slot array is itself randomly partitioned and sent along each edge-disjoint Hamiltonian cycle. The choosing of a slot in the array is equivalent to adding a count value of one to that slot in CPSS, except that in AIDA-H we have an array instead of a scalar.

The length of the slot array S shown in Algorithm 1 should be chosen to be an integer value $N \geq M$, where M is the number of nodes. Indeed, the choice of N depends on the size of M. We discuss the choosing of N as it relates to efficient AIDA-H termination in Section 36.3.

36.2.2 AIDA-D – Decentralized Anonymous Algorithm

AIDA-D, unlike AIDA-H, does not rely on EDHCs. AIDA-D shares the same basic idea of AIDA-H, however AIDA-D assumes a fully connected network, and uses a broadcast operation in its second step which is more efficient than AIDA-H. Just like AIDA-H, AIDA-D has several rounds and each round has two steps.

Step 1: Each p_i has an integer slot array $x_i = [x_{i,1}, \ldots, x_{i,N}]$, $N \geq |P| = M$. Set $x_{i,t_i} = 1$ (t_i is the number that p_i wants to choose as its anonymous ID and ID $= t_i$ must be available for selection) and $x_{i,k} = 0, k \neq t_i$. Now, to provide collusion resistance, each p_i splits its array x_i into M random arrays $x_i^{(j)} = [x_{i,1}^{(j)}, \ldots, x_{i,N}^{(j)}]$, where $x_i = [\sum_{j=1}^{M} x_{i,1}^{(j)}, \ldots, \sum_{j=1}^{M} x_{i,N}^{(j)}]$. Send array $x_i^{(j)}$ to p_j.

Step 2: Each participant p_j adds all the $x_i^{(j)}$ from Step 1, and broadcasts the sum to everyone. It is easy to see that the sum of all these individual broadcasts equals $X = \sum_{i=1}^{M} x_i = [X_1, \ldots, X_N] = [\sum_{i=1}^{M} x_{i,1}, \ldots, \sum_{i=1}^{M} x_{i,N}]$. If $X_t = 1$, it means only 1 participant selected t as its ID and there is no conflict for ID $= t$. Therefore, this ID will be assigned to the one who selected it in Step 1 and is no longer available – for others – in future rounds. If $X_t > 1$ or $X_t = 0$, this means that either none or more than one participant selected t as the ID, and ID t is up for bid in future rounds.

For all p_i, whose ID selection has a conflict in Step 2, repeat Step 1 and 2 until everyone gets an ID without conflict. However, recall that nodes can only bid for IDs that have not already been assigned.

An example of the algorithm is given in Fig. 36.1. After the first round of the algorithm, we have $X = \sum_{j=1}^{M} \sum_{i=1}^{M} x_i^{(j)} = [1, 0, 2]$ which indicates that two participants (p_1 and p_2) wanted ID $= 3$ as their anonymous ID. Therefore, both p_1 and p_2 will enter the next round of the algorithm. However, they can only select ID $= 2$ or ID $= 3$, since ID $= 1$ had no conflict and was already assigned to p_3 in round 1.

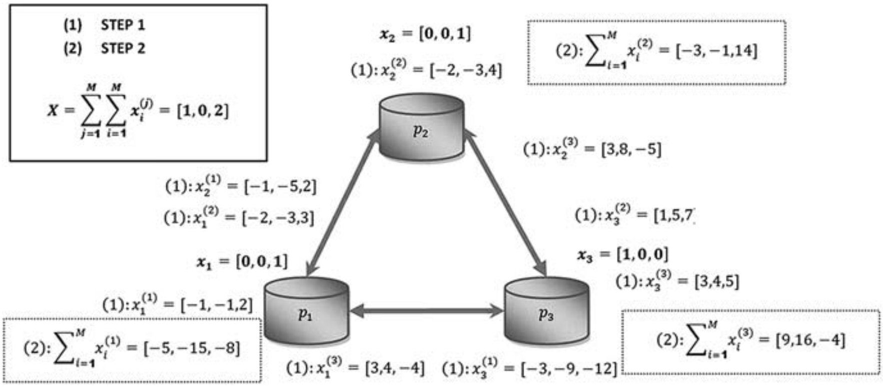

(1) STEP 1
(2) STEP 2
$X = \sum_{j=1}^{M}\sum_{i=1}^{M} x_i^{(j)} = [1,0,2]$
$x_2 = [0,0,1]$
(1): $x_2^{(2)} = [-2,-3,4]$
(2): $\sum_{i=1}^{M} x_i^{(2)} = [-3,-1,14]$
p_2
(1): $x_2^{(3)} = [3,8,-5]$
(1): $x_2^{(1)} = [-1,-5,2]$
(1): $x_1^{(2)} = [-2,-3,3]$
(1): $x_3^{(2)} = [1,5,7$
$x_1 = [0,0,1]$
$x_3 = [1,0,0]$
(1): $x_3^{(3)} = [3,4,5]$
(1): $x_1^{(1)} = [-1,-1,2]$
p_1
p_3
(2): $\sum_{i=1}^{M} x_i^{(1)} = [-5,-15,-8]$
(2): $\sum_{i=1}^{M} x_i^{(3)} = [9,16,-4]$
(1): $x_1^{(3)} = [3,4,-4]$
(1): $x_3^{(1)} = [-3,-9,-12]$

Fig. 36.1 Basic anonymous ID assignment (AIDA-D), $N = 3$, $M = 3$

Like AIDA-H, AIDA-D uses secure sum as a basis for ID assignment. However, there are some important differences. AIDA-D algorithm is completely decentralized. There is no special node that initiates and terminates the computation, but everyone participates equally in the ID assignment process. In Step 1 every node creates a random slot array and splits its array into M partitions sending one to each node. In Step 2 each nodes sums the received partitioned-arrays and broadcasts it to everyone. It is easy to see that the collusion resistant feature of this algorithm is $M-1$, since collusion of $M-1$ or fewer nodes won't let the colluding set discover the ID chosen by any of the non-colluding nodes.

Couple of observations about both of these algorithms are in order. First, the manner in which a node selects a slot in a round is the same for both algorithms. Thus, the number of unique IDs chosen in each round is, and by extension the number of rounds needed, the same no matter which algorithm we use. Second, note that both algorithms have a non-zero probability of running forever. This can happen, if two nodes choose the same slot in each round.

In the remaining sections of this paper, we use AIDA as a synonym for both of our algorithms as the distinction between the two algorithms is *not* relevant to these discussions.

36.3 Performance Analysis

AIDA relies on randomly choosing as many as M slots (the number of nodes) from a slot array of size N for each round until each node has been given a unique, anonymous ID. Unfortunately, as many as M nodes have the possibility of randomly colliding; that is, each node may randomly choose the same slot in single selection round.

An interesting question is how many nodes get a unique ID in a round. Given that M is the number of nodes who need an anonymous ID and N is the size of the slot array (or number of IDs) for AIDA, we wish to compute the probability of completing AIDA in 1 round, and the number of unique IDs assigned in a round.

Problem 1: Probability of finishing the algorithm in one round

In order to finish the algorithm in one round, all participants have to select different IDs regardless. Given fixed N and $M \leq N$, we can easily derive the following equation.

$$Prob_{(\text{AIDA finish in one round})} = \frac{\binom{N}{M}M!}{N^M} = \frac{N!}{(N-M)!N^M} = \prod_{i=1}^{M-1}\left(1-\frac{i}{N}\right) \quad (36.1)$$

An approximation to Eq. (36.1) can be obtained using the approximation $e^x = 1+x$ derived from the Taylor series for e^x.

Replacing $1 - \frac{i}{N}$ in Eq. (36.1) with $e^{-\frac{i}{N}}$ leads to

$$Prob_{(\text{AIDA finish in one round})} \approx \prod_{i=1}^{M-1} e^{-\frac{i}{N}} = e^{-\sum_{i=1}^{M-1} \frac{i}{N}} = e^{-\frac{M(M-1)}{2N}}. \tag{36.2}$$

Problem 2: Number of IDs assigned in one round

With AIDA, if more than one participant selects the same ID, the ID will not be assigned to anyone. Let $C_{(AIDA)}(N', M')$ denote the number of combination of assigning N' IDs to M' participants, where each ID is selected by none, or more than two participants. Then, in terms of multinomial coefficients we have

$$C_{(AIDA)}(N', M') = \sum_{r_1+\cdots+r_{N'}=M', r_i \in \{0,2,3,\ldots\}} \frac{M'!}{\prod_{i=1}^{N'} r_i!}.$$

A term of the sum represents the number of ways in which r_i nodes can be assigned to slot i. Excluding $r_i = 1$ gives the desired sum.

From the exclusion of the term "$\frac{x}{1!}$" from the product (convolution) of exponential generator function, $G(x)$, given below it follows easily that the $C_{(AIDA)}(N', M')$ is $M'!$ times the coefficient of the term $x^{M'}$.

$$G(x) = (1 + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots)^{N'}. \tag{36.3}$$

We now manipulate the generating function $G(x)$ to develop a formula for $C_{AIDA}(N', M')$ that is somewhat easier to compute.
Replace $1 + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ in Eq. (36.3) with $e^x - x$ by applying the Taylor's formula and manipulate to obtain

$$G(x) = \sum_{i=0}^{N'} \binom{N'}{i} (-1)^i x^i \sum_{r=0}^{\infty} \frac{1}{r!} (N' - i)^r x^r. \tag{36.4}$$

Recall that $C_{(AIDA)}(N', M')$ is $M'!\times$(coefficient of $x^{M'}$ in $G(x)$). Thus,

$$C_{AIDA}(N', M') = M'! \sum_{i=0}^{M'} \binom{N'}{i} (-1)^i \frac{1}{(M'-i)!} (N' - i)^{M'-i}.$$

Given N and M, the number of combination of assigning D anonymous IDs to M participants without conflicts is given by

$$\binom{N}{D} \binom{M}{D} D! C_{AIDA}(N - D, M - D). \tag{36.5}$$

And, the corresponding probability is readily given by

$$Prob_{(AIDA)}(D,N,M) = \frac{\binom{N}{D}\binom{M}{D}D!C_{AIDA}(N-D,M-D)}{N^M}. \tag{36.6}$$

Now we can easily compute the expected value of the number of anonymous ID assigned in one round for AIDA.

$$\begin{aligned} Expected_{(AIDA)}(N,M) &= \sum_{d=0}^{M} Prob_{(AIDA)}(d,N,M)\cdot d \\ &= \sum_{d=0}^{M} \frac{\binom{N}{d}\binom{M}{d}d!C_{AIDA}(N-d,M-d)}{N^M}\cdot d. \end{aligned} \tag{36.7}$$

We used Eq. (36.7) to compute the number of rounds for $0 \leq N \leq 50$, $M \leq 30$. For example, our computation reveals that 16.69 IDs are assigned in a round for $N = 50$ and $M = 30$, while 10.5 IDs are assigned for $N = 30$ and $M = 20$.

Problem 3: Termination of AIDA

Note that there is a non-zero probability that no node is assigned a unique ID in a round. If it happens again and again, the algorithm could run forever. So, an interesting question is the number of rounds needed in practice.

We ran a simulation of AIDA varying N from 5 to 30 by 5, and M from 30 to 505 by 25, and performing 10,000 trials. The average number of iterations needed to terminate AIDA for each (M,N) pair is recorded on the Z axis (Fig. 36.2). As is seen, on average AIDA terminates with 4 rounds or less, with the majority of the (M,N) node/slot pairs needing only one to two rounds to terminate.

For each (M,N) we also noted the maximum number of rounds needed to terminate AIDA. For brevity, we omit this plot, but point out that AIDA terminates within a maximum of seven rounds or less, for all parameter values. Given our choice of 10,000 trials we are confident that this data is consistent and indicative of the practical use of AIDA.

36.4 Anonymous ID Assignment and Opt-Out

Now, we turn to the question of global mining and use of anonymous ID for opt-out. We now develop a CPSS algorithm with an opt-out feature satisfying the criteria of Definition 36.1. We term this algorithm: "cycle-partitioned secure sum with opt-out" or CPSOO.

Definition 36.1. An algorithm is "K-collusion resistant for anonymous opt-out" if (a) the global sum is non-viable upon opt-out, (b) any set of K or fewer nodes cannot calculate the value of any node not in the K-node set, and (c) any set of K nodes or fewer cannot determine if some other node not in the K-node set has opted out.

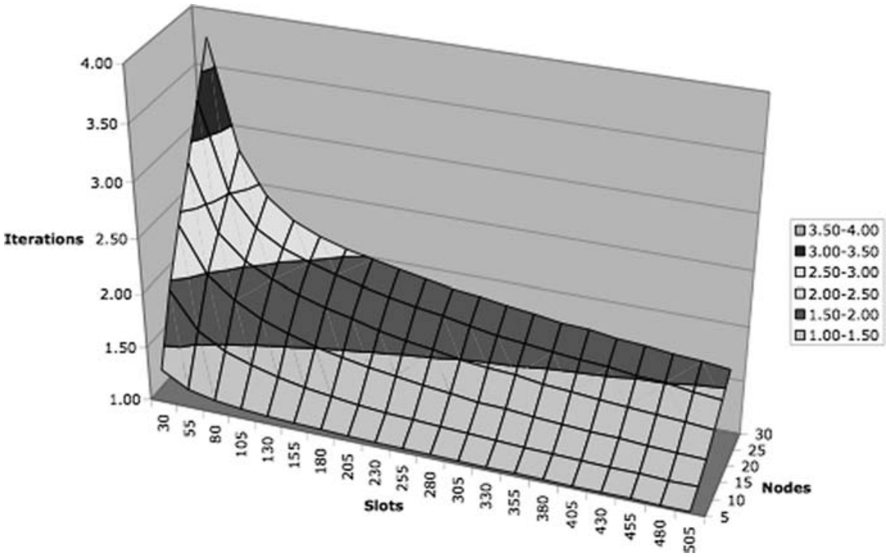


Fig. 36.2 AIDA – average iterations for 10,000 trials

Using AIDA as the initial step, we have seen that we are able to assign each node in our data mining operation a unique ID that is anonymous with collusion resistance (line 1 of Algorithm 2). By integrating either of our AIDA algorithms into CPSOO we may create an array of bits B, where the index of each element in B corresponds to a data mining node's unique ID number. The purpose of the bit array B is to allow nodes to anonymously opt-out without interfering with another node's opt-out and to avoid error by separating the opt-out declaration from the method for making the global sum non-viable (trashing it). Collusion resistance is provided by sending opt-out bits using CPSS. In line 2 of Algorithm 2 the array B is created, on lines 21–25 each node j may choose to opt-out or not *privately* with line 22 in particular being responsible for making the global sum non-viable. The opt-out is recognized on lines 37–42 by computing the value of the opt-out bits and summing them. The anonymous, IDs allow this non-intereference by making sure each node has its own *private* data element that cannot be mistakenly cancelled out by two nodes XORing over the same bit.

In the tradition of CPSS, each element of B is partitioned and sent along each cycle with the traditional sum values. However, this partitioning is a random bit-partition instead of an integer one. For example, suppose again we have M nodes and that for some node N_j with ID #k ($1 \leq k \leq M$ and $j \neq k$) there is a bit value B_k in array B. Then if we bit-partition B_k for C cycles, we have $B_k = \bigoplus_{i=1}^{C} B_k^{(i)}$ where $\oplus$ is the bitwise eXclusive-OR (XOR) operation. (Recall that $1 \oplus 1 = 0$, $1 \oplus 0 = 1$, $0 \oplus 1 = 1$, and $0 \oplus 0 = 0$.)

The opt-out mechanism for CPSOO follows very naturally from this data structure. If a node wishes to opt-out, it simply sets all bit-partitions such that their XOR is equal to 1, that is, its bit-value in B is set to 1. Additionally, the node opting out adds a random number instead of its real count value to the sum in order to ensure that the global sum will be non-viable. More importantly, the true value of the withdrawing node cannot be known by any other nodes. The psuedocode for CPSOO is given in Algorithm 2.

Both protocols employ CPSS as a basis so their collusion resistance capabilities are identical. We assert that

Theorem 36.1. *Given any network with C edge-disjoint Hamiltonian cycles and M nodes such that $M > 4$, CPSOO is "K-collusion resistant for anonymous opt-out" where $K = 2C - 1$.*

For brevity, we do not include the proof of Theorem 36.1 here, see Section 36.5 of [13] for details.

A diagram of some typical inputs and outputs for an arbitrary node "N2" in mid-computation of CPSOO is given in Fig. 36.3. N2's anonymous ID is 3. Here there are $C = 2$ cycles and $M = 5$ nodes. The two edge-disjoint cycles are represented by the dotted and dashed lines, with two distinct sending and two distinct receiving nodes assumed for each ray. The bracketed numbers are the partitioned running subtotals of the count value of the itemset whereas the parenthetical numbers are for the 5 opt-out bits of the bit-array B. A normal and opt-out example are shown.

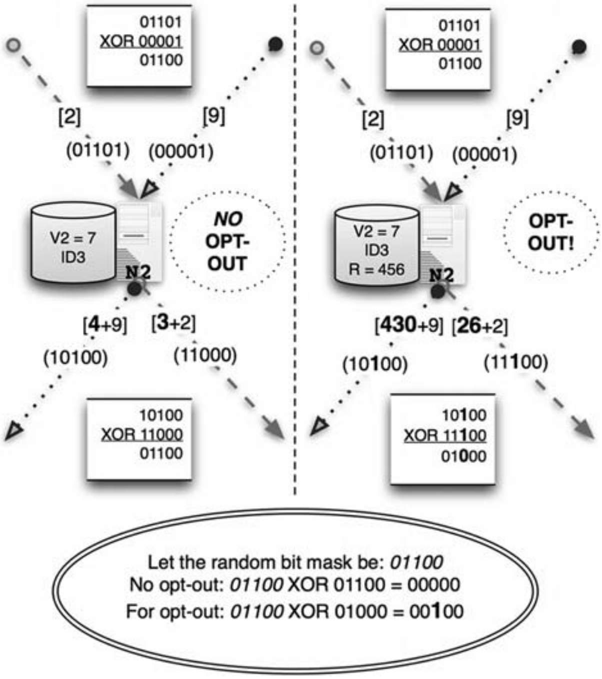


Fig. 36.3 CPSOO: inputs and outputs of an arbitrary node, $M = 5$ nodes and $C = 2$ cycles

In the normal *no opt-out* scenario (left side), the count value 7 is randomly partitioned into 4 and 3 and added to the (incoming) 9 and 2 subtotal respectively. The incoming bit-array partitions (01101) and (00001) are modified to (10100) and (11000), but their XOR is *not* modified, that is: $01101 \oplus 00001 = 01100 = 10100 \oplus 11000$.

However, in the *opt-out* scenario of Fig. 36.3 (right side), a large random number is generated (456), randomly partitioned into 430 and 26, and added to each subtotal to make the global sum non-viable. To signal for opt-out, the node N2 modifies the incoming bit-array partitions (01101) and (00001) such that the XOR of the incoming partitions is different from the XOR of the outgoing partitions, that is: $01101 \oplus 00001 (= 01100) \neq 11100 \oplus 10100 (= 01000)$. Notice also that $01000 \oplus 01100 = 00100$. The third bit is set because N2 has ID3. Since only N2 owns ID3 and the ID is anonymous, the opt-out may occur anonymously and without the possibility of a collision, that is, no opt-out can cancel out another opt-out by both nodes XORing the same bit.

The prime difference between CPSOO and CPSS is the addition of an opt-out feature involving bit partitions. Protocol CPSOO employs AIDA – see Step 1 of Algorithm 2 – as the basis for providing opt-out capabilities. But, an interesting question is whether AIDA is essential to the provision of opt-out capabilities. It turns out that it is not as was shown in [8]; it uses SS as the basis for computing both global sum and for a node to opt out of the mining process (see [8] Section III.B for details).

36.5 Concluding Remarks

Much of the previous work on anonymous protocols are based on specific applications and or encryption. Our protocols of Section 36.2 are a bit simpler in that we seek to assign an anonymous ID to each node of the network. We do not require any information about the node such as its public keys or any individual attributes. We also do not need any centralized servers as would be needed in some of the other protocols, for example [4]. Similar to [16], our work of Section 36.3 used CPSS as a basis. Though the collusion resistance we achieve is similar to [15,16], our contribution is unique because we include an "opt-out" feature to address additional security concerns. In fact, the motivation for our work comes from [11]. They suggested the possibility of using an anonymous opt-out method to prevent statistical disclosure in their secure sum based distributed secure regression.

The use of random numbers is vital to the operation of the secure sum, cycle-partitioned secure sum, and in AIDA algorithms. In secure sum and CPSS a random number is added to the value of the initiating node or site to mask that value. This masking must be done carefully. For a complete discussion on how to easily address this and other numerical issues, please see (Section 4.1 [13]). For brevity, we do not include these details in this chapter.

APPENDIX 36.A: Algorithm Pseudocode

Algorithm 1 AIDA-H: Anonymous ID Assignment, with K-collusion Resistance

```
Require: Given $M$ nodes and $C$ edge-disjoint Hamiltonian cycles.
Ensure: Each node gets a unique, anonymously assigned ID $\in [1, M]$.
{For $N_1$, the initiating node.}
$Len \Leftarrow \mathcal{N}$ {For some choice of $\mathcal{N} \geq M$}
Initialize $S$ array. $\{S_j^{(i)} = 0$, for $1 \leq i \leq C$ and $1 \leq j \leq Len\}$
$NumOnes \Leftarrow 0$
repeat
    $R \Leftarrow ArrayOfRandomNumbers(Len)$ $\{R_j = NewRandomNumber(), \forall j\}$
    $RandomlySelectSlot(S, R, Len, C)$ $\{S = S + R$, increment $S_{selected}$,
    $RandomlyPartition(S_j, C)\ \forall j.\}$
    for $i = 1$ to $C$ do
        Send $S^{(i)}$ to the next node in cycle $i$. $\{S_j^{(i)} \in S^{(i)}$, for $1 \leq j \leq Len\}$
    end for
    {For each node, if necessary, select a random slot.}
    for $k = 2$ to $M$ do
        Intialize $Z$ array. $\{Z_j = 0$, for $1 \leq j \leq Len\}$
        $RandomlySelectSlot(S, Z, Len, C)$ $\{S = S + Z$, increment $S_{selected}$,
        $RandomlyPartition(S_j, C)\ \forall j.\}$
        for $i = 1$ to $C$ do
            Send $S^{(i)}$ to the next node in cycle $i$. $\{S_j^{(i)} \in S^{(i)}$, for $1 \leq j \leq Len\}$
        end for
    end for
    {$N_1$ receives $C$ arrays $S$ & announces all assignments for current round.}
    for $j = 1$ to $Len$ do
        $Sum \Leftarrow 0$
        {Recover the value at $S_j$ and see if it is selected.}
        for $i = 1$ to $C$ do
            $Sum \Leftarrow Sum + S_j^{(i)}$
        end for
        $Sum \Leftarrow Sum - R_j$ {Remove random number from the sum.}
        if $Sum = 1$ then
            $NumOnes \Leftarrow NumOnes + 1$
            Broadcast to all nodes: Node choosing $j$ is ID $NumOnes$.
        end if
    end for
until $NumOnes = M$
```

The pseudocode for the AIDA-H subroutines, *RandomlyPartition* and *RandomlySelectSlot*, are given in Algorithms 2 and 7 of [13]. The basic idea of *RandomlyPartition* can be stated as: divide any integer A into C random non-zero integers A_i such that $\sum_{i=1}^{C} A_i = A$. The function *RandomlyBitPartition* used in Algorithm 2 can also be described easily: given a bit B and some number of partitions Q, the function creates random bit partitions of B named B_i' such that $B = \bigoplus_{i=1}^{Q} B_i'$. It returns the array of these bits called B'.

Algorithm 2 CPSOO: Cycle-partitioned Secure Sum with Opt-out

```
Require: Given $M$ nodes and $C$ edge-disjoint Hamiltonian cycles.
Ensure: $\sum_{k=1}^{M} V_k = \mathit{GlobalSum}$ OR opt-out: $\sum_{k=1}^{M} V_k \neq \mathit{GlobalSum}$
AnonymousIDAssignment() $\{ID_k$ is the unique ID for node $N_k$, $1 \leq ID_k \leq M\}$
{For $N_1$, the initiating node.}
Initialize array $B$. $\{B_j^{(i)} = 0$, for $1 \leq i \leq C$ and $1 \leq j \leq M\}$
$R \Leftarrow \mathit{NewRandomNumber}()$
for $j = 1$ to $M$ do
    $T_j \Leftarrow \mathit{NewRandomBit}()$
    $B_j \Leftarrow \mathit{RandomlyBitPartition}(T_j, C)$ $\{T_j^{(1)} \oplus T_j^{(2)} \oplus \cdots \oplus T_j^{(C)} = T_j\}$
    if $j = ID_1$ AND $N_1$ opts out then
        $V_1 \Leftarrow R$
        $B_j \Leftarrow \mathit{RandomlyBitPartition}(T_j \oplus 1, C)$
    else if $j = ID_1$ then
        $V_1 \Leftarrow V_1 + R$
    end if
end for
$\mathit{RandomlyPartition}(V_1, C)$ $\{\sum_{i=1}^{C} V_1^{(i)} = V_1,\ 0 \neq V_1^{(i)} \in \mathbb{Z}\}$
for $i = 1$ to $C$ do
    Send $V_1^{(i)}$ & $B^{(i)}$ to the next node in cycle $i$. $\{B_j^{(i)} \in B^{(i)}, \forall j \in [1, M]\}$
end for
{For each node, partition & add to subtotals in each cycle.}
for $k = 2$ to $M$ do
    for $j = 1$ to $M$ do
        if $j = ID_k$ AND $N_k$ opts out then
            $B' \Leftarrow \mathit{RandomlyBitPartition}(1, C)$ $\{1 = \bigoplus_{i=1}^{C} B'_i\}$
            $V_k \Leftarrow \mathit{NewRandomNumber}()$
        else
            $B' \Leftarrow \mathit{RandomlyBitPartition}(0, C)$ $\{0 = \bigoplus_{i=1}^{C} B'_i\}$
        end if
        for $i = 1$ to $C$ do
            $B_j^{(i)} \Leftarrow B_j^{(i)} \oplus B'_i$
        end for
    end for
    $\mathit{RandomlyPartition}(V_k, C)$ $\{\sum_{i=1}^{C} V_k^{(i)} = V_k, 0 \neq V_k^{(i)} \in \mathbb{Z}\}$
    for $i = 1$ to $C$ do
        $V_k^{(i)} \Leftarrow V_k^{(i)} + V_{received}^{(i)}$
        Send $V_k^{(i)}$ & $B^{(i)}$ to the next node in cycle $i$. $\{B_j^{(i)} \in B^{(i)}, \forall j \in [1, M]\}$
    end for
end for
$\mathit{GlobalSum} = \sum_{i=1}^{C} V_{received}^{(i)} - R$ $\{N_1$ receives $C$ values.$\}$
$\mathit{BitSum} \Leftarrow \sum_{j=1}^{M}((\bigoplus_{i=1}^{C} B_j^{(i)}) \oplus T_j)$ {reconstruct opt-out bits & add.}
if $\mathit{BitSum} \geq 1$ then
    Broadcast $\mathit{SumCancelled}$ to each other node.
else
    Broadcast $\mathit{GlobalSum}$ to each other node.
end if
```

References

1. Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across private databases. In: SIGMODIC: ACM SIGMOD Interantional Conference on Management of Data (2003)
2. Bottcher, S., Obermeier, S.: Secure set union and bag union computation for guaranteeing anonymity of distrustful participants. JSW **3**(1), 9–17 (2008)
3. Brickell, J., Shmatikov, V.: Efficient anonymity-preserving data collection. In: KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 76–85. ACM, New York (2006)
4. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Commun. ACM **24**(2), 84–88 (1981)
5. Chaum, D.: The dining cryptographers problem: Unconditional sender and recipient untraceability. J. Cryptology **1**(1), 65–75 (1988)
6. Clifton, C., Kantarcioglu, M., Vaidya, J.: Defining privacy for data mining. In: Kargupta, H., Joshi, A., Sivakumar, K. (eds.) National Science Foundation Workshop on Next Generation Data Mining, pp. 126–133, Baltimore, MD (2002a)
7. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, M.Y.: Tools for privacy preserving distributed data mining. SIGKDD Explor. Newsl. **4**(2), 28–34 (2002b)
8. Dong, R., Kresman, R.: Indirect disclosures in data mining. In: The 4th International Conference on Frontier of Computer Science and Technology, Shanghai, China (2009)
9. Ishiyama, M., Kunishi, M., Kohno, M., Teraoka, F.: Secured anonymous ID assignment support for LIN6. In: Lecture Notes in Computer Science, vol. 3090, pp. 297–306. Springer, Germany (2004)
10. Kantarcioglu, M.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. Knowl. Data Eng. **16**(9), 1026–1037 (2004), Senior Member-Chris Clifton
11. Karr, A., Lin, X., Sanil, A., Reiter, J.: Secure regression on distributed databases. Technical Report 141, National Institute of Statistical Science, Research Triangle Park, NC(2004)
12. Neumann, Peter G.: Security criteria for electronic voting. In: 16th National Computer Security Conference, Baltimore, MD (1993)
13. Shepard, Samuel S.: Anonymous opt-out and secure computation in data mining. Master's thesis, Bowling Green State University (2007) / OhioLINK
14. Shepard, Samuel S., Kresman, R., Dunning, L.: Data mining and collusion resistance. In: Ao, S.I., Gelman, L., Hukins, David W.L., Hunter, A., Korsunsky, A.M. (eds.) Proceedings of the World Congress on Engineering 2009, vol. I, pp. 283–288, WCE'09, 1–3 July 2009, London, UK (2009)
15. Urabe, S., Wang, J., Kodama, E., Takata, T.: A high collusion-resistant approach to distributed privacy-preserving data mining. In: Burkhart, H. (ed.) Parallel and Distributed Computing and Networks, vol. 551–803, pp. 326–331. ACTA Press, Innsbruck, Austria (2007)
16. Urabe, S., Wang, J., Takata, T.: A collusion-resistant approach to distributed privacy-preserving data mining. In: Gonzalez, T. (ed.) Parallel and Distributed Computing and Systems, vol. 436-088, pp. 626–631. ACTA Press, MIT Cambridge, USA (2004)
17. von Ahn, L., Bortz, A., Hopper, N.J.: k-anonymous message transmission. In: SIGSAC: 10th ACM Conference on Computer and Communications Security. ACM SIGSAC (2003)

Chapter 37
Clustering Biological Data Using Enhanced k-Means Algorithm

K.A. Abdul Nazeer and M.P. Sebastian

Abstract With the advent of modern scientific methods for data collection, huge volumes of biological data are now getting accumulated at various data banks. The enormity of such data and the complexity of biological networks greatly increase the challenges of understanding and interpreting the underlying data. Effective and efficient Data Mining techniques are essential to unearth useful information from them. A first step towards addressing this challenge is the use of clustering techniques, which helps to recognize natural groupings and interesting patterns in the data-set under consideration. The classical k-means clustering algorithm is widely used for many practical applications. But it is computationally expensive and the accuracy of the final clusters is not guaranteed always. This paper proposes a heuristic method for improving the accuracy and efficiency of the k-means clustering algorithm. The modified algorithm is then applied for clustering biological data, the results of which are promising.

Keywords Data mining · clustering · k-means algorithm

37.1 Introduction

Advances in scientific data collection methods have resulted in the large scale accumulation of biological data at various data sources. Owing to the development of novel techniques such as DNA Microarrays for generating data [1], the rate of growth of scientific databases has become tremendous. Hence it is practically

K.A.A. Nazeer (✉) and M.P. Sebastian
Department of Computer Science and Engineering National Institute of Technology Calicut,
NIT Campus (PO), Kozhikode, India-673601
e-mail: nazeer@nitc.ac.in; sebasmp@nitc.ac.in

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_37,

impossible to extract useful information from them by using conventional database analysis techniques. Effective mining methods are absolutely essential to unearth implicit information from huge databases.

Cluster analysis, as discussed in [2] is one of the major data analysis methods which is widely used for many practical applications. Clustering is the process of partitioning a given set of objects into disjoint clusters. This is done in such a way that objects in the same cluster are similar while objects belonging to different clusters differ considerably, with respect to their attributes. The process of clustering biological data helps to identify interesting patterns and inherent groupings in the underlying data.

The k-means algorithm proposed by [3] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high, especially for large data sets. Moreover, this algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers for improving the performance of the k-means clustering algorithm. This paper discusses a heuristic method for improving the accuracy and efficiency of the k-means clustering algorithm.

37.2 k-Means Clustering Algorithm

This section describes the original k-means clustering algorithm. The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is used as the measure to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may change the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not change anymore. This indicates the convergence criterion for the clustering procedure. Pseudocode for the k-means clustering algorithm as given in [4] is listed as Algorithm 1.

The k-means algorithm is an extensively studied algorithm for clustering and is generally effective in producing good results. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids. Accuracy of the final clusters heavily depends on the selection of the initial centroids. The k-means algorithm is computationally expensive also. Its time complexity is $O(nkl)$ where n is the number of data points, k the number of clusters and l the number of iterations.

Algorithm 1 The k-means Clustering Algorithm

```
Input:
  D = d1, d2,.....dn // set of n data items
  k // Number of desired clusters

Output:
  A set of k clusters

Steps:
  1. Arbitrarily choose k data-items from D as initial centroids;
  2. Repeat
     Assign each item di to the cluster which has the closest centroid;
     Calculate the new mean for each cluster;
     Until convergence criterion is met.
```

37.3 Literature Survey

Several attempts were made by researchers to improve the accuracy and efficiency of the k-means algorithm, as discussed in [5]. A variant of the k-means algorithm is the k-modes method proposed by [6] which replaces the means of clusters with modes. Like the k-means method, the k-modes algorithm also produces locally optimal solutions which are dependent on the selection of the initial modes. The k-prototypes algorithm discussed in [5] integrates the k-means and k-modes processes for clustering the data. In this method, the dissimilarity measure is defined by taking into account both numeric and categorical attributes. As shown in Algorithm 1, the original k-means algorithm consists of two phases: one for determining the initial centroids and the other for assigning data points to the nearest clusters and then recalculating the cluster means. The second phase is carried out repetitively until the clusters get stabilized, i.e., data points stop crossing over cluster boundaries.

Yuan et al. [7] proposed a systematic method for finding the initial centroids. The centroids obtained by this method are consistent with the distribution of data. Hence it produced clusters with better accuracy, compared to the original k-means algorithm. However, this method does not suggest any improvement to the time complexity of the k-means algorithm.

Fahim et al. [8] proposed an efficient method for assigning data-points to clusters. The original k-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose- one similar to the k-means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original k-means algorithm. Hence there is no guarantee for the accuracy of the final clusters.

37.4 Proposed Method

In the method proposed in this paper, both the phases of the k-means algorithm are modified to improve the accuracy and efficiency. The improved method is outlined as Algorithm 2.

Algorithm 2 The Improved Clustering Algorithm

```
Input:
    D = d1, d2,.....dn // set of n data items
    k // Number of desired clusters

Output:
    A set of k clusters

Steps:
    1. Determine the initial centroids of the clusters by using Algorithm 3;
    2. Assign the data points to the clusters by using Algorithm 4;
```

In the first phase, the initial centroids are determined systematically as discussed in [7] so as to produce clusters with improved accuracy. In the second phase of assigning data points to clusters, a variant of the algorithm discussed in [8] is used. It starts by forming the initial clusters based on the relative distance of each data point from the initial centroids. These clusters are subsequently refined by using a heuristic approach. The two phases of the improved method are described below as Algorithms 3 and 4.

Algorithm 3 Finding the Initial Centroids

```
Input:
    D = d1, d2,.....dn // set of n data items
    k // Number of desired clusters

Output:
    A set of k initial centroids

Steps:
    1. Set m = 1;
    2. Compute the distance between each data point and all other data points in the set D;
    3. Find the closest pair of data points from the set D and form a data point set Am (1 <=
       m <= k) which contains these two data points. Delete these two data points from the set D;
    4. Find the data point in D that is closest to the data point set Am. Add it to Am and delete it
       from D;
    5. Repeat step 4 until the number of data points in Am reaches 0.75*(n/k);
    6. If m < k, then m = m + 1. Find another pair of data points from D between which the
       distance is the shortest. Form another data point set Am and delete it from D. Go to Step 4;
    7. For each data point set Am (1 <= m <= k) find the arithmetic mean of the vectors of data
       points in Am. These means will be the initial centroids.
```

Initially, determine the distances between each data point and all other data points in the set. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to A1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold, which is taken to be 0.75*(n/k). At that point, go back to the second step and form another data point set A2. Repeat this till k such sets of data points are obtained. Finally the initial centroids are obtained by taking the arithmetic mean of the vectors in each data point set.

The Euclidean distance is used for determining the closeness of each data point to the cluster centroids. The distance between one vector X = (x_1, x_2,.....x_n) and another vector Y = (y_1, y_2,.....y_n) is obtained as

$$d(X, Y) = \sqrt(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2 \tag{37.1}$$

The distance between a data point X and a data point set D is defined as

$$d(X, D) = min\ (d(X, Y), where\ Y \in D) \tag{37.2}$$

The initial centroids obtained in phase 1 are given as input to the second phase, for assigning data point to the appropriate clusters. The steps involved in this phase are described as Algorithm 4.

Algorithm 4 Assigning data points to clusters

```
Input:
  D = d1, d2,.....dn // set of n data items
  C = c1, c2,.....ck // set of k centroids

Output:
  A set of k clusters

Steps:
  1. Compute the distance of each data point di (1 <= i <= n) to all the centroids cj (1 <=
     j <= k) as d(di, cj);
  2. For each data point di, find the closest centroid cj and assign di to cluster j.
  3. Set ClusterId[i] = j; // j:Id of the closest cluster
  4. Set NearestDist[i] = d(di, cj);
  5. For each cluster j (1 <= j <= k), recalculate the centroids;
  6. Repeat
  7. For each data point di,
     7.1 Compute its distance from the centroid of the present nearest cluster;
     7.2 If this distance is less than or equal to the present nearest distance, the data point stays
         in the cluster; Else
         7.2.1 For every centroid cj (1 <= j <= k) Compute the distance d(di, cj); Endfor;
         7.2.2 Assign the data point di to the cluster with the   nearest centroid cj;
         7.2.3 Set ClusterId[i] = j;
         7.2.4 Set NearestDist[i]= d(di, cj); Endfor;
  8. For each cluster j (1 <= j <= k), recalculate the centroids; Until the convergence criteria
     is met.
```

The first step in Phase 2 is to determine the distance between each data point and the initial centroids of all the clusters. The data points are then assigned to the clusters having the closest centroids. This results in an initial grouping of the data points. For each data point, the cluster to which it is assigned (ClusterId) and its distance from the centroid of the nearest cluster (NearestDist) are noted. Inclusion of data points in various clusters may alter the cluster centroids. For each cluster, the centroids are recalculated by taking the mean of the values of its data points. Up to this step, the procedure is almost similar to the original k-means algorithm except that the initial centroids are computed by a systematic procedure.

The next stage is an iterative process which makes use of a heuristic method to improve the efficiency. During the iteration, the data points may get redistributed to different clusters. The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. At the beginning of the iteration, the distance of each data point from the new centroid of its present nearest cluster is determined. If this distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to compute its distance from other centroids. This results in the saving of time required to compute the distances to k-1 cluster centroids. On the other hand, if the new centroid of the present nearest cluster is more distant from the data-point than its previous centroid, there is a chance for the data point getting included in another nearer cluster. In that case, it is required to determine the distance of the data point from all the cluster centroids. Identify the new nearest cluster and record the new value of the nearest distance. The loop is repeated until no more data points cross cluster boundaries, which signifies the convergence criterion. The heuristic method described above results in significant reduction in the number of computations and thus improves the efficiency.

37.5 Computational Complexity

Phase 1 of the enhanced algorithm requires a time complexity of $O(n^2)$ for finding the initial centroids, as the maximum time required here is for computing the distances between each data point and all other data points in the set D. In the original k-means algorithm, before the algorithm converges the centroids are calculated many times and the data points are assigned to their nearest centroids. Since complete redistribution of the data points takes place according to the new centroids, this takes $O(nkl)$, where n is the number of data points, k is the number of clusters and l is the number of iterations. To obtain the initial clusters, Algorithm 4 requires $O(nk)$. Here, some data points remain in its cluster while the others move to other clusters depending on their relative distance from the new centroid and the old centroid. This requires $O(1)$ if a data point stays in its cluster, and $O(k)$ otherwise. As the algorithm converges, the number of data points moving away from their cluster decreases with each iteration. Assuming that half the data points move from their clusters, this requires $O(nk/2)$. Hence the total cost of this phase of the algorithm is $O(nk)$, not $O(nkl)$. Thus the overall time complexity of the enhanced algorithm (Algorithm 2) becomes $O(n^2)$, since k is much less than n.

37.6 Experimental Results

The improved algorithm was tested with multivariate data taken from the UCI repository of machine learning databases [9]. The iris data, echocardiogram data and the breast cancer data were clustered using the original k-means algorithm and the improved algorithm.

The results of the experiments are tabulated in the Tables 37.1–37.3. For the original k-means algorithm, three experiments each were conducted for the three data sets for different values of the initial centroids. The average values of the accuracy and time taken were then computed and tabulated. For the improved algorithm, the data sets and the value of k are the only inputs required and one experiment each were conducted for the same data sets. The values of accuracy and time taken are then tabulated. Figures 37.1–37.3 illustrate the performance of the improved algorithm compared to the original k-means algorithm. It can be seen that the improved algorithm significantly outperforms the k-means algorithm in terms of accuracy and efficiency.

37.7 Conclusion

The k-means algorithm is widely used for clustering large sets of data. But the standard algorithm do not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids. Moreover, the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times, during every iteration of the loop.

Table 37.1 Performance comparison for iris data

Algorithm	Accuracy (%)	Time Taken (mS)
K-means	78.7	71
Enhanced K-means	88.6	69

Table 37.2 Performance comparison for echocardiogram data

Algorithm	Accuracy (%)	Time Taken (mS)
K-means	53.3	73
Enhanced K-means	70	72

Table 37.3 Performance comparison for breast cancer data

Algorithm	Accuracy (%)	Time Taken (mS)
K-means	86.6	72
Enhanced K-means	95	70

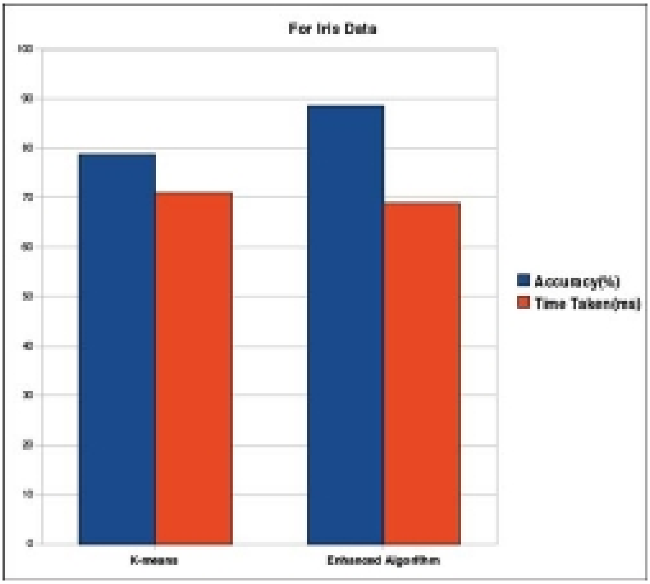


Fig. 37.1 Performance comparison for iris data

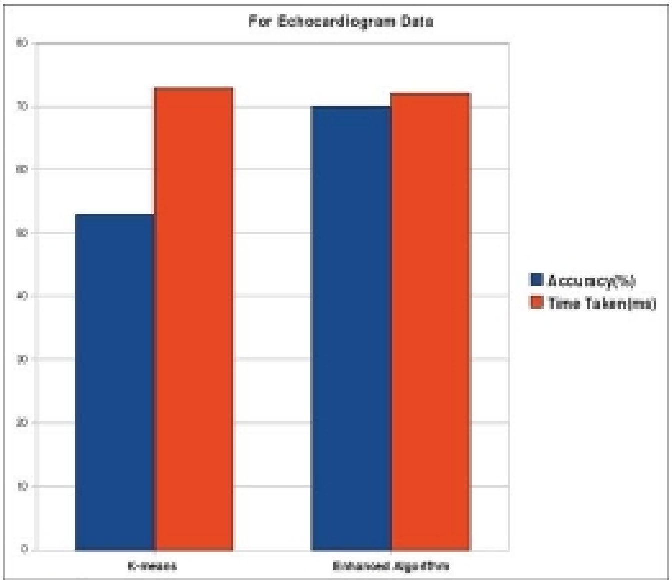


Fig. 37.2 Performance comparison for echocardiogram data

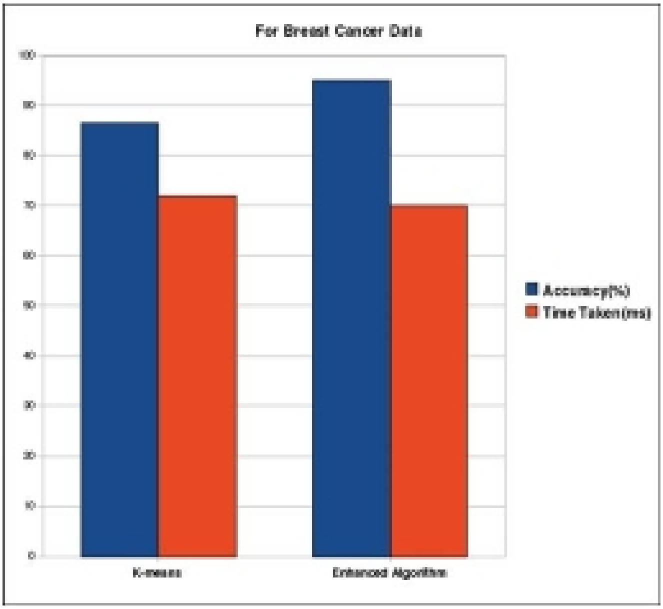


Fig. 37.3 Performance comparison for breast cancer data

This paper presents an enhanced k-means algorithm which combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. This method ensures the entire process of clustering in $O(n^2)$ time without sacrificing the accuracy of clusters. The previous improvements of the k-means algorithm compromise on either accuracy or efficiency.

A limitation of the proposed algorithm is that the value of k, the number of desired clusters, is still required as an input. Evolving some statistical methods to compute the value of k, based on the distribution of data, is suggested for future research. The method for finding the initial centroids may be refined further to improve the time complexity.

References

1. Daxin J., Chum T., Aidong Z.: Cluster analysis for gene expression data. IEEE Trans. Data Knowl. Eng., **16**(11), 1370–1386 (2004)
2. Han, J.: Data mining concepts and techniques. Morgan Kaufmann Publishers, An imprint of Elsevier, San Francisco, CA (2006)
3. McQueen, J.: Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist. Prob. (1), 281–297 (1967)
4. Dunham, M.H.: Data Mining-Introductory and Advanced Concepts. Pearson Education (2006)
5. Huang Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Disc. (2), 283–304 (1998)
6. Chaturvedi, J.C.A., Green, P.: K-modes Clustering. J. Classif. (18), 35–55 (2001)

7. Yuan, F., Meng, Z.H., Zhang, H.X., Dong, C.R.: A new algorithm to get the initial centroids. In: Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004
8. Fahim, A.M., Salem, A.M., Torkey, A., Ramadan, M.A.: An efficient enhanced k-means clustering algorithm. J. Zhejiang Univ. **10**(7), 1626–1633 (2006)
9. Merz, C., Murphy, P.: UCI Repository of Machine Learning Databases. http://archive.ics.uci.edu/ml/

Chapter 38
The Ornstein–Uhlenbeck Processes Driven by Lévy Process and Application to Finance

Ömer Önalan

Abstract In this study we deal with aspects of the modeling of the asset prices by means Ornstein-Uhlenbech process driven by Lévy process. Barndorff-Nielsen and Shephard stochastic volatility model allows the volatility parameter to be a self-decomposable distribution. BNS models allow flexible modeling. For this reason we use as a model the IG-Ornstein-Uhlenbeck process. We calibrate moments of Lévy process and OU process. Finally we fit the model some real data series. We present a simulation study.

Keywords Barndorff-Nielsen and Shephard model · financial market · IG-Ornstein-Uhlenbeck process · Lévy processes

38.1 Introduction

Empirical studies have also shown that the volatility is not constant as postulated by famous Black-Scholes model. In reality the logarithmic return distribution has fatter than normal distribution implies. The characteristic properties of logarithmic returns are high kurtosis and negative skewness. These facts can not explain assumption by a constant volatility. Volatility has a stochastic structure. Therefore mean-reverting dynamics can be suitable candidate for the modeling of volatility. The stock market prices evolve freely but other a lot of real asset, price spreads are observed in the short time, but in the long time, the demand of product is adjusted and the prices move towards around the level of production cost of asset. The stochastic volatility models are driven by Lévy processes is introduced by [1, 2]. The Bates model is simpler but in this model jumps and stochastic volatility are independent. BNS model denotes a connection of jumps and stochastic volatility. A Brownian motion

Ö. Önalan (✉)
Faculty of Business Administration and Economics, Marmara University,
Bahçelievler Campus, 34590, Istanbul, Turkey
e-mail: omeronalan@marmara.edu.tr

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_38,

may be a good model for a particle movement. After a hit the particle does not stop after the changing position, but it moves continuously with decreasing speed. Brownian motion is not differentiable anywhere. Ornstein -Uhlenbeck process was proposed by Uhlenbeck and Ornstein (1930) to improvement the Brownian motion model. The paper is organized as follows. Section 38.2 reviews well known properties of Lévy process. In Section 38.3 we set up OU-processes. We explain estimators. In Section 38.4 we fit the model real data. Finally, the Section 38.5 includes conclusions.

38.2 Levy Processes

Definition (Lévy process). A Lévy processes is a continuous time stochastic processes $(L_t)_{t \geq 0}$ with,

1. $L_0 = 0$
2. Stationary increments. For all $s > 0, t > 0,\ L_{t+s} - L_t$ has the same distribution as L_s
3. Independent increments. For all $0 < t_0 < t_1 < \cdots < t_n = t, L_{t_i} - L_{t_{i-1}}; (i = 0, 1, 2, \ldots, n)$ are independent
4. Cadlag paths

Remark. In a Lévy processes discontinuous occurs at random times. The sample paths of a Lévy process are right continuous and left limits. Brownian motion $B_t \sim N\left(\mu t, \sigma^2 t\right)$ and Poisson process $N_t \sim Pois\,(\lambda t)$ for some density $\lambda \in (0, \infty)$ are levy process. The jumps of Lévy process $\Delta L_t = L_t - L_{t^-}$ are very important to understand structure levy processes. *Levy measure* v is a measure satisfying $v\{0\} = 0$ and

$$\int_{-\infty}^{\infty} \left(|x|^2 \wedge 1\right) v\,(dx) < \infty \tag{38.1}$$

For any Borel subset B of $R - \{0\}$.

$$v\,(B) = E\,[card\,\{t \in [0,1] : \Delta L_t \neq 0\,,\, \Delta L_t \in B\}] \tag{38.2}$$

$v\,(B)$ is expected number per unit time of jumps whose size belong to B. Thus $v\,(dx)$ is intensity of jumps of size x. Let Ψ (u) is the characteristic function of a random variable L, If for every positive integer n, there exist a random variable $L^{(1/n)}$ such that,

$$\Psi_L\,(u) = (\Psi_{L^{(1/n)}}\,(u))^n \tag{38.3}$$

We say that the distribution of L is infinitely divisible. Anyone can define any infinitely divisible distribution a stochastic process $L = (L_t)_{t \geq 0}$ called Lévy process [3, p. 44; 4, p. 8].

Theorem (Lévy Khintchine representation). *Let $(L_t)_{t \geq 0}$ be a Lévy process. The characteristics function $(L_t)_{t \geq 0}$ is of the form, $E\left\lfloor e^{iuL_t} \right\rfloor = e^{t\,\Psi(u)}$*

Where $\Psi(u)$ is cumulant of L_1 given by the Lévy- Khintchine formula,

$$\Psi(u) = ibu - \frac{\sigma^2 u^2}{2} + \int_{-\infty}^{\infty} \left(e^{iux} - 1 - iux 1_{\{|x| \le 1\}}\right) v(dx) \tag{38.4}$$

$\left(b, \sigma^2, v\right)$ is called Lévy triplet. The Lévy–Ito decomposition reveals the path structure of a Lévy process.

Theorem (Levy–Ito Decomposition). *Let $(L_t)_{t \ge 0}$ be a Lévy process and v its Lévy measure and verifies,*

$$\int_{|x| \le 1} |x|^2 v(dx) < \infty \text{ and } \int_{|x| > 1} v(dx) < \infty$$

$$L_t = bt + \sigma B_t + \lim_{\varepsilon \downarrow 0} \{L_t^{\varepsilon}\} \tag{38.5}$$

$$L_t^{\varepsilon} = \sum_{s \le t} \Delta L_s 1_{\{|\Delta L_s| > \varepsilon\}} - t \int_{\varepsilon < |x| \le 1} x v(dx) \tag{38.6}$$

The *subordinators* are special case of Lévy process. All subordinators are pure upward jumping process. It has non- decreasing sample paths (i.e. Poisson and IG Lévy processes are subordinators).

Definition (Self-decomposability). Let Ψ (u) be the characteristic function of a random variable. We call X self- decomposable if

$$\Psi(u) = \Psi(cu)\, \Psi_c(u) \tag{38.7}$$

For all $u \in R$ and all $c \in (0,1)$ and for some family of characteristic functions $\{\Psi_c : c \in (0,1)\}$ [3, p. 47]. Let $v(dx)$ denote the a Lévy measure of infinitely divisible measure P. The form of $v(dx)$ is $v(dx) = u(x)\,dx$ the such $|x|\,u(x)$ is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$ [3, p. 48]. Let $W(dx)$ denotes the Lévy measure of L_1. If the Lévy density u of the self-decomposable law D is differentiable then the Lévy measure W has a density w and u and w are related by

$$w(x) = -u(x) - x\,\frac{du(x)}{dx} \tag{38.8}$$

Theorem. *For any Lévy process $L = (L_t)_{t \ge 0}$ and for a function f, satisfying regularity conditions,*

$$\ln\left[E\left(\exp\left\{iu \int_{R+} f(s)\, dL_t\right\}\right)\right] = \int_{R+} \ln\left[E\left(\exp\{iuf(s)\, L_1\}\right)\right] ds \tag{38.9}$$

For proof, you can look [2, 5].

Theorem. *A random variable X is self-decomposable if and only if it there exist a Lévy process $L = (L_t)_{t\geq 0}$ such that X has representation of the form,*

$$X = \int_0^\infty e^{-s} dL_s \tag{38.10}$$

Let $v\,(dx)$ and $\mu\,(dx)$ are Lévy measures of respectively X and L. They are related by [5, p. 31].

$$v\,(dx) = \int_0^\infty \mu\,(e^s dx)\; ds \tag{38.11}$$

38.3 Ornstein-Uhlenbeck Processes

Ornstein-Uhlenbeck process was proposed by Uhlenbeck and Ornstein (1930) as an alternative to Brownian motion. This process was driven by a Brownian motion with drift that is a Lévy process.

OU Process driven Brownian Motion: A one dimensional Gaussian OU process $X = (X_t)_{t\geq 0}$ can be defined as the solution to the stochastic differential equation,

$$dY_t = -\lambda\, Y_t dt + \sigma\, dW_t \tag{38.12}$$

If X_t is the interest rate at time t and m is a reference value for the rate,

$$dX_t = -\lambda\,(X_t - m)\; dt + \sigma\, dW_t, \qquad X_0 = x_0 \tag{38.13}$$

with $\sigma > 0$ and $\lambda > 0$. Let $Y_t = X_t - m$. We get

$$dY_t = dX_t = -\lambda\, Y_t dt + \sigma\, dW_t \tag{38.14}$$

So, consequently, $d\left(e^{\lambda t} Y_t\right) = \lambda\, e^{\lambda t} dW_t$. Let $Z_t = e^{\lambda t} Y_t$ $(Z_0 = x_0 - m)$

We obtain $Z_t = (x_0 - m) + \int_0^t \sigma\, e^{\lambda s} dW_s$ so $X_t = Y_t + m = e^{-\lambda t} Z_t + m$

$$= e^{-\lambda t}\left((x_0 - m) + \int_0^t \sigma\, e^{\lambda s} dW_s\right) + m \tag{38.15}$$

X_t is unique strong markov solution of (3.2) [6, p. 298]. Finally we obtain that

$$X_t \sim N\left(m + e^{-\lambda t}\,(x_0 - m)\,,\; \left(\sigma^2/(2\lambda)\right)\left(1 - e^{2\lambda t}\right)\right)$$

This distribution as $t \to \infty$ to the stationary distribution $N\left(m, \sigma^2/2\lambda\right)$.

The probability distribution of X_t approach an equilibrium probability distribution called the stationary distribution. This stationary distribution has a stationary density function. For a time changed Brownian motion, another representation is here,

$$X_t = m + e^{-\lambda t}(x_0 - m) + \sigma\, e^{-\lambda t} W_{\left(e^{2\lambda t} - 1\right)/2\lambda} \tag{38.16}$$

$$E(X_t) = e^{-\lambda t} + m\left(1 - e^{-\lambda t}\right) \tag{38.17}$$

$$Var(X_t) = \left(\sigma^2/(2\lambda)\right)\left(1 - e^{2\lambda t}\right) \tag{38.18}$$

Theorem. *The correlation function of X_t is*

$$Corr(X_t, X_{t+k}) = \frac{e^{-\lambda t}\left(1 - e^{-2\lambda t}\right)}{\sqrt{\left(1 - e^{-2\lambda t}\right)\left(1 - e^{-2\lambda(t+k)}\right)}} \tag{38.19}$$

When $t \to \infty$ the correlation of X_t tents to,

$$\lim_{t\to\infty} Corr(X_t, X_{t+k}) = e^{-\lambda k} \tag{38.20}$$

OU Process driven General Lévy Processes: Let $L = (L_t)_{t\geq 0}$ is a time homogeneous Lévy process, for $\lambda > 0$, Ornstein-Uhlenbeck (OU) type process has

$$X_t = e^{-\lambda t} X_0 + \int_0^t e^{-\lambda(t-s)} dL_s = e^{-\lambda t} X_0 + e^{-\lambda t}\int_0^t e^{\lambda s} dL_s \tag{38.21}$$

It is unique strong solution below SDE,

$$dX_t = -\lambda X_t dt + \sigma\, dL_t, \quad X_0 = x_0 \tag{38.22}$$

Where λ denotes the decay rate. The λ enters to the stationary solution of OU process. This leads difficulties in solution of SDE. We can remove this difficulties by a simple change of time in the stochastic integrals [1, p. 75]. We can rewrite OU process as follows

$$X_t = e^{-\lambda t} X_0 + \int_0^t e^{-\lambda(t-s)} dL_{\lambda s} \tag{38.23}$$

If $Y = (Y_t)$ is an OU process with marginal law D, then we say that Y is a D – OU process. When given a one dimensional distribution D there is exist a stationary OU process whose marginal law is D if and only if D is self-decomposable [7]. We have a result that [1],

$$X_t = e^{-\lambda t} X_0 + e^{-\lambda t}\int_0^{\lambda t} e^{s} dL_s$$

Proposition. *For any $t, h > 0$, [6],*

$$e^{-\lambda t}\int\limits_{t}^{t+h} e^{\lambda s} dL_{\lambda s} = \int\limits_{0}^{h} e^{\lambda s} dL_{\lambda s}, \quad L^{*}(h) = \int\limits_{0}^{h} e^{\lambda s} dL_{\lambda s} = \int\limits_{0}^{\lambda h} L_s ds.$$

In the case of a $D - OU$ process, process L_t denotes the background driving Lévy process (BDLP). We can write following relation that between the characteristic functions of the BDLP Ψ_{L_1} and Ψ_X,

$$\ln\left[\Psi_{L_1}(u)\right] = u\left(d \ln\left(\Psi_X(u)\right)/du\right) \tag{38.24}$$

Let us denote by $k_D(u)$ the cumulant function of the self-decomposable law of D and $k_L(u)$ the cumulant function of the BDLP at time $t = 1$ i.e $k_L(u) = \log E\left\lfloor e^{-u\,L_1}\right\rfloor$. In other words $k(u) = k_L(u) = \log E\left\lfloor e^{-uL_1}\right\rfloor$ is the cumulant function of L_1 [1]. We can say the cumulant function of X_t can be directly found from the cumulant function of L_1. If v_L denotes the Lévy measure of L_1, we will assume that $\int_R x^2 v(dx) < \infty$ and we shall write $E(L_1) = \mu$ and $Var(L_1) = \sigma^2$ and we will assume that X_0 is independent of L and that [7, 8, p. 3], $X_0 = \int_0^{\infty} e^{-s} dL_s$ Parameter Estimation of Model: We will use the moment estimation methods to estimate the model parameters $\left(\mu, \sigma^2, \lambda\right)$. We will take discrete spaced observations. Let $L = (L_t)_{t \geq 0}$ is a levy process, then $E(X_0) = \mu$ and $Var(X_0) = \sigma^2/2$ (for detail [8, p. 4]. In this section our aim is match the theoretical moments and empirical moments. Theoretical auto covariance and auto correlation of X_t is given by

1. $\gamma(k) = Cov\left[X_{t+k}, X_t\right] = \dfrac{\sigma^2}{2} e^{-\lambda k}$
2. $\rho(k) = Corr\left(X_{t+k}, X_t\right) = e^{-\lambda k}$

We can be write autocorrelation and empirical moments of time series X

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$
2. $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
3. $\hat{\gamma}_n = \left[\hat{\gamma}(0), \hat{\gamma}(1), \ldots, \hat{\gamma}(d)\right], \quad k \in \{0, 1, 2, \ldots, d\}$

$$\hat{\gamma}(k) = \frac{1}{n}\sum_{i=1}^{n-k}\left(X_{i+k} - \bar{X}\right)\left(X_i - \bar{X}\right)$$

4. $\hat{\rho}_n = \left[\hat{\rho}(0), \hat{\rho}(1), \ldots, \hat{\rho}(d)\right], \quad k \in \{0, 1, \ldots, d\}$
   $\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\rho}(0)}$, Finally strongly consistent estimators of $\left(\mu, \sigma^2, \lambda\right)$ respectively [8, p. 7]

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \sigma_n^2 = \frac{2}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 \tag{38.25}$$

$$\hat{\lambda}_1 = -\log\left(\hat{\rho}\left(1\right)\right),\ \hat{\lambda}_2 = \min{}_\lambda \sum_{k=1}^{d}\left(\hat{\rho}\left(k\right) - e^{-\lambda k}\right)^2,\ \hat{\lambda}^* = \min\left\{0,\hat{\lambda}_2\right\}$$

We can use the as an estimator of λ [8, p. 7].

Likelihood Estimation for IG-OU Process: We can estimate the parameters of IG process using $x_0, x_1, \ldots, x_n$ observations of sample $X_0, X_h, \ldots, X_{nh}$ of X. *The initial estimates of a and b*

- $Y_k = \int_{\lambda(k-1)h}^{\lambda kh} e^s dL_s = e^{\lambda h} X_{kh} - X_{(k-1)h},\ k = 1, 2, \ldots, n$
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_k$ and $s_Y^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\left(Y_k - \bar{Y}\right)^2$

$$\hat{a} = \frac{\hat{b}\,\bar{Y}}{\left(e^{\lambda h} - 1\right)} \text{ and } \hat{b} = \frac{1}{s_Y}\sqrt{\frac{\bar{Y}\ \left(e^{2\lambda h} - 1\right)}{\left(e^{\lambda h} - 1\right)}} \tag{38.26}$$

The simulation of IG-OU Process: We simulate the paths of process by means of the Euler scheme,

$$X_{hk} = e^{-\lambda h} X_{(h(k-1))} + \sum_{k=1}^{100} e^{k\tau} L_\tau \tag{38.27}$$

$L = (L_t)$ is the corresponding BDLP of X and $\tau = 0.0005$ [9, p.103]. We use the IG random number generator proposed by [3, pp. 111–112].

- *Set* $Y = N^2$
- *Set* $X_1 = (a/b) + \left(Y/2b^2\right) - \left[\sqrt{4abY + Y^2}\Big/\left(2b^2\right)\right]$
- Generate a uniform random number U
- If $U \leq a/(a + X_1 b)$ then return $X = X_1$ else return $X = a^2\big/\left(b^2 X_1\right)$

Sample path of an IG process: We simulate a sample path of an IG process $L = \{L_t : t \geq 0\}$. The value of this process ay time points $\{nh : n = 0, 1, \ldots\}$ as follows; Simulate n, i.i.d IG random variable I_n I_n with parameter $IG\left(ah, b\right),\ L_0 = 0$ $L_{nh} = L_{(n-1)h} + I_n,\ n \geq 1$ (Table 38.1).

Table 38.1 Parameter estimation for return of GM

Parameter	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
Estimated val	0.000024	0.000906	3.985	0.02250

38.4 The Modelling of Data

We describe the stock price process,

$$S_t = S_0 \exp\left[\int_0^t \left(\mu + (\beta - (1/2))\, \sigma_t^2\right) ds + \int_0^t \sigma_s d W_s\right] \tag{38.28}$$

$$\int_{t-1}^{t} \sigma_s d W_s \;\cong\; \sigma_s \varepsilon,\ \varepsilon \sim N\,(0, 1) \tag{38.29}$$

Integrated volatility is defined as integral of the spot volatility $IV_t = \int_{t-1}^{t} \sigma_s^2 ds$. A non parametric measure for integrated volatility is realized volatility. The realized volatility can be estimated by the sum of intra daily squared returns as follows, $RV_t = \sum_{i=1}^{M} r_i^2,\ t = 1, 2, \ldots, T$ where M are the number of intra day observations. If we use the discrete version of price process,

$$r_t = \mu + \beta \sigma_t^2 + \sigma_t\, \varepsilon_t \tag{38.30}$$

where μ is the rate of return and β is the skewness parameter of the returns. r_t denotes the return process. If σ^2 has an inverse Gaussian distribution then, $r = \mu + \beta\sigma^2 + \sigma\varepsilon$ It has a normal inverse Gaussian distribution [10, pp. 280–281]. This r is an average of the continuous time volatility on one trading day. We assume that the volatility process is a constant times the number of trades (or volume) on each trading day.

$$\sigma_t^2 = \eta\, v_t \tag{38.31}$$

η is a constant and %95 confidence interval for η is $\left(2.065 \times 10^{-7},\ 2.414 \times 10^{-7}\right)$ We use $\eta = 2.2395 \times 10^{-7}$ constant value.

Application to Real Data: Our data set consist of General Motors stock prices from 1/2/1990 through 10/12/2007. The total number of observations is 4481. The parameters are calculated using with (3.15) formulas (Figs. 38.1–38.5).

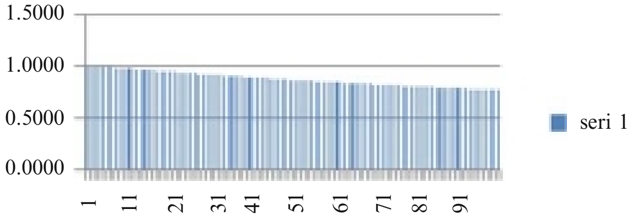


Fig. 38.1 Autocorrelation for closing prices of GM

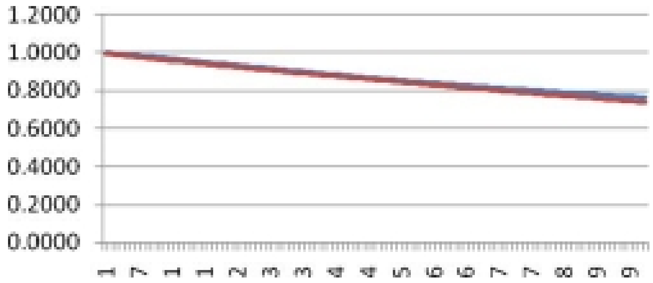


Fig. 38.2 True autocorrelation and estimated autocorrelation

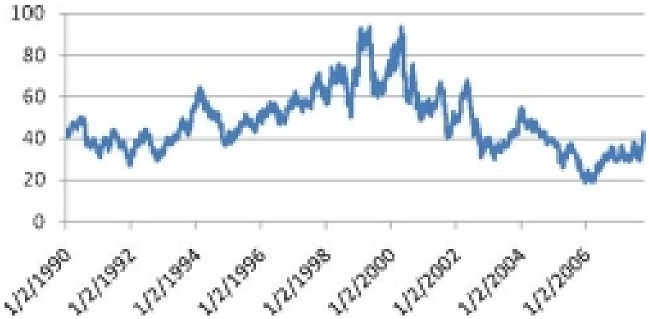


Fig. 38.3 Historical price data for GM

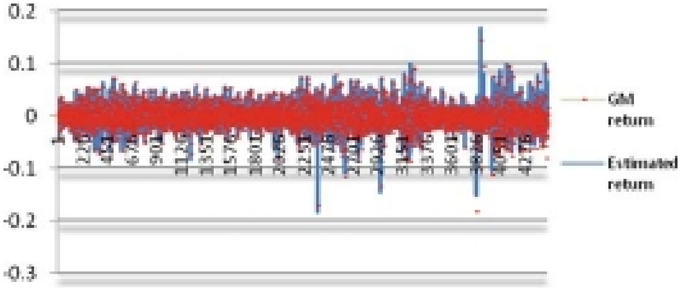


Fig. 38.4 GM returns and estimated returns

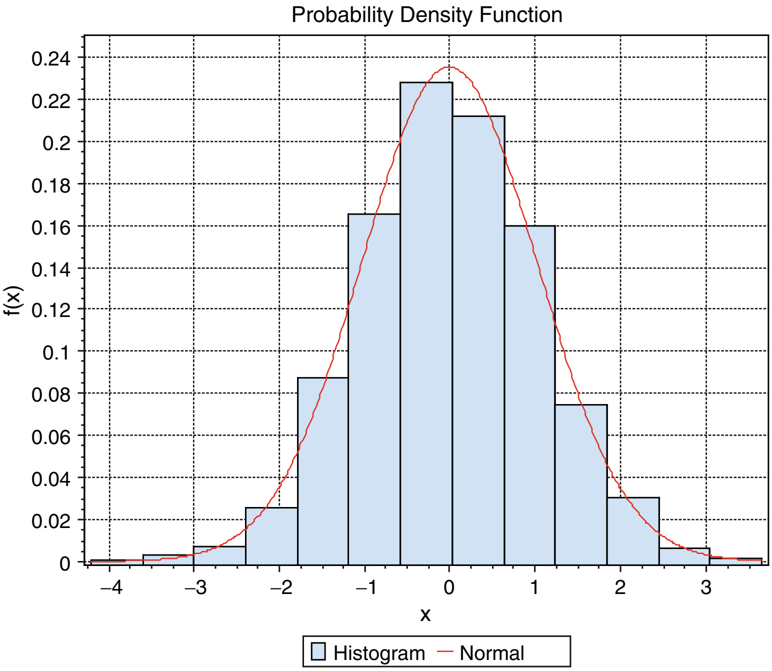


Fig. 38.5 Histogram of the squared residuals of GM actual returns and estimated returns

38.5 Conclusion

In this paper, we investigated an Ornstein-Uhlenbeck process driven by Lévy process for to model stock prices. We shown that we can be use the log returns and stochastic volatility at the same time in a model. The autocorrelation function of an Ornstein-Uhlenbeck process is decreasing as exponential. Exponential autocorrelation function approximates well empirical autocorrelation function of General Motors stock. This result represent that Ornstein-Uhlenbeck process can be a good fit model for describe real data. Furthermore trading intensity (volume) can be use a model for the volatility. Accurate model is important in mathematical finance and risk management.

References

1. Barndorff-Nielsen, O.E., Shephard, N.: Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion). J. Roy. Stat. Soc. Series B **63**, 167–241 (2001).
2. Barndorff-Nielsen, O.E., Shephard, N.: Financial volatility, Lévy processes and pover variation (2000). www.levyprocess.org.

3. Schoutens, W.: Lévy Processes in Finance: Pricing Financial Derivatives. Wiley, England (2003)
4. Papapantoleon, A.: An introduction to Lévy proceses with application to finance (2008). http://arxiv.org/PS_cache/arxiv/pdf/0804/0804.0482v2.pdf
5. Graf, F.: Exotic option pricing in stochastic volatility Lévy models and with fractional Brownian motion. Thesis, Universitat Konstanz (2007)
6. Protter, P.: Stochastic Integration and Differential Equations. Springer, Berlin (1990)
7. Sato, K.I.: Levy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge (1991)
8. Spiliopoulos, K.: Method of moments of Ornstein-Uhlenbeck processes driven by general Lévy processes (2008). http://www.math.umd.edu/∼kspiliop/
9. Valdivieso, L.: Parameter estimation for an IG_OU stochastic volatility model (2005). http://lstat.kuleuven.be/research/seminars_events/files/3afmd/valdivieso.PDF
10. Lindberg, C.: The estimation of the Barndorff-Nielsen and Shephard model from daily data based on measures of trading intensity. Appl. Stochast. Model. Business Indus. **24**, 277–289 (2008)

Chapter 39
Discrete Orthogonality of Zernike Functions and Its Application to Corneal Measurements

A. Soumelidis, Z. Fazekas, M. Pap, and F. Schipp

Abstract The optical aberrations of human eyes – as well as, those of other optical systems – are often characterized with the Zernike coefficients. Although, these coefficients are normally obtained from discrete measurement data via discrete computations, the developers and programmers of these computer programs could not rely on the discrete orthogonality of the Zernike functions despite the orthogonality of the continuous Zernike functions. Recently, meshes of points over the unit disk were found and reported that ensure this property. In the present paper, such meshes of points are used for computing Zernike coefficients in respect of cornea-like test surfaces. Test results are presented concerning the precision of the surface reconstruction from the aforementioned coefficients. The meshes proposed, however, are not exactly like what the developers hoped for. Further work is necessary in two respects, firstly, how to tune the conventional measurements so that the advantages of the proposed meshes can be exploited, secondly, how to design optical sensors that are based on such meshes.

Keywords Corneal topography · discrete orthogonal systems · Zernike functions

39.1 Introduction

The optical behaviour of the human eye is often characterized with Zernike coefficients. Before elaborating on how to compute these coefficients in an efficient manner utilizing the discrete orthogonality of the Zernike system reported recently,

A. Soumelidis (✉) and Z. Fazekas
Computer and Automation Reserch Institute, 13-17 Kende Street, Budapest, Hungary
e-mail: soumelidis@sztaki.hu; zfazekas@sztaki.hu

M. Pap
University of Pecs, 6 Ifjusag Road, Pecs, Hungary
e-mail: papm@ttk.pte.hu

F. Schipp
Eotvos Lorand University, 1/A Pazmany Walk, Budapest, Hungary
e-mail: schipp@ludens.elte.hu

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_39,

firstly the Zernike functions and their parametrizations are touched upon. Then the significance of Zernike functions in ophthalmology – concerning eye-modelling and measurements – is pointed out.

39.1.1 Zernike Functions

The orthogonal system of Zernike functions was introduced in [10] to model symmetries and aberrations of optical systems (e.g., telescopes). In Fig. 39.1, one of the Zernike functions – namely Y_1^2 – is shown as an example. (Note that in this figure – and in several others appearing in this paper – the function values are represented as gray tones.) In Fig. 39.2, the same function is shown from a different viewing angle. (Again, the surface is gray-toned.) A minute version of the previous view appears in the image-pyramid shown in Fig. 39.4. This figure and Fig. 39.3 represent a possible way of indexing (ordering) the Zernike functions. In Fig. 39.5, the

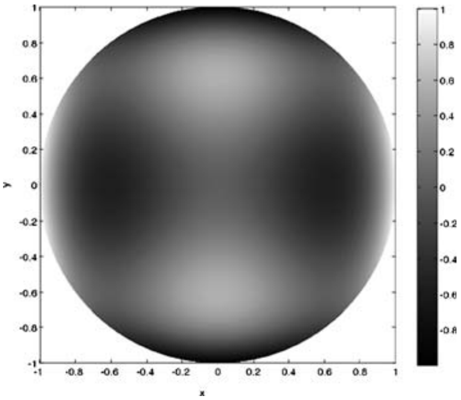


Fig. 39.1 An example of the Zernike functions, namely Y_1^2

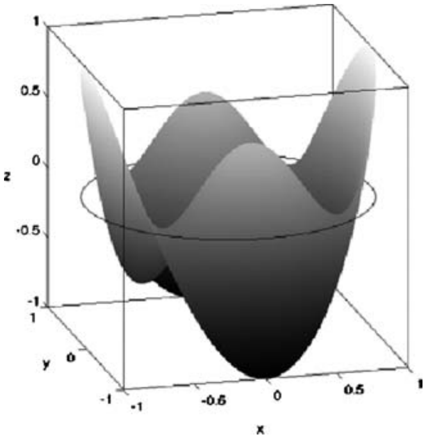


Fig. 39.2 The same Zernike function rendered from a different viewing angle

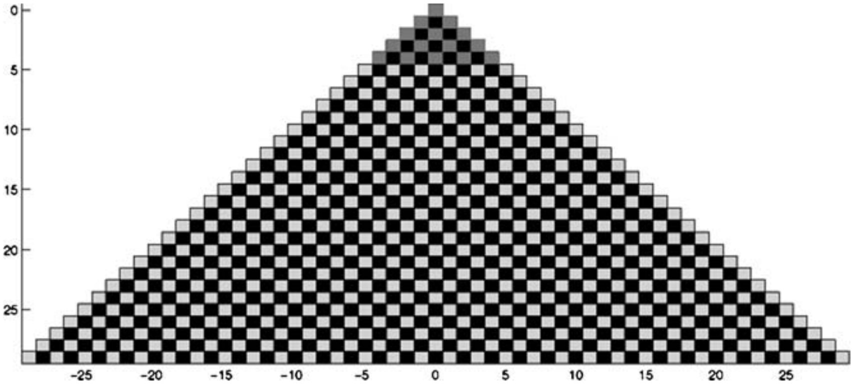


Fig. 39.3 A widely-used index-space of the Zernike function. The radial components of the functions appearing in a row are polynomials of the same degree. The *black squares* represent index-pairs that are not allowed. The index-pairs of those Zernike functions that have been included in Fig. 39.4 are marked with slightly darker *gray squares*

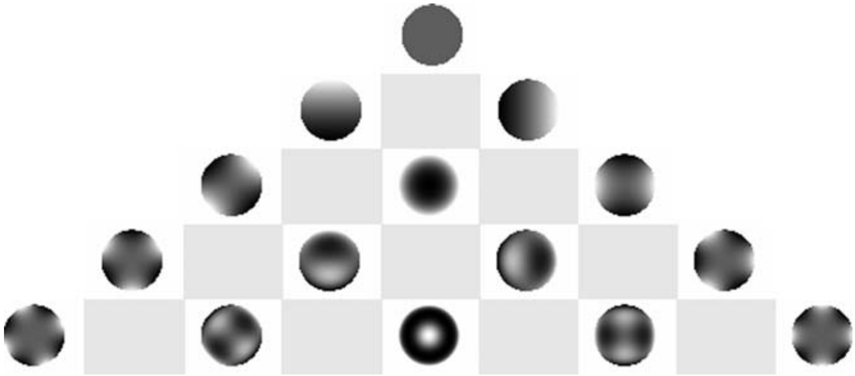


Fig. 39.4 A selection of the Zernike functions

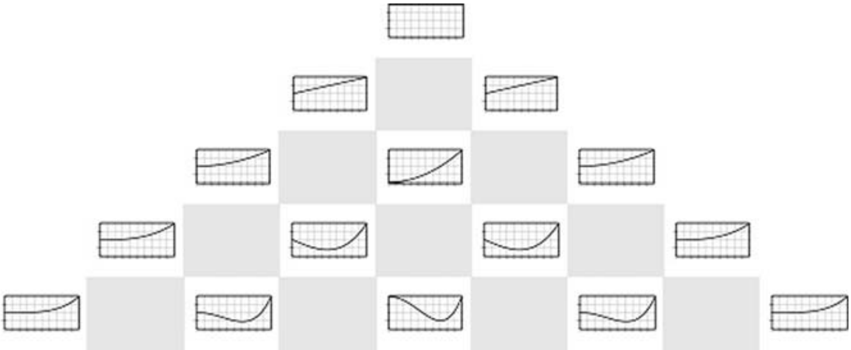


Fig. 39.5 The radial components (i.e., polynomials) of the Zernike functions shown in Fig. 39.4. The components on off-axis positions are repeated for the symmetric positions

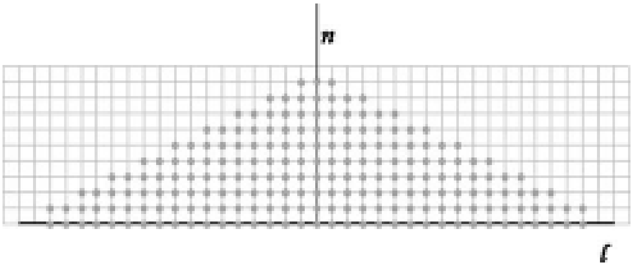


Fig. 39.6 The index-space used in this paper

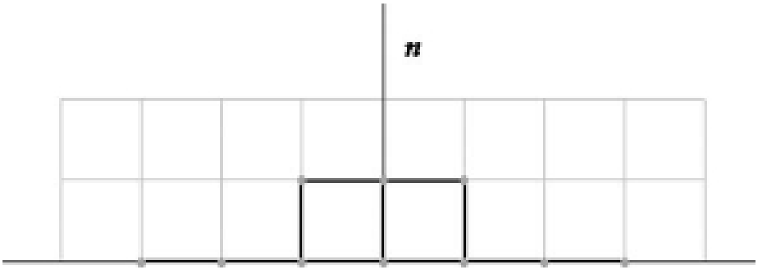


Fig. 39.7 A smaller version of the above index-space

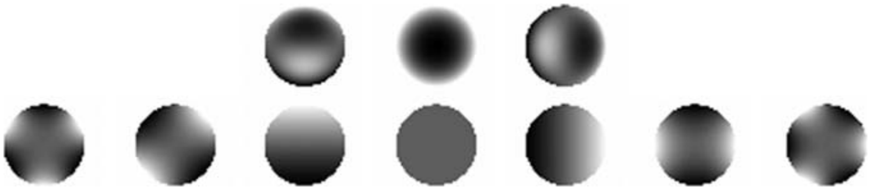


Fig. 39.8 Zernike functions associated with the indices shown in Fig. 39.7

corresponding radial components (polynomials) are presented. The indices of the functions actually included in Figs. 39.4 and 39.5 are marked with slightly darker gray fields in Fig. 39.3.

Another way of indexing of the Zernike functions is also used in this paper; it corresponds to the Y_n^l notation used above. This notation is defined in Section 39.2. The corresponding indexing is presented graphically in Figs. 39.6 and 39.8. The function shown in Fig. 39.1, for example, does not appear in this two-rowed image-pyramid, however, it would appear in an image-pyramid of this sort with more than two rows.

Some of the important and useful properties of the Zernike system are summarized in [5]. The mentioned paper can be used as a pointer to a wider range of relevant publications.

In Section 39.2, the continuous Zernike functions and their indexing used in this paper are presented.

39.1.2 *Zernike Functions in Ophthalmology*

Nowadays, the ophthalmologists are quite familiar with the smoothly waving Zernike-surfaces, such as shown in Fig. 39.2. They use these surfaces exactly in the way as was intended by Zernike, that is, to characterize various symmetries and aberrations of an optical system. In their case, those of human eyes, or more precisely, those of the corneal surfaces – examined with and computationally reconstructed by corneal topographer devices – and those of the eyeballs. In the latter case, Shack-Hartmann wavefront-sensors are used for the optical measurements.

The mentioned characterizations are given partly in the form of Zernike coefficients. As the optical aberrations may cause serious acuity problems, and are significant factors to be considered in planning of sight-correcting operations, wide range of statistical data – concerning the eyes of various groups of people – is available concerning the most important Zernike coefficients, see e.g., [6]. Another interesting use of the Zernike coefficients was reported recently. The optical aberrations of the eye make it difficult or in certain cases impossible to make high-resolution retinal images without compensating these aberrations. However, by properly compensating them, high-resolution retinal imaging can be achieved, as reported in [4].

39.1.2.1 Corneal Measurements and Modelling Corneal Surface

The purpose of a cornea topographic examination is to determine and display the shape and the optical power of the living cornea. Due to the high refractive power of the human cornea, the knowledge of its detailed topography is of great diagnostic importance. The corneal surface is often modelled as a spherical calotte (see Fig. 39.23), though, more complex surface models are also used for various purposes, including testing corneal topographers with non-spherical surfaces, see e.g., [7]. A spherical and two non-spherical test surfaces are shown in Figs. 39.15, 39.17 and 39.19.

The monocular cornea topographers evaluate the virtual image of some measurement pattern that is reflected and – after reflection – somewhat distorted by the corneal surface. Many of the reflection-based corneal topographers use a system of bright and dark concentric rings, called Placido disk, as measurement-pattern. Such a measurement pattern and its more sophisticated variant are shown in Figs. 39.9 and 39.10.

The measurement properties of the conventional Placido disk based topographers are rather problematic, as no point-to-point correspondences are available for the purpose of surface reconstruction [1]. On the other hand, the reflection-based monocular corneal topographers with more elaborate and more distinguishable measurement patterns are still popular, though they often rely on manual positioning and on some means to mark a surface-point.

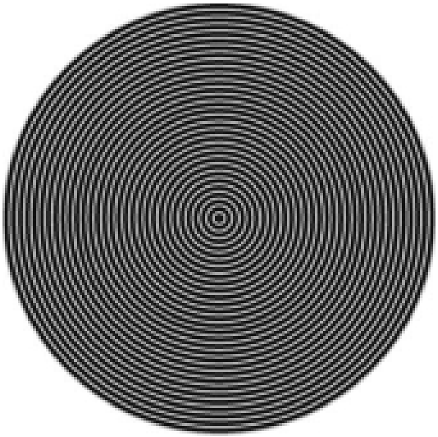


Fig. 39.9 A Placido disk

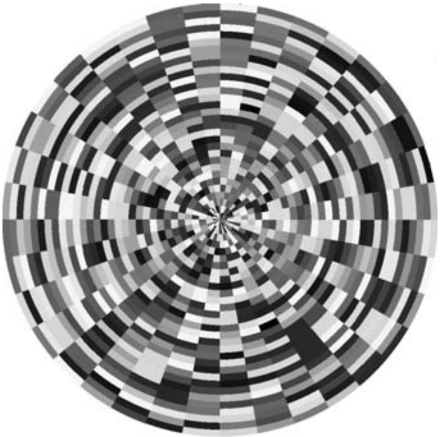


Fig. 39.10 Asophisticated version of the Placido disk

Recently, a multi-camera surface reconstruction method was suggested for the purpose of corneal topography in [2]. The reconstruction is achieved by solving the partial differential equations describing the specular reflections at the corneal surface. Though, the multi-camera corneal measurement approach is expected to be more precise and more robust than the monocular measurements, the monocular corneal topographers nevertheless will remain in use for some time. For this reason, it is worth noting that the discrete mesh of Fig. 39.11 could well be considered for the purpose of monocular corneal measurements.

39.1.2.2 Utilizing Discrete Orthogonality

Although, Zernike coefficients were obtained from measurements at discrete corneal points and via computations using some discretization of the continuous Zernike functions, the developers of these algorithms could not rely on the

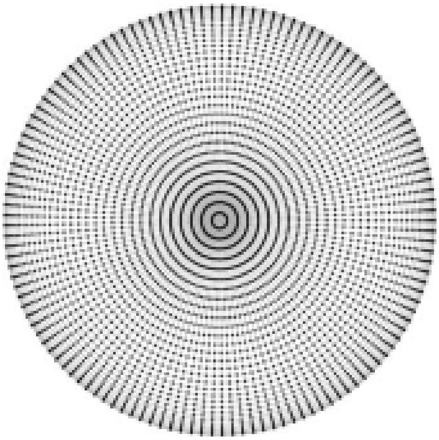


Fig. 39.11 An example of the X_N meshes of points, namely X_{32}

discrete orthogonality – see e.g., [3] – Zernike functions simply because no mesh of points ensuring discrete orthogonality was known. Not surprisingly, the discrete orthogonality of Zernike functions was a target of research for some decades, e.g., [9], and only recently was a mesh of points – ensuring discrete orthogonality of the Zernike functions – found and introduced in [5].

In the present paper, the mesh introduced there is used to calculate the Zernike coefficients for some artificial cornea-like surfaces. Tests were carried out on the precision of the discrete orthogonality obtained via the mentioned discretization, and on the precision of the surface reconstruction from the Zernike coefficients.

In Section 39.3, the discretization approach introduced in [5] is presented briefly. In Section 39.4, an efficient way of computing the Zernike coefficients is presented. Finally, we draw conclusions in Section 39.5.

39.2 Continuous Zernike Functions

A surface over the unit disk can be described by a two-variable function $g(x, y)$. The application of the polar-transform to variables x and y results in

$$x = \rho \cos \vartheta, \quad y = \rho \sin \vartheta, \tag{39.1}$$

where ρ and ϑ are the radial and the azimuthal variables, respectively, over the unit disk, i.e., where

$$0 \leq \rho \leq 1, \quad 0 \leq \vartheta \leq 2\pi. \tag{39.2}$$

Using ρ and ϑ, $g(x, y)$ can be transcribed in the following form:

$$G(\rho, \vartheta) := g(\rho \cos \vartheta, \rho \sin \vartheta). \tag{39.3}$$

The set of Zernike polynomials of degree less than 2N is as follows.

$$Y_n^l(\rho,\vartheta) := \sqrt{2n + |l| + 1} \cdot R_{|l|+2n}^{|l|}(\rho) \cdot e^{il\vartheta},$$
$$(l \in \mathbb{Z}, \quad n \in \mathbb{N}, \quad |l| + 2n < 2N) \tag{39.4}$$

The radial polynomials $R_{|l|+2n}^{|l|}$ can be expressed with the Jacobi polynomials $P_k^{\alpha,\beta}$ in the following manner:

$$R_{|l|+2n}^{|l|}(\rho) = \rho^{|l|} \cdot P_n^{0,|l|}(2\rho^2 - 1). \tag{39.5}$$

Some of these radial polynomials are shown in Fig. 39.5 and listed in Table 39.1. The argument transform used in (39.5) to generate the radial Zernike polynomials can be followed for R_8^0 in Fig. 39.12.

Table 39.1 The radial components – i.e., polynomials in radial variable ρ – of the Zernike functions shown in Fig. 39.8

		$3\rho^2 - 2\rho$	$2\rho^2 - 1$	$3\rho^2 - 2\rho$		
ρ^3	ρ^2	ρ	1	ρ	ρ^2	ρ^3

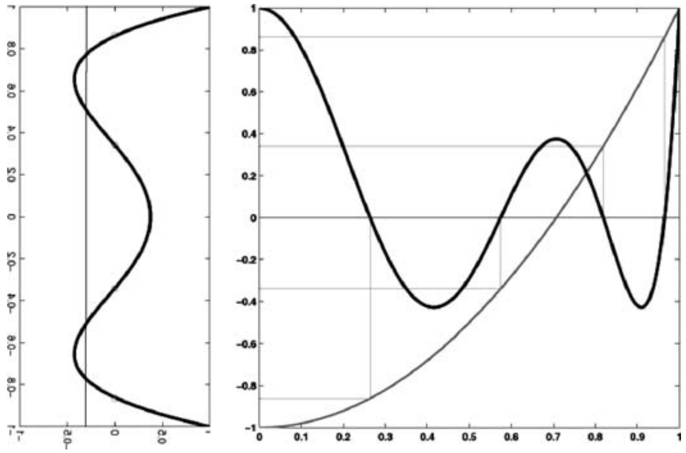


Fig. 39.12 R_8^0 (*right*) is derived – via the indicated argument transformation – from P_4 Legendre polynomial (*left*, *rotated*)

39.3 Discretization of Zernike Functions

The mesh, i.e., the set of nodal points – given in [5] and proven to ensure the discrete orthogonality of Zernike functions over this mesh – is as follows:

$$X_N := \{z_{jk} := \left(\rho_k^N, \frac{2\pi j}{4N+1}\right) : k = 1, \ldots, N, j = 0, \ldots, 4N\}, \tag{39.6}$$

where

$$\rho_k^N := \sqrt{\frac{1+\lambda_k^N}{2}}, \quad k = 1, \ldots, N. \tag{39.7}$$

In (39.7), λ_k^N is the k-th root ($k = 1, \ldots, N$) of the Legendre polynomial P_N of order N. In Fig. 39.13, mesh X_4, in Fig. 39.11 mesh X_{32}, while in Fig. 39.12 the Legendre polynomial P_4 of order 4 are shown as concrete examples. By using the discrete integral of (39.8), the discrete orthogonality of the Zernike functions was proven in [5]. The discrete orthogonality relation is given in (39.9).

$$\int_{X_N} f(\rho, \phi) d\nu_N := \sum_{k=1}^{N} \sum_{j=0}^{4N} f\left(\rho_k^N, \frac{2\pi j}{4N+1}\right) \frac{A_k^N}{2(4N+1)} \tag{39.8}$$

In (39.8), the A_k^N's are the weights that are associated with the discrete circular rings in the mesh. In Fig. 39.14, the weights A_k^4's are shown. These weights are derived for the radial Zernike polynomials from the quadrature formula of Legendre polynomials P_N of order N via the application of the argument transform used in (39.5) and shown in Fig. 39.12. For example, A_k^4's are equal to the corresponding

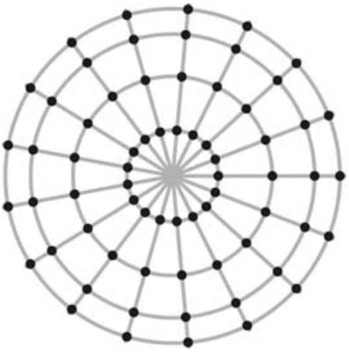


Fig. 39.13 An example of the X_N meshes of points, namely X_4. This mesh contains $N(4N+1) = 4 * 17 = 68$ nodal points

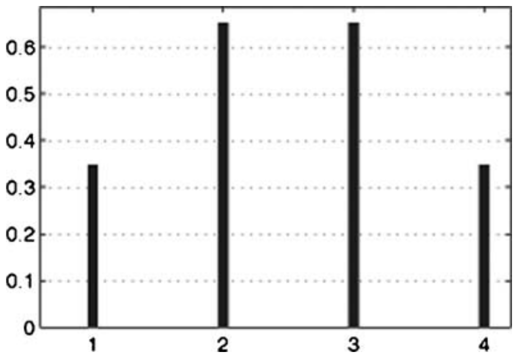


Fig. 39.14 The weights, i.e., the Christoffel numbers, that appear in the quadrature formula for Legendre polynomial P_4

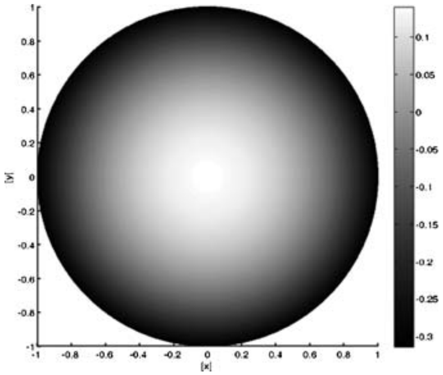


Fig. 39.15 A unit-hemisphere rendered using *gray tones* representing the z-coordinates of the surface points

weights – called Christoffel numbers – associated with the Legendre polynomial P_4 in the mentioned quadrature formula.

$$\int_X Y_n^m(\rho,\phi)\overline{Y_{n'}^{m'}(\rho,\phi)}d\nu_N = \delta_{nn'}\delta_{mm'}. \tag{39.9}$$

In the above orthogonality relation $n + n' + |m| < 2N$ and $n + n' + |m'| < 2N$ is assumed.

The quadrature formulas are significant tools in constructing discrete orthogonal systems. Quadrature formulas are known for some well-researched continuous orthogonal polynomials – of certain importance – of one variable since Gauss's time. See e.g., [8]. The quadrature formulas are expressed in the following way:

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{N} f(\lambda_k^N)A_k^N. \tag{39.10}$$

Interestingly, the integration of function $f(x)$ is much more precise than a numerical integration over some arbitrary (e.g., equidistant mesh). In our case, that is, for the discretization of the radial Zernike polynomials – i.e., the radial component of the Zernike functions – the N roots of Legendre polynomials P_N were used. The exact formula for deriving the A_k^N weights is not given here, for this see [5]. We just note here that the formula is exact for every polynomial f of order less than $2N$.

In Fig. 39.12, a particular Legendre polynomial, namely P_4 is shown. Its 4 roots fall in the interval $[-1, 1]$. The weights $A_1^4, \ldots, A_4^4$ – associated with these four roots – that should be used in the quadrature formula in (39.10) are shown in Fig. 39.14.

39.4 Computing the Discrete Zernike Coefficients

The discrete Zernike coefficients associated with function $T(\rho, \phi)$ can be calculated with the following discrete integral:

$$C_n^m = \frac{1}{\pi} \int_{X_N} T(\rho, \phi) \overline{Y_n^m(\rho, \phi)} d\nu_N. \tag{39.11}$$

It is worth noting that if $T(\rho, \phi)$ happens to be an arbitrary linear combination of Zernike functions of degree less than $2N$, then the above discrete integrals, i.e., for n's and m's satisfying the inequality $2n + |m| < 2N$, result in the exact Zernike coefficients, i.e., which are calculated from the corresponding continuous integrals.

A MATLAB implementation was created for computing the discrete Zernike representation of test surfaces. These test surfaces, such as shown in Figs. 39.15, 39.17, 39.19, 39.20 and 39.23, were selected from the ones suggested in [7].

The mentioned implementation, which generates the discrete Zernike representation, was used to create Figs. 39.21 and 39.22. These figures show the Zernike coefficients for the surfaces shown in Figs. 39.17 and 39.19. Note that the non-zero Zernike coefficients appearing in these figures are in agreement with the parity of the corresponding input surfaces (Fig. 39.19).

In Figs. 39.16, 39.18 and 39.22, the reconstruction errors – in case of using Zernike coefficients computed over X_{32} – are shown. (It is interesting to see characteristic differences between these error surfaces.)

We note here that the discrete integral in (39.11) was implemented via carrying out a circular inverse FFT for each of the discrete circles in the mesh. It is not surprising then to see in Fig. 39.24 that there is only one non-zero value (for each circle of the mesh) that appears in the inverse fast Fourier transform of the centrally positioned spherical calotte (shown in Fig. 39.23).

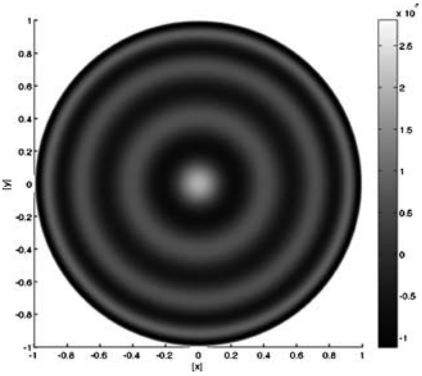


Fig. 39.16 Reconstruction errors – in the order of 10^{-7} – for the unit-hemisphere. The Zernike coefficients used in the reconstruction were computed via discrete integral over X_{32} mesh

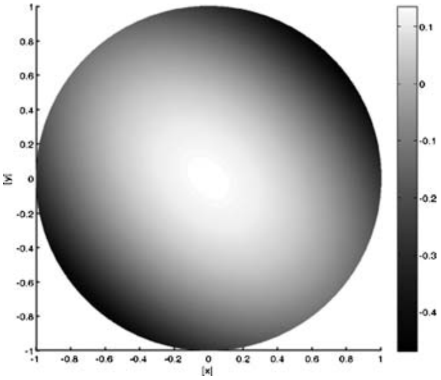


Fig. 39.17 A sphero-cylindrical surface

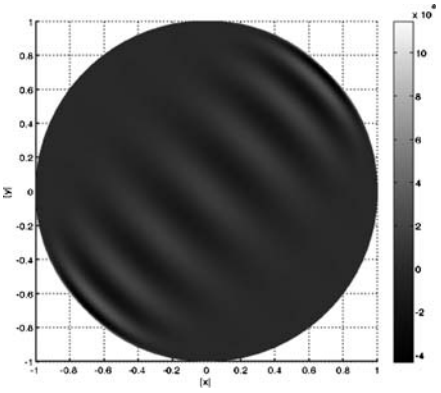


Fig. 39.18 Reconstruction errors for the sphero-cylindrical surface

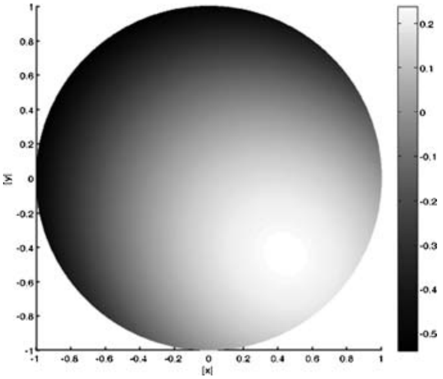


Fig. 39.19 A slant-sphero surface

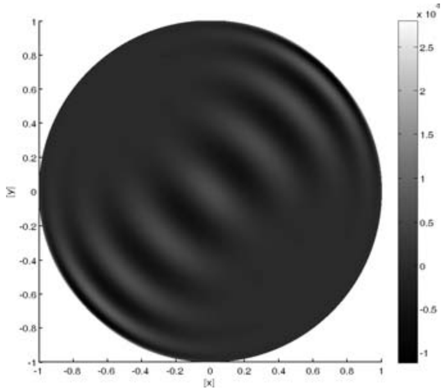


Fig. 39.20 Reconstruction errors for the slant-sphero surface

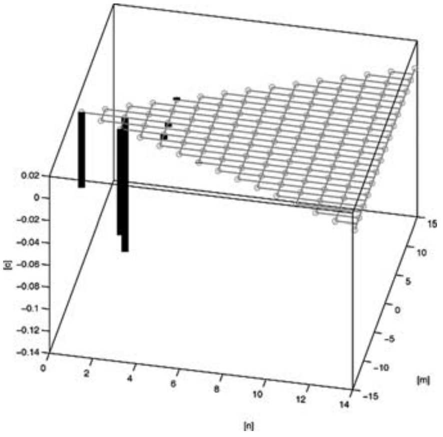


Fig. 39.21 The Zernike coefficients for the spherocylindrical surface shown in Fig. 39.17. Note that off-axis non-zero coefficients appear in the Zernike representation

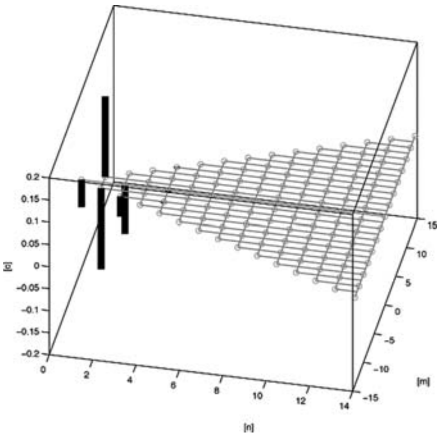


Fig. 39.22 Zernike coefficients for the slanted spherocylindrical surface shown in Fig. 39.19. Note the appearance of Zernike coefficients corresponding to the odd radial functions (see Fig. 39.5)

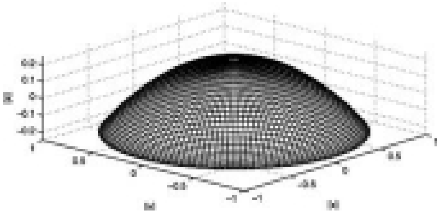


Fig. 39.23 The corneal surface is often modelled with spherical calotte. The parameters of the above calotte were chosen to make it "cornea-like"

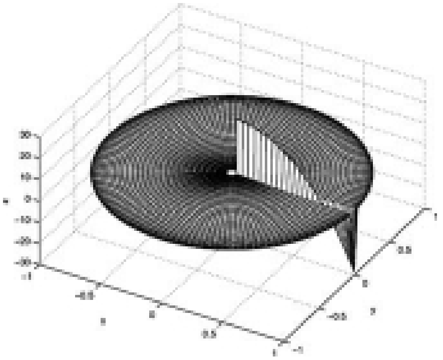


Fig. 39.24 The result of the circular inverse FFT – used in computing the Zernike coefficients – for the discrete circles in the proposed mesh

39.5 Conclusions and Future Work

The discretization used in this paper has relevance to the concrete application field, i.e., corneal measurements, but could also benefit physicists and engineers dealing with optical measurements, or with the design of optical measurement devices. Further work is required – concerning the mentioned meshes and the discrete orthogonality of the Zernike system – in two respects, firstly, how to tune – or make use of the data provided by – the conventional measurements so that the advantages of the proposed meshes can be exploited, secondly, how to design optical sensors that are based on these meshes.

References

1. Corbett, M.C., Rosen, E.S., O'Brart, D.P.S.: Corneal topography: principles and applications. Bmj Publ. Group, London (1999)
2. Fazekas, Z., Soumelidis, A., Bódis-Szomorú, A. Schipp, F.: Specular surface reconstruction for multi-camera corneal topographer arrangements. In: 30th Annual International Conference of IEEE EMBS, pp. 2254–2257, Vancouver, Canada (2008)
3. Iskander, D.R., Collins, M.J., Davis, B.: Optimal modeling of corneal surfaces with Zernike polynomials. IEEE Trans. Biomed. Eng. **48**(1), 87–95 (2001)
4. Ling, N., Zhang, Y., Rao, X., Wang, C., Hu, Y., W., Jiang., Jiang, C.: Adaptive optical system for retina imaging approaches clinic applications. In: Series Springer Proceedings in Physics, pp. 305–315. Springer, Berlin, Germany (2006)
5. Pap, M., Schipp, F.: Discrete orthogonality of Zernike-functions. Mathematica Pannonica **16**(1), 137–144 (2005)
6. Salmon, T.O., van de Pol, C.: Normal-eye Zernike coefficients and root-mean-square wavefront errors. J. Cataract Refr. Surg. **32**(12), 2064–2074 (2006)
7. Soumelidis, A., Csakany, B.: Specification of test corneal surfaces. (Project Report) CORNEA-INT-2M02. MTA-SZTAKI, Budapest, Hungary (2005)
8. Szego, G.: Orthogonal polynomials, 4th edn. AMS, New York (1981)
9. Wyant, J.C., Creath, K.: Basic wavefront aberration theory for optical metrology. In: Applied Optics and Optical Engineering, vol. 11. Academic, New York (1992)
10. Zernike, F.: Beugungstheorie des Schneidenverfahrans und Seiner Verbesserten Form, der Phasekontrastmethode. Physica **1**, 1137–1144 (1934)

Chapter 40
A New Scheme for Land Cover Classification in Aerial Images: Combining Extended Dependency Tree-HMM and Unsupervised Segmentation

Mohamed El Yazid Boudaren and Abdel Belaïd

Abstract An important challenge to any image pixels classification system is to correctly assign each pixel to its proper class without blurring edges delimiting neighboring regions.

In this paper, we present an aerial image mapping approach that advantageously combines unsupervised segmentation with a supervised Markov model based recognition. The originality of the proposed system carries on three concepts: the introduction of an auto-adaptive circular-like window size while applying our stochastic classification to preserve region edges, the extension of the Dependency Tree–HMM to permit the computation of likelihood probability on windows of different shapes and sizes and a mechanism that checks the coherence of the indexing by integrating both segmentations results: from unsupervised over segmentation, regions are assigned to the predominating class with a focus on inner region pixels. To validate our approach, we achieved experiments on real world high resolution aerial images. The obtained results outperform those obtained by supervised classification alone.

Keywords Land cover classification · hidden Markov model · aerial images

40.1 Introduction

Land Cover Classification (LCC) in high resolution aerial images is an important application of remote sensed data. It consists of identifying the natural objects present in a high resolution aerial image given a set of known patterns. In the most general case of aerial images, when the image contains several regions of different patterns,

M.E.Y. Boudaren (✉)
Applied Maths Lab., Military Polytechnic School, P.O. Box 17, Algiers, 16111, Algeria
e-mail: boudaren@gmail.com

A. Belaïd
LORIA Lab., Read Team, P.O. Box 239, Vandoeuvre-lès-Nancy, 54506, France
e-mail: abelaid@loria.fr

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_40,

the aim is to label each pixel with the corresponding texture. Evidently, the labeling process subsumes image segmentation but besides segmenting the image to different regions, it assigns each region to one of the natural objects patterns.

Achieving the classification at pixel level is a big issue in LCC problem. In fact, it is easier to identify an image of a relatively big size than identifying a lonely pixel. In fact, pixel-wise approaches for image classification are not usually suitable to solve problems often found in remote sensing application [1, 2]. They result in a disgusting salt and pepper effect.

Recent researches clearly show the advantages of integrating spatial dimension to spectral features by using segmentation based classification methods and, hence, focusing into image regions instead of pixels [3, 4].

More elaborated approaches use a family of Markov models to model the contextual interactions between labels. However, genuine 2D-Markov modeling of the contextual information is a time consuming iterative process [5].

On the other hand, reasonable complexity approaches identify each pixel by taking into account its neighboring pixels, usually by computing a similarity measure (likelihood probability for instance) on square windows centered at concerned pixels [6]. The drawbacks of such approaches are the following:

- They adopt square windows which may introduce a bias toward rectangular regions. Moreover, corner pixels are more distant than other pixels. Adopting non-square windows is usually unaffordable due to the used model or measure nature.
- The bigger is the window, the more likely the identification is correct. However, adopting a too big window may penalize small regions. A tradeoff is generally made.
- Since a static window size is adopted, the window size is then too small to perform efficient classification for all image pixels and too high to preserve edges since the classification of frontier pixels are biased through introduction of neighborhood pixels.

In this paper, we propose a system that overcomes the previous difficulties by introducing the following:

- Segmentation is achieved through unsupervised segmentation which preserve region edges even if it provides an over-segmented image.
- Each region is identified through stochastic supervised classification.
- Likelihood probability may be computed on windows of different sizes and shapes centered at considered pixels.
- To determine window size and shape, an auto-adaptive distance is computed based on the considered pixel position towards region edges.
- To permit likelihood probability computation on non-rectangular windows, we extended the Dependency Tree-hidden Markov model (DT-HMM) by allowing four directional dependencies instead of two, and adopting the central pixel as root instead of upper-left pixel when dealing with rectangular windows.

The reminder of the paper is organized as follows: in Section 40.2, we introduce our Extended Dependency Tree-HMM (EDT-HMM) that extends DT-HMM. Section 40.3 describes our indexing scheme. Section 40.4 shows experimental results. Conclusion and future works are given in Section 40.5.

40.2 Extended Dependency Tree-Hidden Markov Models

Markov models (Markov Random Fields, Hidden Markov Fields, Hidden Markov Models, Hidden Markov Trees ...) were extensively and successfully used for texture modeling and segmentation [7]. This is majorly due to their ability to model contextual dependencies and noise absorption [8]. However, their performance depends widely on the model architecture: genuine 2D-models yield better results but exhibits much higher computational complexity [8]. In general, the more complex is the model, the better are the performances.

Nevertheless, for computational complexity reasons, several approaches consider linear models like HMM even if such a model is not suited for two-dimensional data [10]. More elaborated approaches resort to 2D-models with simplifying assumption. One simplifying assumption that provides good results with a linear complexity is that assumed in DT-HMM [11, 12]: one site (pixel) may depend on either the horizontal or vertical predecessor, but not on both the same time.

The extension of DT-HMM in this work is motivated by two reasons: the need to compute likelihood probability on non-rectangular shaped windows of different sizes and the need to adopt central pixel (to be labeled) as the dependency tree root since the root shows more interactions with neighbors than other pixels do.

40.2.1 EDT-HMM Overview

Before describing our model principles, let us define the applicability conditions of the EDT-HMM model on a window w with respect to root r.

The window w must fit the following condition:

- For each site s of w, s must have at least one neighbor $v \in N_s$ that belongs to w and fulfills: $\|v, r\| < \|s, r\|$ where N is the 4-neighborhood and $\| \ \|$ is the Euclidean distance.

Let w be a window verifying the condition above, and let r be the center of the window and $Y_r = \{y_s/s(i, j) \in w\}$ be the set of features vectors (RGB for instance) of pixels inside $w \cdot Y_r$ is then the observable process. Let X be the hidden process. The likelihood probability is given by:

$$P(Y_r/\lambda) = \sum_{X} P(Y_r/X, \lambda)P(X/\lambda) \qquad (40.1)$$

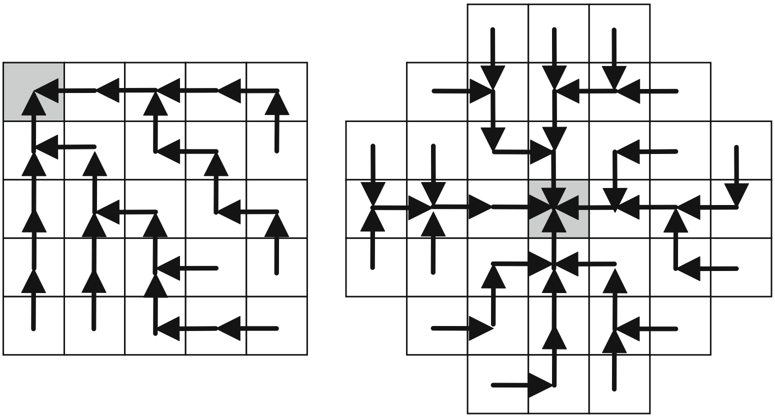


Fig. 40.1 Examples of random dependency trees according to DT-HMM (*left*) and EDT-HMM (*right*) formalisms

Unlike DT-HMM, where each pixel may have a predecessor chosen between two directions, in the EDT-HMM modeling, a pixel s may have a predecessor v chosen randomly from the 4-Neighborhood (up, down, right or left) and verifying the Euclidean distance property. Note that, the neighborhood directions of all pixels of w define a tree structure T like depicted in Fig. 40.1. We note $T(s) = v$.

The likelihood probability to observe Y_r given the parameters of the DT-HMM $\lambda(\pi, A, B)$ can be approximated as follows:

$$\begin{aligned} P(Y_r/\lambda) &\approx \sum_T P(Y_r/T, \lambda) \\ &\approx \sum_T \sum_X P(Y_r/X, \lambda) P(X/T, \lambda) \\ &\approx \sum_T \sum_X \left\{ \prod_{s \in w} P(y_s/x_s, \lambda) P(x_s/T, \lambda) \right\} \end{aligned} \tag{40.2}$$

In this paper, we propose to evaluate the likelihood on a set of random dependency trees τ. The previous equation becomes:

$$\begin{aligned} P(Y_r/\lambda) &\approx \sum_{T \in \tau} P(Y_r/T, \lambda) \\ &\approx \sum_{T \in \tau} \sum_X P(Y_r/X, \lambda) P(X/T, \lambda) \\ &\approx \sum_{T \in \tau} \sum_X \left\{ \prod_{s \in w} P(y_s/x_s, \lambda) P(x_s/T, \lambda) \right\} \end{aligned} \tag{40.3}$$

Thereafter, we remind the definition of the Model parameters π, A and B.

$$b_i(O) = P(y_s = O/x_s = i) \tag{40.4}$$

$$P(x_s = j/x_v = i, \lambda) = \begin{cases} \pi_j & \textit{if } s = r \\ a_{ij} & \textit{otherwise} \end{cases} \tag{40.5}$$

where $i, j = 1, ..N$ represent hidden states.

Note that a_{ij} only depends on i and j and not on the direction (horizontal or vertical).

To compute the likelihood probability of Eq. (40.3), we define the backward function $\beta_i\,(s)$ representing the probability of observing the data contained in the sub-tree of T with s as a root starting from the hidden state i.

$$\beta_i\,(s) = \begin{cases} b_i\,(y_s) & \text{if s is a leaf} \\ b_i\,(y_s) & \prod\limits_{T(v)=s} a_{ij}\beta_j\,(v) \ \text{otherwise} \end{cases} \tag{40.6}$$

Note that the likelihood probability of Eq. (40.3) can be evaluated as follows for each dependency tree T:

$$P(Y_r/T, \lambda) = \sum_{i=1}^{N} \pi_i \beta_i(r) \tag{40.7}$$

This computation exhibits a reasonable complexity (linear with window size).

The extension of the DT-HMM only concerns likelihood probability computation and Viterbi decoding whereas learning is performed the same way as in DT-HMM context.

The Viterbi decoding process can be achieved in a similar way to the likelihood probability computation. On the other hand, learning is performed via an iterative way the same as for the DT-HMM model, since the parameters are the same:

- Initialize model parameters
- Choose a random dependency tree T as described above (respecting the Euclidean distance constraint)
- Perform learning as in a linear framework (like in 1D-HMM)

40.3 Classification Scheme

To produce a class map of a given aerial image, we follow the scheme depicted in Fig. 40.2. In the following paragraphs, we describe each step.

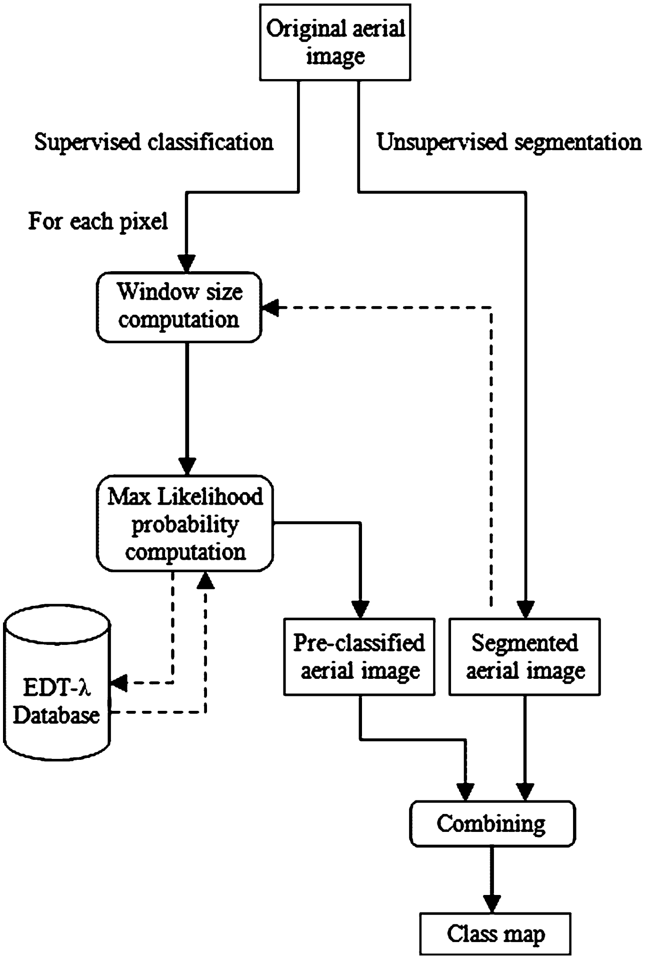

Original aerial image
Supervised classification
Unsupervised segmentation
For each pixel
Window size computation
Max Likelihood probability computation
EDT-λ Database
Pre-classified aerial image
Segmented aerial image
Combining
Class map

Fig. 40.2 Classification scheme

40.3.1 Image Unsupervised Segmentation

Before classifying the image pixels, we need to perform an image unsupervised segmentation that fits the following conditions:

- Image edges are preserved
- Pixels of the same region belong necessarily to the same natural object class, i.e. we may have an over-segmentation but not under-segmentation

This step serves as a pre-processing one, and will guide the rest of steps of the classification process.

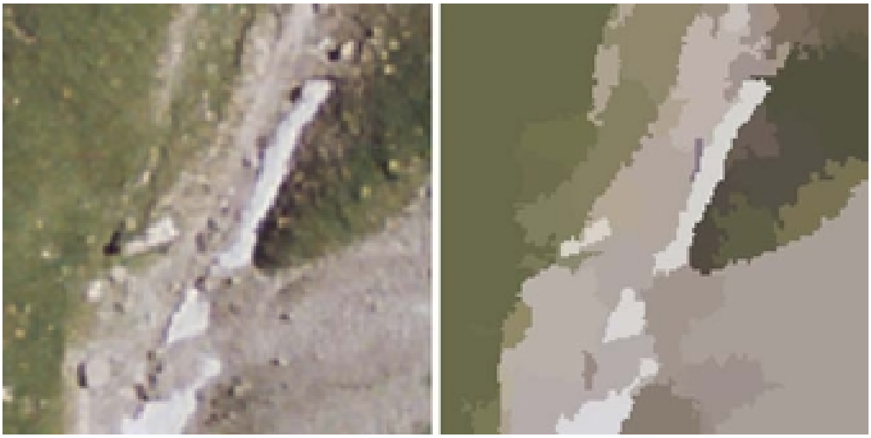


Fig. 40.3 A sample of high resolution aerial image (RGD 73–74, 2008) (*left*) and its corresponding unsupervised segmentation (*right*)

One unsupervised segmentation that has been shown to provide good results is the one produced by the EDISON system [13, 14] which we use in this work. A sample of a high resolution aerial image (50 cm per pixel) and the corresponding unsupervised segmentation via the EDISON system are provided in Fig. 40.3.

40.3.2 Window Size Computation

Since texture is not a local phenomenon, in order to classify a pixel s, we consider it with its neighborhood. More explicitly, we will compute the likelihood of the data inside a window centered at the pixel under consideration.

Let us denote w_s such a window and Y_s the data associated to that window. The class λ_s of the central pixel s is the class that maximizes the likelihood probability:

$$\lambda_s{}^* = \underset{\lambda \in \Lambda}{\arg\max}\ P\,(Y_s/\lambda) \tag{40.8}$$

Most approaches adopt a square window of a fixed size for all pixels. A trade-off is usually made so that the window is enough big to correctly classify the central pixel and enough small to preserve the region edges.

In this work, we propose to dynamically compute the window size to allow our system to deal with a maximum amount of information without distorting region edges. The more the pixel to classify is far from the region boundary, the larger is the window whereas edge pixels are classified without considering their neighborhood.

Hence, window size is chosen so that pixels within the window belong to the same region according to an unsupervised over-segmentation of the image.

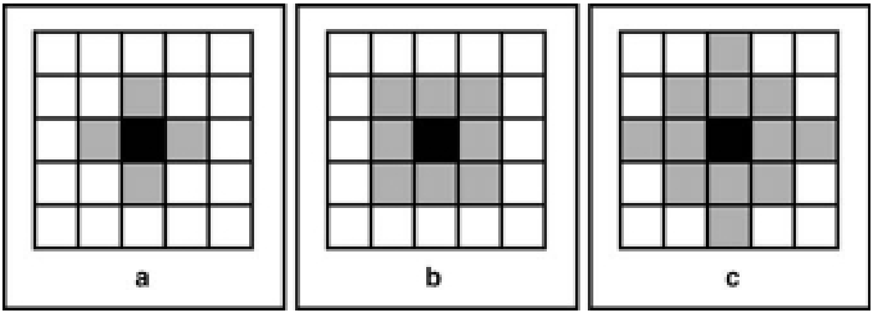


Fig. 40.4 Window shape and size for different values of Ray_s. a-$Ray_s{}^2 = 1$, b-$Ray_s{}^2 = 2$, c-$Ray_s{}^2 = 4$

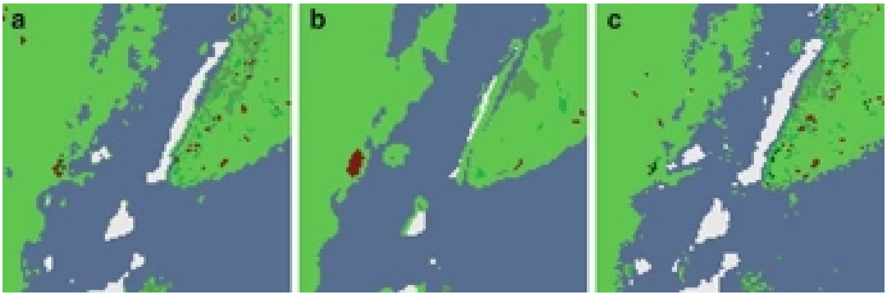


Fig. 40.5 Impact of window size on the pre-classification accuracy. a-$Ray_s{}^2 = 1$, b-$Ray_s{}^2 = 10$, c-auto-adaptative Ray_s

Window shape and size depend on a unique parameter Ray_s that represents the maximum Euclidean distance between neighbors and central pixel s. Figure 40.4 shows samples of windows of different shapes and sizes.

This parameter is obtained from the pre-segmented image. Its value is the maximum value possible so that pixels within the window belong to the same region. A comparative analysis of the window size impact on the accuracy of classification of the aerial image of Fig. 40.3 is shown in Fig. 40.5.

40.3.3 Image Pre-classification

To assign a pixel to a class, we compute the likelihood probability of observing the window centered on that pixel according to the EDT-HMM of each natural object class. The pixel is then allocated to the class that maximizes this probability (Fig. 40.5).

The parameters of the EDT-HMM corresponding to each natural object class are obtained after a learning process achieved on mono-class aerial images. To represent each pixel, we used the classical RGB color space. To estimate the parameters of the DT-HMM of each class, we achieve K-Means clustering on pixels of mono-class image of the corresponding class to divide the image pixels on N sub-classes. Subsequently, we obtain the parameters of N Gaussian functions. These parameters serve as an initialization of our EDT-HMM. The final parameters of the model are then obtained after an iterative process as described in the previous section.

40.3.4 Classification Correction

After the previous steps, the resulted class map suffers from the so called salt and pepper phenomenon. This is majorly to the difficulty to distinguish between several similar textures, especially for pixels near boundaries. To overcome this involvedness we propose to merge pixels of the same region (according to the unsupervised segmentation) into the same natural object class with a focus on inner pixels of the region, since those pixels were classified considering larger windows (Fig. 40.6). Explicitly, each region R is assigned the natural class that fits the following rule:

$$X_R{}^* = \underset{X \in \Lambda}{\arg\max} \sum_{\lambda_s = X,\, s \in R} size \ \ (w_s)$$

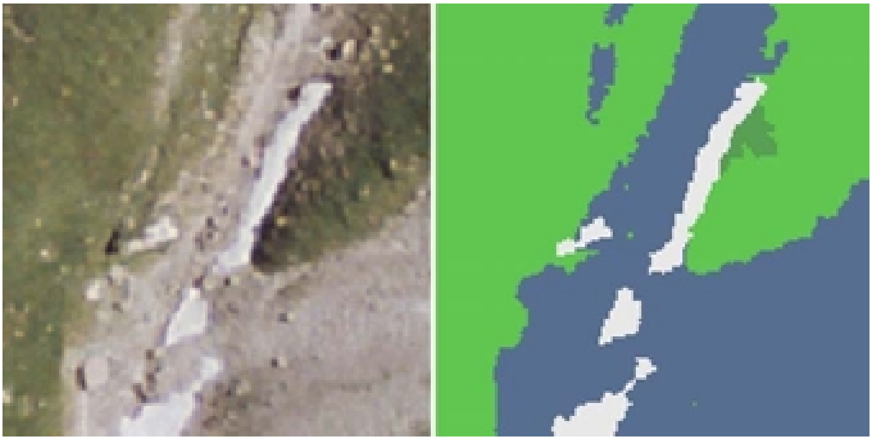


Fig. 40.6 Image classification correction Original aerial image (*left*), class map (*right*)

40.4 Experimentation

40.4.1 Data Overview

For our experimentation, we consider real world aerial images with a resolution of 50 cm per pixel (Fig. 40.6). The images were provided by La Régie de Gestion des Données des pays de Savoie, France ([9]).

The pictures were taken in relatively good light conditions; however, some images suffer from presence of shadow in some parts.

40.4.2 Learning Database

Learning was performed on mono-class images. These images were carefully extracted from the aerial images of the same area of study (Fig. 40.7).

40.4.3 Mono-class Images Generation

To demonstrate the capacity of the DT-HMM to represent natural object textures, we generate mono-class images using the corresponding DT-HMM and compare them to images generated by 1D-HMM and GMM (Fig. 40.8).

40.4.4 Experimental Results

To evaluate the robustness of our aerial images pixels' classification system, we considered three types of test images:

- Mono-class images, for which the classifier is expected to assign all pixels to the corresponding class (Fig. 40.9)

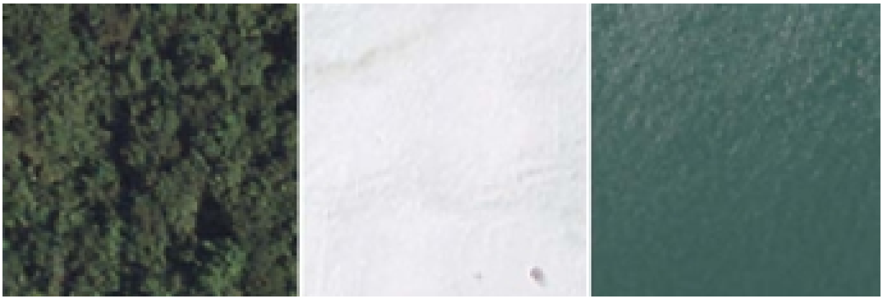


Fig. 40.7 Samples of learning images of classes: tree (*left*), snow (*middle*) and water (*right*)

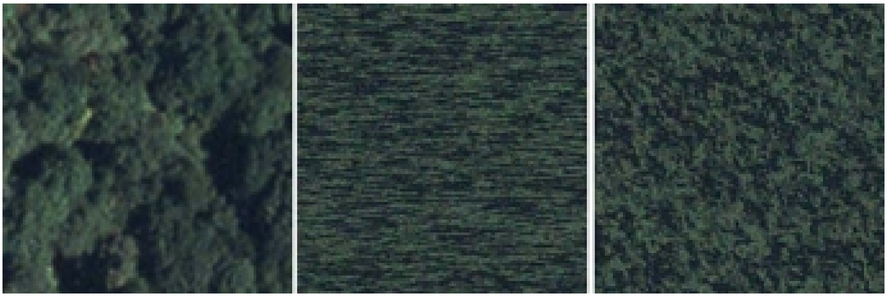


Fig. 40.8 Mono-class image generation of class Tree using: 1D-HMM (*middle*) and DT-HMM (*right*)

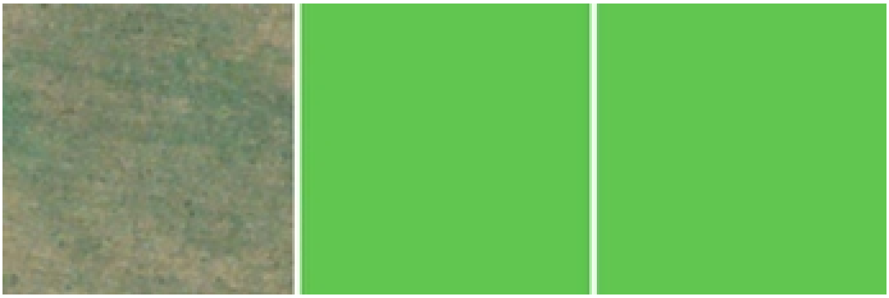


Fig. 40.9 Mono-class image classification. Original aerial image (*left*), ground truth map (*middle*) and class map (*right*)

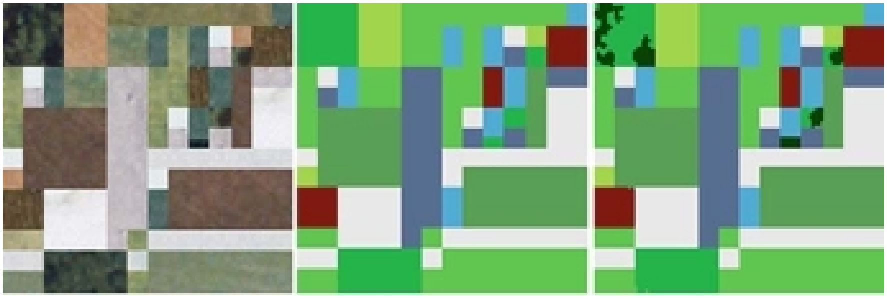


Fig. 40.10 Classification result on mosaic image. Original aerial image (*left*), ground truth map (*middle*) and class map (*right*)

- Mosaic images, assembled by combining different classes into regular boxes so that we can easily produce a corresponding ground truth map (Fig. 40.10)
- Natural aerial images, for which we don't have a precise ground truth map. Thus, only a visual evaluation can be achieved in this case (Fig. 40.6)

To produce the unsupervised segmentation of areal images, we acknowledge the use of EDISON system software [13, 14]. The obtained classification of the mosaic image is very similar to the corresponding ground truth. Notice that the shadow class is not included in the ground truth map since we know to which class the shadowy pixels belong.

40.5 Conclusion

In this paper, we proposed an approach that advantageously combines supervised EDT-HMM modeling and unsupervised segmentation to classify land cover pixels. Instead of achieving our classification using a static window size, we resorted to an auto-adaptive window size depending on the position of pixel under consideration towards region boundaries.

Overall, the experimental results show that our system produces satisfactory class maps in a reasonable time given the linear complexity of the modeling. Note that several textures are so similar to each other that it is sometimes very difficult even for a human to distinguish between them.

As future work, we propose to apply EDT-HMM modeling to other kinds of problems.

References

1. Mesev, V.: Remotely Sensed Cities. Taylor & Francis, London (2003)
2. Thomas, N.C., Congalton, R.: A comparison of urban mapping methods using high-resolution digital imagery. Photogrammet. Eng. Remote Sens. **69**(9), 963–972 (2003)
3. De Jong, S.M., Freek, D.M.: Remote Sensing ImageAnalysis: Including the Spatial Domain. Springer, Berlin (2006)
4. Jensen, J.: Introductory Digital Image Processing. Prentice-Hall, New York (2006)
5. Levin, E., Pieraccini, R.: Dynamic planar warping for optical character recognition. IEEE International Conference on Acoustics, Speech and Signal Processing **3**, 149–152 (1992)
6. Permuter, H., Francos, J., Jermyn, I.H.: Gaussian mixture models for texture and colour for image database retrieval. IEEE International Conference on Acoustics, Speech and Signal Processing **3**, 569–572 (2003)
7. Noda, H., Mahdad Shirazi, N., Kawaguchi, E.: MRF based texture segmentation using wavelet decomposed images. Pattern Recog. **35**, 771–782 (2002)
8. Pieczynski, W.: Markov models in image processing. Traitement de Signal **20**(3), 255–277 (2003)
9. RGD73-74, 2008, Régie de Gestion des Données des Deux Savoies. http://www.rgd73-74.fr
10. Boudaren, M.Y., Labed, A., Boulfekhar, A.A., Amara, Y.: Hidden Markov model based classification of natural objects in aerial pictures. IAENG International Conference on Signal and Image Engineering, pp. 718–721, London, 2–4 July 2008
11. Merialdo, B.: Dependency Tree Hidden Markov Models. Research Report RR-05-128, Institut Eurecom (2005)
12. Merialdo, B., Jiten, J., Galmar, E., Huet, B.: A new approach to probabilistic image modeling with multidimensional hidden Markov models. Adap. Multimed. Retriev. 95–107 (2006)
13. Comaniciu, D., Meer, P.: Mean shift: A robust approach towards feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
14. Meer, P., Georgescu, B.: Edge detection with embedded confidence. IEEE Trans. Pattern Anal. Mach. Intell. **23**(12), 1351–1365 (2001)

Chapter 41
Applying View Models in SOA: A Case Study

Anca D. Ionita and Monica Florea

Abstract This chapter analyzes LD-CAST – a business cooperation framework using business processes and semantic technologies for supporting local development by means of transnational activities. The system, based on Service Oriented Architecture and developed for a European STREP project, had multiple stakeholders, involving the necessity to describe various views, on various levels of abstraction. Both classical view models and new ones, specific for services, are used here for completing the description.

Keywords Software architecture · service oriented architecture · view models · e-business

41.1 Introduction

An important trend in developing distributed systems is the adoption of Service Oriented Architecture (SOA) such as to be able to involve multiple organizations, but at the same time to preserve their independency from the point of view of adopted platforms and implementation languages [1]. Services can also wrap existent legacy systems, allowing a seamless migration to SOA and the integration into existent frameworks, which can be extended by publishing new services.

Service engineering involves cooperation of multiple tools and actors, not only for the front-end business specific functionality, but also for the back-end architecture – based on components for: managing processes and ontologies, maintaining

A.D. Ionita (✉)
Computers and Industrial Informatics Department, University “Politehnica” of Bucharest,
Spl. Independentei 313, 060042, Bucharest, Romania
e-mail: Anca.Ionita@mag.pub.ro

M. Florea
SIVECO ROMANIA SA, Victoria Park, Sos. Bucuresti-Ploiesti 73–81, 013685,
Bucharest, Romania
e-mail: Monica.Florea@siveco.ro

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_41,

smart service repositories, automatically discovering services and supporting security. Thus, these systems become very complex and involve various stakeholders who collaborate for its development and maintenance; consequently, the architecture representation must be done both at a high and at a low level of abstraction, showing more or less details. For instance, the system integrator needs information related to all the subsystems interfaces, while a project manager only has to know the system partitioning and the dependencies, in order to schedule the tasks. They all need descriptions of the system from the points of view which are appropriate to them, such as to be able to specify, develop or evaluate the software, as well as to manage and finance the correspondent projects.

Viewpoint modeling has been used for various purposes: to specify requirements [2], to document architecture [3] and, generally, to help development [4] of both physically and software-intensive systems. This chapter presents a brief overview of existent view models, followed by a case study where some of these approaches are applied for describing a business collaboration framework developed in LD-CAST project [5]. The system focuses on the cooperation and integration of public institutions, for rendering accessible to Small and Medium Enterprises (SMEs) a multitude of trans-border services and applications for business development, in the context of the enlarged Europe.

41.2 View Models

A key issue in software development is to define architecture properly, such as to be understood by all the involved stakeholders and to grasp all the concerns covered by the system. Due to the application complexity, this is only possible with multiple models/diagrams – as we currently see when using the standard Unified Modeling Language [6] (UML). These models correspond to the description of the system from various points of view of various stakeholders interested in the development; so, one creates multiple views of the system, having an important role in its description and documentation. Moreover, for assuring the consistency, there are elements that can be traced from one model to another, showing a tight correspondence between types of diagrams inside or between views.

An early application of views appeared for data modeling, as the ANSI-SPARC Three-level Architecture, defining views that describe the part of the database that is relevant to a particular user [7]. A classical and well-known approach for software-intensive systems is the "4 + 1" View Model of Philippe Kruchten [8], which is generic and not bound to a certain notation, but can be applied, for instance, with UML models [9]. Another model, which is generally dedicated for documenting software architecture, was conceived by Software Engineering Institute (SEI) [3].

For describing software, there is also a standard, called "Recommended Practice for Architectural Description of Software-intensive Systems", which was first adopted as IEEE Std 1471–2000 [10], and then as ISO/IEC 42010:2007 [11]. It defines view-related concepts and their relationships; the architectural description is organized by views, consisting of models. It also identifies the system stakeholders, having various concerns. In order to create views, the description first selects the viewpoints of interest, for each of them specifying the addressed stakeholders and the covered concerns. For each viewpoint, one establishes methods for creating models, thus giving the possibility to realize the system views.

There are also view models dedicated to more specific architectures, like the Reference Model for Open Distributed Processing (RM-ODP) adopted by ISO [4]. Regarding SOA, some approaches try to capture specific elements related to business analysis, business processes, service definition and discovery, quality attributes, taking into account all the potential stakeholders; examples of view models dedicated to SOA systems are:

- Service Views [12] – for modeling SOA in the enterprise
- BDC [13] – composed of three specific views: Business Analysis, Composition and Design
- View-based Modeling Framework (VbMF) [14] – conceived for the integration of process-driven SOA with meta-modeling
- The Reference Architecture for SOA from OASIS [15] – a viewpoint model with three views: Business via Services, Realizing SOA and Owning SOA, each of them containing several models defined with UML diagrams

For the purpose of our study, the view models were compared regarding the scope of their application domains and their stakeholders; a summary is given in Table 41.1.

Table 41.1 A comparison of views models based on their scope and stakeholders

View model	Views/view points	Models	Scope	Target stakeholders
ANSI/SPARC 3 level architecture	3	Multiple	Data modeling	Users, integrators, developers
“4 + 1”	5	Not fixed, possible UML	Software-intensive systems	Users, programmers, integrators, engineers
SEI model	≤12/3	–	Software architecture	Thirteen stakeholders defined
RM-ODP	/5	A language per viewpoint, no notation	Open distributed processing	System developers
Service views	9/3	–	SOA	Nine categories
BDC	3	6	SOA	Analysts, designers, developers
VbMF	3 + specific extensions	Any process model	Process-driven SOA	Reverse engineers
OASIS RA for SOA	11/3	UML diagrams	SOA	Fourteen stakeholders

A survey of software architecture viewpoint models, focusing on the coverage of all the domain elements, related to structures, stakeholders and concerns, was presented by Nicholas May [4].

41.3 Case Study for LD-Cast

Nowadays, the delivery of business services may involve providers from multiple countries all around Europe; SMEs may easier develop their business if they have access to trustful systems that make accessible a multitude of trans-border services and applications for business development. This implies various problems related to the non-homogeneity of national terminologies, processes and technologies. The LD-CAST system, dedicated to local development cooperation actions [5], is designed for overcoming such problems, by introducing a modeling platform that integrates the different services and drives the execution of composite business services. The system uses Web services, whose selection is based on semantic search and discovery and on client preferences related to quality criteria like time and cost.

The subchapters below describe the system using various view models, starting with the classical "4 + 1" View Model (see Section 41.3.1) continuing with some new models specially defined for SOA: Service Views and BDC (see Section 41.3.2) and ending with some useful elements that may be extracted from other, more general, view models: RM-ODP and SEI Model (see Section 41.3.3).

41.3.1 Applying the "4 + 1" View Model

The "4 + 1" View model is composed of: the *Logical View* (representing structural elements, levels of abstraction and separation of concerns); the *Process View* (treating interaction patterns and concurrency); the *Development View* (showing the decomposition in subsystems used in the development environment); the *Physical View* (outlining the distribution of software components on physical nodes, either computers or processors); the *Scenarios* (describing use cases and considered the +1 view, because they make the connections between the other four views). Figure 41.1 presents a representation of this model, applied for the LD-CAST system.

The main use cases supported by LD-CAST are:

- Business Service Information – delivering information related to the business services; it may be performed by any Guest
- Business Service Management – requesting and monitoring business services; it is possible for the registered End-Users (see BS Management in Fig. 41.1)
- Business Process Management – allowing modeling and publishing of specific business processes, whose activities are to be implemented by Web services; it is associated to the Business Process Designer
- Ontology Management – creating and maintaining the ontology and available for the Knowledge Engineer

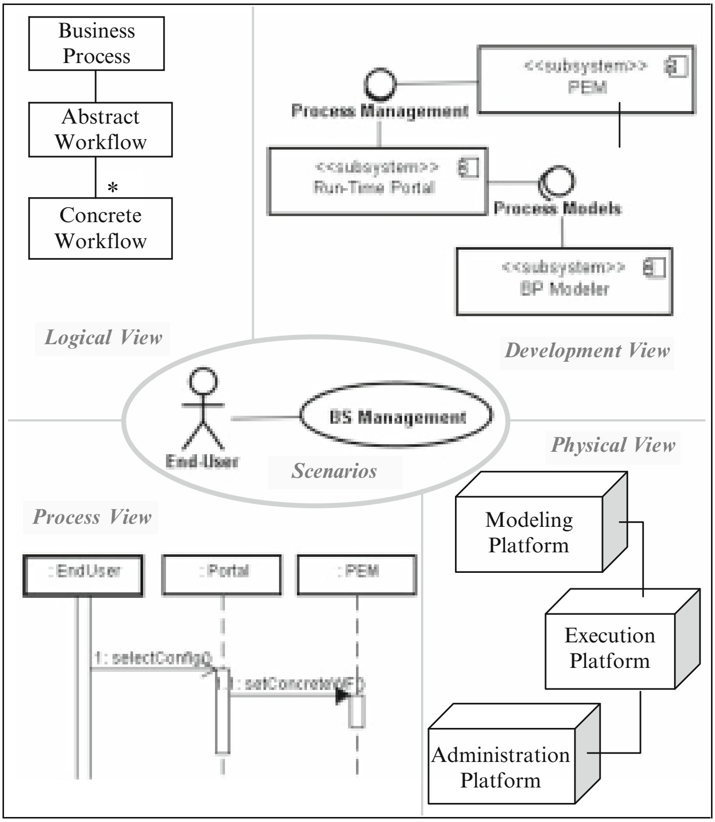


Fig. 41.1 4 + 1 Views for LD-CAST

- Web Service Management – registering and publishing Web services, which can be performed by the Service Provider Administrator
- Monitor Business Service Requests – delivering services that also include manual activities; it is assigned to the Service Provider Clerk
- Security Management – assuring the system security and integrity, which is the responsibility of the LD-CAST Administrator
- Performance Monitoring – supervising the clients and system performance, also effectuated by the LD-CAST Administrator

The logical view outlines the existence of three layers of abstraction, correspondent to business services and their mapping on different technologies (see Fig. 41.1). The most important concept from the top level is the *Business Process* (BP) – a process that is associated to a business service and may render it executable; it pertains to the Modeling Layer and is expressed in a language comprehensible for the business domain experts. In order to automate these processes, they have to be transformed into executable workflows. As LD-CAST flexibility requirement imposes a late binding of Web services, there are actually two layers of execution. First, one obtains an *Abstract Workflow* – a sequence of abstract services, corresponding to a business process, but introducing technical details, to allow the automatic generation of a Concrete Workflow; it stands on the Abstract Execution Layer. Then, one builds a *Concrete Workflow* – a workflow where each abstract service was replaced with a correspondent Web service; it is situated on the Concrete Execution Layer.

The Development View describes the partitioning of LD-CAST in the following subsystems:

- *Run Time Portal* – used by Guests and End-Users for obtaining information and business services from LD-CAST
- *Business Process Modeler* (see BP Modeler from Fig. 41.1) – accessed by the Business Process Designer for defining process models
- *Ontology Management* – used by the Knowledge Engineer to create the reference business ontology
- *Concrete Resource Management* – containing the Web service registry and the support for characterizing and publishing services, available for the Service Provider Administrators
- *Process Execution Management* (PEM) – generating concrete BPEL workflows and executing them
- *Search and Discovery* (S&D) – performing a semantic matchmaking between the set of available Web services and the abstract services corresponding to a requested business process
- *Security* – supporting a federated identity solution for authentication and authorization
- *Performance Monitoring* – supplying a centralized monitoring console for the service execution and for the subsystems performance

The Process View describes the interaction between subsystems, actors and classes for implementing scenarios and executing business processes. For instance, for the BS Management scenario, the End-User initiates an interaction between Run-Time Portal, PEM and S&D subsystems (see Fig. 41.1). Thus, when the End-User selects a process for the desired business service, the system gives him the choice to choose one of the available configurations of concrete Web services found by S&D; this configuration is used by PEM, to generate the concrete workflow and to start the process, whose execution is going to be monitored by the requester and sometimes assisted by a Service Provider Clerk, for performing half-manual activities.

The Physical View emphasizes the geographical distribution of LD-CAST, which has a Core containing the above presented subsystems, but also Local Agencies (responsible for payment and local management of users accessing Core services) and Service Providers (which provide services that will be orchestrated by LD-CAST Core). Moreover, the Core subsystems are deployed on three platforms (see Fig. 41.1):

- *Modeling Platform* – dedicated for system set up and configuration, and containing: Business Process Modeler, Ontology Management and Concrete Resource Management

- *Execution Platform* – supporting the automatic execution of business services, by integrating Run Time Portal, Process Execution Management and Search and Discovery
- *Administration Platform* – for managing users and controlling the efficiency of the system operation, using Security and Performance Monitoring subsystems

41.3.2 Applying SOA Specific Models

41.3.2.1 Service Views

The Service Views model [14] is specialized for SOA and contains 9 views, each one described at two levels: logical and physical; their match on stakeholder categories is variable, in respect with the application. Besides, there are *Quality of Service* (QoS) and *Quality of Management* (QoM) attributes, which crosscut all the views and reflect two important viewpoints: the consumer and the provider respectively. No details are given related to the definition of models.

Use cases, scenarios and business processes are grouped in the *Business View*, which specializes the Use Case View from the "4 + 1" View Model, for explicitly introducing process orchestration.

There are also four more views dedicated to services: Discovery, Interface, Transformation and Invocation; they all fit well for the LD-CAST description. For our case-study, the *Discovery View* outlines the late binding of services, and the fact that they are discovered through the existent matches between the annotations of the process activities and those of the registered Web services, both selected from Business Ontology. The *Interface View* shows the use of standard technologies for internal and external Web services, based on SOAP and WSDL, and describes specific information included in service contracts, like price and delivery time. The *Transformation View* is not necessary for the services from external providers, but it is useful for matching the dynamically generated input forms, on the inputs of services called by a certain concrete workflow, and also for matching service outputs on the presentation of the results seen by the End-Users through the portal. The *Invocation View* is characterized by the use of BPEL4WS (Business Process Execution Language for Web Services) [16] for instantiating processes and for executing them on the Active BPEL Engine.

For *Component*, *Data* and *Infrastructure Views*, there are good correspondences on Development, Logical and Physical views from the "4 + 1" View model, which was applied in 3.1. An interesting contribution is brought by *Test View*, due to the difficulties generally encountered in testing services, related to versioning standards, payment models, simulating a heavy loading and compensation activities [1]. In this separate view, one should describe the business processes specially defined for testing the integration of Execution and Modeling platforms in LD-CAST. Then, for validation, other attributes mentioned in the Service Views model were also considered: *security, reliability, availability, usability* and *performance*.

The Service View model outlines that maintenance is an important concern related to a SOA system. In our case-study, *manageability* was treated by monitoring and tracing the execution of services, using the Run-Time Portal (by the End-Users and the Service Provider Clerks) and the Performance Monitoring (by the LD-CAST Administrator). Moreover, *extensibility* was one of the most important objectives during the entire development, consisting in:

- Designing new processes for business services
- Registering new web services
- Annotating processes and web services
- Enriching the ontology
- Adding new Local Agencies and Service Providers
- Adapting the multilingual content of the LD-CAST Portal

The Composite Service View includes all the views of the component services. However, more details would be necessary in LD-CAST to describe the way it aggregates services provided by several e-government suppliers into new services [17]. The complexity stands in the fact that there is no limitation to pre-arranged providers, but the selection is dynamic, such as to fulfill a given end user request.

41.3.2.2 The BDC View Model

The BDC View model defines three views: Business Analysis, Service Design and Service Composition, dedicated for service oriented software [13]. They correspond to three kinds of stakeholders: business and system analysts, designers and developers, service integrators. This subchapter describes how we can apply this model for LD-CAST.

The *Business Analysis View* consists of two models:

- The business domain definition model – represented, in our case, with class diagrams (similar to the Logical View from the "4 + 1" Views) and also with the reference ontology for chambers of commerce.
- The business process model – described with UML activity diagrams and a template for documenting business variability; Table 41.2 shows the application of this template for the process "Certificate with Company Data", defined on LD-CAST platform.

The *Service Design View* consists of two models:

- Service Classification model – making the difference between internal and external services, which is also done in LD-CAST; the description template for external services is similar to the way one describes services for Concrete Resource Manager, where the service type is always a Web service and the vendor is usually a chamber of commerce. Moreover, for supporting service discovery, LD-CAST also uses a semantic description, based on the business ontology.
- Service Design model – with a template based on WSDL concepts, consequently fitting the WSDL service description from LD-CAST.

Table 41.2 Business variability in LD-CAST

Template element	Description	Value for the case-study
[id]name	Business process name	Certificate with company data
Cv_property	Variability type	Common
Who	Business entity description	Chambers of commerce
What	Business entity target	Deliver trustful information related to a company
When	Time information	When requested by a registered End-User
Action	Actions description	Upload certificate to the portal
Variation point	Where variations may occur	Filling-in forms, payment and supply certificate
Variant	The value of variant	Open
Cardinality	Number of variants	Not limited
Binding time	Domain time – for static binding/ application time – for dynamic binding	Application time

The Service Composition View consists of an interaction model (e.g. a UML activity diagram) and a description containing: the service invocation sequence, inputs and outputs of the composite service, the composition method (e.g. orchestration for LD-CAST) and the invocation sequence of the interface.

41.3.3 Useful Elements from Other View Models

41.3.3.1 RM-ODP, the Perspectives for Distributed Processing

The RM-ODP framework defines concepts and languages for the following viewpoints [4]:

- Enterprise (describing the organization objectives and business processes)
- Information (related to information management) – from this point of view, a specificity of LD-CAST is the use of national ontologies mapped on a reference business ontology
- Computational (presenting the system components and their interfaces)
- Engineering (outlining interactions between components) – in LD-CAST, this view is characterized by SOA, with statically bound internal Web services and dynamically bound external services
- Technology (describing the choices made for the implementation)

Among these viewpoints, the Technology one is important for our case-study, as it captures preoccupations that are not present in other view models. Here there are some technology choices from LD-CAST: the *Run Time Portal* was implemented with JEM (JBoss Enterprise Middleware); the *Business Process Modeler* uses ADOeGov [18] modeling language; the *Ontology Management* is based on

OPAL (Object, Process, Actor Modeling Language) [19]; the *Process Execution Management* (PEM) subsystem includes an ActiveBPEL engine; the *Security* subsystem adopted the federated identity solution of Shibboleth.

41.3.3.2 The SEI Viewpoint Model: "Views and Beyond"

The SEI model, which is loosely called "views and beyond", is based on the idea that the number of views for documenting architecture should not be fixed, but it should be decided in respect with the system. There are three viewtypes: *Module*, *Component-and-connector* and *Allocation*. For each of them, one can elaborate a style guide, representing the main design approaches [3]. The views are styles applied to the described system and they can be documented with more or less details, in respect with each stakeholder necessity.

For our case-study, a view that is useful and was not present in other models corresponds to the *Work assignment* style, from the *Allocation* viewtype. Even if for small projects it can be combined with *Implementation* or *Module decomposition* views, it is very useful in our case, especially for the stakeholders related to funding and management. The *Project coordinator* and the *Project manager* for each partner need detailed information related to the work assignment. So do various kinds of developers and maintainers of the system. The *Project officer* from the European Community also needs some details related to this view.

Acknowledgement We would like to thank the LD-CAST Consortium, including both technical partners and chambers of commerce, who inspired our study related to the views of various stakeholders in a system based on SOA.

References

1. Sommerville, J.: Software Engineering. Addison-Wesley, Reading, MA (2006)
2. Sommerville, J., Sawyer, P.: Viewpoints: principles, problems and a practical approach to requirements engineering. Annal. Softwar. Eng. **3**, 101–130 (1997)
3. Clements, P., Bachmann, F., Bass, L., Ivers, J., Garlan, D., Little, R., Nord, R., Stafford, R.: Documenting Software Architectures: Views and Beyond. Addison-Wesley, Reading, MA (2003)
4. May, N.: A survey of software architecture viewpoint models. Proceedings of the Sixth Australasian Workshop on Software and System Architectures, pp. 13–24, Melbourne, Australia (2005)
5. Local Development Coordination Actions enabled by Semantic Technology (LD-CAST) Project. http://www.ldcastproject.com
6. Object Management Group, UML 2.0 Superstructure Specification (2005). http://www.omg.org
7. West, M.: Developing High Quality Data Models. In: Fowler, J. (ed.) EPISTLE (1999)
8. Kruchten, P.: Architectural blueprints—The "4 + 1" view model of software architecture. IEEE Software **12**(6), 42–50 (1995)
9. FCGSS White Paper, Applying 4 + 1 view architecture with UML2 (2007)

10. IEEE Std 1471–2000 IEEE Recommended Practice for Architectural Description of Software-Intensive Systems (2000)
11. ISO/IEC 42010:2007, Systems and software engineering – Recommended practice for architectural description of software-intensive systems (2007)
12. Ibrahim, D., Misic, V.B.: Service views: a coherent view model of the SOA in the enterprise. In: IEEE International Conference on Services Computing SCC'06, pp. 230–237 (2006)
13. Park, J., Moon, M., Yeom, K.: The BCD view model: business analysis view, service composition view and service design view for service oriented software design and development. In: Proceedings of the 12th IEEE International Workshop on Future Trends of Distributed Computing Systems FTDCS, pp. 37–43 (2008)
14. Tran, H., Zdun, U., Dustdar, S.: View-based reverse engineering approach for enhancing model interoperability and reusability in process-driven SOAs. In: Proceedings of the 10th International Conference on Software Reuse ICSR'08, Springer LNCS (2008)
15. OASIS. Reference Architecture for Service Oriented Architecture Version 1.0, Public Review Draft 1 (2008). http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/soa-ra-pr-01.pdf
16. OASIS. Web Services Business Process Execution Language for Web Services version 2.0 (2007)
17. Ionita, A.D., Catapano, A., Giuroiu, S., Florea, M.: Service oriented system for business cooperation. In: Proceedings of ICSE International Workshop on Systems Development in SOA Environments SDSOA '08, pp. 13–18. ACM, New York (2008)
18. Palkovits, S., Orensanz, D., Karagiannis, D.: Process modelling in Egovernment – living process modelling within a public organisation. In: IADIS International e-Society, pp. 3–10 (2004)
19. D'Antonio, F., Missikoff, M., Taglino, F.: Formalizing the OPAL eBusiness ontology design patterns with OWL. Business episodes and workflow integration: a use case in LD-CAST. In: Third International Conference on Interoperability for Enterprise Applications and Software, I-ESA 2007, pp. 345–356 (2007)

Chapter 42
Optimal Sample Number for Autonomous and Central Wireless Sensor Actuator Network

Amir M. Jafari and Walter Lang

Abstract Applying distributed system concept with wireless communication in sensor and actuator network for process automation applications is led to developing autonomous wireless sensor actuator network. One of the requirements for this development is determining the sample number. Because of the hardware and power supply limitations, wireless sensors are applied to discrete control system. In wireless sensor network the highest sample number in comparison to the wired network is restricted. Moreover, the lowest sample number is also constrained by the limitations imposed by the control limits values. In this chapter, firstly the central structure is reviewed and then autonomous structure is introduced. Afterwards relations between sample number and actuator frequency drift in discrete domain is formulated and presented. In addition, ways to trade off the sample number with the actuator frequency and control limits value are acknowledged. One approach to find the optimal sample rate for each network structure is proposed. It is shown when the sensor network becomes larger, autonomous network can partly compensate communication number and energy consumption increase by adding the sample number whereas central network does not support this feature.

Keywords Autonomous network · central network · sample number · Wireless Sensor Actuator Network (WSAN)

A.M. Jafari (✉)
Institute for Microsensors, Actuators and Systems (IMSAS) at the University of Bremen,
Bibliothekstrasse, 1 – D-28359 Bremen – Germany
e-mail: jeffrey@imsas.uni-bremen.de

W. Lang
Institute for Microsensors, Actuators and Systems (IMSAS) at the University of Bremen,
Bibliothekstrasse, 1 – D-28359 Bremen – Germany
e-mail: wlang@imsas.uni-bremen.de

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_42,

42.1 Introduction

Integration of wireless communication in Sensor Actuator Network (SAN) led to the new research in realm of network structure and technical approaches of network establishment. In order to implement Wireless Sensor Actuator Network (WSAN) in an automation process application, two structures are considered: Central structure and Autonomous structure. As example of WSAN implementation is in Heating, Ventilation and Air Conditioning (HVAC) system [1, 2].

In a central network, sensors measure the environmental parameters and send their data to center of network. The center processes the data and makes decision for actuators. Afterwards the instruction will be sent to actuators. None of the network nodes i.e. sensor or actuator do not have the role to make decision. They are just subordinate to the center and follow the instructions [3, 4]. The central structure is identified by two topology types. The first one is called by Tanenbam [5] as Star and the same is called by Elazar [6] Center-Periphery model. The second topology called hierarchical [3, 6]. Figure 42.1 shows which part of system model reside inside the center, the sensor and actuator.

One limitation with the central network is scalability. When network size in terms of nodes numbers and applications increases, the center can go out of the peripherals or resources. In such case the first solution could be substituting the center with more powerful one. This solution is not only sometimes impossible but also it is expensive. Another solution is adding new centers and connecting them. The coordination can be done by a higher level center. Such system structure is basically called "distributed system" [4, 5, 7, 8] although there is no exact definition for distributed system [4, 5]. In distributed structure each subsystem has its own center which is responsible for local decisions then they are connected to each other to establish a system. This system should look like a single system [5]. In such configuration the decision which requires the information from different subsystems is made in the system center which locates over the subsystems. Besides the scalability, hardware distribution, reliability and connectivity as advantages are counted for distributed system [4, 5, 7].

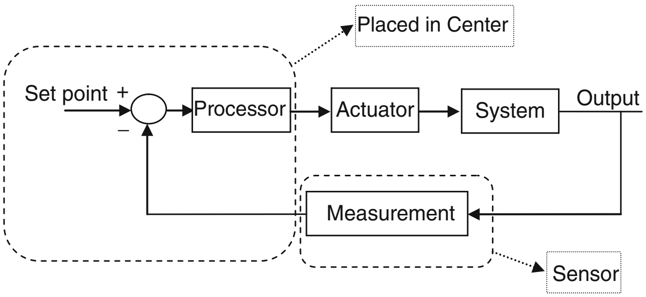


Fig. 42.1 Controller of the system resides in center

Following distributed system, in industrial automation the Distributed Control System (DCS) is developed. A bunch of sensors and actuators are grouped by a controller in star or bus topology [9–11]. The controllers are networked in controller level. Above all of them a supervisory level is placed which called the Supervisory Control And Data Acquisition (SCADA) [10, 11]. Autonomous Wireless Sensor Actuator Network (AWSAN) is targeted the SAN and is considered to be alternative for Central WSAN (CWSAN). By applying the distributed concept over the SAN, the advantages of distributed system will be inherited for Sensor Actuator Network (SAN).

Autonomous entity is an entity makes decision for itself. On the other word by autonomous we mean "self-decision making". In order to make decision it may need information from other entities, therefore entities should be networked to exchange the data. In this sense the nodes are "interdependent". This data exchange is interpreted as "cooperation". By networking, the "connectivity" advantage of distributed system between applications and resources will be preserved too. The proposed topology for AWSAN is mesh topology because of the reliability feature of distributed system.

The reason for calling such system as autonomous and not distributed is emphasizing on the self-decision making character and the consideration of the communication and decision-making structure together. Versus Hierarchy for central network, this structure is called Heterarchy [3, 4].

In order to develop Autonomous network with above considerations a routing algorithm is needed to establish direct communication between nodes. In case of AWSAN, a Sequential Coordinate Routing Algorithm (SCAR) is developed [12]. Two advantages, lower energy consumption [13] and more robustness [14] are derived for autonomous network. The question is what the sampling frequency should be for reading the control variable or how often the sensor should measure the control variable. This chapter offers a method for finding an optimal sample number for further development of WSAN. Moreover it shows that how the autonomous network is capable to deal with scalability in comparison to the central network.

In a central network each sensor sends its measured parameter on each sample time periodically, but in autonomous network the sensor makes decision when it should send the information to its correspondent actuator. For example in On-off control method, the sensor compare the measured parameter with the limits value when it goes over them, it sends a message to its correspondent actuator. In this sense the sensors are autonomous entities because they decide when to send a message. The actuators make their own decision based on the sent data from sensors and perhaps other parameters from other nodes [4]. In fact the control tasks are performed by the actuators. In this sense actuators are autonomous entity as well. In Fig. 42.3, it is shown that which part of the control system is located in which node. The process unit is separated as representative of decision-maker (controller) and it is placed in actuator (Fig. 42.2).

In this work, WSAN consists of nodes equipped with a wireless transceiver (CC2420), a tiny ultra low energy consumer microcontroller MSP 430 family and two batteries for power supply. "Tmote sky" from "Moteiv" [15] is taken as sample of such nodes. The wireless transceiver is IEEE 802.15.4 standard compliant

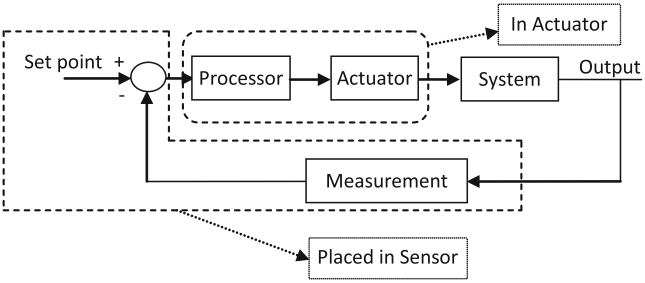


Fig. 42.2 Autonomous control system model

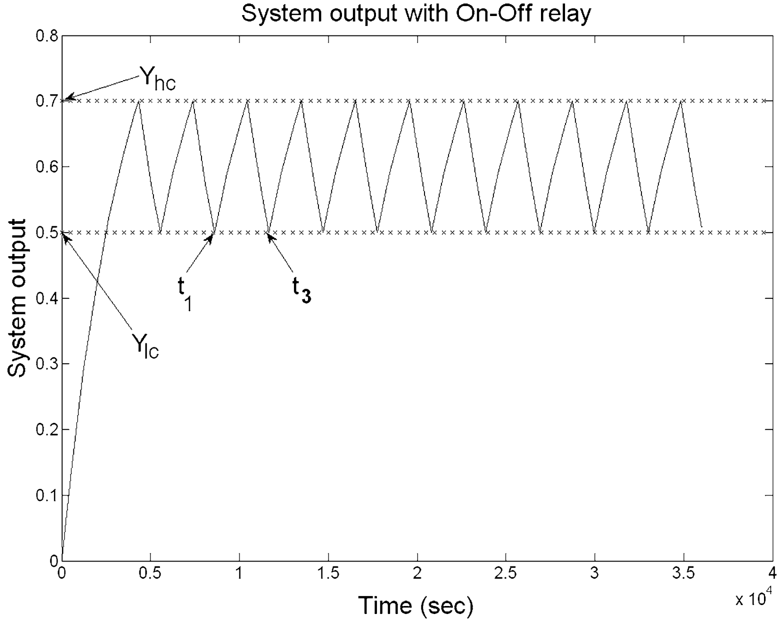


Fig. 42.3 System output with relay

and its radio range is limited. The routing protocol is not real-time protocol [8] and wireless communication in mesh networking especially for fast process faces with problems. For this reason and because of the hardware and power supply limitations, wireless nodes are preferably used for On-off or discrete event control system.

42.2 Sample Frequency Calculation

Applying wireless communication in industrial automation is challenging. One of the challenges shows up is the real-time property of the system. In digital control system, the control variables are sampled and processed. The sampling period depends on the system time constant which defines the natural frequency of the system. Nyquist-Schannon's sampling theorem states that the sampling frequency should be greater than double of the system natural frequency. Ideally the sample frequency is ten times larger [11]. During a sample interval the sampled variable should be transmitted and processed before next sample arrives. If the control task cannot be achieved during the sample intervals, it means that the control system cannot follow the system variations and it can lead to instability [11]. By choosing the high sample frequency the signal get closer to continuous signal form and frequency spectrum will have better resolution. In wired network it is possible to increase the sample frequency as high as enough.

In wireless network increasing the sample frequency causes more message transmissions and higher network traffic which leads to other consequences like more delay in message transmission and higher network energy consumption. On the other hand delay time in transmission can lead to instability of the system. The sample number can be compromised with the transmission number or nodes energy consumption.

It is assumed that the transfer function of the system in Fig. 42.3 is first order and its Laplace transform is represented by H(s) which is expressed in Eq. (42.1). The set point value is assumed to be Y_0 and the limit values are Y_{hc} and Y_{lc} with equal distances from Y_0. The On-off relay is implemented as actuator (Fig. 42.3). Figure 42.5 shows the system step response for $Y_{hc} = 0.7$ and $Y_{lc} = 0.5$ for 10 h. The actuator On-off frequency in a continuous domain is calculated by Eq. (42.2).

$$H(s) = 1/(T_n \times s + 1) \tag{42.1}$$

$$f_c = \frac{1}{t_3 - t_1} = \frac{1}{T_n \times \ln\left(\dfrac{Y_{hc} \times (1 - Y_{lc})}{Y_{lc} \times (1 - Y_{hc})}\right)} = \frac{f_n}{c} \tag{42.2}$$

Equation (42.3) shows the recursive equation in a discrete domain when $H(s)$ is mapped to the z-plane with sample time T_s and a normalized output. In this equation N is the sample number during T_c which is equal to inverse of f_c computed in Eq. (42.2). T_s is the sampling period (f_s sampling frequency) and c is defined in Eq. (42.2). Assuming that the last sampling occurs just before the limit values; then the system output goes beyond the limit values up to the next sample time. This incident is counted as an error. In order to avoid such errors the new limit values are defined for discrete time system. These new values are equal to the samples of the output value on one sample period time before the limits Y_{hc} and Y_{lc}. These limits are called Y_{hd} and Y_{ld} in Eq. (42.4). Considering this definition by lower sample number $(Y_{hc} - Y_{hd})$ becomes greater; consequently the limits interval $(Y_{hd} - Y_{ld})$ becomes smaller. Smaller limit intervals lead to a greater actuator frequency f_d. In other words, moving to discrete domain with a lower sample rate leads to higher actuator frequency which should be confronted (Fig. 42.4).

$$\begin{aligned} y(n) &= (1 - \exp(-T_s/T_n)) + \exp(-T_s/T_n) \times y(n-1) \\ T_c &= N \times T_s \end{aligned} \tag{42.3}$$

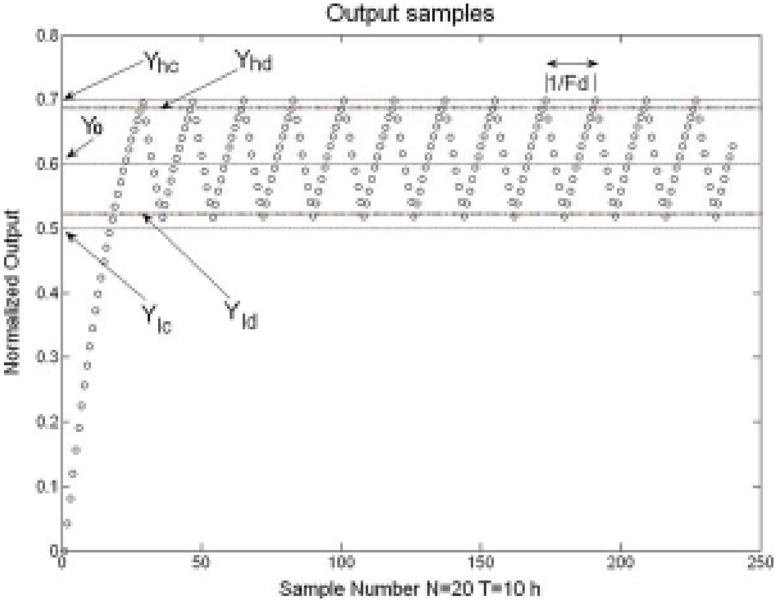


Fig. 42.4 Digitized system output with relay

$$f_s = N \times f_c = (N \times f_n)/c$$
$$y(n) = (1 - \exp(-c/N)) + \exp(-c/N) \times y(n-1)$$

$$Y_{hd} = (Y_{hc} + \exp(-c/N) - 1)/ + \exp(-c/N)$$
$$Y_{ld} = Y_{lc}/\exp(-c/N) \tag{42.4}$$

$$N > c/(-\ln(1 + Y_{lc} - Y_{hc})) \tag{42.5}$$

Since Y_{hd} should always be greater than Y_{ld}, a boundary limit exists for sampling number N which is defined in Eq. (42.5). This is our first criteria for choosing sampling number. As an example for $Y_{hc} = 0.7$ and $Y_{lc} = 0.5$, N should be strictly greater than 3 ($f_s \geq 4^* f_c$). In Fig. 42.6 the digitized output for the above system with $N = 20$ is depicted. The time axis is for 10 h. In comparison with Fig. 42.5 it can be seen that the actuator state changes 20% more than its value in a continuous domain of the control system.

In AWSAN when the system output reaches limits, sensor sends a message to the actuators. In Fig. 42.6 it can be seen that the number of message transmissions is double the number of the actuator status changes (i.e. one message for on-off and one message for off-on transient states). It denotes that the message transmission number is proportional to the actuator frequency. Since reduction of the sample number decreases the discrete limits interval and it leads to amplifying the actuator

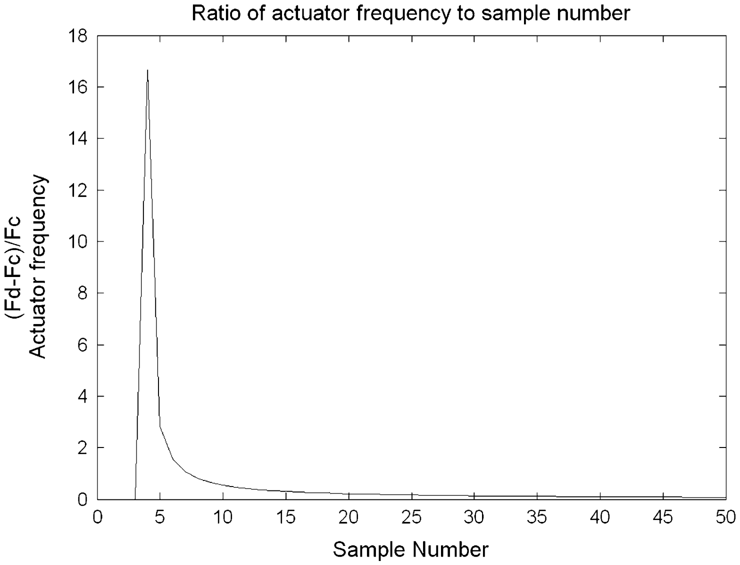
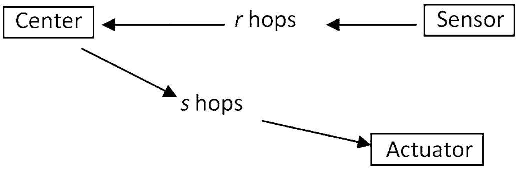


Fig. 42.5 An example of actuator frequency ratio versus sample number

Fig. 42.6 Central network communication path



frequency, consequently the number of message transmissions increases. By raising the sample number, the microcontroller occupancy and energy consumption increases too. This phenomenon causes losing more messages during the routing of other sensor messages in addition to increasing the process energy consumption. Therefore the sample number should be compromised in a way that it is neither very small that causes the increase in the actuator frequency and message transmission nor so large that the microcontroller becomes too occupied and the process energy consumption increases highly.

In a central WSAN, sensor sends message to the center at each sample time (Fig. 42.6). Increasing the sample number, raises the message transmissions number directly which causes more transmission energy consumption and high network traffic. Moreover high frequency is not beneficial for actuator life time as well. Reduction of the sample number leads to the rising of actuator frequency which means the center should send more messages to the actuator.

Increasing the sample number in central network causes more transmission energy while in autonomous network it leads to more process energy consumption. In addition process energy consumption is much smaller than transmission energy consumption. Therefore sample number in an autonomous network can be greater than its value in a central network which implies that with the same energy consumption, lower actuator frequency and better control quality can be achieved with autonomous configuration.

$$\Delta f/f_c = (f_d - f_c)/f_c = \frac{c}{\ln\left(\dfrac{(Y_{hc} + \exp(-c/N) - 1)(\exp(-c/N) - Y_{lc})}{Y_{lc} \times (1 - Y_{hc})}\right)} - 1 \tag{42.6}$$

The normalized difference between actuator frequencies in continuous and discrete domain is calculated in Eq. (42.6). By this equation the sample number and actuator frequency can be compromised. The graph of Eq. (42.6) is depicted for $Y_{hc} = 0.7, Y_{lc} = 0.5$ and $T_n = 3{,}600\,\mathrm{s}$ in Fig. 42.7. Based on Eq. (42.5), N must be greater than 3. For $N = 4$ the actuator frequency increases 16.67 times (1,667%) of its frequency in continuous domain (Fig. 42.7). This oscillation is not reasonable for actuator. Moreover it indicates that 16 times more messages should be sent to the actuator to turn on or off. When the sample number is equal to 20 the drift will be 20% which is more acceptable. By increasing N from $N = 30$ to $N = 50$, the actuator frequency decreases just about 6.5% while this increase (from $N = 30$ to $N = 50$) leads 66.66% increase of the process energy consumption, the node microcontroller occupancy in autonomous network and message number in central network. This increase (from $N = 30$ to $N = 50$) sounds not very helpful.

Now, we assume that we have a control task with no restriction in the limit values, so that the set point Y_0 is given and the upper and lower limits are in equal distance from Y_0 in the continuous domain. Rewriting Eq. (42.6) results in Eq. (42.7). In these equations by choosing two arbitrary parameters, the third parameter can be computed. For example if 20 samples number ($N = 20$) and maximum 20% drift ($\Delta f/f_c \leq 0.2$) is acceptable for the actuator, ΔY will be 0.18. Utilizing Eq. (42.7) offers the trade off option between three parameters: limits, actuator frequency and sample number.

$$\begin{aligned} p(N) &= \Delta f/f_c = (f_d - f_c)/f_c \\ &= \frac{c}{\ln\left(\dfrac{(Y_0 + \Delta y + \exp(-c/N) - 1)(\exp(-c/N) - Y_0 + \Delta y)}{(Y_0 - \Delta y) \times (1 - Y_0 - \Delta y)}\right)} - 1 \\ c &= \ln(((Y_0 + \Delta y) \times (1 - Y_0 + \Delta y))/((Y_0 + \Delta y) \times (1 - Y_0 - \Delta y))) \end{aligned} \tag{42.7}$$

42.3 Sample Number Selection

In central and autonomous networks, message transmission number is related to the sample number and actuator frequency. The sample number can be selected in compromise with the transmission number in central network and energy consumption in autonomous network.

42.3.1 Central Network

In central network the communication path is considered. The sensor measures the environment parameter in each sample time and sends it to the center through r hops. The center checks the sensor value; if it is over the limits it sends a message to change the actuator status. With the sample number of N, the number of message transmissions from the sensor to the center during time T is equal to $T/T_s \times r = ((T \times N)/T_c) \times r$. At the same time interval T, the number of instruction message transmissions from the center to the actuator is equal to $(T/T_d) \times 2 \times s = (T \times (p(N)+1)/T_c) \times 2 \times s$. By adding these two values the total number of messages in unit time equals to Eq. (42.8).

$$g(N,r,s) = (N \times r + (p(N)+1) \times 2 \times s) \times f_c \tag{42.8}$$

The function is minimal at N = 6 with the assumed system parameters. Considering the minimum of the transmission numbers, the best sample number is equal to 6. It indicates that the sample should be taken at every $\mathrm{T_s} = \mathrm{T_c}/\mathrm{N} \approx 508$ s.

Table 42.1 shows that for s from 1 to 10, the N increases from 6 to 12. But suppose that the sample number remains equal to 6 when $s = 10$, then the transmission number will be $g(6,1,10) \approx 0.1884$. Now suppose that $N = 12$ when $s = 10$, the transmission number will be $g(12,1,10) \approx 0.0132$. Now we compare the $g(6,1,10)/g(6,1,1) \approx 4.91$ with $g(12,1,10)/g(6,1,1) \approx 3.55$. It shows that by increasing sample number when ops number increase the message number is relatively reduced 27% which counted as an advantage.

From another angle we hold the $s = 1$ and start to increase r one unit at a time. Table 42.1 shows that when $r$increases, N does not change significantly. N should be reduced but its value is limited by Eq. (42.5). This claim can be verified by inequality Eq. (42.9). This inequality compares the increase of hop number between sensor-center and center-actuator. It shows that the message number increases less center-actuator wing. From this observation it is concluded that in central network it is more efficient to choose the center closer to the sensor than the actuator. Practically it is more reasonable to consider the center closer to the node which has higher loads to deliver. Finally, if there are r hops from the sensor to the center and s hops

Table 42.1 Sample number corresponding to hops number center to actuator and from sensor to center

s	Minimum g	N	r	Minimum g	N
1	0.0036	6	1	0.0036	6
2	0.005	8	2	0.0056	6
3	0.0062	8	3	0.0074	5
4	0.0073	9	4	0.0091	5
5	0.0084	10	5	0.0107	5
6	0.0094	10	6	0.0123	5
7	0.0104	11	7	0.014	5
8	0.0113	11	8	0.0156	5
9	0.0123	12	9	0.0173	5
10	0.0132	12	10	0.0189	5

from the center to the actuator, the proper sample number is where g is minimal. As an example for r =3 and s =7, N corresponding to the minimum g is equal to 8.

$$[g(5,10,1)/g(6,1,1)] > [g(12,1,10)/g(6,1,1)] \cong 5.25 > 3.66 \qquad (42.9)$$

42.3.2 Autonomous Network

For autonomous network we consider the communication path like Fig. 42.7. The sensor in this figure measures the environment parameter at each sample time. Then the sensor compares it with the limit values; if it is beyond them, it sends a message to the actuator to inform. In this section it is assumed that the average of the process energy consumed by wireless node microcontroller is fixed and it is considered as the unit for energy consumption measurement. Another assumption is that the transmission energy from one node to another is equal to $e = 10$ times of process energy.

$$h(N,r) = (p(N)+1) \times 2 \times r \times f_c \qquad (42.10)$$

In Fig. 42.7 the number of transmissions for N in time T is equal to $(T/T_d)\times 2\times r = (T\times(p(N)+1)/T_c)\times 2\times r$ and in time unit it is equal to the function h in Eq. (42.10) $(h(N,r) = g(N,0,r))$.

As N increases from $N = i$ to $N = i+1$, $p(N)$ decreases. This reduction causes the reduction of h, the message transmission number. Equation (42.11) formulates the reduction of the transmission numbers which is equivalent to $\Delta h^{*}e$ of the process energy consumption reduction. Decreasing of transmission numbers also causes reduction of the process energy for forwarding messages in intermediate nodes. We call the summation of these two energy consumption reductions as *saved energy*. From another side by increasing the N, the process energy consumption increases in order to take more samples. We look at this energy consumption increase as *cost energy*. Therefore N can be increased as much as the *saved energy* is still greater than the *cost energy*, which is formulated in Eq. (42.12). Optimal sample number is maximum N so that the inequality 12 becomes valid.

$$\begin{aligned}\Delta h_i^{i+1} &= h(i,r) - h(i+1,r)\\ &= (p(i) - p(i+1)) \times 2 \times r \times f_c = \Delta p_i^{i+1} \times 2 \times r \times f_c \qquad (42.11)\end{aligned}$$

For $Y_{hc} = 0.7, Y_{lc} = 0.5$ and $T_n = 3{,}600$ s, Table 42.2 shows N corresponding to each r. For example when $r = 2$ then $N = 15$ and $T_s = T_c/N \approx 190$ s is the

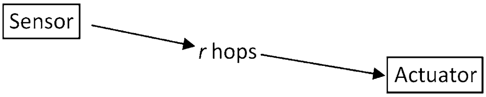


Fig. 42.7 Autonomous network structure

Table 42.2 Maximum N for which Eq. (12) is valid for different r

r	1	2	3	4	5	6	7	8	9	10
N	12	16	19	21	23	25	27	28	30	31

optimum sample period.

$$\begin{aligned}\Delta h \times e + \Delta h/r \times (r-1) - (i+1-(i)) \times f_c \\ = \Delta p_i^{i+1} \times ((e+1) \times r - 1) \times f_c \times 2 - f_c > 0\end{aligned} \qquad (42.12)$$

Considering these conditions for N, if there are other criteria as well, N could also be compromised with them. For example in Table 42.1 when r is 2, $N = 6$ but with respect to (Eq. 42.6) and Fig. 42.7 the actuator frequency increases about 150%. If this oscillation is not acceptable as a criterion, N can be increased to 8 and actuator frequency drift reduces to about 80%.

References

1. Yamaji, M., Ishii, Y., Shimamura, T., Yamamoto, S.: Wireless sensor network for industrial automation. 5th International Conference on Networked Sensing Systems, 2008, pp. 253–253, 17–19 June 2008
2. Oesterlind, F., Pramsten, E., Roberthson, D., Eriksson, J., Finne, N., Voigt, T.: Integrating building automation systems and wireless sensor networks. SICS Technical Report T2007, 4 May 2007
3. Neil, A.: Duffie, challenges in design of heterarchical controls for dynamics logistic systems. First International Conference on Dynamics Logistic, LDIC 2007, pp. 3–24, August 2007
4. Dressler, F.: Self-Organization in Sensor and Actor Networks. Wiley, Hoboken, NJ (2007)
5. Tanenbaum, A., Van steen, M.: Distributed Systems Principles and Paradigms, 1st edn. Prentice-Hall, New York (2002)
6. Elazar, DJ.: Exploring Federalism. University of Alabama Press, Tuscaloosa (1987)
7. Jie, W.: Distributed System Design. CRC Press (1999)
8. Fei-Yue, W., Derong, L.: Networked Control Systems: Theory and Applications. Springer-Verlag, London (2008)
9. William, S.: Leviene, The Control Handbook, vol. I. CRC Press & IEEE Press, New York (1999)
10. Zhang, P.: Industrial Control Technology. William Andrew Inc, Norwich, NY (2008)
11. Gregory, K.: McMillan, Douglas M. Considine, Process/Industrial Instruments and Controls Handbook, 5th edn. McGraw-Hill, New York (1999)
12. Jafari, A.M., Sklorz, A., Lang, W.: Target-oriented routing algorithm based on sequential coordinates for autonomous wireless sensor network. J. Networ. Acad. Publ. **4**(6):421–427 (August 2009)
13. Jafari, A.M., Sklorz, A., Lang, W.: Energy consumption comparison between autonomous and central wireless sensor network **6**:166–170 (April 2009). In: Communications of SIWN, ISSN: 1757–4439 (Print) ISSN: 1757–4447 (CD-ROM)
14. Jafari, A.M., Hentschel, D., Lang, W.: Robustness in autonomous and central wireless sensor network: the orchard example. The Fourth International Conference on Systems and Networks Communications (ICSNC 2009), pp. 242–247, 20–25 September 2009
15. MoteivCorporation.tmote-sky-datasheet-02 (2006), www.moteiv.com

Chapter 43
WI-FI Point-to-Point Links: Performance Aspects of IEEE 802.11a, b, g Laboratory Links

J.A.R. Pacheco de Carvalho, H. Veiga, P.A.J. Gomes, C.F. Ribeiro Pacheco, N. Marques, and A.D. Reis

Abstract Wireless communications using microwaves are increasingly important, e.g. Wi-Fi. Performance is a very relevant issue, resulting in more reliable and efficient communications. Laboratory measurements are made about several performance aspects of Wi-Fi (IEEE 802.11 a, b, g) point-to-point links using two types of access points from Enterasys Networks (RBT-4102 and RBTR2). Through OSI levels 3, 4 and 7, detailed results are presented and discussed, namely: latency, ICMP packet loss, TCP throughput, jitter, percentage datagram loss and FTP transfer rate.

Keywords WLAN · Wi-Fi · IEEE 802.11a · IEEE 802.11b · IEEE 802.11g · point-to-point links · wireless network laboratory performance measurements

43.1 Introduction

Wi-Fi (Wireless Fidelity) is a wireless communications technology whose importance and utilization have been growing for complementing traditional wired networks. Wi-Fi has been used both in ad hoc mode, for communications in temporary situations e.g. meetings and conferences, and infrastructure mode. In this case, an AP (Access Point) is used to permit communications of Wi-Fi devices with a wired based LAN (Local Area Network) through a switch/router. In this way a WLAN, based on the AP, is formed which is known as a cell. A WPAN (Wireless Personal Area Network) arises in relation to a PAN (Personal Area Network).

Point-to-point and point-to-multipoint configurations are used both indoors and outdoors, requiring directional and omnidirectional antennas. There are detailed

J.A.R. Pacheco de Carvalho (✉), C.F. Ribeiro Pacheco, and A.D. Reis
Unidade de Detecção Remota, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
e-mail: pacheco@ubi.pt; a17597@ubi.pt; adreis@ubi.pt

H. Veiga, P.A.J. Gomes, and N. Marques
Centro de Informática, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
e-mail: hveiga@ubi.pt; pgomes@ubi.pt; nmarques@ubi.pt

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_43,

studies about wireless communications, wave propagation [1,2] and WLAN practical implementation [3]. Studies have been made about long distance Wi-Fi links in rural areas [4,5].

Wi-Fi uses microwaves in the 2.4 and 5 GHz frequency bands and IEEE 802.11 a, 802.11 b and 802.11 g standards [6]. In ETSI (European Telecommunications Standards Institute) countries IEEE 802.11b and 802.11 g are used both in indoors and outdoors through 13 channels in the 2,400–2,485 MHz frequency band, permitting nominal transfer rates up to 11 and 54 Mbps, respectively. IEEE 802.11 a permits nominal transfer rates up to 54 Mbps. It is available in most European countries for indoor applications through four channels in both the 5,150–5,250 MHz and the 5,250–5,350 MHz frequency bands. In the same countries 11 channels are available in the 5,470–5,725 MHz frequency bands for both indoors and outdoors. As the 2.4 GHz band is increasingly used, leading to higher interferences, the 5 GHz band is interesting given lower interferences, in spite of larger absorption and shorter ranges. The standards mentioned use CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) as the medium access control. The 802.11 architecture has been studied in detail, including performance analysis of the effective transfer rate [7]. An optimum factor of 0.42 was determined for the effective transfer rate in 11 Mbps point-to-point links, giving an effective transfer rate of 4.6 Mbps. Studies are available about Wi-Fi performance in indoor crowded environments having significant obstacles to signal propagation [8].

Several measurements have been made at OSI levels 1, 3, 4 and 7 for point-to-multipoint and point-to-point configurations in the 2.4 GHz band [9, 10]. Results have also been reported for WiMAX [11], very high speed FSO [12] and Wi-Fi [13]. In the present work further results are presented and discussed for laboratory performance measurements of IEEE 802.11 a, b, g point-to-point links using different access points. We consider latency, ICMP packet loss, TCP throughput, jitter, percentage datagram loss, and FTP transfer rate.

The rest of the paper is organized as follows: Section 43.2 presents the experimental details i.e. the measurement setup and procedure. Results and discussion are presented in Section 43.3. Conclusions are drawn in Section 43.4.

43.2 Experimental Details

Two types of experiments were carried out, which are referred as Exp-A and Exp-B. In the measurements of Exp-A we used Enterasys RoamAbout RBT-4102 level 2/3/4 access points (mentioned as AP-A), equipped with 16–20 dBm IEEE 802.11 a/b/g transceivers and internal dual-band diversity antennas [14], and 100-Base-TX/10-Base-T Allied Telesis AT-8000S/16 level 2 switches [15]. The access points had transceivers based on the Atheros 5213A chipset, and firmware version 1.1.51. They were parameterized and monitored through both the console using CLI (Command Line Interface) and a HTTPS (Secure HTTP) incorporated server. The configuration was for minimum transmitted power and equivalent to point-to-point, LAN to LAN mode, using the internal antenna. For the measurements of Exp-B we used Enterasys RoamAbout RBTR2 level 2/3/4 access points (mentioned as AP-B), equipped with 15 dBm IEEE 802.11 a/b/g cards [14], and 100-Base-TX/10-Base-T Allied Telesis AT-8000S/16 level 2 switches [15]. The access points had RBTBH-R2W radio cards similar to the Agere-Systems model 0118 type, and firmware version 6.08.03. They were parameterized and monitored through both the console and the RoamAbout AP Manager software. The configuration was for minimum transmitted power i.e. micro cell, point-to-point, LAN to LAN mode, using the antenna which was built in the card.

Interference free channels were used in the communications. WEP (Wired Equivalent Privacy) encryption was not activated. No power levels above the minimum were required as the access points were very close. The results obtained in the present work were insensitive to AP emitted power level.

Both types of experiments, Exp-A and Exp-B, were made using the laboratory setup shown in Fig. 43.1.

For 7-echo UDP traffic injection (OSI level 4) the WAN Killer software was available [16]. Packet size was set to the default of 1,500 bytes. The traffic injector was the PC (Personal Computer) with IP 192.168.0.1, having the PC with IP 192.168.0.5 as destination. Latency was measured as the round trip time of ICMP (Internet Control Message Protocol) packets (OSI level 3) involving the PCs having IPs 192.168.0.2 and 192.168.0.6. Percentage packet loss was also measured for different ICMP packet sizes (32 and 2,048 bytes) through the same two PCs.

In addition, measurements were made for TCP connections and UDP communications, using Iperf software [17], permitting network performance results to be recorded. For a TCP connection, TCP throughput was obtained. For a UDP communication with a given bandwidth parameter, UDP throughput, jitter and percentage loss of datagrams were obtained. TCP packets and UDP datagrams of 1,470 bytes size were used. A window size of 8 kbytes and a buffer size of the same value were

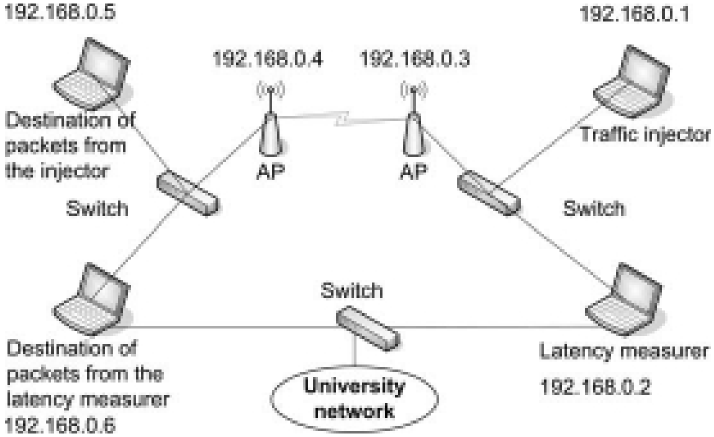


Fig. 43.1 Experimental laboratory setup scheme

used for TCP and UDP, respectively. A setup scheme similar to the one given in Fig. 43.1 was used, but having two PCs only. One, with IP 192.168.0.2 was the Iperf server and the other, with IP 192.168.0.6, was the Iperf client. Jitter, which can be seen as the smooth mean of differences between consecutive transit times, was continuously computed by the server, as specified by RTP (Real Time Protocol) in RFC 1889 [18]. This scheme was also used for FTP measurements, where FTP server and client applications were installed in the PCs with IPs 192.168.0.2 and 192.168.0.6, respectively.

Batch command files were written to perform the TCP, UDP and FTP tests. The results were obtained in batch mode and stored as data files in the client PC disk. Each corresponding PC had a second Ethernet network adapter to permit remote control from the official University network, via an additional switch.

43.3 Results and Discussion

Both access points AP-A and AP-B were configured, for every one of the standards IEEE 802.11a, b, g, with various fixed transfer rates. For every fixed transfer rate measurements were made, for both Exp-A and Exp-B, of latency and percentage packet loss for ICMP packet sizes of 32 and 2,048 bytes, with percentages of injected traffic varying from 0% to maximum values. In this way, data were obtained for comparison of the laboratory performances of IEEE 802.11 b (namely at 5.5 and 11 Mbps), 802.11 g and 802.11 a (namely at 12, 36 and 54 Mbps in both cases) links, measured at OSI levels 3 and 4 using the scheme of Fig. 43.1.

At OSI level 1 in Exp-B, for every one of the cases, the local and remote values of the signal to noise ratios SNR were recorded. The best SNR levels were observed for 802.11 g and 802.11a. The lowest noise levels were for 802.11 a. Similar trends were observed in Exp-A.

In each experiment type the measurements at OSI levels 3 and 4 permitted determination, for every standard and fixed transfer rate, of the maximum percentage of network utilization under quality conditions i.e. for values of latency and percentage packet loss less than 10 ms and 2%, respectively. Some sensitivity to AP type was observed. The average values obtained from Exp-A and Exp-B are shown in Table 43.1. We found that, for every standard, the maximum percentage of network utilization under quality conditions decreases with increasing fixed transfer rate. The best performance was, on average, for 802.11 a. The results obtained in Exp-A and Exp-B are illustrated for 802.11 a and 802.11 g at 36 Mbps in Figs. 43.2 and 43.3, respectively, where the data points were joined by straight lines.

Further measurements, using TCP connections and UDP communications at OSI level 4, were carried out for both Exp-A and Exp-B. In each experiment, for every standard and nominal fixed transfer rate, an average TCP throughput was determined. This value was used as the bandwidth parameter for every corresponding UDP test, giving average jitter and average percentage datagram loss. The results are shown in Figs. 43.4–43.6. In Fig. 43.4, polynomial fits were made for each AP

Table 43.1 Average maximum percentages of network utilization under quality conditions versus IEEE 802.11 a, b, g standards and fixed transfer rate (Mbps)

IEEE standard/ fixed transfer rate (Mbps)	5.5 (Mbps)	11 (Mbps)	12 (Mbps)	36 (Mbps)	54 (Mbps)
802.11b	70%	45%			
802.11g			60%	50%	40%
802.11a			65%	55%	40%

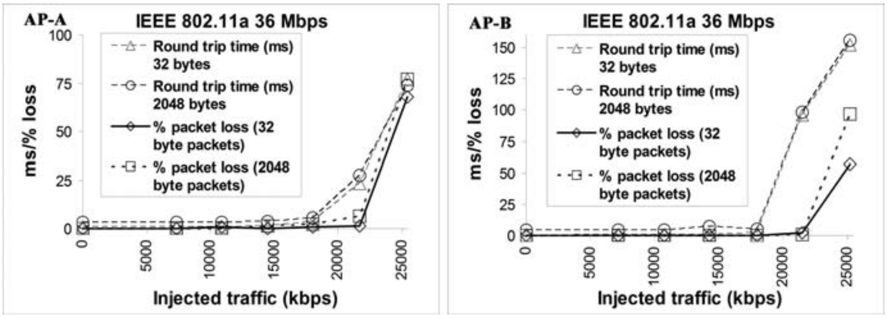


Fig. 43.2 Results for 802.11a, 36 Mbps; Exp-A and Exp-B

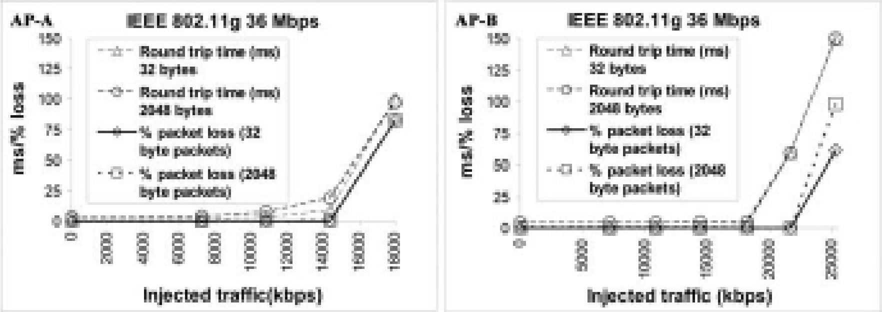


Fig. 43.3 Results for 802.11g, 36 Mbps; Exp-A and Exp-B

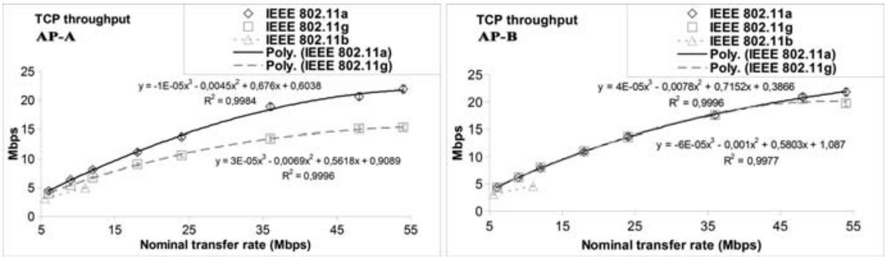


Fig. 43.4 TCP throughput results versus technology and nominal transfer rate; Exp-A and Exp-B

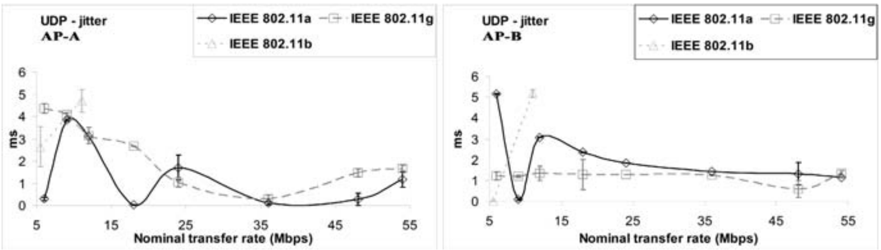


Fig. 43.5 UDP – jitter results versus technology and nominal transfer rate; Exp-A and Exp-B

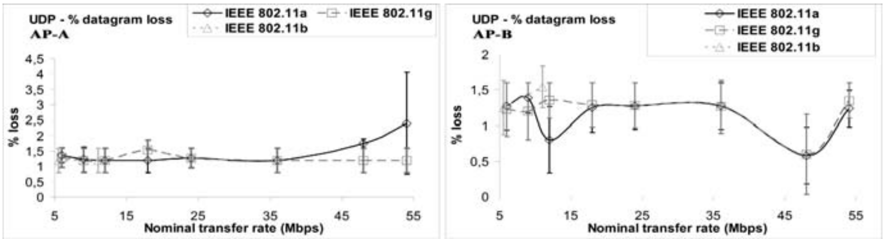


Fig. 43.6 UDP – percentage datagram loss results versus technology and nominal transfer rate; Exp-A and Exp-B

implementation of IEEE 802.11 a, g, where R^2 is the coefficient of determination. It follows that the best TCP throughputs are, by descending order, for 802.11 a, 802.11 g and 802.11 b. In Exp-A (Fig. 43.4), the data for 802.11 a are significantly higher (12–43%) than for 802.11 g. In Exp-B (Fig. 43.4), the data for 802.11 a are on average 2.5% higher than for IEEE 802.11 g. The best throughput performance was for AP-B. In Figs. 43.5 and 43.6, the data points were joined by smoothed lines. In Exp-A (Fig.43.5) the jitter data are rather scattered: jitter is, on average, higher for IEEE 802.11b (3.7 ms), g (2.3 ms). In Exp-B (Fig. 43.5), the jitter data show some fluctuations; jitter is, on average, higher for IEEE 802.11b (2.6 ms), a (2.1 ms). In both Exp-A and Exp-B (Fig. 43.6), generally, the percentage datagram loss data agree rather well for all standards. They are 1.3% and 1.2%, on average, respectively.

At OSI level 7, FTP transfer rates were measured versus nominal transfer rates configured in the APs for the IEEE 802.11b, g, a standards. Every measurement was the average for a single FTP transfer, using a binary file size of 100 Mbytes. The results from Exp-A and Exp-B are represented in Fig. 43.7. Polynomial fits to data are given. It was found that in both cases the best performances were, by descending order, for 802.11 a, 802.11 g and 802.11 b: the same trends found for TCP throughput. The FTP transfer rates obtained in Exp-A, using IEEE 802.11 b, were better than in Exp-B. The FTP performances obtained for Exp-A and IEEE 802.11a were slightly

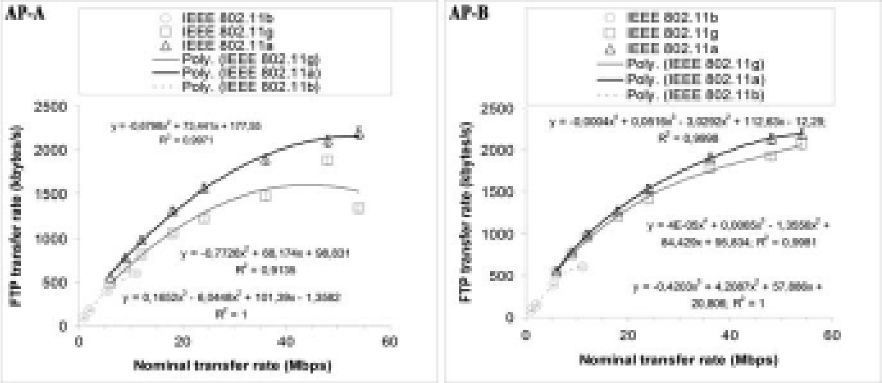


Fig. 43.7 FTP transfer rate results versus technology and nominal transfer rate; Exp-A and Exp-B

better in comparison with Exp-B. On the contrary, for Exp-A and IEEE 802.11 g, FTP performances were significantly worse than in Exp-B, suggesting that AP-B had a better FTP performance than AP-A for IEEE 802.11 g.

43.4 Conclusions

In the present work a simple laboratory arrangement was implemented that permitted systematic performance measurements of available access point equipments in IEEE 802.11 a, b, g point-to-point links. At OSI level 1, the best values of SNR were for 802.11 g and 802.11 a, while the lowest noise levels were for 802.11 a. At OSI levels 3 and 4, the measurements permitted to find the maximum percentages of network utilization under conditions of communications quality. Some sensitivity to AP type was observed. It was found that, for every standard, the maximum percentage of network utilization under quality conditions decreases with increasing fixed transfer rate. The best performance was, on average, for 802.11 a.

Through OSI level 4 the best TCP throughputs were found, by descending order, for 802.11 a, 802.11 g and 802.11 b. TCP throughputs were also found sensitive to AP type. The lower values of jitter were, on average, found for IEEE 802.11 a, and 802.11 g. For the percentage datagram loss, a reasonably good agreement was found for all standards. At OSI level 7, the measurements of FTP transfer rates have shown that the best performances were, by descending order, for 802.11 a, 802.11 g and 802.11 b: the same trends found for TCP throughput. FTP performances were also found sensitive to AP type. Additional performance measurements either started or are planned using several equipments, not only in laboratory, but also in outdoor environments involving, mainly, medium range links.

Acknowledgment Supports from University of Beira Interior and FCT (Fundação para a Ciência e a Tecnologia)/POCI2010 (Programa Operacional Ciência e Inovação) are acknowledged. We acknowledge Enterasys Networks for their availability.

References

1. Mark, J.W., Zhuang, W.: Wireless Communications and Networking. Prentice-Hall, Upper Saddle River, NJ (2003)
2. Rappaport, T.S.: Wireless Communications Principles and Practice, 2nd edn. Prentice-Hall, Upper Saddle River, NJ (2002)
3. Bruce III, W.R., Gilster, R.: Wireless LANs End to End. Hungry Minds, NY, (2002)
4. Chebrolu, K., Raman, B., Sen, S.: Long-distance 802.11b links: performance, measurement and experience. Proceedings of 12th Annual International Conference on Mobile Computing and Networking, pp. 74–85, Los Angels, CA, 23–29 September 2006
5. Raman, B., Chebrolu, K.: Experiences in using WiFi for rural Internet in India. IEEE Commun. Mag. **45**(1), 104–110 (2007)
6. IEEE Std 802.11-2007. IEEE standard for local and metropolitan area networks-specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. http://standards.ieee.org/getieee802. Accessed 10 October 2007
7. Schwartz, M.: Mobile Wireless Communications. Cambridge University Press, Cambridge (2005)
8. Sarkar, N.I., Sowerby, K.W.: High performance measurements in the crowded office environment: a case study. Proceedings of ICCT'06-International Conference on Communication Technology, pp. 1–4, Guilin, China, 27–30 November 2006
9. Pacheco de Carvalho, J.A.R., Gomes, P.A.J., Veiga, H., Reis, A.D.: Wi-Fi and very high speed optical links for data and voice communications. Proceedings of CISTI 2007-2^a Conferência Ibérica de Sistemas e Tecnologias de Informação, pp. 441–452, UFP, Universidade Fernando Pessoa, Porto, Portugal, 21–23 de Junho 2007
10. Pacheco de Carvalho, J.A.R., Gomes, P.A.J., Veiga, H., Reis, A.D.: Development of a University networking project. In: Putnik, G.D., Cunha, M.M. (eds.) Encyclopedia of Networked and Virtual Organizations, pp. 409–422. IGI Global, Hershey, PA, 2008
11. Pacheco de Carvalho, J.A.R., Veiga, H., Costa, R., Gomes, P.A.J., Reis, A.D.: A contribution to experimental performance evaluation of point-to-point WiMAX links. Proceedings of ISSPIT 2008-8th IEEE International Symposium on Signal Processing and Information Technology, pp. 150–153, Sarajevo, Bosnia and Herzegovina, 16–19 December 2008
12. Pacheco de Carvalho, J.A.R., Veiga, H., Gomes, P.A.J., Cláudia, F.F.P., Pacheco, R., Reis, A.D.: Experimental performance study of very high speed free space optics link at the University of Beira interior campus: a case study. Proceedings of ISSPIT 2008-8th IEEE International Symposium on Signal Processing and Information Technology, pp. 154–157, Sarajevo, Bosnia and Herzegovina, 16–19 December 2008
13. Pacheco de Carvalho, J.A.R., Gomes, P.A., Veiga, H., Ribeiro Pacheco, C.F., Marques, N.A., Reis, D.: A contribution to laboratory performance evaluation of IEEE 802.11 a, b, g point-to-point links: a case study. Proceedings of WCE 2009-World Congress on Engineering 2009, vol. I, pp. 884–888, London, 1–3 July 2009
14. Enterasys Networks. Roam About R2, RBT-4102 Wireless Access Points. http://www.enterasys.com. Accessed 20 December 2008
15. Allied Telesis. AT-8000S/16 Layer 2 Managed Fast Ethernet Switch. http://www.alliedtelesis.com. Accessed 20 December 2008
16. SolarWinds. Engineer's Toolset Network Traffic Generator: WAN Killer; http://www.solarwinds.net.
17. NLANR. http://dast.nlanr.net.
18. Network Working Group. RFC 1889-RTP: A Transport Protocol for Real Time Applications; http://www.rfc-archive.org.

Chapter 44
A Subnet Handover Scheme Based Communication System of Subway

Bing Chen and Xuefeng Yan

Abstract The wireless LAN based communication system of subway consist of a network of mobile nodes moving together in high velocity. We propose the idea of the new handover scheme based on mobile IP. The station (STA) taking the access point (AP) as the gateway, implements the multiple-hop subnet mobile IP by cooperation of the home agent, the foreign agent, interior network routing agent and reused tunneling. We implement a new subway communication system based on the above scheme, which includes four subsystem, vehicle subsystems, trackside subsystem, central control subsystem and roaming subsystem.

Keywords Mobile communication · roaming scheme · mobile IP · tunnel

44.1 Introduction

The traditional subway signal systems are simplex communications systems, which use the subway track circuit to transmit the information from the ground control center to the train. With the extensive application of wireless technology, wireless communication-based train control systems access to a vigorous development. That is, through the radio stations deployed on the train and next to the track, a continuous duplex communication between the train and the ground control system can be established.

The challenge we face is the moving WLANs demanding wireless access to heterogeneous networks and smooth handoff without disconnection, while they are moving. The mobile devices access the networks by the access points, and the corresponding requirement of these devices is to provide reliable service with high QOS. The mobile entity is a unit named subnet which includes many devices, not one device alone, linked by the networks, and is moving quickly as a whole.

B. Chen (✉) and X. Yan
The College of Information Science and Technology, Nuaa, No. 29, Yudao Street,
Nanjing, Jiangsu 210016, China
e-mail: cb_china@263.net; yxf@nuaa.edu.cn

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_44,

Mobility management is one of the most important issues for seamless service. Many researchers have put forward some technologies in this area and the IP-based mobility, such as Mobile IP (MIP) [1] and Cellular IP (CIP) [2], is the most important one. A considerable amount of solutions based on MIP, for example Mobile Regional Registration (MIP-RR) [3], Hierarchical Mobile IP(HMIP) [4], Paging Mobile IP (PMIP) [5] etc., are proposed with different focuses. However, none of them is a satisfactory solution for roaming many terminals in a subnet as a whole. Because when all mobile nodes (MNs) roam together form current AP to the next one, these technologies would process all the handoffs of massive links at a short time. when we have much handovers, so called handover strength will happen.

The above constraints suggest the design of a new roaming scheme to roam all the devices as a subnet in heterogeneous networks. The key issues are to ensure the mobile subnets serve continuous and high-speed data services regardless their movements. The new scheme is a network layer (L3) scheme which can support the subnet roaming with a high speed (more than 50 km/h). The traditional Mobile IP supports only one-hop between the MN and the exit of the tunnel, while in the new scheme can be multi-hops.

The rest of the paper is organized as follows. After introduction, we describe the related works in Section 44.2. Section 44.3 describes the proposed scheme, including the principle and some key issues. Section 44.4 introduced the communication system of our city based on the new scheme. The performance analysis and results of proposed scheme is described in Section 44.5. Finally, Section 44.6 summarizes the paper and future works.

44.2 Related Work

The roaming of mobile node is a prolific research field in wireless and mobile communications. The typical representatives which realized roaming in data link layer include the cellular technology such as GPRS and 3G, Wireless LAN based on the IEEE802.11 and IEEE802.16 etc. In [6], the transfer rate at MAC layer of GPRS and 3G is less than 64 and 144 Kbps respectively when moving at high speed. In [7], WiMAX, IEEE 802.16e, support roaming, which have the range of 5 km and transfer rate of 15 Mbps at physical layer. WLAN, IEEE802.11b/a/g, can achieve the speed of 11/54 Mbps transfer rate [8]. But WLAN can only support low-velocity moving. In 802.11f the devices can't roam among different networks [9].

Roaming at the network layer is independent on the networking technologies [10], the primary solutions include changing the IP address of the mobile host, specific host roaming, Mobile IP(MIP) and cellular IP (CIP) [2] etc. The first two of them don't fit for large-scale network. HAWAII [11] and cellular IP (CIP) are some kinds of micro-mobility protocol based on IP routing. The region registered MIP (MIP-RR) [3], Hierarchy MIP (HMIP), and intra-domain Management Protocol (IDMP) are the tunneling-based micro-mobility protocols, which mainly focus on the register, management and smooth handoff intra-networks [10]. The main technology of roaming inter-networks is mobile IP (MIP) [12–14]. However, MIP still suffers from some shortcomings, such as latency, triangle routing etc.

44.3 Proposed Subnet Handover Scheme

44.3.1 Principle of Subnet Handover Scheme

According to the loosely coupled principle, we proposed and designed a subnet handover scheme that allows a set of MNs to roam across different types of wireless networks with high velocity while providing seamless connectivity and high-speed communication.

Figure 44.1 shows the network configuration and the mobile subnets which includes several MNs and a STA. In the scenario the subnets move fast, keeping high-speed duplex communication with the fixed network F_net. In the Fig. 44.1, M_net1 and M_net2, two mobile subnets with different topologies, are Ethernet and wireless ad-hoc respectively. M_net1 and M_net2 establish wireless connections with AP through STA1 and STA2 respectively, and then communicate with nodes in F_net, which contains several subnets. Because M-net will connect to different AP of various subnets through STA when it is moving, we need manage the inter-network handover of it.

We propose a subnet roaming scheme based on extended mobile IP. Figure 44.2 illustrates how to roaming the mobile network in heterogeneous networks. Tthe node MN connects the remote node RN in F_net by passing through STA, AP, foreign agent, and home agent, and the distance from outlet of the tunnel to MN is more than one hop. The main idea of the scheme is to implement multiple hops mobile IP based on home agent, foreign agent, intra-network routing agent, reused tunneling,

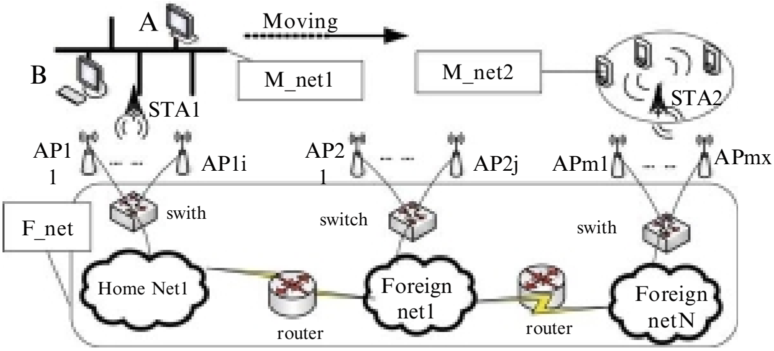


Fig. 44.1 Network configuration of proposed subnet roaming scheme

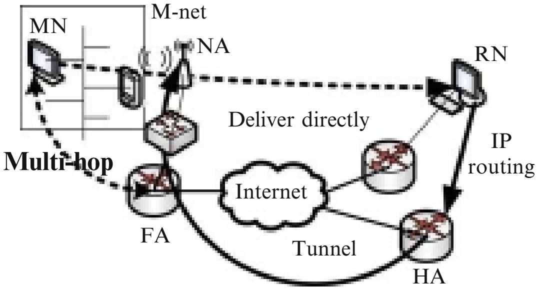


Fig. 44.2 Roaming the subnet in heterogeneous networks

and STA/AP gateway. By extending mobile IP protocol, we can breakthrough the limitation of roaming of single node, make all the nodes of a subnet move simultaneously while keeping duplex communications. The details are as follows:

1. Each MN in the M_net has a fixed IP address and the STA has two IP address, one is for the wireless and the other for Ethernet.
2. The STA and AP are promoted as network layer devices to provide the functionality of gateways. The data packets can be delivered from any node in M_net to any one in F_net and vice versa.
3. When the mobile subnet moving, a MAC algorithm named Well-time is adopted to choose the trigger time when STA and M_net moves in the same network, then the STA switching to next AP. A network layer handover occurs when M_net moves between different networks, the STA will get a its CoA and then the routing agent on AP will send registration information to foreign agent FA or home agent HA according to current position, so as to maintain the routing information from F_net to M-net.
4. A reused tunnel model for massive MNs is used to avoid establishing a tunnel for every MN separately. A references count is added to manage the reuse of the tunnel. When a MN connects to RN, the count will be added by one and reduced by one the other way round. Only if reference count decreases to zero, the tunnel will be deleted.
5. The HA decides how to deal with the packets according to the status of M_net. If M_net is not in the home network, HA will intercept all the packets send to M_net and forward them to the current FA of M_net.
6. The STA receives data packet from FA, whose MAC address of destination hardware is the same as itself or a broadcast address, and then removes the header of L2 to analyze the header of IP protocol. If the destination IP address is the IP address of one of the nodes in M_net (e.g., node MN), the IP data packets will be encapsulated into a Ethernet data packets and then deliver to MN. Conversely, the packets sent to F_net will be captured by STA through Ethernet port, and reassembled and delivered after analysis of the L3 header.

44.3.2 Gateway Model on L3 STA

The STA of mobile subnet, a wireless communication device, and the MNs in M_net which connects with STA, are treated as a moving unit called STAPC. The nodes in F_net access MNs by STA. Figure 44.3 shows the scenario that mobile subnet enters one region of F_net. The packets from PC2 to STAPC are received by the wireless port of STA, and then transmitted to its destination by Ethernet port. The data to PC2 are sent to STA, and then to wireless media by its wireless port, finally to PC2 by the routing of F_net. Therefore, STA should transmit packets between the wireless network and the Ethernet. For this purpose, STA, as a network layer device, should have the functionality of gateway to transfer packets between wireless network and wired network. The principles of L3 STA are as following.

44.3.3 Dynamic NAP Scanning and Finding Model

When the subnet is moving at a high velocity, one of the key issues of successful handover is find the neighbor APs and choose the right one as the next AP (NAP) with a shortest time. Usually, the STA scans neighbor APs in its region and put the result into the neighboring list, which is used to save ESSID and strength of the signals of neighboring APs, and sort them descending by their strength. After that, we can find the AP with the largest strength. If its strength is larger than that of current AP and the differences is more than a threshold, it will be the NAP, to which STA

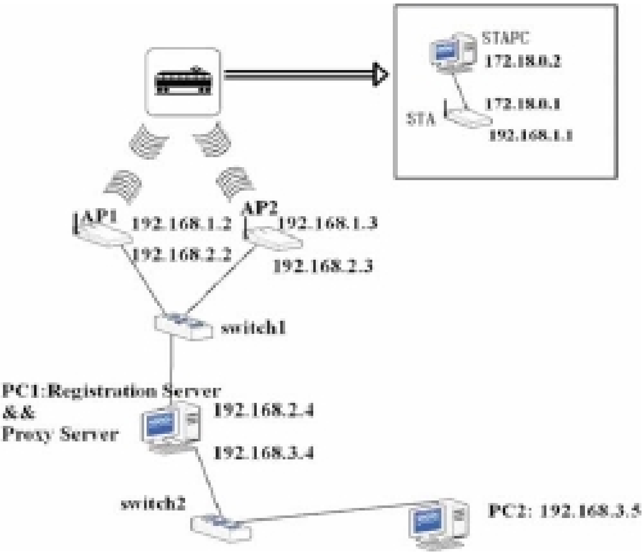


Fig. 44.3 The gateway model on STA

sends registration request information and waits for response. If the AP's response is received before the time threshold, the STA asserts that this AP can provide stable access services and we set ESSID of the STA as that of this AP; otherwise, it will discard it and choose the second highest AP as NAP. From that we can see that the time to choose the right NAP can be defined as:

$$T_{total} = N_{channels}{}^{*}T_{each} + T_{Process} \tag{44.1}$$

Where $N_{channels}$ is the number of channels, T_{each} represents the maximum scan waiting time of single channel, and $T_{process}$ denotes other process time needed. If the scanning time is larger than T_{each}, we should stop scanning and deal with the next channel. We can adopt the following ways to reduce the scanning time:

1. Reduce the number of channels should be scanned
2. Decrease the waiting time of single channel. If the possible APs of each channel are known, we stop scan it when all its APs are scanned
3. If the next candidate AP can be determinate in advance, we need not to scan the channels. If the moving of subnet is regularity, or down a fixed path like a train, we can use switching queues to avoid channels scan

44.3.4 The Reused Tunnel Model

The reused tunnel model avoids creating a tunnel for every MN. With the moving of MNs, the reused model maintains the tunnels by:

1. Creation of the tunnels, which includes real creation and virtual creation. Considering the situation shown in Fig. 44.2, the mobile node A enters the foreign network. After the foreign agent obtains the information of A, it will send registration request to HA. Next, HA processes the registration request and creates a tunnel to the FA. For each tunnel a reference counter is defined also, which counts the number of the processed requests. When the registration finished, the counter is set a value by one.
2. When mobile node B enters the foreign network. FA obtains B's information and registers to HA. HA will find that there exists a tunnel, therefore, the creation of real tunnel is no longer needed, and increasing of reference counter is enough. This is so called virtual creation. When multiple moving units come in, they share the existing tunnels similarly.
3. Deletion of tunnels also includes real deletion and virtual deletion. When mobile node moves to another foreign network, after the new FA registers it to HA, it decreases its reference counter. This is a virtual deletion of tunnel because if the reference counter is not 0, we will not delete the tunnel really, therefore, the packets to B are still transmitted through it. When B is moving to another network, the reference is decreased similarly. If the reference is 0, the tunnel will be deleted.

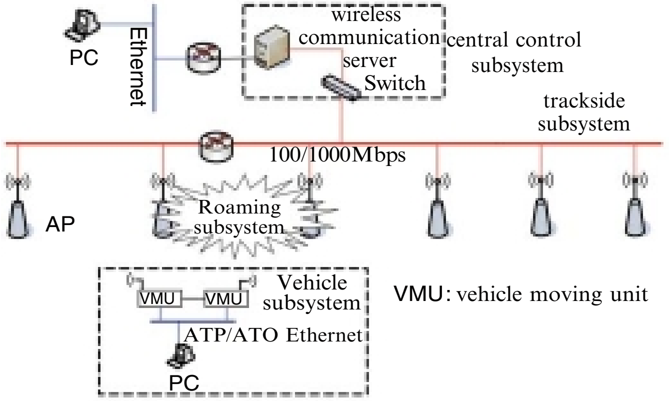


Fig. 44.4 Subnet-based subway communication system

44.4 The Subnet Based Communication System

Based on the subnet handover scheme proposed above, we designed and implemented a new subway communication system. As shown in Fig. 44.4. The system mainly consists of the following four components: vehicle subsystem, trackside subsystem, central control subsystem and roaming subsystem.

44.4.1 The Vehicle Subsystem

The vehicle subsystem cooperates with the trackside subsystems to transmit the information. It consists of two separate STA, which is installed in the driver's cabs at the head and tail of the train respectively. The two STAs communicate with each other by the Ethernet. Figure 44.5 is the hardware architecture of the STAs.

44.4.2 The Trackside Subsystem

Trackside subsystem is the bridge of vehicle subsystem and central control subsystem. It receives control information and data from the vehicle subsystem with different wireless cards, and forwards them to the wired network through the Ethernet interface. At the same time, the trackside subsystem also accepts control information from the wired network, and forwards to the vehicle subsystems if needed. The wireless interfaces of the trackside subsystems support the standard of IEEE802.11b/g, and the wired one is 100 Mbps Ethernet interfaces. As shown in Fig. 44.6, Trackside subsystem consists of AP, communication media, switches, routers etc.

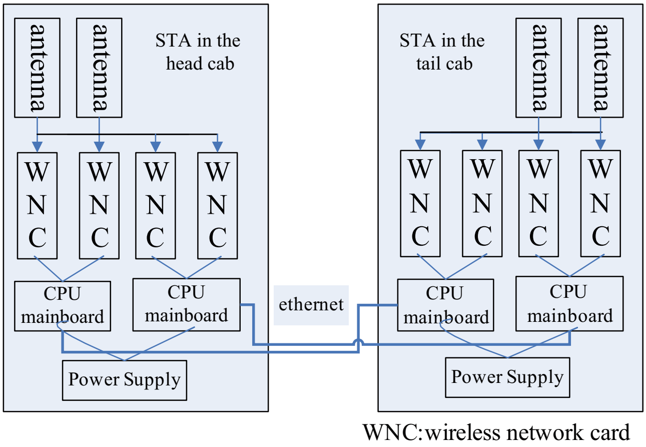


Fig. 44.5 The hardware architecture of the STAs

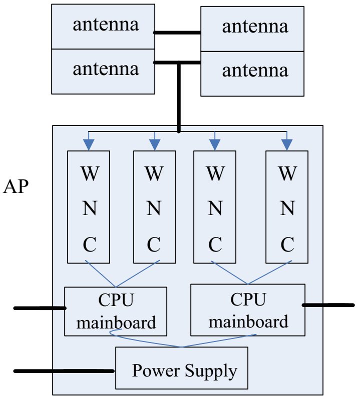


Fig. 44.6 The hardware architecture trackside subsystem

44.4.3 The Central Control Subsystem

The central control subsystem includes wireless communication server, core switches and routers. The wireless communication server is a centralized management and control equipment of the communication system of subway. The routers are the ground interface devices of a variety of applications, and the core switches are used in connecting the trackside wireless AP with the communication server. The wireless communications server has topology of the whole system as well as the location APs which are deployed next to the track, its main features are as following.

1. Radio frequency (RF) management. To ensure an optimal performance of the whole system, the wireless communication server can set RF working status of each AP automatically, and it also monitors the running status of radio waves of each AP.

2. Security policy. When the vehicle subsystem accesses to trackside wireless network, multi-identifier based authentication identities are provided, include 802.1, WEB authentication, MAC, SSID etc. All the STAS can handover between any APs without re-authentication.
3. Network management. All the configuration of the wireless network can be finished on the communication server. For example, open and maintain all the APs and vehicle subsystem, including radio spectrum, wireless security, access authentication, handover management.
4. Intrusion detection. Cooperating with the trackside APs, The wireless communication server can find locate and track the illegal intrusion. The illegal data stream can be filtered and alarm information can be reported to the user at the same time. The server can also take some countermeasures automatically, for example, changing the ratio power.

44.4.4 The Roaming Subsystem

The roaming subsystem requires STA, AP, home agent and the foreign agent cooperating with each other. In our realization, the STA, AP already contain the corresponding processing modules, which are described in Section 44.3. Figure 44.7 describes the procedure of the new communication system. After the initialization of the system, the MN (M-node) of mobile subnet sets up a link connecting with STA. After STA finishing the registration to FA or HA through AP and notifying M_node that the link is ready, the data transmission can be started.

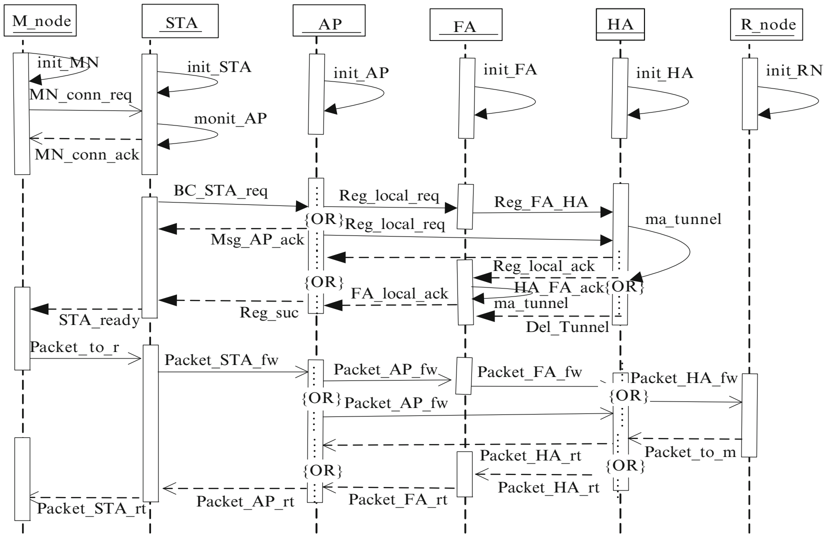


Fig. 44.7 The procedure of the new communication system

44.5 Performance Analysis

We carried out different types of test for the developed protocol. Figure 44.8 shows the data transfer rate of static test. According to the statistics from many experiments, the rate keeps 2.66 Mbps in the static situation, and the peak rate can reach 3 Mbps. Ninety five percentage or above of all these rates fall into the range form 2.4 Mbps to 3.0 Mbps.

The intra-region moving test and inter-region moving test were carried out outdoors. The distances among APs are 90 m. We tested the situations where the car moved in 10, 30 and 50 km/h respectively and continues for 2 h. The data transmission result of the third experiment (moving with 50 km/h) is shown in Fig. 44.9.

Figure 44.10 shows the data transmission rate when M_net mvoes between two subnets. The results show that the peak data transmission rate is up to 3.0 Mbps, and

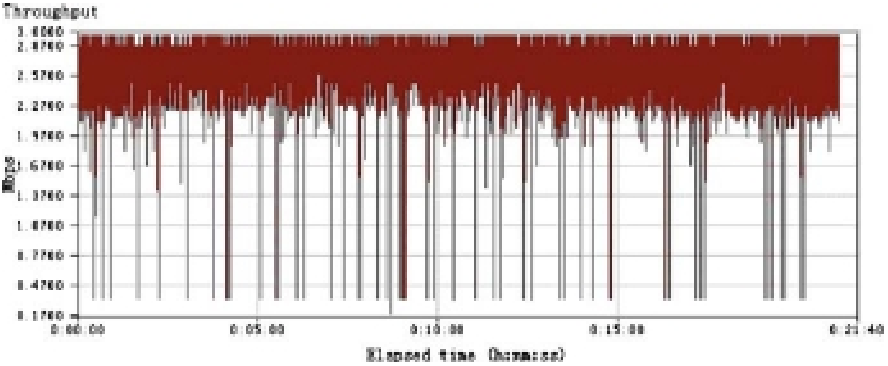


Fig. 44.8 Data transfer rate of UDP in static test (time = 0.5 h, and file size = 1,024 bytes

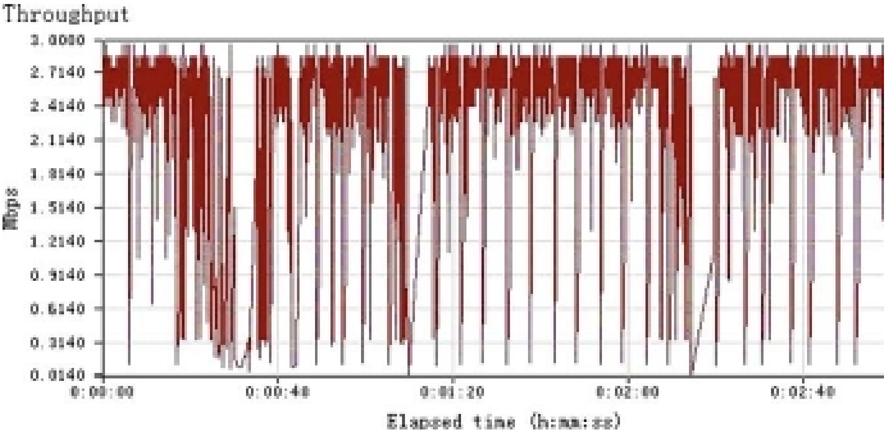


Fig. 44.9 The throughput of M_net1 moving in networks

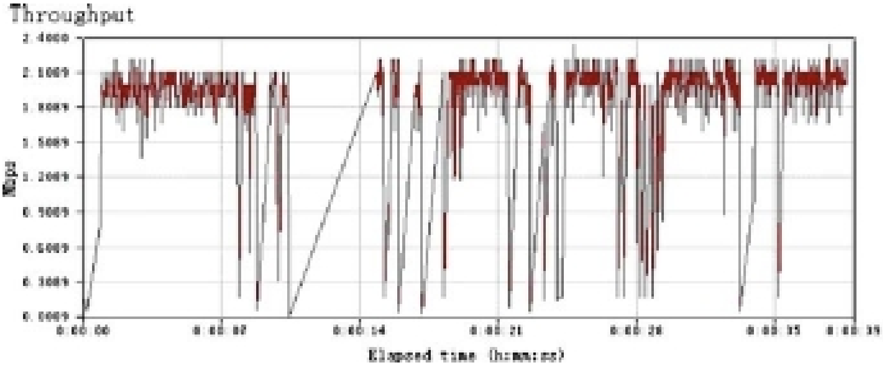


Fig. 44.10 The real time data transmission rate when STA moving from APli to AP2

the average rate roughly equals to 1.99 Mbps. Also, 95% of all the values of the rate are between 2.2 and 3.0 Mbps. From Fig. 44.10 we know the peak rate can reach 2.4 Mbps, and the average rate is 1.26 Mpbs.

From all of the tests results above, we can see that the wireless signals are influenced by the terrains, which will play a side effect on the switch, thus to connect the wrong AP.

44.6 Conclusion

The subnet handover scheme is a new scheme based on the Mobile IP which can roam a wireless subnet as a whole when moving with a high velocity. The distance between MN and the exit of tunnel can be multi-hop. The scheme is applied to the subway communication system, and it is promising for many emerging wireless application, such as data transmission and video service for the passengers on the train or plane. The future work will be focus on decreasing the handover latency and improving performance of the scheme. Double STA can be deployed to the mobile subnet as the gateways connecting M_net with AP, one for the data transmission and the other for scanning and finding the next AP.

References

1. Charles, E.: Perkins, Mobile IP: Design Principles and Practices. Addison-Wesley, Reading, MA (1998)
2. Campbell, A.T.: Design, implementation, and evaluation of cellular IP [J]. IEEE Person. Commun. **8**,42–49 (2000)
3. Gustaffon, E.: Mobile IPv4 Regional Registration [EB/OL]. http://www1.ietf.org/mail-archive/ietf-announce/Current/msg20878.html

4. Soliman, H., Castelluccia, C., El-Malki, K., Bellier, L.: Hierarchical mobile IPv6 mobility management. IETF (2005, http://www.ietf.org/rfc/rfc4140.txt)
5. Zhang, X., Castellanos, J.G., Cambell, A.T.: P-MIP:Paging extensions for mobile IP. IEEE Mob. Networ. App. **7**:127–141 (2002)
6. Etoh, M.: Next Generation Mobile Systems 3G and Beyond [M]. Wiley, New York (2005)
7. IEEE Std. For Local and Metropolitan Area Networks, IEEE 802.16e: Air Interface for Fixed and Mobile Broadband Wireless Access Systems[S]. IEEE Press, USA (2005)
8. IEEE 802.11b-1999. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band[S]. IEEE-SA Standards Board, USA (2000)
9. IEEE Std. 802.11f, IEEE trial-use recommended practice for multi-vendor access point interoperability via an inter-access point protocol across distribution systems supporting[S]. IEEE 802.11 Operation IEEE Press (2003)
10. Akyildiz, I.F., Jiang, X., Mohanty, S.: A survey of mobility management in next-generation all-IP-based wireless systems. IEEE J. Wireless Commun. **8**, 16–28 (2004)
11. Ramjee, R.: IP-Based access network infrastructure for next-generation wireless data networks [J]. IEEE Personal Commun. **8** (2000)
12. Perkins, C, : IP mobility support for IPv4[S], RFC3344 (August 2002.8)
13. Perkins, C, : IP Encapsulation within IP[S], RFC2003 (October 1996.9)
14. Lam, P.P., Liew, S.C., Lee, J.Y.B.: Cellular universal IP for nested network mobility[J]. Comput. Networ. **51**:3617–3631 (2007)

Chapter 45
PROMESPAR*: A High Performance Computing Implementation of the Regional Atmospheric Model PROMES

Juan E. Garrido, Enrique Arias, Diego Cazorla, Fernando Cuartero, Iván Fernández, and Clemente Gallardo

Abstract This paper describes the parallelization process of the code PROMES. The parallel code, called PROMESPAR, has been carried out under a distributed platform (cluster of PCs) and using Message Passing Interface (MPI) communication subroutines.

Keywords Regional atmospheric model · parallelization · message passing interface

45.1 Introduction

Climate change induced by human activities is one of the topics to which more attention is devoted to scientific research today. This is due, not only by the great complexity involved in the processes affecting the climate, but also to the threat involved in the serious impact that occurs on the economics and the environment in many parts of the planet. Three or 4 decades ago, it was believed that the oceans would be able to absorb the pollutants emitted by human activities; but today, maritime degradation is undeniable. Even more recently, the idea that humanity could induce a change in climate was a hypothesis that received little scientific support.

J.E. Garrido (✉), E. Arias, and D. Cazorla
Inst. Investigación en Informática. University of Castilla-La Mancha. 02071 Albacete, Spain
e-mail: Juanenrique.Garrido@uclm.es; Enrique.Arias@uclm.es; Diego.Cazorla@uclm.es

F. Cuartero
Inst. Investigación en Informática. University of Castilla-La Mancha. 02071 Albacete, Spain
e-mail: Fernando.Cuartero@uclm.es

I. Fernández and C. Gallardo
Inst. de Ciencias Ambientales. University of Castilla-La Mancha. 45071 Toledo, Spain
e-mail: Ivan.Fernandez@uclm.es; Clemente.Gallardo@uclm.es

*This work has been supported by National Project CGL2007-66440-C04.

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_45,

However, there is now a broad consensus among scientists, about the evidence of anthropogenic climate change and the need for better knowledge about likely developments in the following decades.

To simulate the climate, we use numerical models reproducing the main processes occurring in the five components of the climate system: Atmosphere, hydrosphere, geosphere, and biosphere, and the exchange of mass and energy between them. The results obtained by the models are evaluated and compared with the observed features of the climate in recent decades. Once it is found the quality level of the climate model is correct, we apply it to simulate potential changes in the climate, considering various scenarios of anthropogenic emissions of greenhouse gases and aerosols. Since this information, we can deduce the potential impact of climate change produced in such a hypothesis.

The history of weather forecasting is intimately associated to development of high performance and parallel computing [9].

In fact, thanks to the parallelization of weather prediction models, it is provided to scientists the ability to deal with longer simulations, to increase the spatial resolution, etc.

Throughout the last decade, several parallel approaches have been developed. Among them [10], we remark the based on vectorial multiprocessors such as CCM2 and its scalable variants [11, 17, 18], massively parallel computers (adaptation from the spectral model of the National Meteorology Centre) [25], distributed memory multiprocessors [5] (integrated prediction system) and passing messages [28], MM5 model [16] and another MM5 versions [23, 24], and the application of Grid technology (IrisGRID [3], CrossGrid [2], Climateprediction.net [1] program).

The paper is organized as follows. Section 45.2 introduces the regional atmospheric model PROMES, and in Section 45.3 the parallelization of PROMES is presented. The experimental results are outlined in Section 45.4. Finally, the conclusions and future work are commented in Section 45.5.

45.2 The Regional Atmospheric Model PROMES

PROMES is a regional atmospheric numerical model used in meteorological and climate research and also for weather forecasting. PROMES has been developed by MOMAC (MOdelizacin para el Medio Ambiente y el Clima) research group at the University of Castilla-La Mancha (UCLM) and the Complutense University of Madrid (UCM), and it was originally described in [7]. It is a hydrostatic limited-area model with sigma levels as vertical coordinates [26] and Lambert conic projection for the horizontal coordinates [4]. The spatial arrangement of variables follows the so called Arakawa-C grid [22]. The studied region is divided on a set of vertical columns and each column in turn is divided into several levels. Thus, the state of the atmosphere is defined at each time in a finite number of grid-boxes arranged in a mesh (Fig. 45.1)

PROMES model uses a split–explicit time integration scheme, based on [13]. The different terms of the primitive equations that govern atmospheric dynamics are

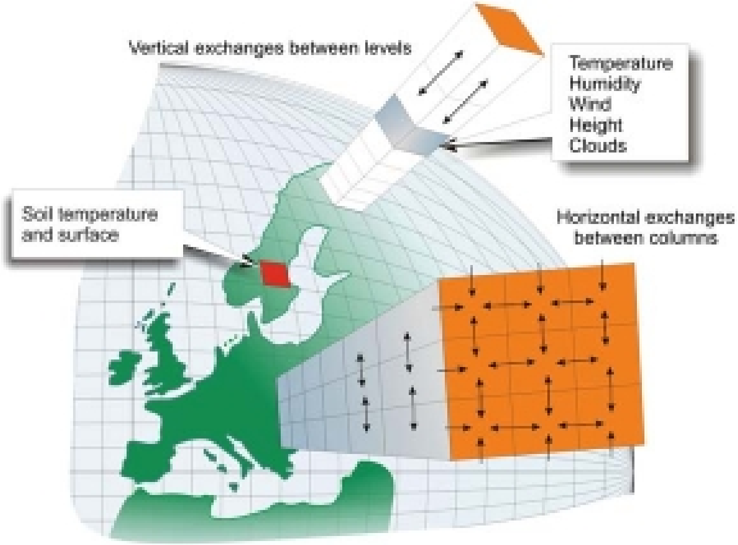

Vertical exchanges between levels
Temperature
Humidity
Wind
Height
Clouds
Soil temperature and surface
Horizontal exchanges between columns

Fig. 45.1 Grid of calculus

integrated with time steps depending on their typical time-scale. Numerical schemes based on finite differences are used for solving those terms. A forward–backward scheme is applied to the gravity waves terms, a cubic–spline upstream method is used to solve the advection terms and a fourth-order explicit scheme is employed for horizontal diffusion. The needed lateral boundary values are updated from analysis or Global Circulation Model data. The vertical interpolation of the large scale variables to model levels follows the method described in [14]. The model variables are relaxed to the external information in a contour band following [8].

The physical parameterizations included in the version of PROMES used in this study were those described below. The absorption and scattering of shortwave radiation is based on [4], and longwave radiation processes are parameterized according to [15, 27]. Shortwave heating and infrared cooling is calculated according to [27]. Turbulent vertical exchange of the prognostic variables in the planetary boundary layer (PBL) is modelled as proposed by [29] by using four regimes: stable, mechanical turbulence, forced convection and free convection. For the first three cases a local K-theory parameterization is applied [6]. In the case of free convection, a non-local scheme is used. Outside the PBL the vertical diffusion is also computed using K-theory. PROMES takes into account the exchanges of energy and water between soil, vegetation and atmosphere by using the landsurface scheme called SECHIBA [12]. SECHIBA gets the atmospheric forcings from PROMES and calculates sensible and latent heat fluxes. At each grid-box, bare soil and up to seven vegetation types are permitted to be present in different proportions. The soil water content is calculated in two layers, meanwhile soil temperature is computed in seven layers [20]. The resolved-scale cloud formation and its associated precipitation processes are modelled according to [19]. Sub-grid scale convective clouds and

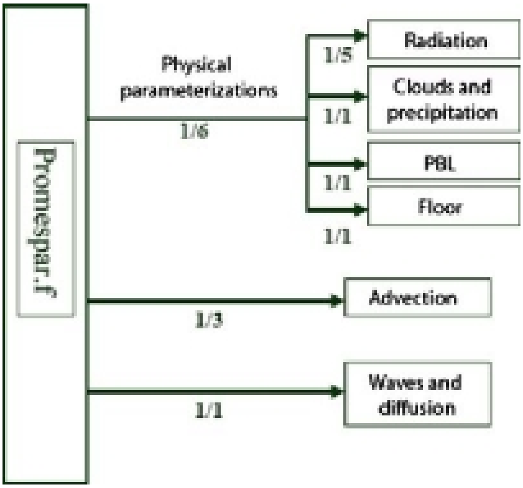


Fig. 45.2 General squeme of PROMES code

their precipitation are parameterized using the [21] method. As a summary of this PROMES description, a scheme of the model code is shown in Fig. 45.2.

As can be seen above, PROMES solved a set of equations and several complex parameterizations that involved a huge number of calculations. Therefore, whether an accurate solution is needed, parallel platforms to solve the problem are essential in order to obtain the results in a reasonable time.

45.3 PROMESPAR: A Distributed Memory Implementation of PROMES

As it was previously commented, in order to obtain a very accurate solution in a reasonable time, it is necessary the use of parallel platforms. In this paper, a distributed memory implementation of PROMES code, called PROMESPAR, is presented.

The parallelization of PROMES consists on dividing the domain on a set of subdomains getting out the work to carried out into the different processors (see Fig. 45.3). Once the domain has been divided the processors just exchange the frontier information.

In order to obtain an equally load balancing, a constrain is applied to the size of the subdomain and the number of processor to be used. This constrain is given by Eq. (45.2)

$$ProcXBlockSize = \left(\frac{OrXmatSize}{XsizeProc}\right) \pm XBorderSize \tag{45.1}$$

$$ProcYBlockSize = \left(\frac{OrYmatSize}{YsizeProc}\right) \pm YBorderSize \tag{45.2}$$

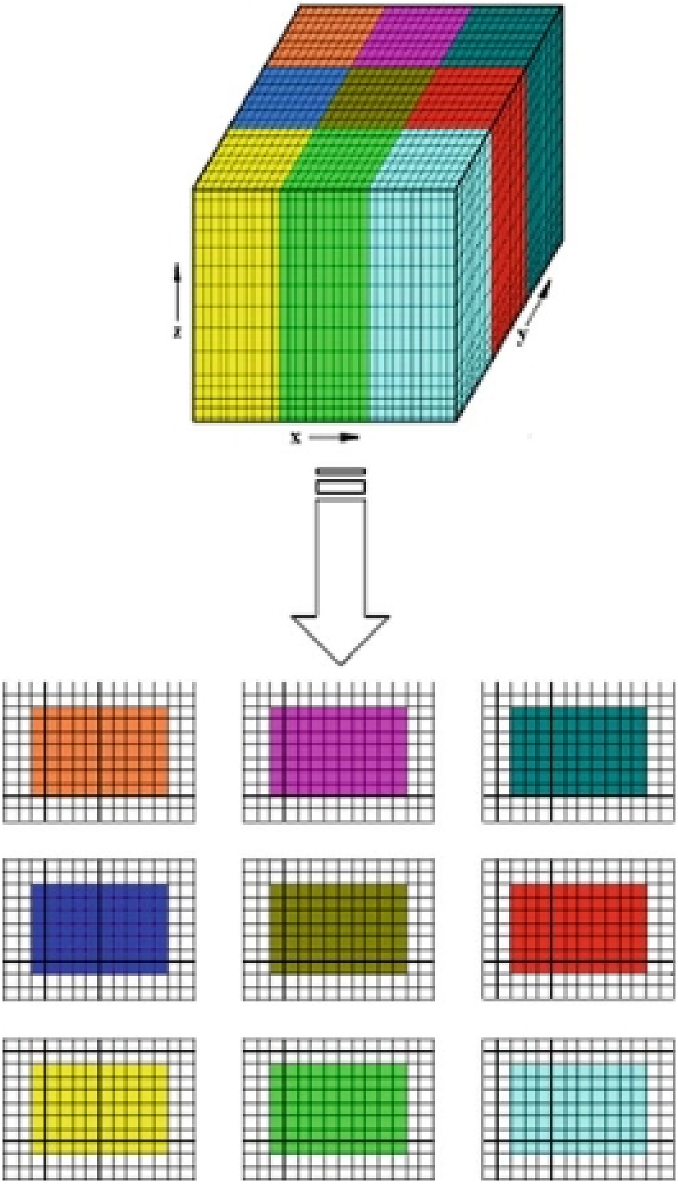


Fig. 45.3 Squeme of splitting the domain into subdomains

where *ProcXBlockSize* and *ProcYBloclSize* mean the size of blocks for each processor at X or y coordinate, respectively, which is computed from the original dimension of the matrix (*OrXmatSize* and *OrYmatSize*) and the number of processors by each coordinate (*XsizeProc* and *YsizeProc*), and taking into account the boundary conditions (*XBorderSize* and *YBorderSize*).

However, processor 0 has additional tasks due to the fact that it acts as master reading initial conditions, boundary values for the domain, etc from files.

In any case, the good load balancing could be affected mainly by two factors:

- **Static imbalance**. Those processors whose subdomains contain maritima zones have less computational load. This circumstance is due to the fact that the computations needed for solving the forecasting model are simplest in this kind of cells (some physical phenomena as the effect of orography, heat exchange with the masses of plants, etc are not taken into account).
- **Dynamic imbalance**. This kind of imbalance is devoted by the initial conditions. For instance, the effect of solar radiation could vary if a cloudy day or a sunny day is considered. These effects are unpredictable. However, other effects as the solar radiation during the night are predictable.

In the implementation of PROMESPAR the following libraries have been considered:

- MPI: Messing Passing Interface use for communications purpose. This library supports the communication between the different processors of the distributed memory platform.
- NETCD: NetCDF (network Common Data Form) is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.
- IOPSL: Library for input/output operations with meteorological data.
- Other physical libraries: computation of solar radiation, heat exchange ground-atmosphere, etc.

Figure 45.4 represents the workflow of the parallel implementation of PROMES, PROMESPAR.

The workflow in Fig. 45.4 is followed by each processor, and the barriers on Fig. 45.4 mean communication or synchronization takes amount the different processors.

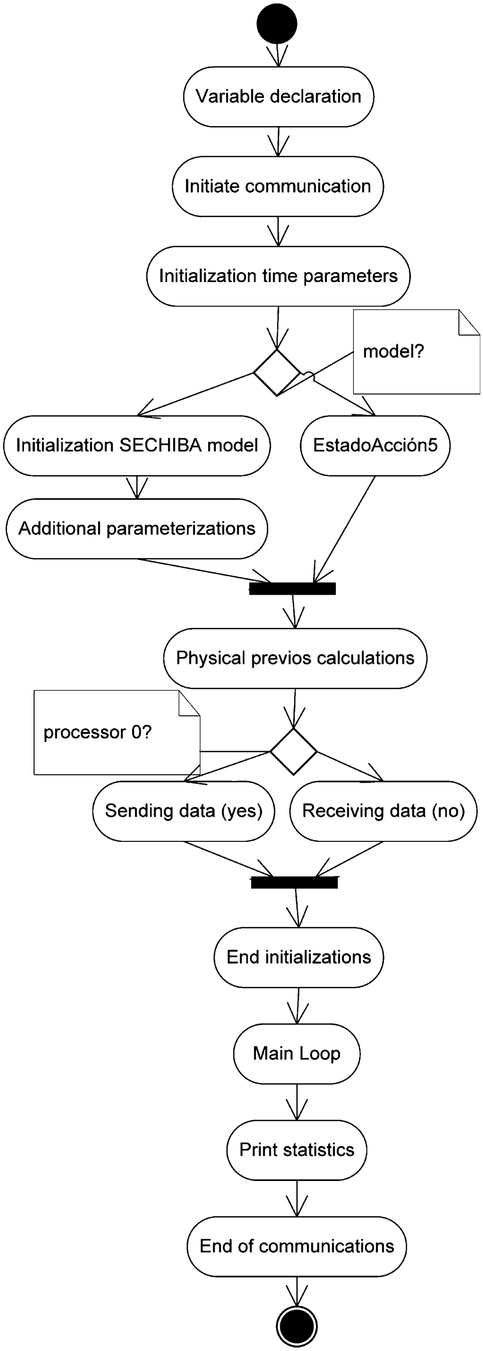
45.4 Experimental Results

The experimental results have been obtained taken into account 24 h of simulation. The distributed memory implementation has been run into a cluster of PCs with 16 Intel processors at 1.8 GHz, each one with 512 MB of main memory and interconnected by a Myrinet Network using NFS file system.

The performance obtained in the parallel implementations are evaluated in terms of

- Execution time: Time spent in order to solve the problem.
- Speed-up: The ratio of the time taken to solve a problem on a processor to the time required to solve the same problem on a parallel computer with p identical processors.

Fig. 45.4 Workflow of PROMESPAR



- Efficiency: A measure of the fraction of time for which a processor is usefully employed; it is defined as the ratio of the speed-up to the number of processors.

Most time consuming has been spent at main loop where are contained the most computational cost operations. In particular, apart from send and receive operations for communication purpose, physical operations are invoked. These operations are shown at Figs. 45.2 and 45.4.

The experimental results considered in this section take into account a 24 h simulation, which is equivalent to carry out 2881 iterations of main loop.

Figures 45.5–45.7 show the results of the previous experiment (24 h simulation) in terms of execution time, speed-up and efficiency.

From the experimental results, the main conclusion is that the best results, in terms of execution time has been obtained considering eight processors. However, in terms of speed-up and efficiency best results are obtained for two processors. This is a normal circumstance due to the influence of the communications. However, for this particular applications the main goal is to reduce the execution time.

As it was previously commented, the most time consuming of PROMESPAR code is spend on main loop. Figure 45.8 show a detailed study of the time spend on main loop. It is possible to observe that *fisicapal*, *Coriolis* and *Diffusion* functions spent the most quantity of time, and obviously the parallelization approach allows to reduce this execution time, overall from one to two processors. Anyway, the reduction of execution time results quite good.

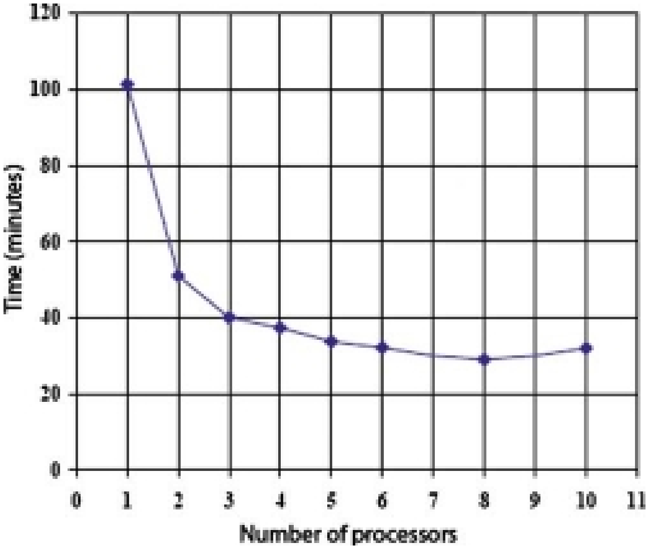


Fig. 45.5 Execution time of PROMESPAR

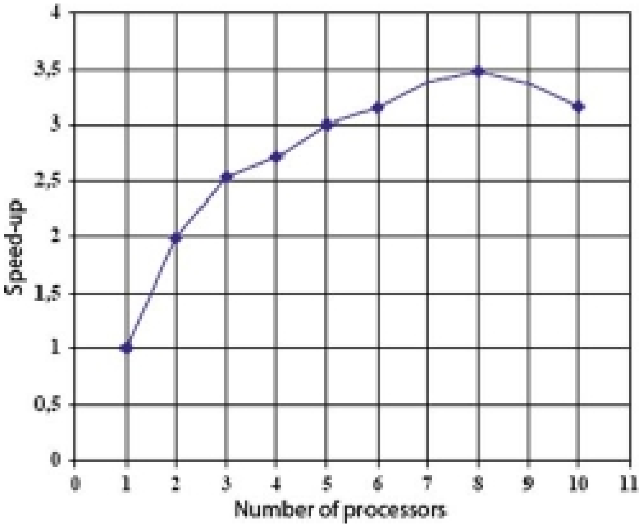


Fig. 45.6 Speed-up of PROMESPAR

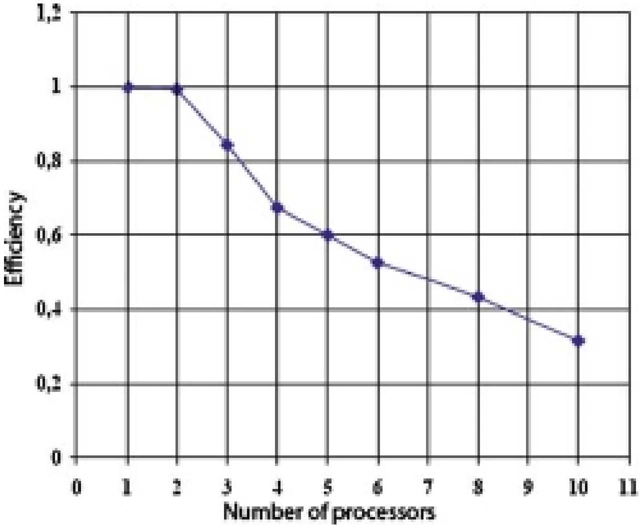
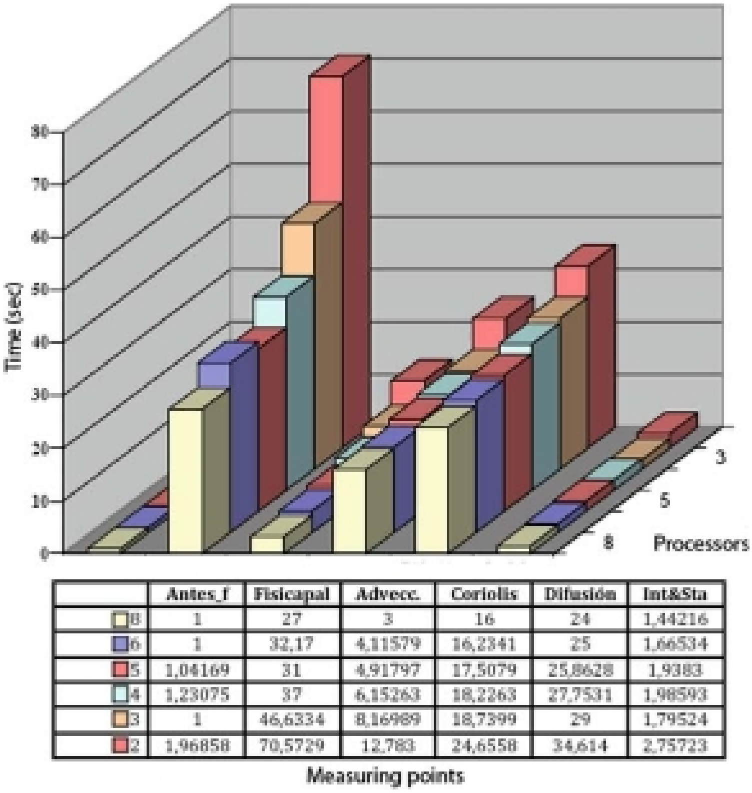


Fig. 45.7 Efficiency of PROMESPAR

45.5 Conclusion

PROMES is a mesoscale regional atmospheric model developed, among others, by some of the authors of this paper. However, due to the high time consuming by PROMES code and the necessity of having more accurate results, both circumstances justify the used of parallelism. In this paper, a distributed memory



	Antes_f	Fisicapal	Advecc.	Coriolis	Difusión	Int&Sta
🟨8	1	27	3	16	24	1,44216
🟦6	1	32,17	4,11579	16,2341	25	1,66534
🟥5	1,04169	31	4,91797	17,5079	25,8628	1,9383
⬜4	1,23075	37	6,15263	18,2263	27,7531	1,98593
🟧3	1	46,6334	8,16989	18,7399	29	1,79524
🟥2	1,96858	70,5729	12,783	24,6558	34,614	2,75723

Fig. 45.8 Execution time of the main loop of PROMESPAR for an hour simulation

implementation of the regional atmospheric model PROMES has been carried out. This parallel implementation is called PROMESPAR.

The experimental results show a dramatically execution time reduction by means of the use of a parallel platform considering the same configuration that the original PROMES code. These results leads to think that either longer or more accurate simulations could be carried out spending the same time, or more complex models could be considered. In fact, the authors are extending PROMES code in order to be able of making climate change studies. Climate change studies consider 100 years simulations spending, obviously, lot of time and then if the researchers want to provide conclusions from these studies the use of parallelism becomes essential.

References

1. Climateprediction.net. http://www.climateprediction.net/.
2. Crossgrid. http://www.crossgrid.org/.
3. Irisgrid. http://www.irisgrid.es.

4. Anthes, R.A., Hsie, E.-Y., Kuo, Y.-H.: Description of the Penn State/NCAR Mesoscale Model Version 4 (MM4). NCAR Technical Note 282. NCAR, Boulder, CO 80307 (1987)
5. Barros, S.R.M., Dent, D., Isaksen, L., Robinson, G., Mozdzynsky, G., Wollenweber, F.: The ifs model: a parallel production weather code. Parallel Comput (21), 1621–1638 (1995)
6. Blackadar, A.K.: Modeling the nocturnal boundary layer. In: American Meteorological Society, Boston (ed.) Proceedings of the Third Symposium on Atmospheric Turbulence, Diffusion and Air Quality, pp. 46–49 (1976)
7. Castro, M., Fernańdez, C., Gaertner., M.A.: Descrition of a meso-scale atmospheric numerical model. Mathematics, climate and environment, pp. 230–253 (1993)
8. Davies, H.C.: A lateral boundary formulation for multilevel prediction models. Quart. J. Roy. Meteor. Soc. 102, 405–418 (1976)
9. Drake, J., Foster, I.: Introduction to the special issue on parallel computing in climate and weather modelling. Parallel Comput. (21), 1539–1544 (1995a)
10. Drake, J., Foster, I.: Special issue on parallel computing in climate and weather modelling. Parallel Computing (21) (1995b)
11. Drake, J., Foster, I., Michalakes, J., Toonen, B., Worley, P.: Design and performance of a scalable parallel community climate model. Parallel Comput. (21), 1571–1591 (1995)
12. Ducoudr, N., Laval, K., Perrier, A.: SECHIBA, a new set of parameterizations of the hydrologic exchanges at the landatmosphere interface within the LMD atmospheric general circulation model. J. Climate **6**, 248–273 (1993)
13. Gadd, J.J.: A split explicit integration scheme for numerical weather prediction. Q. J. R. Meteorol. Soc. **104**, 569–582 (1978)
14. Gaertner, M.A., Castro, M.: A new method for vertical interpolation of the mass field. Mon. Wea. Rev. **124**, 1596–1603 (1996)
15. Garand, L.: Some improvements and complements to the infrared emissivity algorithm including a parameterization of the absorption in the continuum region. J. Atmos. Sci. **40**, 230–244 (1983)
16. Grell, D.O., Dudhia, J., Stuffer, D.R.: A description of the fifth-generation penn state/ncar mesoscale model (mm5). Ncar/tn-398+str, National Center for Atmosphere Research, Boulder, CO (1994)
17. Hack, James J., Rosinski, James M., Williamson, David L., Boville, Byron A., Truesdale, John E.: Computational design of the ncar community climate model. Parallel Comput. (21), 1545–1569 (1995)
18. Hammond, S., Loft, Richard D., Dennis, John M., Sato, Richard K.: Implementation and performance issues of a massively parallel atmospheric model. Parallel Comput. (21), 1593–1619 (1995)
19. Hsie, E.Y, Anthes, R.A., Keyser, D.: Numerical simulation of frontogenesis in a moist atmosphere. J. Atmos. Sci **41**, 2581–2594 (1984)
20. Jacobsen, I., Heise, E.: A new economic method for the computation of the surface temperature in numerical models. Beitr. Phys. Atmosph. **55**, 128–141 (1982)
21. Kain, J.S., Fritsch, J.M.: Convective parameterization for mesoscale models: The KainFritsch scheme. The Representation of Cumulus Convection in Numerical Models. Meteor. Monogr. No. 46, Amer. Meteor. Soc. 165–170 (1993)
22. Mesinger, F., Arakawa, A.: Numerical methods used in atmospheric models. GARP Publication Series, 17:1–64 (1976)
23. Michalakes, J.: MM90: A scalable parallel implementation of the Penn State/NCAR Mesoscale Model (MM5). Parallel Computing, 23:14, 2173–2186 (1997)
24. Michalakes, J., Canfield, T., Nanhyndiah, R., Hammond, S., Grell, G.: Parallel implementation, validation and performance of mm5. In: Proceedings of the Sixth ECMWF Workshop on the Use of Parallel Processors in Meteorology, World Scientific, River Edge, NJ pp. 266–276 (1995)
25. Sela, J.G.: Weather forecasting on parallel architectures. Parallel Comput (21), 1639–1654 (1995)
26. Shuman, F.G., Hovermale, J.B.: An operational sixlayer primitive equation model. J. Appl. Meteor, **7**, 525–547 (1968)

27. Stephens, G.L.: Radiation profiles in extended water clouds ii: parameterizaton schemes. J. Atmos. Sci. **35**, 2123–2132 (1978)
28. Wehner, M.F., Mirin, A.A., Eltgroth, P.G., Dannevik, W.P., Mechoso, C.R., Farrara, J.D., Spahr, J.A.: Performance of a distributed memory finite difference atmospheric general circulation model. Parallel Comput. (21), 1655–1675 (1995)
29. Zhang, D.L., Anthes, R.A.: A high resolution model of the planetary boundary layer sensitivity test and comparisons with sesame79 data. J. Appl. Meteor. **21**, 1594–1629 (1982)

Chapter 46
Transparent Integration of a Low-Latency Linux Driver for Dolphin SCI and DX

Rainer Finocchiaro, Lukas Razik, Stefan Lankes, and Thomas Bemmerl

Abstract High-speed interconnects like Dolphin's SCI and DX fulfil even high communication performance requirements. One of the prerequisites, though, is that the communication software must be either based on IP sockets or specifically adapted to the interconnect. Software written directly for Ethernet, arguably the most widespread interconnect today, cannot profit from this fast hardware. In this article, we present a Linux driver that fills this gap by allowing transparent usage of Dolphin hardware. ETHOM provides an Ethernet interface and makes use of the lowest message passing layer of Dolphin's driver stack in order to exchange Ethernet frames. It enhances the functionality of SCI and DX by offering an Ethernet and with that an IP interface.

Keywords Ethernet · SCI · Dolphin DX · Linux · TIPC

46.1 Introduction

Computational power has always been a scarce resource and prognoses predict that this situation will not change any time soon. While computer performance increases, the *demand* for more computational power increases at least at the same pace.

Until very recently, CPUs as the main component of a computing system grew more powerful by raising the clock frequency. Today parallelism in the form of additional cores per die adds to the performance increase. From a hardware point of view, the next level of parallelism is the gathering of single computers to form a cluster. Traditionally, the single computers – called nodes – in these clusters were connected by Ethernet in one of its incarnations. Concerning the software, the predominant protocol used on top of Ethernet is the TCP/IP stack. With software running on

R. Finocchiaro (✉), L. Razik, S. Lankes, and T. Bemmerl
Chair for Operating Systems, RWTH Aachen University, Kopernikusstr. 16,
52056 Aachen, Germany
e-mail: finocchiaro@lfbs.rwth-aachen.de; razik@lfbs.rwth-aachen.de;
lankes@lfbs.rwth-aachen.de; bemmerl@lfbs.rwth-aachen.de

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_46,

the cluster that communicates intensively, the network more and more becomes the limiting factor of overall cluster performance.

So, there are two problems to cope with: (1) Networking hardware in the form of Gigabit Ethernet is too slow for several purposes; 10 Gigabit Ethernet is still in the beginnings and rather expensive. (2) Then, TCP/IP is a protocol suite designed for communication in wide area networks, offering elaborate mechanisms for routing, to deal with even extensive packet loss, etc. It is not so well suited for clusters.

To overcome these problems, there are mainly two approaches in order to allow faster communication (latency and bandwidth wise):

1. Usage of high-speed networks, each having their own low-level programming interface (API), most providing an implementation of the POSIX socket API, and some offering an IP interface. Examples of these networks include InfiniBand [10], Myrinet [15], QsNet [16], SCI [3], and Dolphin DX [4]. An IP interface for Dolphin DX has been presented in [12].
2. Replacing the software layer TCP, UDP – and sometimes IP as well – while keeping the Ethernet hardware. Examples of these replacement protocols include SCTP (Stream Control Transmission Protocol [7]), DCCP (Datagram Congestion Control Protocol [11]), UDP-Lite [13], AoE (ATA over Ethernet [9]), and TIPC (Transparent Interprocess Communication Protocol [14, 17]).

Being developed originally at Ericsson, the abovementioned TIPC has its origin in the telecommunication sector, but provides some characteristics making it suitable for high performance computing (HPC) with clusters, such as an addressing scheme supporting failover mechanisms and the prospect of less overhead for exchanging data within a cluster. TIPC is the transport layer of choice of the Kerrighed project [18], where it is used for kernel to kernel communication. Currently it cannot make use of high-speed networks like InfiniBand, SCI, or DX, as neither do they provide an Ethernet interface, nor does TIPC provide a specialised "bearer", which is the adaptation layer between TIPC and a native network interface.

A first approach to enable TIPC to make use of high-speed networks is described in [5], where we elaborate on ETHOS, an Ethernet driver built using Linux kernel-space UDP sockets to send and receive data. ETHOS therefore directly supports almost all high-speed interconnects. Measurements with ETHOS on top of SCI and InfiniBand show significantly higher bandwidth and lower latency than Gigabit Ethernet.

In order to further reduce communication latency, we decided to sacrifice compatibility with other high-speed interconnects and use the next lower software layer available in the Dolphin Express stack, the *Message Queue Interface*. Using this interface, *ETHOM (ETHernet Over Message-Queue driver)* provides an Ethernet interface for SCI and Dolphin DX hardware. Therefore, in addition to the TCP/UDP-Sockets already provided by the Dolphin software stack, ETHOM offers an Ethernet interface, enabling interface bonding, bridging and other layer 2 kernel features, as well as (IP-)Routing for the SCI and Dolphin DX interconnects. Furthermore, TIPC is enabled to make use of these two network technologies leveraging its Ethernet bearer.

In the next section, we give a short overview about the hardware that we enable to be used as Ethernet replacement. Section 46.3 elaborates on the design and the architecture of our driver and in Section 46.4, we provide some basic experimental results. We conclude with the current status in Section 46.5.

46.2 Dolphin's High-Speed Interconnects

46.2.1 Scalable Coherent Interface (SCI)

The Scalable Coherent Interface [1, 8] is an established interconnect technology for transparent communication on the memory access level and/or the I/O read/write level. It maps (parts of) the physical address spaces of the connected nodes into one global address space, which allows to export and import memory and access it transparently via programmed input/output (PIO), or explicitly using direct memory access (DMA) transfers. Cache coherency between the nodes is supported by the standard, but not via I/O interfaces like PCI. The nodes are connected in multidimensional torus topologies without a central switch, as each host adapter also switches packets between its multiple links.

The current SCI hardware generation (D352) achieves remote store latencies starting at 220 ns and a maximum bandwidth of 334 MiB/s per channel.

46.2.2 Dolphin DX

The Dolphin DX interconnect [4] is based on the protocols for the Advanced Switching Interface (ASI). As such, it also couples buses and memory regions of distributed machines, but is designed for PCI Express and not for coherent memory coupling. Also, it does not use distributed switching like SCI; instead, all nodes connect to a central switch. Current switches offer 10 ports, and can be scaled flexibly.

Nevertheless, DX offers many of the same features as SCI from a programmers perspective, namely transparent PIO and DMA access to remote memory and remote interrupts. This makes it possible to integrate it into the existing software stack for SCI, offering the same APIs as for SCI.

The performance of DX has significantly improved compared to SCI for both, PIO and DMA transfers. The latency to store 4 bytes to remote memory is 40 ns, while the bandwidth reaches about 1.397 GiB/s already at 64 bytes transfer size.

46.2.3 Dolphin Software Stack

The SISCI API [2] is the basic and most efficient possibility to use SCI or DX as high-speed interconnect. SISCI is a shared-memory programming interface that makes the features of the SCI and DX interconnects accessible from user space. It consists of a user-space shared library (`libsisci`), which communicates with the SISCI kernel driver via `ioctl()` operations to create and export shared memory segments, map remote memory segments to the address space of the calling process, send and wait for remote interrupts, and perform DMA transfers from and to remote memory segments.

These means allow processes running on different machines to create common, globally distributed shared memory regions and read and write data from and to there either via PIO or DMA operations. Synchronisation can be performed via shared memory or via remote interrupts.

To obtain optimal communication performance, data transfers need to be aligned to suitable SCI packet and buffer sizes (16, 64 and 128 bytes), and remote read operations should be avoided except for very small data sizes.

SISCI does not provide means to pass messages between processes except for writing to some shared memory location and synchronising via either shared memory or remote interrupts. Therefore, based on this shared memory interface, Dolphin supplies a thin software layer for communication via message queues (`MBox/Msq`). It allows to establish uni-directional communication channels between machines which can be operated via simple `send()` and `recv()` operations. This software layer takes care of alignment, data gathering, error checking and so forth, and offers different optimised protocols for small, medium, and large data sizes.

It is also the basis for Dolphin's SuperSockets, which in user space offer a Berkeley API compliant sockets interface via `libksupersockets`.

46.3 Architecture of ETHOM

In order to bring together the two worlds of Ethernet-based software and Dolphin's high-speed networks, we inserted a thin layer of indirection below the Ethernet interface (see Fig. 46.1). This layer passes the Ethernet frames to the *SCI Message Queues*, which represent the lowest message passing layer of the Dolphin software stack (compare Section 46.2). Compared with ETHOS, we sacrifice compatibility with other high-speed interconnects for better performance at the additional cost of higher system load. At the lowest level, SCI or DX cards physically deliver the data to the peer nodes.

46.3.1 Configuration

ETHOM is configured in three phases: at compile time, at loading, and at run time of the driver. For simplicity reasons, basic configuration is rather static; number of peers in the network and their *ETHOM host_id* to *SCI node IDs* mapping have to be specified at compile time. At load time, most importantly the ETHOM `host_id`

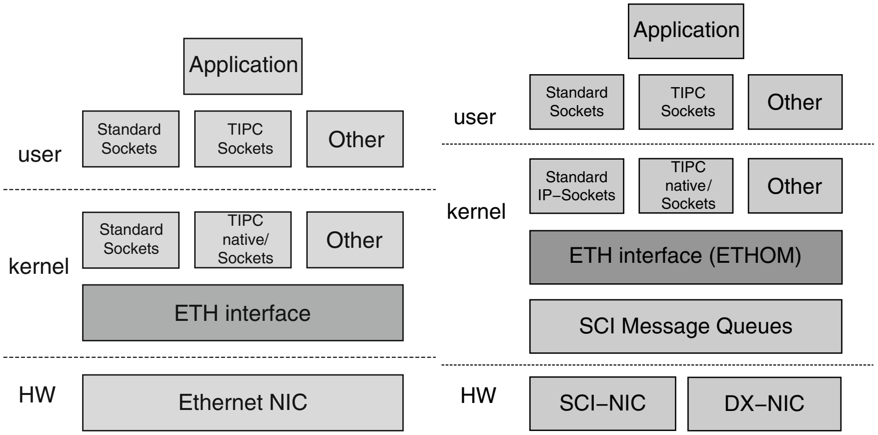


Fig. 46.1 Network architecture of ETHOM (*right*) in Comparison to Standard Architecture (*left*)

has to be passed as a parameter allowing to use one binary for all hosts in the network. Optionally, *direct flushing* after each call to `send_msg()` can be enabled for the sender side, *dynamic polling* for the receive thread. A transmit timeout can be specified that tells the kernel after which period of time to drop packets. With the above mentioned parameters, the Ethernet interface is set up and ready to go. The IP address, MTU, etc. can be assigned at run time with `ifconfig`. All module parameters specified at load time can be changed at run time.

46.3.2 Connection Establishment

As shown in Fig. 46.2, after loading the driver, on each node two unidirectional message queues are created for every peer node in the network (e.g. 14 message queues on each node in case of 7 peer nodes). Message queue IDs are calculated from the local and the peer node number as

$$\begin{aligned}\mathrm{ID}_{\mathrm{ReceiveQueue}} &= \#\mathrm{hosts} \times \mathrm{peer} + \mathrm{local}\\ \mathrm{ID}_{\mathrm{SendQueue}} &= \#\mathrm{hosts} \times \mathrm{local} + \mathrm{peer}\end{aligned}$$

This way they are guaranteed to be unique throughout the cluster.

For each peer node, two threads are started (e. g. 14 threads on each node in case of 7 peers), one trying to connect the local send to the distant receive queue and one waiting for a connection on the local receive queue. As soon as the first of the threads waiting on the local receive queues has accomplished its connection, this thread becomes the *master thread* that polls on *all* connected receive queues. All the other send and receive threads terminate as soon as their connection is established,

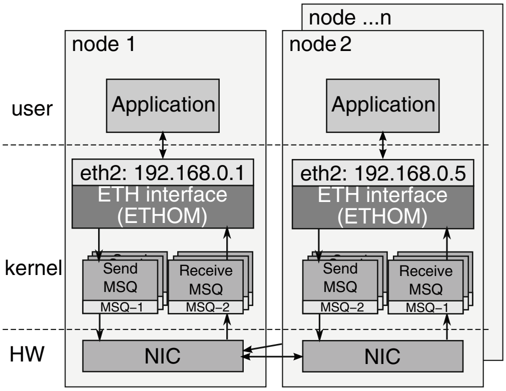


Fig. 46.2 Implementation of ETHOM

effectively reducing the number of remaining threads to one. On the occasion that a peer node does not connect directly, a new connection attempt is made periodically.

In case that IP communication is performed on top of ETHOM, IP addresses can be specified arbitrarily, they do not have to correspond to node numbers. Just like with hardware Ethernet devices, the *Address Resolution Protocol* (ARP) is used at first contact to find the node that provides the sought-after IP address. For this purpose, the kernel sends so called *ARP requests*, Ethernet frames with the hardware address `ff:ff:ff:ff:ff:ff`, that ETHOM forwards to all hosts in the network. The interface providing the missing IP address, which is encapsulated in the request, answers with its hardware address and after that the correct mapping between destination's IP address and Ethernet hardware address is known at the sending kernel.

46.3.3 Communication Phase

Exchanging data between two nodes in a network is described on the basis of Fig. 46.3: An application on ETHOM host 1 on the left sends data through a TCP socket to an application on ETHOM host 4 on the right.

Sending. When an application on host 1 writes data to a TCP socket connected to a receiver on host 4, this data is passed to the kernel networking stack. The kernel then splits it into packets fitting into the previously specified MTU (Fragmentation) – if necessary – and equips each packet with an Ethernet header. This newly constructed *Ethernet frame* is passed to ETHOM by calling its `ethom_tx()` function. There, the minimum length of the packet is checked and if needed padding bytes are added, before the Ethernet frame is given to `ethom_tx_action()`. In `ethom_tx_action()`, the last byte of the destination hardware address (indicating `dest_host`, here "04") encapsulated in the Ethernet header is used to find the send (TX) message queue which is connected to the receive (RX) message queue

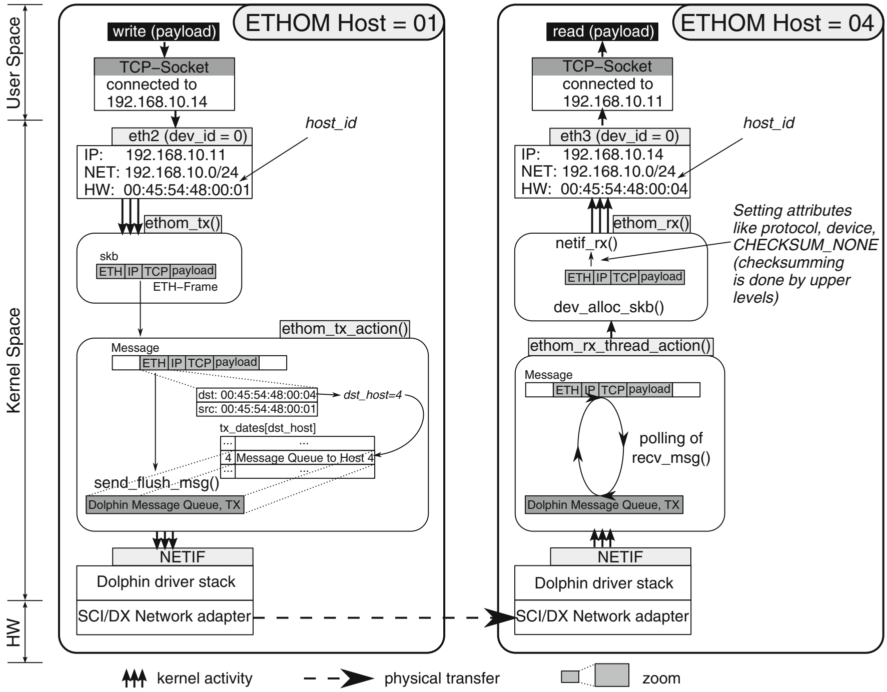


Fig. 46.3 Data transfer through ETHOM from sender to receiver

on the destination host. Depending on the `flush` parameter either `send_msg()` or `send_flush_msg()` is called to forward the message to the Dolphin driver stack and finally the hardware. `send_msg()` should be beneficial for data throughput, while `send_flush_msg()` – which we chose for our measurement and general operation – should reduce latency.

Receiving. Arriving on host 4 (compare Fig. 46.3), the data is directly written to the message queue's data space in main memory by the Dolphin hardware; no interrupt is called to signal the arrival of data. As mentioned before, a thread is started executing the function `ethom_rx_thread_action()` that either dynamically or not polls on the receive message queue. This thread, repeatedly calling Dolphin's `recv_msg()` function fetches the data shortly after arrival and passes it upwards to `ethom_rx()`. In `ethom_rx()`, an skb structure is allocated with `dev_alloc_skb()`, attributes like `dev`, `protocol`, `ip_summed` are set, so that the kernel level above ETHOM accepts the skb, and the Ethernet frame is passed upwards with `netif_rx()`. Here, the IP packets are reassembled from several Ethernet frames (if they were fragmented before), IP and TCP headers are stripped off again, and the user data reaches its final destination, the application on host 4.

In case of a node failure or shutdown, all other nodes continue working as before. Reconnection of message queues as soon as a node comes up again is not yet implemented, though.

46.4 Performance Evaluation

In this section, we briefly present basic experimental results. For a more detailed analysis, please refer to [6].

The measurements were performed on two different clusters: (1) *PD* consisting of nodes equipped with Pentium D 820 processors from Intel, on-board Gigabit Ethernet Controllers (BCM5721), a D352 SCI card from Dolphin, and an MHGS-18 DDR InfiniBand adapter from Mellanox (20 Gb/s). (2) *Xeon* consisting of nodes equipped with Xeon 5355 processors from Intel, on-board Gigabit Ethernet Controllers (Intel 82563EB), a DX510H adapter from Dolphin, and the same InfiniBand adapter as PD.

We chose NPtcp from the widely used NetPIPE benchmark suite in version 3.7.1 in order to generate easily comparable and reproducible low-level latency and bandwidth data.

46.4.0.1 Latency

Figure 46.4 shows the round-trip latency (RTT/2) for messages of varying sizes on the vertical axis and the message size on the horizontal axis.

The upper curve starting at 50 μs represents Gigabit Ethernet, the reference that ETHOS and ETHOM compete with. The lowest latencies are delivered by ETHOM on SCI, followed by ETHOM on DX; the highest times are the Ethernet times. A dramatic decrease in latency can be seen for Ethernet with message sizes between 16 and 48 B, which indicates polling for new messages on the receiving side. For larger messages, the high raw bandwidths of InfiniBand and DX lead to lower latency as for SCI. Comparing ETHOM on SCI with ETHOS on SCI, an improvement in latency of around 10 μs for small messages and around 15 μs for larger ones can be observed.

All in all, ETHOM on SCI provides an improvement in latency by a factor of two and above on our measurement platform over Gigabit Ethernet and about a 30% improvement over its companion ETHOS.

46.4.0.2 Bandwidth

Figure 46.5 shows the bandwidth for varying message sizes measured with NPtcp.

Gigabit Ethernet delivers for all message sizes the lowest bandwidth (excluding the aforementioned interval between 16 and 48 B). For small messages, ETHOM on

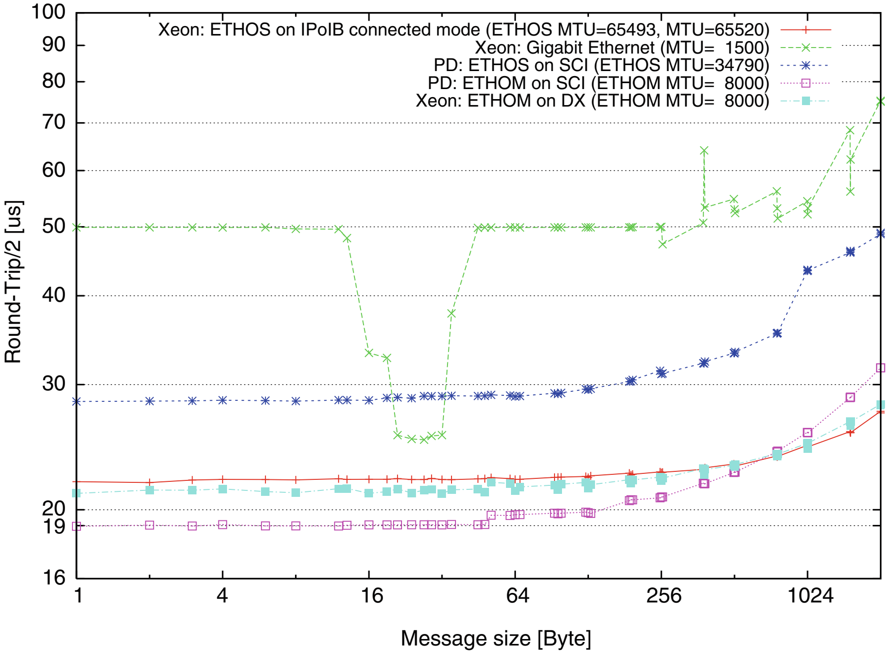


Fig. 46.4 Latency measured with NPtcp

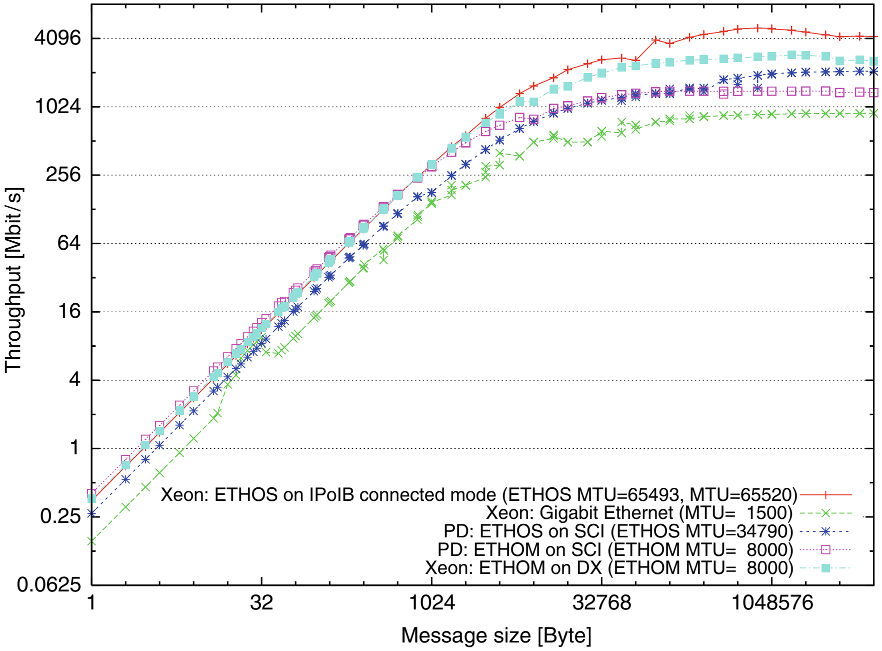


Fig. 46.5 Throughput measured with NPtcp

SCI performs best with ETHOM on DX and ETHOS on IPoIB close by. At about 1 KB, the three curves split again each gradually approaching its maximum, which is at 1.5 Gb/s for SCI and 3 Gb/s for DX (with their current limitation to an MTU of 8 KB) and about 5 Gb/s for ETHOS on InfiniBand. Comparing ETHOM with ETHOS on SCI, it can be noticed that for small messages (up to 8 KB) ETHOM provides a 50% increase in bandwidth. For large messages (256 KB and above) ETHOS benefits from the support for larger low-level packets and maybe additional buffering in the sockets layer.

To sum up, ETHOM on SCI exhibits a twofold increase in bandwidth for messages up to 1 KB over Gigabit Ethernet and about a 50% increase over ETHOS.

46.5 Conclusions

The performance evaluation presented in Section 46.4 and more detailed in [6] shows that ETHOM – making use of a high-speed interconnect like either SCI or Dolphin DX – is a solution that offers better performance than Gigabit Ethernet, latency wise and bandwidth wise. Regarding the different price range of Gigabit Ethernet and these high-speed interconnects, this comparison is only reasonable, when low-latency (and maybe high-bandwidth) Ethernet interfaces are required, which cannot be provided by Gigabit Ethernet.

Comparing the results with ETHOS [5], which implemented an Ethernet interface using kernel-level UDP sockets as its lower interface, we observe a 30–70% improvement in bandwidth for small to medium-sized messages and about a 30% decrease in latency, when SCI is used.

The advent of *many cores* should have a twofold positive effect: (1) The network should become an even bigger bottleneck for communicating applications, as the connection is shared by a bigger number of cores, so better communication performance is highly appreciated. (2) Having a smaller ratio between the one core sacrificed for communication and the number of cores still available for computation reduces the relative communication overhead.

Currently, ETHOM fulfils our main aim to enable TIPC – and any other software communicating via Ethernet frames – to use SCI and DX. Besides ETHOS, it provides the only Ethernet interface for SCI and DX; as a side effect, support for IP-routing is now offered using the standard kernel IP stack on top of ETHOM.

On the other hand side, porting software to the native interfaces of high-speed interconnects almost always provides better performance and efficiency at runtime – obviously at the cost of porting effort. As usual, it remains to the user to balance the pros and cons.

References

1. ANSI/IEEE Std. 1596-1992, Scalable Coherent Interface (SCI): IEEE (2007)
2. SISCI Interface Specification 2.1.1 (1999). Dolphin Interconnect Solutions.
3. Dolphin Interconnect Solutions: The Dolphin SCI Interconnect (1996) http://www.dolphinics.com
4. Dolphin Interconnect Solutions The Dolphin DX Interconnect (2007). http://www.dolphinics.com/products/pent-dxseries-dxh510.html
5. Finocchiaro, R., Razik, L., Lankes, S., Bemmerl, T.: ETHOS, a generic Ethernet over Sockets Driver for Linux. In: Proceedings of the 20th International Conference on Parallel and Distributed Computing and Systems (PDCS) (2008)
6. Finocchiaro, R., Razik, L., Lankes, S., Bemmerl, T.: ETHOM, an Ethernet over SCI and DX Driver for Linux. In: Proceedings of 2009 International Conference of Parallel and Distributed Computing (ICPDC 2009), London, UK (2009)
7. Fu, Shaojian and Atiquzzaman, M.: SCTP: state of the art in research, products, and technical challenges. In: Proceedings of the IEEE 18th Annual Workshop on Computer Communications, CCW 2003, pp. 85–91 (2003)
8. Hellwagner, H., Reinefeld, A. (eds.): SCI: Architecture and Software for High Peformance Compute Clusters, Lecture Notes in Computer Science, vol. 1734. Springer-Verlag, Berlin, Germany (1999)
9. Hopkins, S., Coile, B.: AoE (ATA over Ethernet) (2006) http://www.coraid.com/site/co-pdfs/AoEr10.pdf
10. InfiniBand Trade Association: Infiniband Architecture Overview (2002). http://www.infinibandta.org/events/past/it_roadshow/overview.pdf
11. Kohler, E., Handley, M., Floyd, S.: Datagram Congestion Control Protocol (DCCP) (2006) http://ietfreport.isoc.org/rfc/PDF/rfc4340.pdf
12. Krishnan, V.: Towards an Integrated IO and Clustering Solution for PCI Express. In: Proceedings of IEEE International Conference on Cluster Computing (CLUSTER'07), Austin, TX (2007)
13. Larzon, L.-A., Degermark, M., Pink, S., Jonsson, L.-E., Fairhurst, G.: The Lightweight User Datagram Protocol (UDP-Lite) (2004). http://ietfreport.isoc.org/rfc/PDF/rfc3828.pdf.
14. Maloy, Jon (2004). TIPC: Providing Communication for Linux Clusters. In Proceedings of the Ottawa Linux Symposium, pages 347–356. http://www.linuxsymposium.org/proceedings/LinuxSymposium2004_V2.pdf
15. Myricom Inc.: Myrinet 2000 Product List (2008). http://www.myri.com/myrinet/product_list.html
16. Quadrics Ltd.: Quadrics QsNetII (2003). http://www.quadrics.com
17. Stephens, A., Maloy, J., Horvath, E.: TIPC Programmer's Guide (2008). http://tipc.sourceforge.net/doc/tipc_1.7_prog_guide.pdf
18. The Kerrighed Team Kerrighed: a Single System Image operating system for clusters (2008). http://www.kerrighed.org

Chapter 47
Effect of Dyslipidemia on a Simple Morphological Feature Extracted from Photoplethysmography Flow Mediated Dilation

M. Zaheditochai, E. Zahedi, and M.A. Mohd Ali

Abstract Dyslipidemia is considered to be one of the main heart risk factors, affecting the endothelial vascular function, which can be non-invasively investigated by ultrasound flow-mediated dilation (US-FMD). However, US-FMD comes at a high-cost and is operator-dependent. In this paper, the effect of dyslipidemia on the photoplethysmogram (PPG) signal recorded from collateral index fingers is investigated following a previous study where it was shown that results similar to that of US-FMD can be replicated by the PPG. Two groups, consisting of 30 healthy subjects free from any risk factors and 30 subjects who have only dyslipidemia as risk factor were respectively considered. The percent change in the AC (peak-to-peak) values versus time of the PPG after flow release following 4 min of brachial artery blockage (reactive hyperemia) was obtained (PPG-FMD). Results indicate that a very simple morphological feature allows for a significant ($p < 0.00001$) discrimination between the control and pathologic groups.

Keywords Vascular characterization · brachial artery blockage · vascular dilation

47.1 Introduction

Atherosclerosis is one of the main major causes of cardiovascular disease (CVD). Several risk factors such as smoking, high blood pressure, diabetes and high level of the cholesterol in the blood (dyslipidemia) may lead to developing atherosclerosis. An index which is capable to assess atherosclerosis using vascular endothelial dysfunction would therefore be beneficial to early diagnosis.

M. Zaheditochai (✉) and M.A. Mohd Ali
Department of Electrical, Electronic & System Engineering,
Universiti Kebangsaan Malaysia Bangi, Selangor, Malaysia
e-mail: mojganzahedi@yahoo.com; mama@vlsi.eng.ukm.my

E. Zahedi
School of Electrical Engineering, SHARIF University of Technology, Tehran, Iran
e-mail: zahedi@sharif.edu

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_47,

47.1.1 Vascular Endothelial Dysfunction

Atherosclerosis causes a dysfunction in the endothelium cells action; therefore evaluation of the endothelial function plays an important role in its early detection. Anatomically, the vascular wall is composed of three layers: the Intima, which is the nearest to the lumen and affected by blood flow. The second layer (Media) includes smooth muscle cells and the third layer is Adventitia. Endothelial cells are located in the Intima, the interior surface of the vascular wall. Because of their strategic location (contact with blood), endothelial cells play a critical role in various vascular functions such as controlling the thrombosis by their anticoagulant and antithrombotic surface, interactions of leukocyte and platelet with vessel wall and regulation of vascular tone and growth [1].

The endothelium maintains vascular homeostasis by establishing a balance between endothelium derived relaxing (such as Nitric Oxide – NO) and constricting factors. Any change in this balance makes the vasculature susceptible to vasoconstriction as well as leukocyte adherence, platelet activation, mitogenesis, pro-oxidation, thrombosis, impaired coagulation, vascular inflammation, and atherosclerosis [2]. The measurement of flow mediated dilation (FMD) has been established as an effective method to evaluate endothelial dysfunction [3, 4]. This technique (described in the next section) evaluates the ability of the vascular bed to self-regulate its tone and control the blood flow in response to either a physical or pharmacologic stimuli.

47.1.2 Flow Mediated Dilation

In this technique, a sphygmomanometric cuff is used to occlude the blood flow in (usually) the brachial artery (BA) by cuff inflation to suprasystolic pressure causing ischemia in the lower arm. Consequently, the vessels below the occlusion dilate through a self-regulated function. After 4–5 min of occlusion, the pressure is suddenly released leading to reactive hyperemia which causes shear stress on the inner wall (Intima) of the blocked BA. Subsequently, the BA is dilated due to the release of vasodilators. The diameter of the BA is measured from high-resolution longitudinal ultrasound B-mode images. As various factors affect FMD (such as temperature, food, drugs and sympathetic stimuli), subjects have to fast for at least 8 h without using any vasoactive medications for at least four half-lives, caffeine, high fat foods, vitamin C and tobacco and finally should be studied in a temperature controlled room [5, 8].

Unfortunately the above method proves to be expensive due to the equipment involved, prone to errors requiring an experienced operator [9].

More recently, another non-invasive method based on photoplethysmography [9] (PPG) has been developed. PPG is a signal reflecting pressure waves generated by the function of peripheral vascular arteries due to heart contraction. Dynamic blood volume changes due to the pulses are detected by a photoelectric probe placed on the finger tip (finger PPG) or ear lobes, among other sites. During systole, blood volume is increased due to the heart ventricle's contraction and as a result, light transmission through the peripheral vasculature is reduced and vice-versa during the diastole [11].

PPG pulse amplitude changes due to flow-mediated dilation in the index finger are recorded following a 4 min brachial artery occlusion (PPG-FMD). As PPG reflects blood volume changes in the micro vascular bed of tissue [10], this means that PPG is affected by the occlusion.

The investigated risk factors were obesity, diabetes, hypertension and hypercholesterolemia [12]. Considering the effects of those risk factors, it was found that PPG-AC curves followed closely the US-FMD responses (good correlation) except for subjects who had more than one risk factor. In the present study the focus is on one risk factor only: dyslipidemia [9], and how PPG-FMD data is influenced by this single risk factor is investigated in the next sections.

47.2 Methods

47.2.1 Data Acquisition

The source of the raw database is from Universiti Kebangsaan Malaysia [9] (UKM). As factors such as temperature, food, drug, caffeine and sympathetic stimuli affect vascular flow-mediated vasodilation, each subject had adhered to a strict diet protocol before experimentation.

There are three steps for data acquisition: before occlusion, during occlusion and after cuff deflation as shown in Fig. 47.1.

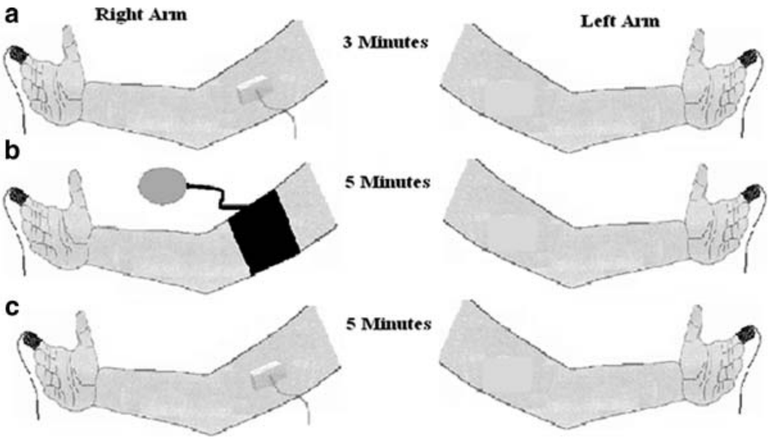


Fig. 47.1 Data acquisition: **(a)** before occlusion to establish the baseline **(b)** occlusion which lasts 4 min **(c)**, after cuff deflation (reactive hyperemia) 4 min (Adapted from study [9])

In the first step (a), PPG were simultaneously recorded from index fingers of both right and left hand for 3 min without any blockage in order to establish the baseline.

In the second step (b), the left hand PPG was the reference, and a blood-pressure cuff was used to create flow blockage (stimulus) in the brachial artery (BA) of the right arm. The cuff was inflated to the suprasystolic pressure (50 mmHg above the subject's systolic blood pressure) inducing total arterial occlusion.

In the third step (c), after 4 min occlusion the flow was established again by rapid cuff deflation, followed by reactive hyperemia in the BA and subsequent dilation.

PPG signals were recorded from both hands [5, 9] during the whole process and the final sampling rate was 50 Hz. In this study, two groups of healthy and pathologic subjects consisting of 30 subjects in each group were considered. Healthy groups are without any history of heart diseases as well as heart risk factors. The pathological group consists of subjects have only dyslipidemia as a risk factor.

47.2.2 Signal Processing

MATLAB (The Math Works Inc.) and SPSS (SPSS Inc.) software were used for signal processing. PPG signals contain both DC and AC components: the DC is related to respiration, sympathetic nervous system activity and thermoregulation while AC parameter is referred to the cardiac synchronous changes with each heart beat in skin micro-vascular blood volume [13].

In this study, the focus is on the AC component of the PPG signal. PPG AC refers to the difference between the amplitudes of the valley and the peak of the same heart cycle in the PPG signal. For the first step the DC of the signal was removed and the AC value was extracted. As the PPG-AC curves were subject dependent, the next step was normalization.

It is assumed that the difference between the mean value (M) over 3 min of PPG AC of the left (L) and right (R) index fingers are equal before occlusion and after it. Therefore, the effect of hyperemia is removed from the baseline by Eq. (47.1) which computes the new baseline (M (R_{after})).

$$\begin{aligned} &M(L_{Before}) - M(R_{Before}) = M(L_{After}) - M(R_{After}) \\ &M(R_{After}) = new\ baseline \end{aligned} \tag{47.1}$$

Equation (47.2) shows the normalized value for each subject based on its new baseline which we used as PPG %AC change.

$$PPG\ \%AC = \frac{T_{old} - new\ baseline}{new\ baseline} \times 100\% \tag{47.2}$$

After normalization based on the new baseline, features were extracted to compare two groups of healthy and pathologic subjects.

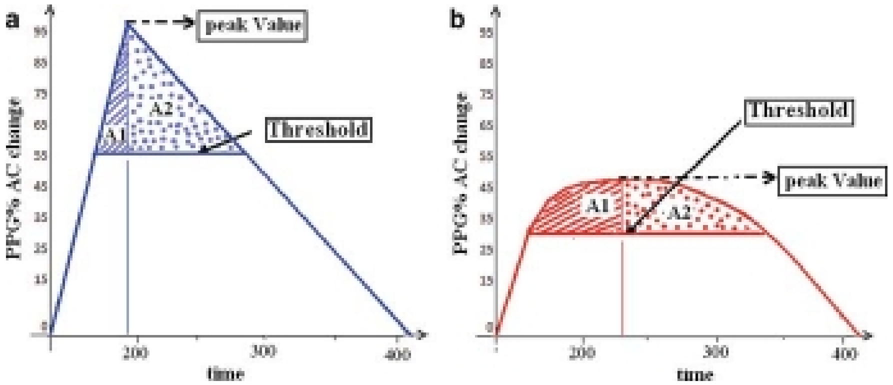


Fig. 47.2 A schematic of FMD PPG %AC change after cuff deflation for a typical healthy subject **(a)** and a pathologic subject **(b)** after cuff deflation

A common feature was needed to compare the plots with each other, to this end; a threshold was defined for each subject. It was referred to the percent of the peak value which is shown in Fig. 47.2. During processing threshold levels from 15% to 95% of the peak value by interval of 10% were tested.

Figure 47.2 shows a typical schematic of this behavior for both healthy and pathologic subjects, after normalization. It is clear that as a common behavior, the PPG %AC change rises to its peak value and then falls towards the baseline. For the healthy subjects, the graph shows a sharp slope in both directions whereas for the pathologic subjects the response is rather blunted.

47.2.3 Feature Definition

Two features were extracted from the PPG %AC. In a previous study [14] the threshold was placed at 35% of the graph's peak, the time period between two cross points of the graph with threshold was measured as Time feature. The results of the t-test showed [14] that the mean of the two groups were significantly different (p-value is 4.4%). The present study was extended to extract other features as below:

- Area under the curve (AUC) before the peak time but above the threshold line: depicted in Fig. 47.2 as A1. This feature reflects the increase in the micro vessels diameter.
- AUC above the threshold after the peak but above the threshold line: depicted in Fig. 47.2 as A2. This feature reflects the decrease in the micro vessels diameter towards the baseline.

47.3 Results

Figure 47.3 shows a typical PPG %AC change for a subject in the healthy group whereas the origin of time indicates cuff deflation. The response increases and reaches the peak value during hyperemia at about 210 s, as a consequence of the stimulus on the vessel wall. Then the diameter goes back to its normal size. This phenomenon is similar to the very well known US-FMD response corresponding, but is obtained here through PPG %AC change. The two threshold values (15% and 85% of peak) are also shown as examples.

Figure 47.4 shows the same data for a subject who has dyslipidemia. As it can be seen, the curve reaches the peak but goes down with some delay compared to the healthy subject (Fig. 47.3). In other words the diameter of the vascular bed increases due to hyperemia but it does not return to its normal size as sharply as it occurred in healthy subjects. It was observed that in some subjects the trace did not even reach the baseline again.

For the first feature (A1: AUC starting from the cuff release point to the peak value), the mean value for pathologic subjects is generally higher than for healthy ones. This could be explained by the sharp changes of the PPG %AC change for the healthy subjects. It was found that by increasing the threshold to 85% of the peak value, a significant difference between the two groups could be seen (Fig. 47.5).

On the opposite, for the second feature (A2: AUC starting from the peak time up to the end), the value for healthy subjects is much more than pathologic subjects. A very significant difference ($p < 10^{-4}$) is obtained when the threshold is at 15% of the peak value. When the threshold is increased towards 85%, the difference becomes less (Fig. 47.6).

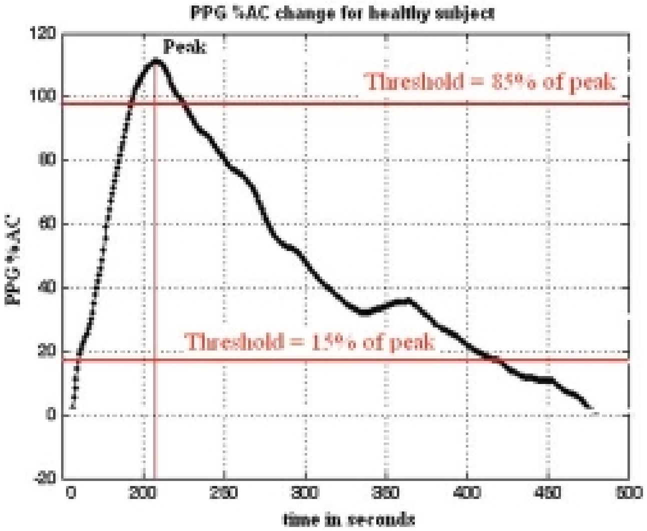


Fig. 47.3 PPG %AC change for a typical healthy subject after cuff deflation

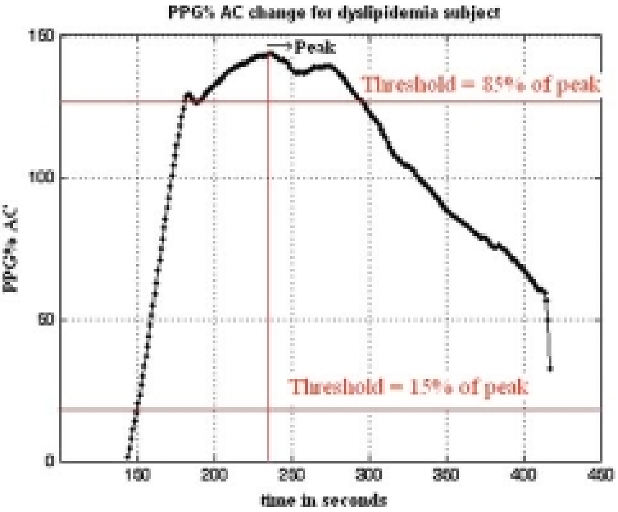


Fig. 47.4 PPG %AC change for a typical dyslipidemic subject after cuff deflation

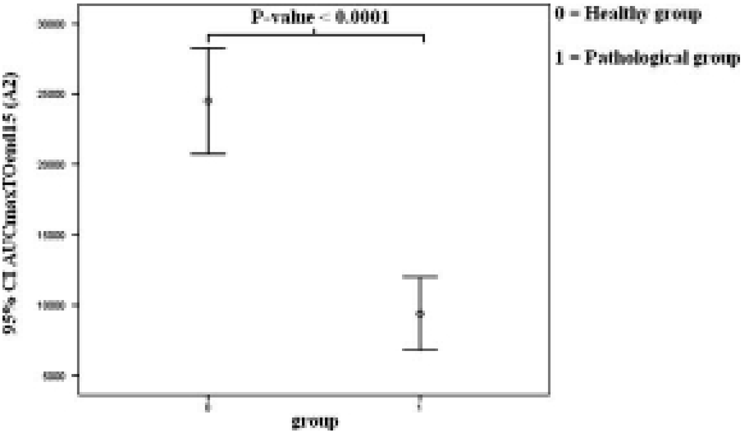


Fig. 47.5 Error plot of AUC (A1) where the threshold is 85% of the peak value including healthy group (0) and pathological group (1)

Figure 47.7 shows the ROC curve for both features A1 (with 85% threshold) and A2 (with 15% threshold).

The mean of the two groups were significantly different (T-test p-value of less than 0.3%) and the 95% confidence interval of the difference was from 1,058 to 4,723 for A1. Figure 47.5 shows the error plot for two groups for A1.

The results show that the mean of the two groups were significantly different (T-test p-value of less than 0.01%) and 95% confidence interval of the difference was from −19,589 to −10,610 for A2. Figure 47.6 shows us the error plot for two groups.

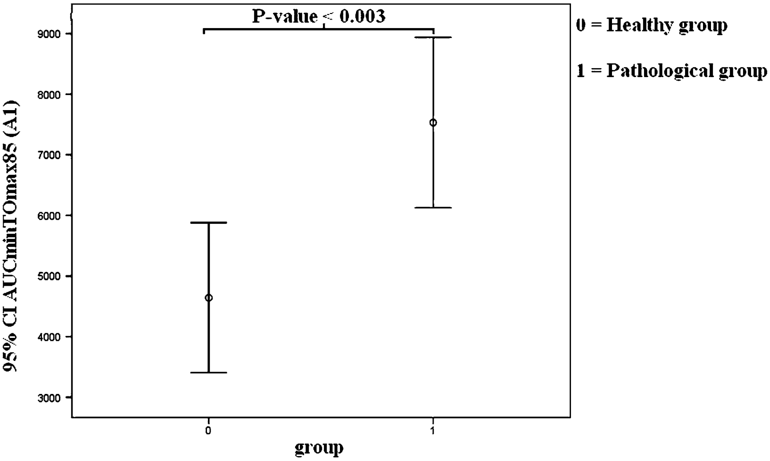


Fig. 47.6 Error plot of AUC (A2) where threshold is 15% of the peak value including healthy group (0) and pathological group (1)

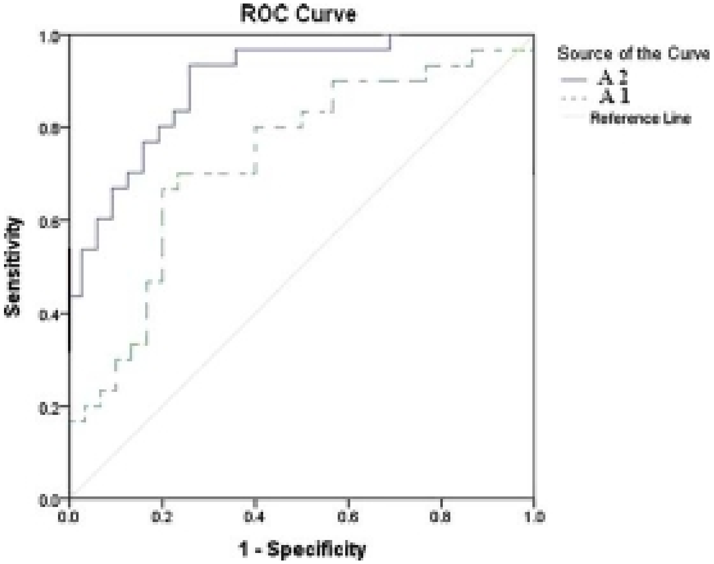


Fig. 47.7 ROC curve for the feature of A1 where threshold is 85% of the peak value (AUCminTOmax85) and the other is A2 where threshold is 15% of the peak value (AUCmax-TOend15)

The result of the last two features shows that for the pathologic subjects the vessels diameter cannot come back to its normal size compared to healthy subjects. Table 47.1 lists out the characteristics of the test variables.

Table 47.1 Characteristics of the investigated features

			Mean	Standard deviation	Mean	Standard deviation		95% Confidence interval of the diff.	
Feature	Threshold	Area ROC	Healthy (N = 30)		Pathologic (N = 30)		P-value	Lower	Upper
AUC start to peak (A1)	85%	0.733	4,642	3,311	7,533	3,764	<0.003	1,058	AUC start to peak (A1)
AUC peak to end (A2)	15%	0.893	24,503	10,073	9,403	6,975	<0.00001	−19,589	AUC peak to end (A2)

47.4 Conclusion

Following brachial arterial blockage, the percent AC change in the PPG signal undergoes through a very typical cycle. Dyslipidemic subjects show a significant difference in the shape of this signal which was investigated in this paper. A simple morphological feature and a low-cost setup have allowed a significant discrimination to take place.

The physiological reason for this observed phenomenon requires further study. Evidence suggests that in healthy subjects – with a normal vascular function control – the behavior of the vessel diameters after hyperemia differs from pathologic subjects where this function seems to be altered. A larger population and other risk factors should be studied to eventually reach a useful index assisting physicians in their diagnostic of the vascular health.

Acknowledgement The authors would like to thank Universiti Kebangsaan Malaysia for supporting this work under the University Research Grant UKM-AP-TKP-07–2009.

References

1. Celermajer, D.S.: Endothelial dysfunction: Does it matter? Is it reversible. J. Am. Coll. Cardiol. **30**(2):325–33 (August 1997)
2. Verma, S., Anderson, T.J.: Fundamentals of endothelial function for the clinical cardiologist. J. Am. Heart Assoc. Circulat. **105**:546–549 (2002)
3. Meyer, B., Mörtl, D., Strecker, K., Hülsmann, M., Kulemann, V., Neunteufl, T., Pacher, R., Berger, R.: Flow-mediated vasodilation predicts outcome in patients with chronic heart failure. J. Am. Coll. Cardiol. **46**(6):1011–1018 (2005)
4. Anderson, E.A., Mark, A.L.: Flow-mediated and reflex changes in large peripheral artery tone in humans. J. Am. Heart Assoc. Circulat. **79**:93–100 (1989)
5. Corretti, M.C., Anderson, T.J., Benjamin, E.J., Celermajer, D., Charbonneau, F., Creager, M.A., Deanfield, J., Drexler, H., Gerhard- Herman, M., Herrington, D.: Guidelines for the ultrasound assessment of endothelial-dependent flow-mediated vasodilation of the brachial artery – a report of the international brachial artery reactivity task force. J. Am. Coll. Cardiol. **39**: 257–265 (2002)
6. Su, W., Cheng, S., Hsu, T., Ho, W.: Abnormal flow-mediated vasodilation in normal-tension glaucoma using a noninvasive determination for peripheral. Invest. Ophthalmol. Visual Sci. **47**(8):3390–3394 (August 2006)
7. Kao, Y.H., Mohler, E.R., Arger, P.H., Sehgal, C.M.: Brachial artery: measurement of flow-mediated dilation with cross sectional US-technical validation. J. Radiol. **228**(3):895–900 (2003)
8. Leeson, P., Thorne, S., Donald, A., Mullen, M., Clarkson, P., Deanfield, J.: Non-invasive measurement of endothelial function: effect on brachial artery dilatation of graded endothelial. J. Heart **78**:22–27 (1997)
9. Zahedi, E., Jaafar, R., Mohd Ali, M.A., Mohamed, A.L., Maskon, O.: Finger photoplethysmogram pulse amplitude changes induced by flow mediated dilation. Physiol. Measure. **29**(5):625–637 (2008)
10. Allen, J.: Photoplethysmography and its application in clinical physiological measurement. J. Physiol. Measure. **28**:1–39 (2007)
11. Crabtree, V.P., Smith, P.R.: Physiological models of the human vasculature and photoplethysmography. Electron. Syst. Con. Division Res. (2003)

12. Zaheditochai, M., Jaafar, R., Zahedi, E.: Non-invasive techniques for assessing the endothelial dysfunction: ultrasound versus photoplethysmography. Proceedings of International Conference of Biomedical Engineering **23**:65–68 (2009)
13. Allen, J., Oates, C.P., Lees, T.A., Murray, A.: Photoplethysmography detection of lower limb peripheral arterial occlusive disease: a comparison of pulse timing, amplitude and shape characteristics. J. Physiol. Measure. **26**:811–821 (2005)
14. Zaheditochai, M., Zahedi, E., Mohd Ali, M.A.: Effect of dyslipidemia on photoplethysmography flow mediated dilation. World Congress Engineering, pp. 1881–1884 (2009)

Chapter 48
Study of the General Solution for the Two-Dimensional Electrical Impedance Equation

Marco Pedro Ramirez Tachiquin, Andrei Gutierrez Solares, Victor Daniel Sanchez Nava, and Octavio Rodriguez Torres

Abstract We study a method for introducing the general solution of the two-dimensional Electrical Impedance Equation, based onto Quaternionic Analysis and Pseudoanalytic Function Theory, for the case when the electrical conductivity can be represented as a separable-variables function only once derivable. We also discuss the contribution of this results into the field of Electrical Impedance Tomography.

Keywords Electrical impedance tomography · quaternions · pseudoanalytic

48.1 Introduction

The Argentine mathematician Alberto Pedro Calderon [3] posed an interesting inverse problem for the Electrical Impedance Equation

$$\operatorname{div}(\sigma \operatorname{grad} u) = 0, \tag{48.1}$$

which is also known as the Inhomogeneous Laplace equation [6], or Poisson equation [8]. Here σ denotes the electrical conductivity and u is the electric potential. This inverse problem is cited in medical engineering as Electrical Impedance Tomography.

Basically, it is a Dirichlet boundary value problem where the values of the potential u are given in a surrounding surface Γ of some domain Ω, and the goal is to approach the conductivity function σ inside Ω using the values of u in such boundary Γ.

The two-dimensional case of Calderon problem (named in honor to A. P. Calderon) quickly received special attention from many researchers, and also almost immediately they realized that the mathematical methods required for

M.P.R. Tachiquin (✉), A.G. Solares, V.D.S. Nava, and O.R. Torres
Benjamin Franklin 47, Col Hipodromo Condesa, Mexico D.F., 06140, Mexico
e-mail: mramirez@lci.ulsa.mx; audreiev@gmail.com; skyplane10@hotmail.com; otorres@lci.ulsa.mx

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*, Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_48,

approaching its solution were not simple at all. Indeed, it is very difficult to find into the specialized literature, mathematical models searching for the solution of the Calderon problem in analytic form. Most of them are based onto a wide class of interesting numerical methods. Nevertheless, the images obtained using such techniques, in some sense, do not provide the required resolution to consider them useful tools for medical diagnosis.

An alternative to the purely numerical approaches for solving the Electrical Impedance Tomography problem, is to use a wide class of analytic solutions of (48.1), considering a set of conductivity functions, and comparing their corresponding solutions u, valued in the boundary points of the domain Ω, with the experimental collected data, iterating this procedure till the error can be considered minimum.

Nonetheless, the mathematical complexity for solving (48.1) without using numerical analysis, provoked that many experts considered impossible to obtain its general solution in analytic form [5], even for the simplest cases of σ (exception done for the constant case).

This point of view prevailed until the Finland researchers Kari Astala and Lassi Päivärinta [1] gave the answer for the two-dimensional case of the Calderon problem by virtue of relating the Electrical Impedance Equation (48.1) with a Vekua equation [14]. This is, they did not only notice the one-to-one relation between the solutions of such Vekua equation and the two-dimensional case of (48.1), but they also proved that this relation warranties the existence and uniqueness of the solution for the Calderon problem in the plane.

A short time latter, the Ukranian-Mexican mathematician Vladislav Kravchenko and Hector Oviedo, also from Mexico, used the elements of Pseudoanalytic Function Theory [2] in order to represent the general solution of (48.1) in terms of Taylor series in formal powers for a certain classes of σ [12]. In fact, this can be considered the first explicit general solution of (48.1).

Applying the elements of Quaternionic Analysis, we will analyze an alternative way for transforming the two-dimensional Electrical Impedance Equation (48.1) into a Vekua equation [13], and we will express its general solution in terms of Taylor series in formal powers, for the case when σ is a separable-variables function, at least once derivable.

We will close the text with a discussion about the contribution of these results within the theory of Electrical Impedance Tomography.

48.2 Preliminaries

48.2.1 Elements of Applied Quaternionic Analysis

We will denote the algebra of real quaternions (see e.g. [7] and [9]) by $\mathbb{H}(\mathbb{R})$. The elements q belonging to $\mathbb{H}(\mathbb{R})$ have the form $q = q_0 + q_1 i_1 + q_2 i_2 + q_3 i_3$, where

q_k, $k = 0, 1, 2, 3$ are real-valued functions depending upon the spacial variables x_1, x_2 and x_3; whereas i_k are the standard quaternionic units, satisfying the relations

$$i_1 i_2 = i_3 = -i_2 i_1, \qquad i_2 i_3 = i_1 = -i_3 i_2, \qquad i_3 i_1 = i_2 = -i_1 i_3$$
$$i_k^2 = -1, \; k = 1, 2, 3.$$

It will be also useful to employ the notation $q = q_0 + \overrightarrow{q}$, where $\overrightarrow{q} = \sum_{n=1}^{3} q_k i_k$ is often named the *vectorial part* of quaternion q, and q_0 is the *scalar part*. It shall be noticed that the subset of purely vectorial quaternions $q = \overrightarrow{q}$ can be identified with the set of three-dimensional vectors $\mathbb{R}^3$. In other words, for every $\overrightarrow{E} = (E_1, E_2, E_3) \in \mathbb{R}^3$ there exist a purely vectorial quaternion $\overrightarrow{E} = E_1 i_1 + E_2 i_2 + E_3 i_3$. The reader can easily verify that this relation is biunique.

In virtue of the last relation, it is possible to represent the multiplication of two quaternions q and p as

$$q \cdot p = q_0 p_0 + q_0 \overrightarrow{p} + p_0 \overrightarrow{q} - \left\langle \overrightarrow{q}, \overrightarrow{p} \right\rangle + \left[\overrightarrow{q} \times \overrightarrow{p} \right], \tag{48.2}$$

where $\left\langle \overrightarrow{q}, \overrightarrow{p} \right\rangle$ represents the standard scalar product $\left\langle \overrightarrow{q}, \overrightarrow{p} \right\rangle = \sum_{k=1}^{3} q_k p_k$ and $\left[\overrightarrow{q} \times \overrightarrow{p} \right]$ denotes the vectorial product in the quaternionic sense [9]. It is evident that in general $q \cdot p \neq p \cdot q$. Thus it is necessary to use the notation $M^p q = q \cdot p$ in order to indicate the multiplication by the right-hand side of the quaternion q by the quaternion p.

The Moisil–Theodoresco partial differential operator D, some times called Dirac operator, is defined as $D = i_1 \partial_1 + i_2 \partial_2 + i_3 \partial_3$, where $\partial_k = \frac{\partial}{\partial x_k}$, and it acts on the set of at least once-derivable quaternionic-valued functions. Using the classic vectorial notation once more, we can write

$$Dq = \text{grad} q_0 - \text{div} \overrightarrow{q} + \text{rot} \overrightarrow{q}, \tag{48.3}$$

where $\text{grad} q_0$ and $\text{rot} \overrightarrow{q}$ are written also in the quaternionic sense [9].

48.2.2 Elements of Applied Pseudoanalytic Function Theory

According to [2], let F and G be a pair of complex-valued functions fulfilling the inequality

$$\text{Im}\left(\overline{F} G\right) > 0, \tag{48.4}$$

where $\overline{F}$ denotes the complex conjugation of $F : \overline{F} = \text{Re}F - i\text{Im}F$, and i is the standard complex unit $i^2 = -1$ (indeed, (48.4) is a special way for expressing F and G to be linearly independent). Therefore any complex-valued function W can be expressed as the linear combination of F and G: $W = \phi F + \psi G$, where ϕ and ψ are purely real-valued functions.

A pair of complex-valued functions satisfying (48.4) is called a *Bers generating pair.*

L. Bers also introduced the concept of a (F, G)-*derivative* of a complex-valued function W (some times cited as the derivative in the sense of Bers). It is defined as

$$\frac{d_{(F,G)}W}{dz} = (\partial_z\phi)\, F + (\partial_z\psi)\, G \tag{48.5}$$

where $\partial_z = \partial_1 - i\,\partial_2$, and it exists iff

$$(\partial_{\overline{z}}\phi)\, F + (\partial_{\overline{z}}\psi)\, G = 0 \tag{48.6}$$

where $\partial_{\overline{z}} = \partial_1 + i\,\partial_2$ (usually the operators ∂_z and $\partial_{\overline{z}}$ are introduced with the factor $\frac{1}{2}$, nevertheless it will result more convenient for us to work without it).

Let us introduce the following functions

$$\begin{aligned} A_{(F,G)} &= -\frac{\overline{F}\,(\partial_z G) - \overline{G}\,(\partial_z F)}{F\overline{G} - \overline{F}G}, \qquad & B_{(F,G)} &= \frac{F\,(\partial_z G) - G\,(\partial_z F)}{F\overline{G} - \overline{F}G}, \\ a_{(F,G)} &= -\frac{\overline{F}\,(\partial_{\overline{z}} G) - \overline{G}\,(\partial_{\overline{z}} F)}{F\overline{G} - \overline{F}G}, \qquad & b_{(F,G)} &= \frac{F\,(\partial_{\overline{z}} G) - G\,(\partial_{\overline{z}} F)}{F\overline{G} - \overline{F}G}. \end{aligned} \tag{48.7}$$

These functions are known as *the characteristic coefficients* of the generating pair (F, G). Using the new notations, the (F, G)-derivative (48.5) of W can be expressed as

$$\frac{d_{(F,G)}W}{dz} = \partial_z W - A_{(F,G)}W - B_{(F,G)}\overline{W}, \tag{48.8}$$

and the condition (48.6) will turn into

$$\partial_{\overline{z}} W - a_{(F,G)}W - b_{(F,G)}\overline{W} = 0. \tag{48.9}$$

This last equation is called *the Vekua equation* [14], and the complex-valued functions W that fulfill (48.9) are named (F, G)-*pseudoanalytic*. We shall notice that the Vekua equation we have cited in the above paragraphs corresponds to a particular case of (48.9), as we will explain further.

The following statements were originally posed in [2].

Remark 48.1. The complex-valued functions of the generating pair (F, G) are (F, G)-pseudoanalytic, and according to (48.8), their (F, G)-derivatives satisfy $\frac{d_{(F,G)}F}{dz} = \frac{d_{(F,G)}G}{dz} = 0$.

Definition 48.2. Let (F, G) and (F_1, G_1) be two generating pairs, and let their characteristic coefficients (48.7) satisfy

$$a_{(F,G)} = a_{(F_1,G_1)} \quad \text{and} \quad B_{(F,G)} = -b_{(F_1,G_1)}. \tag{48.10}$$

Hence the generating pair (F_1, G_1) will be called the *successor* pair of (F, G), as well (F, G) will be the *predecessor* pair of (F_1, G_1).

Theorem 48.3. *Let W be a (F,G)-pseudoanalytic function, and let (F_1,G_1) be a successor pair of (F,G). Then the (F,G)-derivative (48.8) of W will be (F_1,G_1)-pseudoanalytic.*

Definition 48.4. Let (F,G) be a generating pair. Its *adjoint* pair (F^*,G^*) will be defined by the formulas

$$F^* = -\frac{2\overline{F}}{F\overline{G}-\overline{F}G}, \qquad G^* = \frac{2\overline{G}}{F\overline{G}-\overline{F}G}. \tag{48.11}$$

L. Bers also introduced the (F,G)-*integral* of a complex-valued function W:

$$\int_\Gamma W d_{(F,G)}z = F(z_1)\operatorname{Re}\int_\Gamma G^* W dz + G(z_1)\operatorname{Re}\int_\Gamma F^* W dz,$$

where Γ is a rectifiable curve going from z_0 till z_1.

In particular, when $W = \phi F + \psi G$ is (F,G)-pseudoanalytic, then

$$\int_{z_0}^{z} \frac{d_{(F,G)}W}{dz} d_{(F,G)}z = W(z) - \phi(z_0)F(z) - \psi(z_0)G(z), \tag{48.12}$$

but as we have shown, the derivatives in the sense of Bers of F and G is equal to zero, thus the integral expression (48.12) represents the *antiderivative in the sense of Bers* of $\frac{d_{(F,G)}W}{dz}$.

Beside, a continuous complex-valued function w is said to be (F,G)-*integrable* iff

$$\operatorname{Re}\oint G^* w dz + i \operatorname{Re}\oint F^* w dz = 0.$$

Theorem 48.5. *The (F,G)-derivative of a (F,G)-pseudoanalytic function W will be (F,G)-integrable.*

Theorem 48.6. *Let (F,G) be a predecessor pair of (F_1,G_1). A complex-valued function $\mathcal{E}$ will be (F_1,G_1)-pseudoanalytic iff it is (F,G)-integrable.*

Definition 48.7. Let $\{(F_m,G_m)\}$, $m = 0,\pm1,\pm2,\pm3,...$ be a sequence of generating pairs. If every (F_{m+1},G_{m+1}) is a successor pair of (F_m,G_m) we say that $\{(F_m,G_m)\}$ is a *generating sequence*. If $(F_0,G_0) = (F,G)$ we say that (F,G) is *embedded* in the generating sequence $\{(F_m,G_m)\}$.

Let W be a (F,G)-pseudoanalytic function, and let $\{(F_m,G_m)\}$, $m = 0,\pm1,\pm2,\pm3,...$ be a generating sequence in which (F,G) is embedded. We can express the higher derivatives in the sense of Bers of W as

$$W^{[0]} = W; \qquad W^{[m+1]} = \frac{d_{(F_m,G_m)}W^{[m]}}{dz};\ m = 0,1,2,3...$$

We will expose now the definition and properties of the so-called formal powers.

Definition 48.8. The formal power $Z_m^{(0)}(a, z_0; z)$ with center at z_0, coefficient a and exponent 0 is defined as

$$Z_m^{(0)}(a, z_0; z) = \lambda F_m(z) + \mu G_m(z),$$

where the coefficients λ and μ are real constants such that $\lambda F_m(z_0) + \mu G_m(z_0) = a$. The formal powers with exponents $n = 1, 2, 3, ...$ are defined by the formulas

$$Z_m^{(n)}(a, z_0; z) = n \int_{z_0}^{z} Z_{m+1}^{(n-1)}(a, z_0; \varsigma)\, d_{(F_m, G_m)}\varsigma,$$

where all this integral operators are antiderivatives in the sense of Bers (48.12).

Following [2], it is possible to verify that formal powers posses the following properties:

1. $Z_m^{(n)}(a, z_0; z)$ are (F_m, G_m)-pseudoanalytic.
2. If a_1 and a_2 are real constants, then
$$Z_m^{(n)}(a_1 + i a_2, z_0; z) = a_1 Z_m^{(n)}(1, z_0; z) + a_2 Z_m^{(n)}(i, z_0; z). \qquad (48.13)$$
3. The formal powers satisfy the differential relations $\frac{d_{(F_m, G_m)} Z_m^{(n)}(a, z_0; z)}{dz} = Z_{m+1}^{(n-1)}(a, z_0; z)$.
4. The formal powers satisfy the asymptotic formulas
$$\lim_{z \to z_0} Z_m^{(n)}(a, z_0; z) = a(z - z_0)^n.$$

As the reader can immediately notice, this last property illustrates why L. Bers called the functions $Z_m^{(n)}(a, z_0; z)$ “formal powers”.

Remark 48.9. As it has been proved in [2], any complex-valued function W, solution of (48.9), accepts the expansion

$$W = \sum_{n=0}^{\infty} Z^{(n)}(a_n, z_0; z), \qquad (48.14)$$

where the missing subindex m indicates that all formal powers belong to the same generating pair. This is: *expression (48.14) is an analytic representation of the general solution of (48.9).*

The expression (48.14) was called Taylor series in formal powers of W by L. Bers.

48.3 Quaternionic Reformulation of the Electrical Impedance Equation, and Its Relation with the Vekua Equation

Let us consider the Electrical Impedance Equation (48.1)

$$\operatorname{div}(\sigma \operatorname{grad} u) = 0.$$

Indeed, the electric field vector $\overrightarrow{E}$ for the static case is defined as

$$\overrightarrow{E} = -\operatorname{grad} u, \tag{48.15}$$

so we can write

$$\operatorname{div}\left(\sigma \overrightarrow{E}\right) = 0. \tag{48.16}$$

Beside, from (48.15) we immediately obtain

$$\operatorname{rot} \overrightarrow{E} = 0. \tag{48.17}$$

Following [13], let us consider $\overrightarrow{E}$ as a purely vectorial quaternionic-valued function

$$\overrightarrow{E} = E_1 i_1 + E_2 i_2 + E_3 i_3.$$

Introducing the notations

$$\overrightarrow{\mathcal{E}} = \sqrt{\sigma} \overrightarrow{E}, \tag{48.18}$$

and

$$\overrightarrow{\sigma} = \frac{D\sqrt{\sigma}}{\sqrt{\sigma}}, \tag{48.19}$$

the equations (48.16) and (48.17) turn into

$$\left(D + M^{\overrightarrow{\sigma}}\right) \overrightarrow{\mathcal{E}} = 0, \tag{48.20}$$

which is a quaternionic reformulation of (48.1).

48.3.1 The Two-Dimensional Case

Consider now the particular case when

$$\overrightarrow{\mathcal{E}} = \mathcal{E}_1 i_1 + \mathcal{E}_2 i_2 \tag{48.21}$$

and σ depends upon only two spatial variables $\sigma = \sigma(x_1, x_2)$. Thus, the expression (48.19) takes the form

$$\overrightarrow{\sigma} = \sigma_1 i_1 + \sigma_2 i_2,$$

where

$$\sigma_1 = \frac{\partial_1 \sqrt{\sigma}}{\sqrt{\sigma}}, \quad \sigma_2 = \frac{\partial_2 \sqrt{\sigma}}{\sqrt{\sigma}}. \tag{48.22}$$

Substituting (48.21) and (48.22) into (48.20) we obtain the system

$$\partial_1 \mathcal{E}_1 + \partial_2 \mathcal{E}_2 = -\mathcal{E}_1 \sigma_1 - \mathcal{E}_2 \sigma_2, \qquad \partial_1 \mathcal{E}_2 - \partial_2 \mathcal{E}_1 = \mathcal{E}_3 \sigma_1 - \mathcal{E}_1 \sigma_2,$$
$$\partial_3 \mathcal{E}_1 = \partial_3 \mathcal{E}_2 = 0.$$

Multiplying the second equation by $-i$ and adding to the first, it yields

$$\partial_{\overline{z}} (\mathcal{E}_1 - i\mathcal{E}_2) + (\sigma_1 - i\sigma_2)(\mathcal{E}_1 + i\mathcal{E}_2) = 0, \tag{48.23}$$

but according to (48.22) it is possible to see that

$$\sigma_1 - i\sigma_2 = \frac{\partial_z \sqrt{\sigma}}{\sqrt{\sigma}}.$$

Taking this into account and introducing the notation

$$\mathcal{E} = \mathcal{E}_1 - i\mathcal{E}_2, \tag{48.24}$$

the equation (48.23) becomes

$$\partial_{\overline{z}} \mathcal{E} + \frac{\partial_z \sqrt{\sigma}}{\sqrt{\sigma}} \overline{\mathcal{E}} = 0. \tag{48.25}$$

which is a special kind of Vekua equation [14].

We shall mention that in [4], the authors obtained a bicomplex Vekua equation similar to (48.25), starting from a quaternionic equation with the same structure that (48.20), but related to the Dirac equation with different classes of potentials.

In order to analyze the general solution of (48.25), it will be convenient to establish its relation with another Vekua equation of the form [12]

$$\partial_{\overline{z}} W - \frac{\partial_{\overline{z}} \sqrt{\sigma}}{\sqrt{\sigma}} \overline{W} = 0. \tag{48.26}$$

Let $F = \sqrt{\sigma}$ and $G = \frac{i}{\sqrt{\sigma}}$. It is easy to verify these functions satisfy (48.4), so they constitute a generating pair. According to (48.7), a simple calculation will show us that their characteristic coefficients are $A_{(F,G)} = a_{(F,G)} = 0$, $B_{(F,G)} = \frac{\partial_z \sqrt{\sigma}}{\sqrt{\sigma}}$, and $b_{(F,G)} = \frac{\partial_z \sqrt{\sigma}}{\sqrt{\sigma}}$.

Notice that, in concordance with *Definition 2*, the characteristic coefficients corresponding to a successor pair (F_1, G_1) of the pair $(\sqrt{\sigma}, \frac{i}{\sqrt{\sigma}})$ must verify the relations $a_{(F_1,G_1)} = 0, \ \ b_{(F_1,G_1)} = -\frac{\partial_z\sqrt{\sigma}}{\sqrt{\sigma}}$. Beside, a (F_1, G_1)-pseudoanalytic function must fulfill Eq. (48.25).

Remark 48.10. By *Theorem 3*, the $(\sqrt{\sigma}, \frac{i}{\sqrt{\sigma}})$-derivative of any solution of (48.26) will be a solution of (48.25).

We have now established the relation between the Vekua equation (48.25) and the Vekua equation (48.26). Moreover, since the general solution of (48.26) can be represented by means of a Taylor series in formal powers (48.14), once we have succeed to construct a generating sequence where the pair $(\sqrt{\sigma}, \frac{i}{\sqrt{\sigma}})$ is embedded, we will be able to express the general solution of (48.25) by means of the $(\sqrt{\sigma}, \frac{i}{\sqrt{\sigma}})$-derivative of the general solution of (48.26). It is important to mention that, in general, it is not clear how to build a generating sequence in which an arbitrary generating pair is embedded. Although, using new results in Applied Pseudoanalytic Function Theory [11], we are able to write an explicit generating sequence for the case when the desired *embedded generating pair* belongs to a special class of functions that, without loss of generality, fulfill the requirements of the Electrical Impedance Tomography.

48.3.2 Explicit Generating Sequence for the Case When σ Is a Separable-Variables Function

Since the early appearing of Bers Pseudoanalytic Function Theory [2], the development of methods for introducing explicit generating sequences, in which a specific generating pair is embedded, have represented a very interesting challenge. We shall remark that an explicit generating sequence is required if we desire to express the general solution of a Vekua equation in terms of Taylor series in formal powers.

When considering the Electrical Impedance Equation (48.1), to represent the conductivity function σ as a separable-variables function

$$\sigma\left(x_1, x_2\right) = U^2(x_1)V^2\left(x_2\right),$$

has shown to be a very useful approach for the problem of Electrical Impedance Tomography (see e.g. [6]). Here the exponents of the functions are 2 because it will be more comfortable to work with them in this form, but it has also a physical meaning when considering a non polarized media, since the functions will always take real and positive values, assuming they have no zeros in the domain of interest.

For this case, an explicit generating sequence can be constructed adapting to our study the results presented in [10].

Theorem 48.11. *[10] Let (F, G) be a generating pair of the form*

$$F = \sqrt{\sigma} = U(x_1)V(x_2), \qquad G = \frac{i}{\sqrt{\sigma}} = \frac{i}{U(x_1)V(x_2)}. \tag{48.27}$$

Then, it is embedded in the generating sequence

$$\{(F_m, G_m)\}, \qquad m = 0, \pm 1, \pm 2, \pm 3, ...$$

defined as

$$F_m = 2^m U(x_1)V(x_2), \qquad G_m = i\frac{2^m}{U(x_1)V(x_2)};$$

when m is an even number, and

$$F_m = 2^m \frac{V(x_2)}{U(x_1)}, \qquad G_m = i2^m \frac{U(x_1)}{V(x_2)};$$

when m is odd.

Remark 48.12. Given an explicit generating sequence where the generating pair (48.27) is embedded, we are in the possibility of building the Taylor series in formal powers in order to approach the general solution of the Vekua equation (48.26). According to *Theorem 3*, the $\left(\sqrt{\sigma}, \frac{i}{\sqrt{\sigma}}\right)$-derivative of such solution will be the general solution of the Vekua equation (48.25). Hence, the real and the imaginary components of the solution of (48.25) will constitute the general solution for the two-dimensional case of the quaternionic equation (48.20). Finally, using (48.18), it immediately follows we are able to write the general solution for the two-dimensional Electrical Impedance Equation (48.1).

48.4 Conclusions

Since the study of Eq. (48.1) is the base for the Electrical Impedance Tomography problem, the possibility of expressing the general solution of (48.1) by means of Taylor series in formal powers, opens a new path for improving the convergence speed of many numerical methods designed for medical image reconstruction. For example, our proposal could well be used for introducing a hybrid algorithm which employs numerical techniques based onto Neural Networks, whose efficiency have been already proven when dealing with Electrical Impedance Tomography, focusing the Neural Network to propose and to adjust the conductivity inside the domain of interest, while the mathematical methods posed in this work are used to approach the analytic solution of the two-dimensional Electrical Impedance Equation (48.1) for such conductivity, in terms of Taylor series in formal powers. In the opinion of the authors, such hybrid method would yield high quality information for a better estimation of the error between the analytic calculations and the collected data, and this would have direct repercussion in the accuracy of the images obtained by Electrical Impedance Tomography.

Of course, it is still necessary to search for the best methods in order to approach the conductivity σ by means of a separable-variable function, a critical matter for the posed techniques. This will constitute an important question for future researches.

Concerning to the results posed in the present work, we should notice that the suggested mathematical methods impose minimal restrictions to the conductivity function σ. Indeed, it is only necessary for σ to be a separable-variables function in the Cartesian plane, and to be at least once derivable. This is a very general case which includes most part of mathematical approaches for real situations in Electrical Impedance Tomography (see, e.g. [5, 6, 8]).

We shall also notice the numerical methods that might be used, belong almost exclusively to the evaluation of the integral operators of the formal powers. This task can be accomplished by quite standard numerical procedures, hence we can lead our further discussions to approach the constants for the Taylor series at the moment of solving the problem of Electrical Impedance Tomography.

Acknowledgments The authors would like to acknowledge the support of CONACyT project 106722.

References

1. Astala, K., Päivärinta, L.: Calderon's inverse conductivity problem in the plane. Ann. Math. **163**, 265–299 (2006)
2. Bers, L.: Theory of pseudoanalytic functions. IMM (1953)
3. Calderon, A.P.: On an inverse boundary value problem. In: Meyer, W.H., Raupp, M.A. (eds.) Seminar on Numerical Analysis and its Applications to Continuum Physics, pp. 65–73. Rio de Janeiro (1980)
4. Castaneda, A., Kravchenko, V.V.: New applications of pseudoanalytic function theory to the dirac equation. Institute of Physics Publishing, J. Phys. A: Math. Gen. **38**, 9207–9219 (2005)
5. Cheney, M., Isaacson, D., Newell, J.C.: Electrical impedance tomography. J. Soc. Ind. Appl. Math. Rev. **41**(1), 85–101 (1999)
6. Demidenko, E.: Separable laplace equation, magic toeplitz matrix and generalized ohm's law. Appl. Math. Comput. 181, 1313–1327 (2006)
7. Gurlebeck, K., Sprossig, W.: Quaternionic analysis and elliptic boundary value problems. Akademie-Verlag, Berlin (1989)
8. Kim, J., Webster, G., Tomkins, W.J.: Electrical impedance imaging of the thorax. Microwave Power **18**, 245–257 (1983)
9. Kravchenko, V.V.: Applied quaternionic analysis, vol. 28. Researches and Exposition in Mathematics (2003)
10. Kravchenko, V.V.: Recent developments in applied pseudoanalytic function theory. Some Topics on Value Distribution and Differentiability in Complex and p-adic Analysis, Science Press, Beijing (2008)
11. Kravchenko, V.V.: Applied pseudoanalytic function theory. Birkhäuser Verlag AG, Series: Frontiers in Mathematics (2009)

12. Kravchenko, V.V., Oviedo, H.: On explicitly solvable vekua equations and explicit solution of the stationary schrödinger equation and of the equation div $(\sigma\nabla u) = 0$. Comp. Var. Ellip. Eq. **52**(5), 353–366 (2007)
13. Tachiquin, M.P.R., Nava, V.D.S., Jaso, A.F., Torres, O.R.: On the advances of two and three-dimensional electrical impedance equation. 5th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico, IEEE Catalog Number: CFP08827-CDR, ISBN: 978-1-4244-2499-3, Library of Congress: 2008903800, 978-1-4244-2499-3/08 (2008)
14. Vekua, I.N.: Generalized Analytic Functions. Pergamon (1962)

Chapter 49
Biological Application of Widefield Surface Plasmon Resonance Microscope to Study Cell/Surface Interactions and the Effect of TGF-β3, HCL and BSA/HCL on Cell Detachment Assay of Bone Cells Monolayer

Farshid Sefat, Mansour Youseffi, and Morgan Denyer

Abstract Widefield Surface Plasmon Resonance (WSPR) microscope was used to investigate cell surface interactions under two different culture conditions: bone cells cultured on SPR substrate with transforming growth factor β3 (TGF-β3) and without as control. Trypsinisation was carried out in order to investigate its effect on cell detachment, in the presence of TGF-β3, HCl or BSA/HCl solutions. Trypsin was therefore added to four groups of bone cells with addition of TGF-β3, HCl, HCl/BSA solutions and one additional flask as control. These results further confirmed that application of TGF-β3, HCl and HCl/BSA decreased the degree of cell attachment on surface of culture flasks. HCl and BSA/HCl were tested as they are the carriers and solvents for TGF-β3. Cell detachment in control was about 43% after 6 min, which is slow. Bone cells in the presence of BSA/HCl started detaching from the surface about 4–5 min and cell detachment was about 63% after 6 min which was faster as compared to the control. Bone cells in the presence of HCl alone started detaching from the surface about 2 min (after applying trypsin) and cell detachment was about 69% after 6 min which was faster compared to the BSA/HCl and control. Trypsinisation experiments for bone cells cultured with TGF-β3 (50 ng/ml) showed that cells started to detach from the surface about 1 min after application of trypsin and cell detachment was about 85% after 4 min which was faster as compared to the control, HCl and BSA/HCl.

F. Sefat (✉)
Institute of Pharmaceutical Innovation (ipi), University of Bradford, Bradford,
West Yorkshire, BD7 1DP, UK
e-mail: F.Sefat@bradford.ac.uk

M. Youseffi
School of Engineering, Design and Technology-Medical Engineering,
University of Bradford, Bradford, West Yorkshire, BD7 1DP, UK
e-mail: M.Youseffi@bradford.ac.uk

M. Denyer
School of Life Sciences, University of Bradford, Bradford, West Yorkshire, BD7 1DP, UK
e-mail: M.Denyer@bradford.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_49,

Keywords WSPR imaging MG63 bone cell line · BSA/HCl · HCl · TGF-β3 · trypsinisation

49.1 Introduction

Cells are not found in isolation and they usually adhere to other cells or surrounding extracellular matrix (ECM) environment in vivo and substrate or a surface in vitro [1, 2]. The interaction of bone cells with their surrounding ECM environment influence some physiological function and pathological processes [3]. These interactions are mediated by integrins which are capable of transducing the signals from ECM to the cells resulting migration, differentiation and specific protein synthesis. To determine which integrins are involved flow cytometric analysis and immunoprecipitation need to be carried out.

The TGF-β superfamily have different proteins, including bone morphogenetic proteins (BMPs), Mullerian inhibiting substances, nodals, activins, inhibins, as well as TGF-β. These proteins play different roles in a variety of different processes, from development to disease [4]. TGF-β is a cytokine produced by several different cell types in our body, including osteoblast, keratinocytes, etc. This cytokine exists in three isomers in mammals: TGF-β1, 2 and 3 and acts in either a paracrine or autocrine manner. The effect of TGF-β on cell activity can be either positive or negative, depending on the physiological state of the target cell and its environment. For example, when fibroblasts are grown in monolayers in the presence of epidermal growth factor, TGF-β causes a decrease in proliferation. Conversely, when these cells are grown in semi solid medium, TGF-β allows cells to grow [5]. As with all TGF-β superfamily proteins, the effects of TGF-β 1, 2 and 3 are mediated via a variety of signaling pathways [6]. Activation of the signaling pathways is the result of ligand-receptor interactions at the cell surface. The TGF-β receptor complex is made up of serine-threonine kinase receptors type I and II. These bind to a third receptor, type III, also known as betaglycan. This membrane anchored proteoglycan binds to the TGF-β ligand and presents it to the type I/II receptor complex allowing signal transduction to occur. Interestingly, this betaglycan has been shown to act as a dual modulator of TGF-β activity. The soluble form of this protein, no longer anchored to the cell membrane, acts as a inhibitor of TGF-β signaling [7].

There are various techniques to enhance tissue regeneration and applying growth factor to the site of regeneration is the simplest method of inducing cells to proliferate, differentiate and regenerate. Generally, direct application of growth factors has little effect [8] because the growth factor diffuses out from the site of regeneration very quickly. This is a problem that can be solved by a controlled release of growth factor at the site of action over a long period of time by use of a bioabsorbable scaffold. Growth factors mediate cell actions in response to different environmental cues. They may be produced in different ways with different effect such as autocrine effect, paracrine effect, juxtacrine effect, and endocrine effect [8]. Growth factors also have important roles in growth, development, day-to-day maintenance, mobilised remodelling and injury.

Tissue regeneration can be achieved if growth factor gene is transferred into the cells at the site of regeneration and cause cells to secrete growth factor. Bone cells can be isolated from the recipient (bone marrow/connective tissue), expanded in tissue culture, exposed to bioactive factors, combined with the scaffold and finally implanted into the donor site in order to persuade regeneration.

In this work, an entirely new WSPR system has been used [9] to enable analyses of interfacial interactions and eliminate the need for immunostaining. We aim to demonstrate the potential biological applications of this new technology by imaging cell/surface interface and their interactions. These images allow analyses of the physical interactions between bone cells and ECM environment with presence of TGF-β3, HCl, BSA/HCl and without, which influence cell shape and function. These interactions are mediated by integrins which bind the ECM protein to the cell and hence cell surface coupling and adhesion processes occur which can also be investigated via WSPR microscope.

49.2 Aims and Objectives

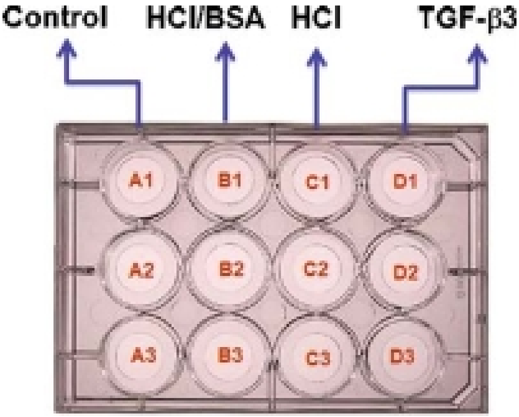
The aims of this study were to investigate the effect of TGF-β3, HCl and BSA/HCl on bone cell detachment via Trypsinization process with better understanding of cell/ECM interactions using WSPR imaging.

49.3 Materials and Methods

Trypsinisation was carried out to investigate the effect of TGF-β3 and its carriers on cell detachment. To establish the appropriate dilution at which to plate cells, a 1 in 3, 1 in 6 and 1 in 12 dilution was plated into the three rows of a 12 well plate. For TGF-β3 to have sufficient time to influence cells in culture, cells were grown for at least 2 days prior to the attachment assay. The 1 in 3 dilution was confluent by day 3 and was therefore chosen for this assay. In order to reconstitute the vile containing TGF-β3, a solution of HCl (4 mM), BSA (1 mg/ml) and distilled water was prepared. Trypsin was added to four groups of cultured bone cells with four different solutions including TGF-β3, HCl, HCl/BSA and bone cell only as control to study the effect of these solutions on cell detachment. HCl and HCl/BSA solutions were used, as they are the carriers for TGF-β3.

Bone cells were cultured in a 12 well petridish and left for 3 days to become confluent with three different cell dilution. Three wells were labeled as A1-A3 on the left of culture dish which was seeded with 1:3 ratio of cell to DMEM known as control. Three other wells were labeled as B1-B3 and seeded with 1:3 ratio of cell to DMEM with addition of 50 ng/ml BSA/HCl. Another three wells were labeled as C1-C3 seeded with 1:3 ratio of cell to DMEM with addition of 50 ng/ml HCl. The other wells were labeled as D1-D3 on the right of culture dish which was

Fig. 49.1 Twelve well culture dish for trypsinisation with four different cell dilutions



seeded with 1:3 ratio as before and with addition of 50 ng/ml TGF-β3. Cells were checked after 3 days to observe their confluency. Cells were imaged every 20 s for a duration of 15 min (45 frames in total). The 12 well cultured dishes were placed under microscope. Old media aspirated and cells were washed with Hank's balanced salt solution (HBSS) and microscope was focused. Trypsin (0.5 ml) was added and recording was carried out for 15 min for duration of 20 s each.

This method was repeated for groups A, B, C and D. The speed of cells detaching from surface is important and it was possible to find out which group (A, B, C or D) detached faster. Figure 49.1 shows the schematic drawing of the 12 well. Trypsin was added to three groups of cultured bone cells with four different solutions including TGF-β3, HCl, HCl/BSA and bone cell only as control to study the effect of these solutions on cell detachment.

49.4 Results and Discussion

Figure 49.2a is the control trypsinisation process for the period of 14 min, showing that cells started detaching from the surface about 6 min after applying trypsin. Cell detachment of about 43% took place after 6 min and hence, it became clear that cells detached very slowly for the control. Figure 49.2b shows the trypsinisation process of bone cells in the presence of BSA/HCl. Cells started detaching from the surface after about 4–5 min and cell detachment was about 63% after 6 min which was faster than control.

Figure 49.2c shows the trypsinisation experiments for bone cells cultured with HCl alone which shows that cells started to detach from the surface about 2 min after application of trypsin. Cell detachment was about 69% after 6 min which was faster as compared to the BSA/HCl and control. Cell detachment was about 85% after 4 min, which was faster as compared to the control and BSA/HCl. Figure 49.2d

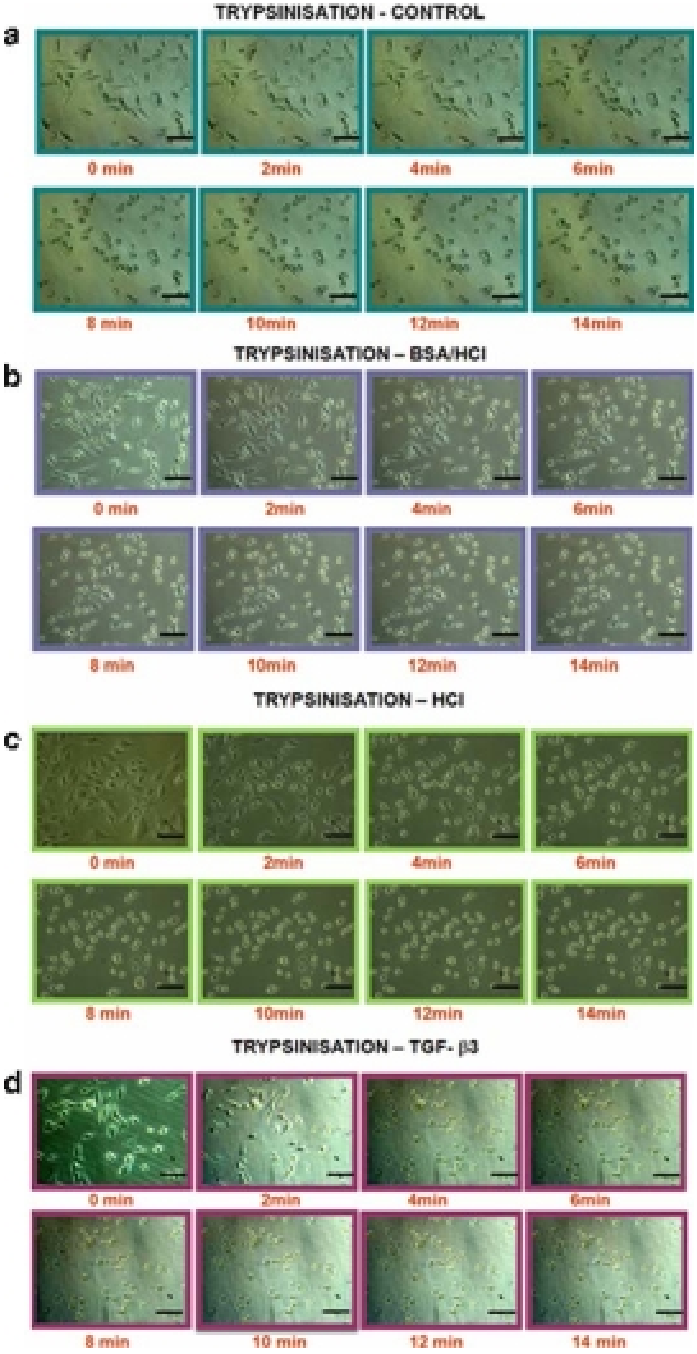


Fig. 49.2 Trypsinisation process: (**a**) Control, (**b**) BSA/HCl, (**c**) HCl, and (**d**) TGF-β3 (scale bar = 100 μm)

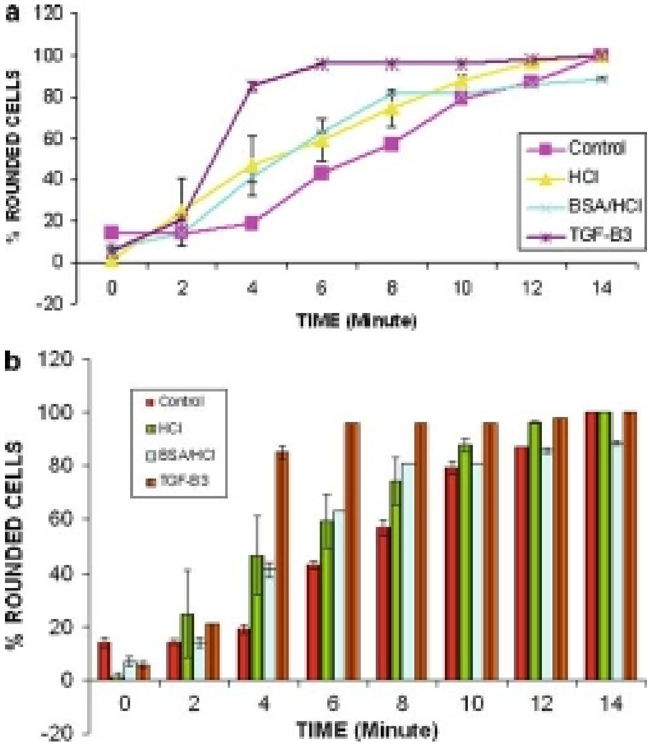


Fig. 49.3 (**a**) Comparison between percentages rounded cells during trypsinisation process for control, HCl, BSA/HCl and TGF-β3 additions. (**b**) Bar chart showing the comparison between percentages rounded cells during trypsinisation process for control, HCl, BSA/HCl and TGF-β3 additions

shows the trypsinisation experiments for bone cells cultured with TGF-β3 which shows that cells started to detach from the surface about 1 min. after application of trypsin and were completely detached by the third minute.

For comparison, as shown in Figure 49.3a and b, a completely different response was recorded with the bone cells plated without TGF-β3, and that the rate of detachment was much slower in control even after 6–8 min. These results further confirmed that application of TGF-β3 at 50 ng/ml decreased the degree of cell attachment on surface.

49.4.1 WSPR Image Analyses

Imaging with the WSPR system revealed that bone cells attached to the surface with high contrast concentrically arranged band like structures at their focal contacts as shown in Fig. 49.4a–c. The highest contrast band like components were localized at

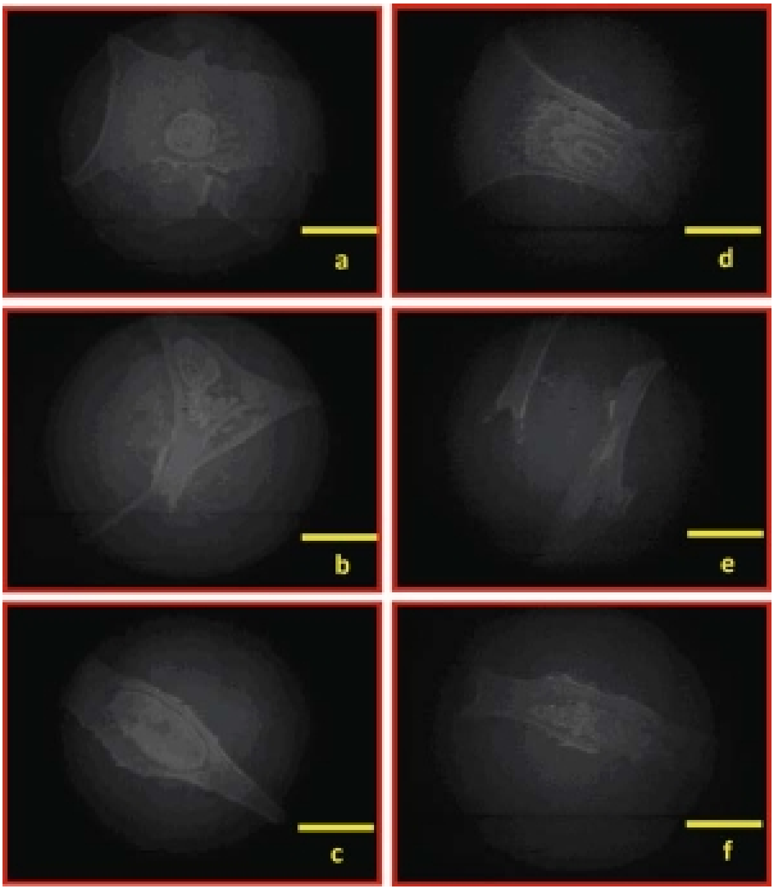


Fig. 49.4 (**a–c**) WSPR images of untreated MG63 bone cell (control); (**d–f**) WSPR images of MG63 bone cell treated with TGF-β3 (scale bar = 25 μm)

the cell periphery and that the un-treated cells had lengths ranging between 25 and 35 μm. On the other hand, the treated cells with TGF-β3 had a very different peripheral structure, only visible via the WSPR images as shown in Figure 49.4d–f. The cells acquired a high contrast band like structure at the boarder of the elongated features along with less concentrically formed band like rings across the cell body. This caused the cells to extend between 50–70 μm in length which is significantly larger than the un-treated bone cells (Control) indicating that the growth factor induces significant changes in cell activity such as cell signaling resulting in up-regulation of proteins associated with the cells acquiring a more migratory phenotype. WSPR imaging may also help us to understand the un-answered questions about the role of integrin/ECM interactions and gene product up-regulation in induced cell migration during wound healing.

49.5 Statistical Analysis

HCl and HCl/BSA showed similar results with significant differences ($P < 0.05$) compared to the control ($P < 0.05$). TGF-β3 showed significant difference ($P < 0.05$) compared to the other treatments. Multiple comparisons with respect to different treatments and at each time point were performed. Comparative results were given as means ±SE. The P values were obtained by one-way ANOVA followed by the Bonferroni adjustment. TGF-β3 appeared to detach faster with more stimulatory action than HCl and HCl/BSA. TGF-β3, HCl and HCl/BSA were significantly different ($P < 0.05$) compared to the control. Error bars = 95% CI (for Fig. 49.3a and b).

49.6 Conclusions

Trypsinisation was carried out in order to investigate its effect on cell detachment, in the presence of TGF-β3, HCl, BSA/HCl and control solutions. The results confirmed that application of TGF-β3 at 50 ng/ml decreased the degree of cell attachment on surface of culture flasks. Also, HCl and HCl/BSA additions enhanced the rate of cell detachment in relation to the negative controls, indicating perhaps that TGF-β3 does not act alone in the trypsinisation, but instead functions synergistically with signalling pathways that are dependent on the availability of hydrogen ions. WSPR images clearly showed the bone cell focal contact points without the need for immunostaining. These images allowed analyses of the physical interactions between cells and the ECM environment. The un-treated (control) bone cells attached to the surface with concentrically arranged band like structures at their focal contacts, with the highest contrast at the periphery of the cell. TGF-β3 treated bone cells acquired band like structure at the border of a more elongated cell with less concentrically formed band like features across the cell body.

References

1. Critchlow, M.A., Blands, Y.S., Ashhurst, D.E.: The effect of exogenous transforming growth factor beta-2 on healing fractures in the rabbit. Bone **6**, 521 (1995)
2. Freshney, R.I.: Culture of Animal Cells, A manual of Basic Technique, 3rd edn. Wiley-Liss, New York (1994)
3. Erlebacher, A., Derynck, R.: Increased Expression of TGF-Beta2 in osteoblast results in an osteoprosis-like phenotype. J. Cell Biol. **132**, 195–210 (1996)
4. Gordon, K.J., Blobe, B.C.: Role of transforming growth factor-β superfamily signaling pathways in human disease. Biochimica et Biophysica Acta (BBA). Molecul. Basis Disease **1782**(4), 197–228 (2008)
5. Border, W.A., Nobel, A.A.: Transforming growth factor β in tissue fibrosis. New Engl. J. Med. **331**, 1286–1292 (1994)

6. Itoh, S., Itoh, F., Goumans, M.J., Ten, D.P.: Signaling of transforming growth factor-β family members through Smad proteins. FEBS J. **267**(24), 6954–6967 (2000)
7. Lopez-Casillas, F., Payne, H.M., Andres, J.L., Messague, J.: Betaglycan can act as a dual modulator of TGF-beta access to signaling receptors: mapping of ligand binding and GAG attachment sites. J. Cell Biol. **124**(4), 557–567 (1994)
8. Khan, S.N.: Bone growth factors. Orthop. Clin. **31**, 375 (2000)
9. Abdul Jamid, M.M., Denyer, M.C.T., Youseffi, M., Britland, S.T., Liu, S., See, C.W., Somekh, M.G., Marlafeka, S.: Monitoring immunostaining process of proteins by using a widefield surface plasmon microscope. J. Anat. **206**, 515 (2005)

Chapter 50
Application of a Novel Widefield Surface Plasmon Resonance Microscope in Cell Imaging and Wound Closure Properties of TGF-β3, BSA/HCl and HCl in Cultured Human Bone Cell Monolayer

Farshid Sefat, Mansour Youseffi, and Morgan Denyer

Abstract A newly developed Widefield Surface Plasmon Resonance (WSPR) Microscope was used to investigate the morphology of MG63 bone cells and their interfacial interactions with ECM proteins. This allowed detailed imaging of cell surface coupling at lateral resolution down to ∼500 nm. In this work, bone repair was investigated and modulated by different stimulus including growth factors. TGF-β3 is a cytokine known to be associated with the scarless healing of skin and it is highly probable that it may play a role in the repair of other tissues. Thus the aim of this study was to investigate the effect of TGF-β3 on closure of a model wound in cultured monolayers of the MG63 human bone cells. This in vitro work examined and compared the wound closure properties of TGF-β3, and its dosage carriers HCl and BSA/HCl. The wound healing response was investigated in TC grade culture flasks by creating a wound (with average scratch width of 300 ± 10–30 μm SD, 1.7–5 μm SEM) on confluent monolayers of MG63 human bone cells. After wounding, cultures were then treated with 50 ng/ml TGF-β3 at concentration of 4 mM HCl and 1 mg/ml BSA and distilled water. Also, the same method was applied for cell cultured monolayers with no growth factor as control and with HCl/BSA and HCl only solutions. After wounding, wound width was measured every 5 h over a 30-h period. The results showed that TGF-β3 enhanced the rate of wound repair in a monolayer of MG63 bone cells. It was found that after 20 h all the culture flasks treated with TGF-β3 (with 15.5% of wound remained open), HCl (with 16% of

F. Sefat (✉)
Institute of Pharmaceutical Innovation (ipi), University of Bradford, Bradford,
West Yorkshire, BD7 1DP, UK
e-mail: F.Sefat@bradford.ac.uk

M. Youseffi
School of Engineering, Design and Technology-Medical Engineering, University of Bradford,
Bradford, West Yorkshire, BD7 1DP, UK
e-mail: M.Youseffi@bradford.ac.uk

M. Denyer
School of Life Sciences, University of Bradford, Bradford, West Yorkshire, BD7 1DP, UK
e-mail: M.Denyer@bradford.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_50,

wound remained open) and finally BSA/HCl (with 17.7% of wound remained open) had resulted in faster wound healing compared to control (with 85% of wound remained open). These results indicated that wound closure in model MG63 wound with TGF-β3 was higher than the control.

Keywords Widefield surface plasmon resonance microscope · bone cell engineering · BSA/HCl · HCl · TGF-β3 · wound healing

50.1 Introduction

Surface Plasmon Resonance (SPR) occurs at an interface between a dielectric and a conductor when p-polarized light at a specific angle strikes the conductor thus exciting oscillation of free electrons. Changes in the SPR excitation angle can be induced by the binding of bio-molecular species to the metallised layer and this change is directly related to the refractive index and thickness of that molecular species. Surface plasmon systems are therefore extremely useful for measuring the thickness of molecules on surfaces down to the subnanometric level. However, the configuration of the standard surface plasmon microscope is based upon the Kretschmann [1] configuration in which surface plasmons propagate and spread and hence degrade the image resolution. Also, the numerical aperture is restricted by the range of excitation angles and these factors lead to standard surface plasmon microscopes having average lateral resolutions of about 20 μm.

An entirely new WSPR system has been developed [2] to enable analyses of interfacial reactions at sub nanometric vertical resolutions and sub-micron lateral resolutions. The main difference between the standard surface plasmon microscope and the newly developed WSPR system is the method of exciting surface plasmons (SPs). In the standard SPs based system a prism coated with a thin gold layer is normally used to excite SPs, but in the WSPR system a high numerical aperture lens was used to excite SPs on thin coverslips coated with a 50 nm thick gold layer.

In this study we aim to demonstrate the potential biological applications of this new technology by imaging cell/protein interactions and cell/surface interface for better understanding of the processes involved in cell guidance. These images allow analyses of the physical interactions between cells and ECM proteins which influence cell shape and function. These interactions are mediated by integrins which bind the ECM protein to the cell and hence cell surface coupling and adhesion processes occur which were also investigated via WSPR images. WSPR imaging of cell cultured bone cells on protein patterned substrates may therefore provide the un-known details of interfacial interactions between integrin and ECM and hence gene product up-regulation in induced cell migration during wound healing.

Bone tissue engineering is a promising field in the area of medicine and involves principles of biology and biomedical engineering with the aim of developing a viable tissue substitute that can restore the function of human tissue. Despite healing of soft tissues, bone healing has features of degeneration, and usually no scar can be find after healing. As soon as the fracture has been bridged by new bone, the remodeling process begins [3]. Bone repair can be manipulated by different stimulus such as growth factors, distraction osteogenesis and electrical stimulation [3]. TGF-β3 is a cytokine produced by different cell types inside the body and influences a number of cell activity such as differentiating, stimulating mesenchymal stem cell (MSC) growth, acting as a chemotactic factor and enhancing bone cells and extracellular matrix (ECM) product secretion [4–6]. Our aims in this part were to investigate the role of TGF-β3 and its dosage carriers HCl and BSA/HCl on wound repair of MG63 bone cell monolayers.

50.2 Aims and Objectives

Firstly to investigate the interfacial interactions between cultured bone cells and substrate by WSPR microscope and secondly to investigate the effect of TGF-β3 on wound closure of cultured MG63 bone cell monolayers. The lab-based experimental work investigated and compared the wound closure properties of TGF-β3, HCl and BSA/HCl in cultured monolayers of human bone cells. Other cellular responses such as proliferation, differentiation and detachment have also been investigated along with different stages of cell behaviour and morphology during wound healing.

50.3 Materials and Methods

50.3.1 Cell Culture

MG63 bone cells were cultured in standard 25 cm^2 culture flasks with 1:5 ratio cell suspension (50,000 cell/ml) in Dulbecco's Modified Eagle Medium (DMEM, SIGMA). Cells were incubated at 37 °C and were split upon reaching confluency, usually every 3–4 days.

50.3.2 SP Substrate Preparation

A prefabricated glass slide (0.18 mm thick and 22 mm diameter), coated with 48 nm gold on top of 1 nm chromium (for better adherence) were stamp patterned with fironectin protein by using a 50 μm stamp. The substrate was then plated with MG63 bone cells. After fixing the cells, the substrates were dehydrated in serial alcohol and were imaged using light microscope. The substrate were then mounted in the sample holder of the WSPR microscope and imaged further.

50.3.3 Wound Healing

Bone cells were cultured in a low glucose culture medium known as Dulbecco's Modified Eagle Medium (DMEM, from SIGMA) containing various supplements such as L-glutamine (4 mM), Penicillin-Streptomycin (5 ml), Amphotericin or a fungizone (1 ml), HEPES buffered culture medium and 'fetal calf serum' (50 ml). The bone cells were cultured inside culture flasks and bathed in the culture media. The cells attached to form a layer at the bottom of the culture flasks. A 'wound' was made using a disposable long nosed plastic pipette. The tip was bent downwards so that it could be inserted into the flasks. The tip was then drawn across the cells on the cultured surface creating the wound. The scratch markings facilitated orientation while imaging and another wound was later applied at 90° angle to the initial scratch and pictures were also taken at the cross points. Thus the same points were always photographed which gave more accurate data for analysis.

A 'test experiment' was performed in order to determine the time frame of wound closure in the wounded models. This was similar to another experiment on wound closure using NIH/3T3 fibroblast monolayers [5] for which wound closure was achieved after ∼300 min. In their study, cell monolayers were wounded with "the corner of a piece of Mylar film which is commonly used in copy machines". An average scratch of 300 μm was produced which is equal to three to four times bigger than cell widths and wound closure was completed after ∼300 min. During our 'test experiment' with TGF-β3, it became clear that complete wound closure was not achieved during this time frame and thus the time frame for our experiment was set to 30 h with data collection every 5 h.

This *in vitro* work examined and compared the wound closure properties of TGF-β3, and its dosage carriers, HCl and BSA/HCl. The wound healing response was investigated in TC grade culture flasks by creating a wound (with average scratch width of 300 ± 10–30 μm SD, 1.7–5 μm SEM) on confluent monolayer of MG63 human bone cell. After wounding cultures were then treated with 50 ng/ml TGF-β3 at concentration of 4 mM HCl and 1 mg/ml BSA and distilled water. Also the same method was applied for cell cultured monolayer with no growth factor as control and with HCl/BSA and HCl only solutions. The culture flasks were then stored inside the incubator and wound width was imaged and measured every 5 h over a 30-h period. Image J software was used in order to measure the distance between the wound edges. Six vertical lines at semi-random horizontal distances were drawn and the distances between the intersections of the lines with the wound edges were measured.

Image analysis consisted of a set of six measurements of the wound width for each image. There were some images in which the wound was completely closed, but in few others there were still some gaps where cells had moved across to cover the wound. For these cases measurements of the gaps were also taken into account. For each image six separate sets of photos were taken along the marker line. This meant that two averages were needed. The first average was the average wound width within an image. The second was the average of the ‘wound width averages’ for each of the culture flasks. Measurements were taken every 5 h for 30 h. These wound width averages were plotted against time.

50.4 Results and Discussion

50.4.1 Cell Culture

The laboratory experiments performed during this study showed that TGF-β3 does indeed speed up the wound closure process in a pure cell culture environment. It was observed that after 20 h of wounding all the culture flasks treated with TGF-β3, HCl and BSA/HCl had resulted in faster wound healing than the control. Fastest wound closure was achieved with TGF-β3, closely followed by the HCl and BSA/HCl in descending order with very little difference (see Figs. 50.1–50.3). These findings proved that in in-vitro cultured human bone cell, TGF-β3 acts as stimulant enhancing wound repair.

Figure 50.1 shows all four wounded bone cell monolayers together during 30 h of healing with addition of HCl, BSA/HCl, TGF-β3 and control. Images for the wound closure process in the control culture flask show that there is no change in wound width even after 10 h. Cell morphology changes can be observed only in the cells at the wound edges after 15 h. These cells have elongated but without migration to the wound site. After 20 h there were still gaps between the cells in the wound site. The cells in the control flask did not seem to form bridges with cells from the opposite wound margin to stabilize the wound site until about 25 h. Cells met each other after ∼25 h and started making bridges but without complete closure. It was clear that wound remained open even after 25 h and hence wound healing occurred very slowly in control culture flasks and ∼61.1% of wound remained open after 30 h (see Fig. 50.3).

Images for the wound closure process in the HCl culture flask (see Fig. 50.1) show that after 5 h the morphology of the cells at the wound edges had changed and that the cells had elongated, spread, replicated and migrated into the wound site perpendicularly to the wound axis. The vertically elongated cells look different (more elongated) to the cells a distance away from the wound edges (more rounded). The cells seem to be less dense and more spread out at the wound edges. After 15 h the cells had migrated into the wound site and elongated in order to meet cells from the opposite side of the wound. These cells then formed bridges to connect the wound edges and close the wound. Soon after bridge formation in the wound site, the wound was closed quite rapidly.

During bridge formation the cells a distance away from the wound edge aligned behind those bridges to organize the cell structure. After ∼25 h the wound was closed and the cell morphology at the wound site was quite different from the morphology of the other cells. At the wound site the cells were more elongated whereas the cells away from the wound were more rounded and denser. There were

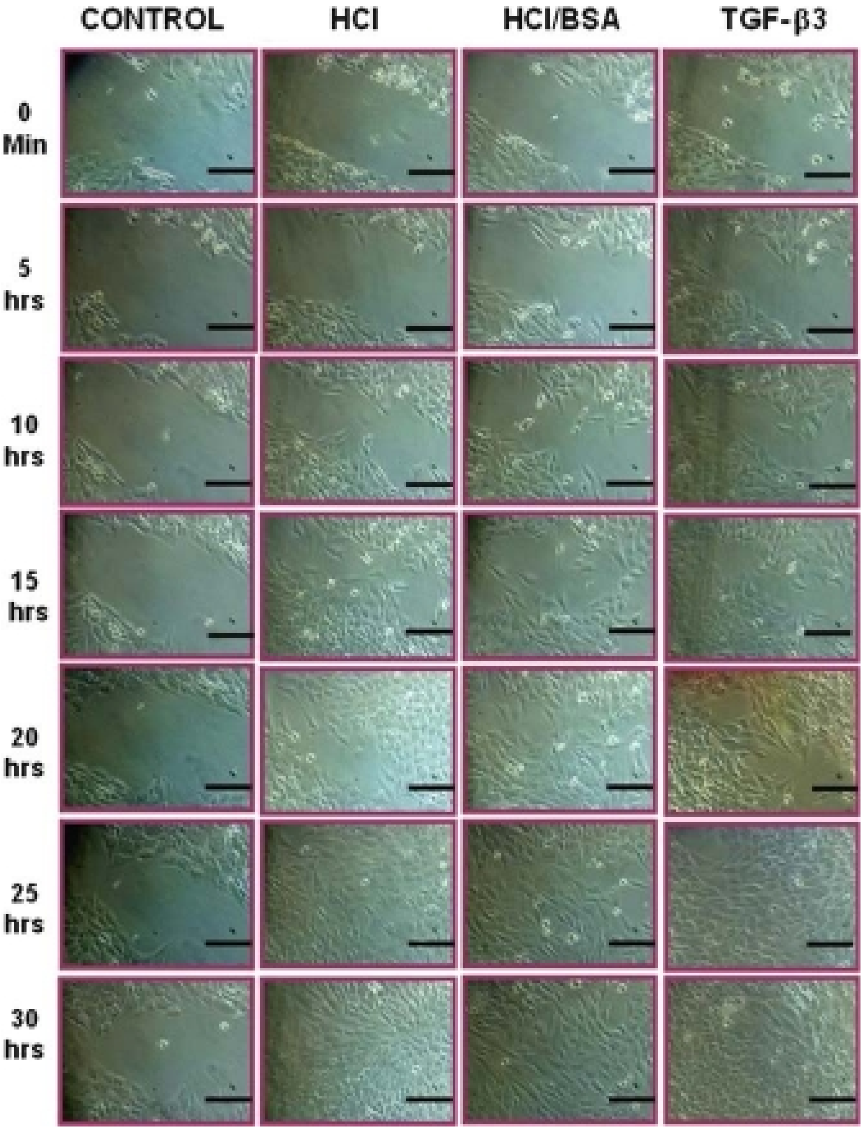


Fig. 50.1 Images of the wound healing process for the bone cell monolayers with HCl, BSA/HCl, TGF-β3 and control for the period of 30 h (scale bar = 100 μm)

still small gaps between cells at the wound site, which will eventually be closed. It was observed that after 20 h the culture flasks with HCl showed high percentage of wound closure (only 15.5% remained open, see Figs. 50.2 and 50.3). By comparison between HCl and control, it became clear that after 30 h the wound in the control remained open by ∼87% whereas HCl addition showed almost complete closure (see Fig. 50.3).

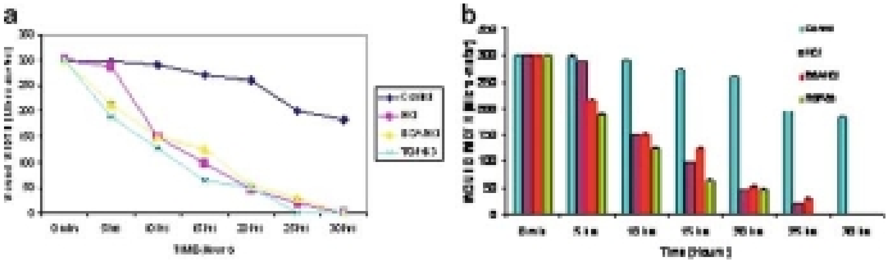


Fig. 50.2 (**a**) Graph of wound closure width against time for wounded bone cell monolayers with addition of HCl, BSA/HCl, TGF-β3 and control. (**b**) Bar chart showing the comparison in wound width with time for control, HCl, BSA/HC and TGF-β3 additions

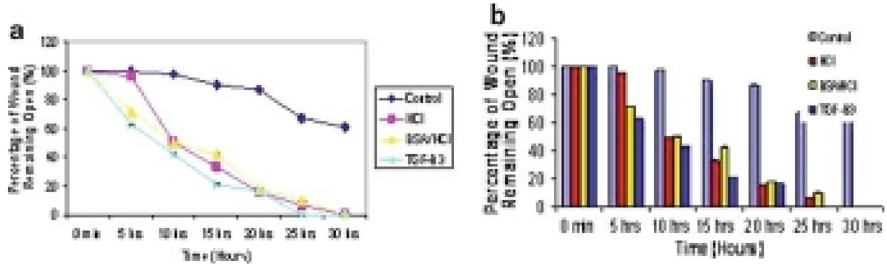


Fig. 50.3 (**a**) Graph of percentage wound closure with time for wounded bone cell monolayers with addition of HCl, BSA/HCl, TGF-β3 and control. (**b**) Bar chart showing the comparison in percentage wound closure with time for control, HCl, BSA/HCl and TGF-β3

Images for the wound closure process in the BSA/HCl culture flask (see Fig. 50.1) show that after ∼5 h the morphology of the cells at the wound edges had changed and that the cells had elongated, spread, replicated and migrated into the wound site perpendicularly to the wound axis. The vertically elongated cells looked different (more elongated) to the cells a distance away from the wound edges. The cells looked less dense and more spread out at the wound edges. After 15 h, the cells looked migrated into the wound site and elongated in order to meet cells from the opposite side of the wound similar to HCl culture flask. These cells then formed bridges to connect the wound edges and closed the wound.

For BSA/HCl, after ∼25 h (see Fig. 50.1) the wound was closed and the cell morphology at the wound site was quite different from the morphology of the other cells. At the wound site the cells were more elongated whereas the cells away from the wound were more rounded and denser. There were still small gaps between cells at the wound site, which are expected to close eventually. HCl and BSA/HCl culture flasks showed very similar results with high percentage of wound closure (only ∼15–18% of wound remained open, see Fig. 50.3). Fast healing occurred in BSA/HCl culture flasks and only 10% of the wound remained open after 25 h (see Fig. 50.3). For BSA/HCl, after ∼25 h (see Fig. 50.1) the wound was closed and the cell morphology at the wound site was quite different from the morphology of the other cells. At the wound site the cells were more elongated whereas the cells away from the wound were more rounded and denser. There were still small gaps between cells at the wound site, which are expected to close eventually.

HCl and BSA/HCl culture flasks showed very similar results with high percentage of wound closure (only ∼15–18% of wound remained open, see Fig. 50.3). Fast healing occurred in BSA/HCl culture flasks and only 10% of the wound remained open after 25 h (see Fig. 50.3). Images for the wound closure process in the TGF-β3 culture flask (see Fig. 50.1) showed very similar results, i.e. after ∼5 h the cells had already elongated and started to migrate perpendicularly to the wound bed. Thus the surface area of the cells at the wound edge had increased due to cells replicating and spreading into the wound site. After ∼10 h the cells had already covered a great part of the wound site. Cells from both wound edges had elongated in order to contact the opposite cells and form bridges. Cells had migrated along these bridges and a distance away from the wound edge they were aligned to reorganize the cell structure. These cells had elongated in order to fill the gaps created by the migrated and elongated cells at the wound edges.

After ∼20 h, cells covered the wound site, but there were still some gaps between cells. It was observed that after 20 h all the culture flasks with TGF-β3 showed high percentage of wound closure (only ∼16% remained open, see Fig. 50.3). Elongation and bridging of the cells inside and outside the wound site was also observed and wound healing had almost completed after 25 h. We believe that both HCl and BSA/HCl did not act as a growth stimulant in a pure cell culture environment but they seem to enhance wound closure as compared to the control. TGF-β3 clearly helped to speed up the wound healing process as seen in Figs. 50.2 and 50.3, and almost similar to HCl and BSA/HCl additions.

Our results suggest that wound healing took place in all four culture flasks but at different rates. TGF-β3 additions caused model wounds to heal fully after 25 h whilst control wound remained open even after 30 h. Figures 50.3a and b compare the percentage wound remaining open against time for all four conditions, with control showing the minimum healing, whereas growth stimulators resulted in faster and almost complete healing particularly with TGFβ3 addition.

50.4.2 Cell Behaviour and Morphology

Three different stages in cell behaviour and morphology could be distinguished for the TGF-β3 treated and control flasks: spreading, migration and bridge formation of the cells. This confirmed earlier observations in bone cell cultures [6]. In the TGF-β3, HCl and BSA/HCl flasks cells proliferated, replicated and grew well and rapidly whereas in the control flasks the cells did not replicate as fast as other cultures supplemented with stimulants. Image analysis revealed that in the TGF-β3, HCl and BSA/HCl flasks proliferation seemed to take place not directly at wound edges, but further away from them. This could be due to certain chemicals released by the cells during wound infliction including inflammatory mediators, and various cytokines such as nitric oxide, prostaglandins, etc, which may inhibit cell signalling and hence prevent cell proliferation at wound margins. However, this is only a hypothesis as the scratching had mostly just torn the cell membrane away rather than cutting them in two pieces. Applied wound scratches to the cell monolayers were not straight but uneven which resulted in great variations in wound width and throughout different culture flasks, therefore, normalization of the data was performed to analyze and compare the results.

50.4.3 WSPR Image Analyses

Imaging with the WSPR system revealed that bone cells attached to the surface with high contrast via concentrically arranged band like structures at their focal contacts as shown in Fig. 50.4a–c. The highest contrast bands like components were localized at the cell periphery and that the un-treated cells had lengths ranging between 25 and 35 μm. On the other hand, the treated cells with TGF-β3 had a very different peripheral structure, only visible via the WSPR images as shown in Fig. 50.4d–f.

The cells acquired a high contrast band like structure at the boarder of the elongated features along with less concentrically formed band like features across the cell body. This caused the cells to extend between 50–70 μm in length which are significantly bigger than the un-treated bone cells (Control) indicating that the growth factor induced significant changes in cell activity such as cell signalling resulting in up-regulation of proteins associated with the cells acquiring a more migratory phenotype. WSPR imaging may also help us to understand the un-answered questions about the role of integrin/ECM interactions and gene product up-regulation in induced cell migration during wound healing.

50.5 Statistical Analysis

The wound showed different rate of healing depending on the type of treatment and this was significantly higher for TGF-β3 ($P < 0.001$), BSA/HCl ($P < 0.001$), HCl ($P < 0.05$) as compared to control ($P < 0.05$). As the incubation time increased the rate of wound closure became similar to each other (after ∼20 h of incubation) in the case of TGF-β3 ($P < 0.05$), HCl ($P < 0.05$) and HCl/BSA ($P < 0.001$). The P values were obtained by one-way ANOVA followed by the Bonferroni adjustment.

50.6 Conclusions

The laboratory experiments showed that TGF-β3 does indeed speed up the wound closure process in a pure cell culture environment. After ∼20 h all the culture flasks treated with TGF-β3, HCl and BSA/HCl had resulted in similar but faster wound

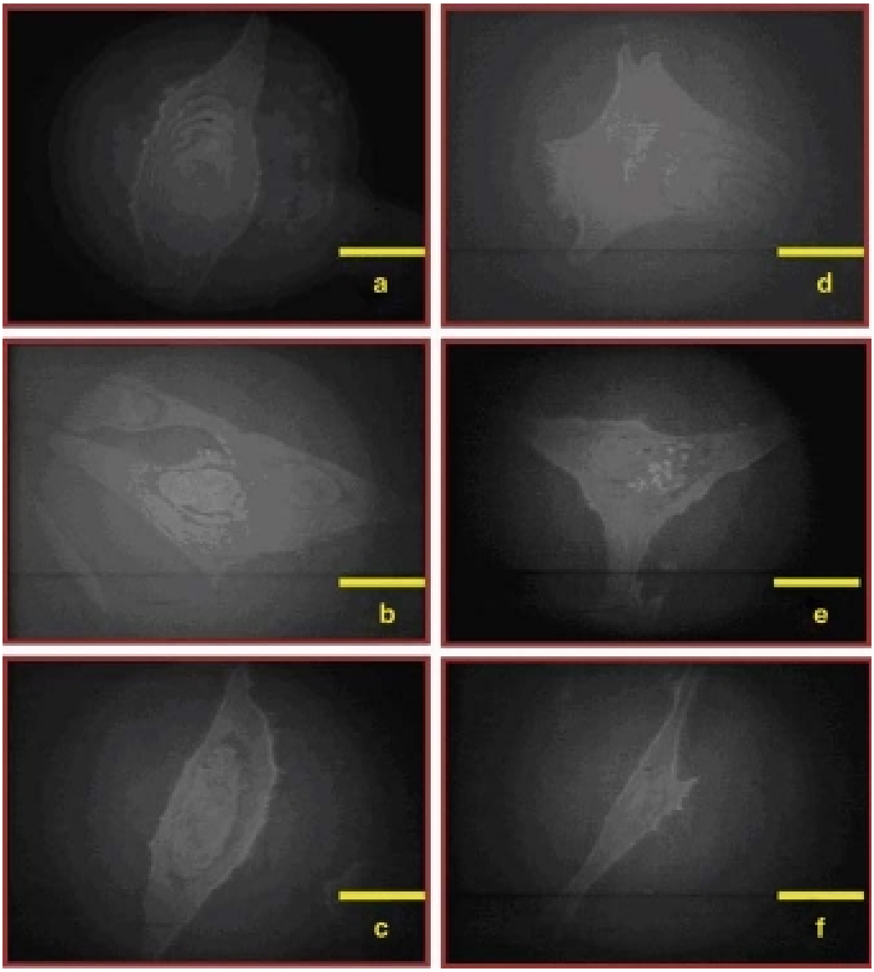


Fig. 50.4 (**a–c**) WSPR images of untreated MG63 bone cell (control); (**d–f**) WSPR images of MG63 bone cell treated with TGF-β3 (scale bar = 25 μm)

healing rate than the control. TGF-β3, HCl alone and HCl/BSA all enhanced the rate of wound repair indicating that TGF-β3 does not act alone in the wound repair system, but instead functions synergistically with signalling pathways that are dependent on the availability of hydrogen ions. Such a mechanism would depend on signalling molecules undergoing a conformational change on binding hydrogen ions. This is not a new concept, one only has to think of haemoglobin's affinity for oxygen as a prime example, but it is potentially a concept that has been overlooked in the wound repair system. This result is important as it shows that healing occurs in control flask but in longer period of time. No clinical research has yet been undertaken on TGF-β3 thus this work is a first step to evaluate TGF-β3, HCl and

HCl/BSA presence in relation to wound closure and the healing process for bone cell monolayers. Imaging of the guided bone cells with WSPR revealed that these cells acquired a high contrast band like structure at the periphery of the elongated cells along with less concentrically formed band like features across the cell body. These images also showed that the bone cells attached to the un-patterned surface with high contrast via concentrically band like structures at their focal contacts but with less elongation as compared to the guided cells. Imaging with WSPR, therefore, allowed observations of the focal contacts without immunostaining.

References

1. Kretschmann, E.H.R.: Radioactive decay of non radiation surface plasmon excited by light. zeitschrift fur naturforschung **23a**, 2135–2136 (1968)
2. Abdul Jamil, M.M.: Application of a novel high resolution widefield surface plasmon microscope in cell engineering, wound healing and development of new binding assays. Ph.D. thesis, University of Bradford, Bradford, UK (2007)
3. Hollinger, J.O.: Bone Tissue Engineering, 1st edn. CRC Press, Boca Raton, FL (2004)
4. Khan, S.N.: Bone growth factors. Orthop. Clin **31**, 375 (2000)
5. Green, J.A., Stockton, R.A., Johnson, C., Jacobson, B.S.: 5-Lipoxygenase and cyclooxygenase regulate wound closure in NIH/3T3 fibroblast monolayers. Am. J. Physiol. Cell Physiol. **287**, C373–C383 (2004)
6. Vooijs, D.P.P., Walboomers, X.F., Parker, J.A.T.C., Von Den Hoff, J.W., Jansen, J.A.: Transforming growth factor Beta3 loaded microtextured membranes for skin regeneration dermal wounds. J. Biomed. Mat. Res. Part A **70A**(3), 402–411 (2004)

Chapter 51
Speech Rehabilitation Methods for Laryngectomised Patients

Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi

Abstract Rehabilitation of the ability to speak in a natural sounding voice, for patients who suffer larynx and voice box deficiencies, has long been a dream for both patients and researchers working in this field. Removal of, or damage to, the voice box in a surgical operation such as laryngectomy, affects the pitch generation mechanism of the human voice production system. Post-laryngectomised patients thus exhibit hoarse, whisper like and sometimes less intelligible speech – it is obviously different to fully phonated speech, and may lack many of the distinctive characteristics of the patients normal voice. However these patients often retain the ability to whisper in a similar way to normal speakers.

This chapter firstly discusses how the laryngectomy affects speech before briefly reviewing the three common methods of speech rehabilitation in such patients. It then presents, as a fourth method, a engineering approach to providing laryngectomy patients the capacity to speak with a more natural sounding voice. As a side effect, this allows them to conveniently use a mobile phone for communications. The approach is non-invasive and uses only auditory information, performing analysis, formant insertion, spectral enhancements and formant smoothing within the reconstruction process. In effect, natural sounding speech is obtained from spoken whispers, without recourse to surgery. The method builds upon our previously published works of using an analysis-by-synthesis approach for voice reconstruction.

Keywords Bionic voice · CELP codec · electrolarynx · laryngectomy · oesophageal speech · rehabilitation · speech processing · TEP · whispered speech

H.R. Sharifzadeh (✉), I.V. McLoughlin, and F. Ahmadi
Block N4, School of Computer Engineering, Nanyang Technological University,
Nanyang Avenue Singapore 639798
e-mail: hami0003@ntu.edu.sg; mcloughlin@ntu.edu.sg; ahmadi@ntu.edu.sg

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_51,

51.1 Introduction

The speech voicing process relies upon modulated lung exhalation passing into the larynx where a taut glottis creates a varying pitch excitation which then resonates through the vocal tract, nasal cavity and out of the mouth. Within the vocal, oral and nasal cavities, the velum, tongue, and lip positions play crucial roles in shaping speech sounds; these are referred to collectively as vocal tract modulators [1].

Corey [2] notes that, "the larynx is the second most common site for cancer in the upper aerodigestive tract" and furthermore continues with "Laryngeal cancers account for approximately 1.2% of all new cancer diagnoses in the United States". National Cancer Institutes SEER data reveals that around four cases of larynx cancer appeared per 100,000 from 1973 to 2000. Corey continues "Squamous cell carcinoma (SCC) is the most common histopathologic diagnosis, accounting for more than 95% of all laryngeal malignancies. Surgery, radiation, or both are the primary treatments for these cancers. Although organ preservation protocols and conservation laryngeal surgeries are in use today, patients with advanced or recurrent SCC of the larynx continue to undergo total laryngectomy in the course of their treatment."

Total laryngectomy patients will have lost their glottis and also the ability to pass lung exhalation through the vocal tract in many cases. Partial laryngectomy patients, by contrast, may still retain the power of controlled lung exhalation through the vocal tract. Despite loss of their glottis, both classes of patient retain the power of vocal tract modulation itself and therefore by controlling lung exhalation (or similar), they have the ability to whisper [3]. In other words, they maintain control of most of the speech production apparatus. Therefore, the novel approach in this chapter is to reconstruct natural speech from the sound created by those remaining speech articulators. However since the major missing component is the pitch-generating glottis, this quest in effect is that of regenerating voiced speech from (pitch-less) whispers.

Various speech rehabilitation techniques exist such as oesophageal speech [4], transoesophageal puncture (TEP) [5], and the electrolarynx [6] (a brief review of these is presented in Section 51.2); but each suffers from weaknesses that range from learning difficulties to clumsy usage and heightened risk of infection. Furthermore, all of these produce speech that is at best unnatural or monotonous. Concentrating on the electrolarynx as the main voice rehabilitation device adopted among laryngectomees [7], valuable efforts [8,9] have recently been made to enhance the quality of the resulting speech by decreasing background and radiated device noise as well as to simplify its usage (i.e. producing a hands free variant). Despite these efforts, however, there has not been any effective method reported to resolve the mechanical sounding (robotized) generated voice characteristic.

The novel speech processing approach discussed in this chapter, by contrast, aims to produce higher quality speech characterised by a more natural sound, utilising a modified code excited linear prediction (CELP) codec to analyse, modify and reconstruct the missing elements in whispered speech, and based in part upon our previously published work in [10–12]. The proposed system works to reconstruct normal speech from whispers in real time, or near-real time.

The chapter is organized as follows: Section 51.2 outlines current methods of speech rehabilitation while Section 51.3 describes whispered speech features regarding the source-filter model and also in terms of their acoustic and spectral features. Section 51.4 points out the proposed approach for the objective of natural speech regeneration along with a novel method for the spectral enhancement during speech reconstruction and finally Section 51.5 summarizes the chapter.

51.2 Current Methods of Speech Rehabilitation

Existing methods of returning speech to post-laryngectomised patients are categorized under three different techniques as briefly reviewed below:

51.2.1 Oesophageal Speech

In this method, the patient is taught to use the oesophagus to expel air by means of stomach contraction rather than lung contraction [4]. The tongue must remain pressed against the roof of the mouth during this procedure to maintain an oesophageal opening.

Oesophageal speech can provide a harsh voice of low pitch, and loudness that is adequate for communication in small groups and quiet settings. Exceptional esophageal speakers may have sufficient versatility and dynamic vocal range to approximate a normal voice, whereas some are unfortunately unable to master this method of communications rehabilitation [13].

Although quite difficult to learn, and often sounding unnatural, oesophageal speech is surprisingly intelligible. However a study by Hillman et al. [14] revealed that only 6% of total laryngectomy patients develop usable oesophageal speech (although five times as many do use or attempt to use it). The current status of oesophageal speech is that it has largely been eclipsed by tracheoesophageal puncture procedures and electrolarynxes [15].

51.2.2 Tracheoesophageal Puncture (TEP)

Surgical operations such as TEP [5] can produce higher quality speech and are particularly suited for those who have had a total laryngectomy and who breathe through a stoma. The TEP procedure creates a small hole to rejoin the oesophagus and trachea. This is then fitted with a one-way valve so that air from the lungs can enter the mouth through the trachea when the stoma is temporarily closed.

Since the introduction of the TEP technique in 1980, numerous clinical and research studies have been published from technique modifications to studies of quality and ease of speech production [15]. While TEP speech is considered by speech–language pathologists as the best method in terms of quality, only around 30% of post-laryngectomised patients use this method of alaryngeal speech [14].

The relatively good speech quality compared to the other voice rehabilitation techniques, and the high success rate of achieving usable voice requiring limited teaching are the main advantages of this method while the daily maintenance of the prosthesis by the patient, the recurrent leakage of the prosthesis after a period of time and the consequent need of replacement by the clinician (including the cost of replacement), are the disadvantages of this method. Furthermore, the prosthesis is clumsy in use and is a potential risk area for infection.

51.2.3 Electrolarynx

The electrolarynx is a razor sized device that needs to be pressed against the side of the throat to resonate the vocal tract [6]. there are two types of electrolarynx: the neck and the intra-oral types, of which the former is the most widely used among the laryngectomees. During phonation, the hand-held device is held against the neck approximately at the level of the former glottis to insert a buzzing vibration into the oral and pharyngeal cavities by means of a built-in electromechanical vibrator. This sound source is transmitted through the neck tissues to resonate the vocal cavity. The user modulates this resonance to create speech by movements of articulators such as the lips, teeth, tongue, jaw and velum.

Speech generated by the electrolarynx is mechanical sounding and monotonous, although some modern units have a hand control to vary pitch. It has been found that the use of the electrolarynx is one of the easier methods of speech rehabilitation, and is a more effective for communication in many situations [16]. Although oesophageal speech and transoesophageal speech are common in voice rehabilitation, electrolarynx phonation is the most commonly adopted method [15], with more than 55% of post-laryngectomised patients currently using it [14].

By and large, these techniques suffer from a common weakness: they produce unnatural monotonous ‘robotized’ speech. The approach discussed in Section 51.4, by contrast, aims to produce higher quality speech by utilising a modified code excited linear prediction (CELP) codec to analyse, modify and reconstruct speech.

51.3 Whispered and Phonated Speech

Whispered speech as opposed to normally phonated (pitched) speech forms the main focus of the research regarding speech regeneration for laryngectomy patients since they, particularly partial laryngectomy patients, can often produce whispered speech with little effort.

Soft whispers (or ‘quiet whispers’) are produced by normally speaking people to deliberately reduce perceptibility, such as in a library, usually done in a relaxed manner [17]. They are produced without vocal fold vibration. and have similar characteristics to whispers from laryngectomised persons (although some patients may be capable of partial phonation).

The absence of vocal cord vibration leads in turn to the absence of fundamental frequency and lack of harmonics [18]. Using a source filter model [19], exhalation can be identified as the source of excitation in whispers, with the shape of the pharynx adjusted so that the vocal cords do not vibrate [20]. Turbulent aperiodic airflow is therefore the source of sound for whispers.

The spectral characteristics of whispered speech sounds still exhibit some spectral peaks at similar frequencies to those for normally phonated speech [21]. These ‘pseudo-formants’ occur within a flatter power frequency distribution, and lack obvious harmonics in the spectra corresponding to the fundamental frequency [18]. Figure 51.1 shows this feature by contrasting the spectra of the vowel /a/ spoken in a whisper and in a normal voice.

Whispered vowels also differ from normally voiced vowels. All formant frequencies (including the important first three formants) tend to be higher [22], particularly the first formant which shows the greatest difference between two kinds of speech.

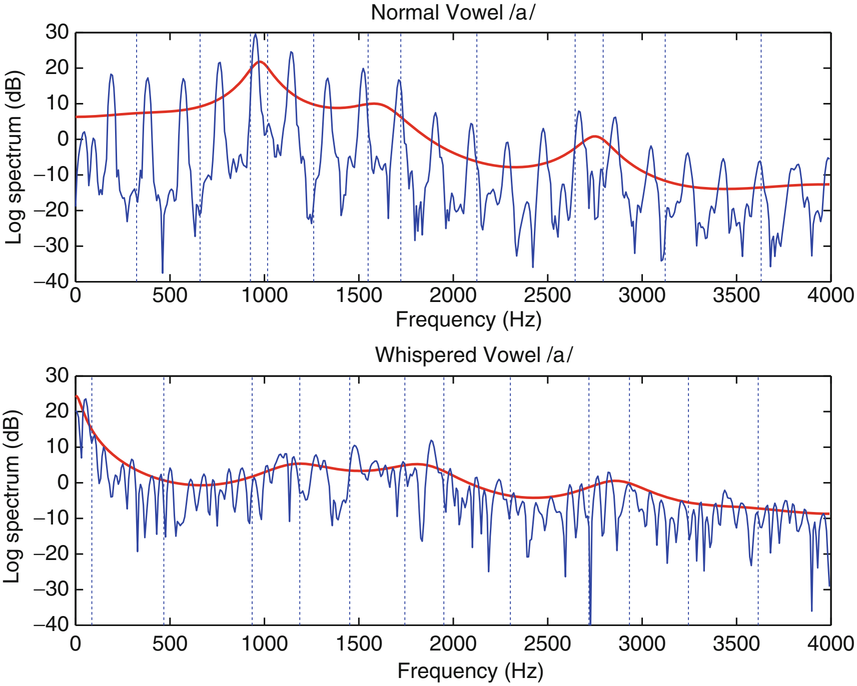


Fig. 51.1 The spectra for vowel /a/ in normal speech (*top*) and whisper (*bottom*) for a single speaker. A smoothed spectrum is overlaid, showing formant peaks

Lehiste [22] reported that F1 is approximately 200–250 Hz higher, whereas F2 and F3 are approximately 100–150 Hz higher in whispered vowels.

Another observed consequences of a glottal opening is an acoustic coupling to the subglottal airways. The subglottal system has its own resonances, which can be defined as its set of natural frequencies with closed glottis. The average values of the first three of these natural frequencies have been estimated to be about 700, 1,650, and 2,350 Hz for an adult female and 600, 1,550, and 2,200 Hz for an adult male [23], but of course substantial person-to-person variation exists.

The effect of these subglottal resonances is to introduce additional pole-zero pairs into the vocal tract transfer function from the glottal source to the mouth output. The most obvious acoustic manifestation of these pole-zero pairs is the appearance of additional peaks in the output spectrum. The influence of zeros can sometimes also be seen as minima in the spectrum [24]. In the next section, the proposed approach for whisper-voice conversion with consideration of these features is discussed.

51.4 Speech Regeneration

The approach described in this chapter utilises a CELP codec to adjust whisper speech to sound more like fully phonated speech. Within CELP, excitation is selected from a codebook of zero-mean Gaussian sequences which are then shaped by an LTP (longterm prediction) filter to convey the pitch fundamental of the speech. Amongst the variants of analysis-by-synthesis LPC (linear predictive coding) schemes, CELP is one of the more popular, especially for low-bit rate coding [25].

Within most CELP codecs, linear prediction coefficients are transformed into line spectral pairs (LSPs) [26] which are used to convey the resonance characteristics of an interconnected tube model of the human vocal tract. Two states in the model describe the vocal tract being either fully open or fully closed at the glottis respectively. Since the human glottis is actually opened and closed rapidly during normal speech, the true resonances lie somewhere between the two extreme conditions (and this can clearly be seen in any analysis of speech LSPs [26]). However for whispered speech, since the glottis does not vibrate, this model no longer holds true: it is therefore necessary to adjust the LSP model slightly [10].

Since altering the frequency of lines may result in unintentional peak formation through narrowing the gap between two irrelevant pairs, it is important to choose the pair of lines corresponding to likely formants. As mentioned, this might be done by choosing the three narrowest LSP pairs which works well when the signal has fine peaks, but in case of the expansion of formant bandwidths (common in whispered speech), which leads to the increase of distance between the corresponding LSPs, the choice of the narrowest LSPs may not identify the correct formant locations, particularly for vowels. Although a strengthened LSP-based method has been described in [11], the enhancement procedure is further modified to perform effectively on all whispered vowels and diphthongs. Section 51.4 describes this effective technique and presents the corresponding results.

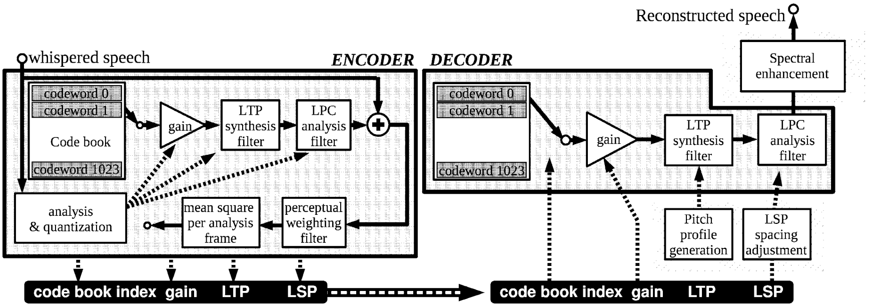


Fig. 51.2 Block diagram of the reconstruction codec, showing a typical CELP encoder on the *left* and decoder on the *right* augmented with additional processing units to adjust LSPs, generate LTP coefficients and spectrally enhance output speech. Note that the LTP coefficients generated by the encoder are not used in the decoder, since these primarily relate to pitch information which is absent in whispered speech

A block diagram of the CELP codec as implemented in this research is shown in Fig. 51.2, with the modifications for whisper-speech reconstruction identified. In comparison with the standard CELP codec, we have added a "pitch template" corresponding to the "pitch estimate" unit while "adjustment parameters" in this model are used to generate pitch factors as well as to apply necessary LSP modifications.

The pitch estimation algorithm implemented in this research is based on extraction parameters from normally phonated speech which are then re-applied in the CELP excitation [11] as a reconstructed pitch signal. A selection algorithm judges the underlying phoneme type from detected parameters. Since the current focus is not on this detector, its decision in this case was manually assessed and, if necessary, appropriately overridden to ensure accuracy.

Reconstruction of phonated speech from whispered samples involves a critical stage of spectral enhancement, in part due to the much lower SNR of recorded whispers compared with normally phonated speech: estimates of vocal tract parameters for such speech have a much higher variance than those of normal speech. As mentioned in Section 51.3, the vocal tract response for whispered speech is noise excited and this differs from the expected response when the vocal tract is excited with pulses as in normally phonated speech.

For the regeneration of vowels, such differences become exacerbated due to the instability of the resonances in the vocal tract (peaks of frequency response of the vocal tract, i.e. formants) being quite strong. To prepare a whispered speech signal for pitch insertion, consideration is therefore required for the enhancement of the spectral characteristics, especially for disordered or unclear formants caused by the background noise and excitation feed-through evident in whispers. A novel approach for this kind of enhancement has been described in [12] which provides new smoothed formant frequencies of whispered speech.

Having obtained modified formant frequencies, it is necessary to apply a proportionate improvement in their corresponding bandwidths. This should be done in such a way that not only should formant frequencies be retained but also their energy improved to prevail over attenuated whispers. For this purpose, a suggestion by Hsiao and Childers [27] has been customised for the present objective. This utilises the different spectral energy between whispered and normal speech, as well as maintaining necessary considerations for whispered speech.

As a matter of fact, the bandwidth modification process tries to form the new spectrum for whispered speech to be as similar as possible to normally phonated speech in terms of formant shapes and spectral tilt. This can be descried through an adjustment to the linear predictive description of a frame of whispered speech. If a pole has a transfer function and power spectrum as follows:

$$H(z) = \frac{1}{1 - re^{j\theta}z^{-1}} \tag{51.1}$$

$$|H(e^{j\phi})|^2 = \frac{1}{1 - 2r\cos(\phi - \theta) + r^2} \tag{51.2}$$

and thus when there are N poles together describing the speech:

$$|H(e^{j\phi})|^2 = \prod_{i=1}^{N} \frac{1}{1 - 2r_i\cos(\phi - \theta_i) + r_i^2} \tag{51.3}$$

then the radii of the poles may be modified such that the spectral energy of the formant polynomial is equal to a specified spectral target value. showing the spectral energy difference between normal and whispered speech (according to [28], whispered speech has 20 dB lower power than its equivalent phonated speech).

Suppose there is a formant pole with given radius and angle, now by using (51.3) the spectral energy of the formant polynomial, $H(z)$, at the modified angle θ_i^M is given by:

$$H(e^{j\theta_i^M})|^2 = \frac{1}{1 - r_i^2} \prod_{j \neq i}^{N} \frac{1}{1 - 2r_j\cos(\theta_i^M - \theta_j^M) + r_j^2} \tag{51.4}$$

where $|H(e^{j\theta_i^M})|^2$ is the specific power spectrum value at angle θ_i^M and N is the total number of modified formant poles. There are two spectral components on the RHS of the equation: one is produced by the pole itself and the other is the effect from remaining poles with modified angles. By solving (51.4), we can find a new radius for the i^{th} pole while retaining its modified corresponding angle, θ_i^M. Furthermore, to maintain stability, if r_i exceeds unity, we use its reciprocal value. Thus, the modified radius, r_i^M, for each pole is obtained through (51.5):

$$r_i^M = 1 - \left(\frac{1}{H_i^M} \prod_{j \neq i}^{N} \frac{1}{1 - 2r_j\cos(\theta_i^M - \theta_j^M) + r_j^2} \right)^{1/2} \tag{51.5}$$

where H_i^M represents $|H(e^{j\theta_i^M})|^2$, the target spectral energy for each pole.

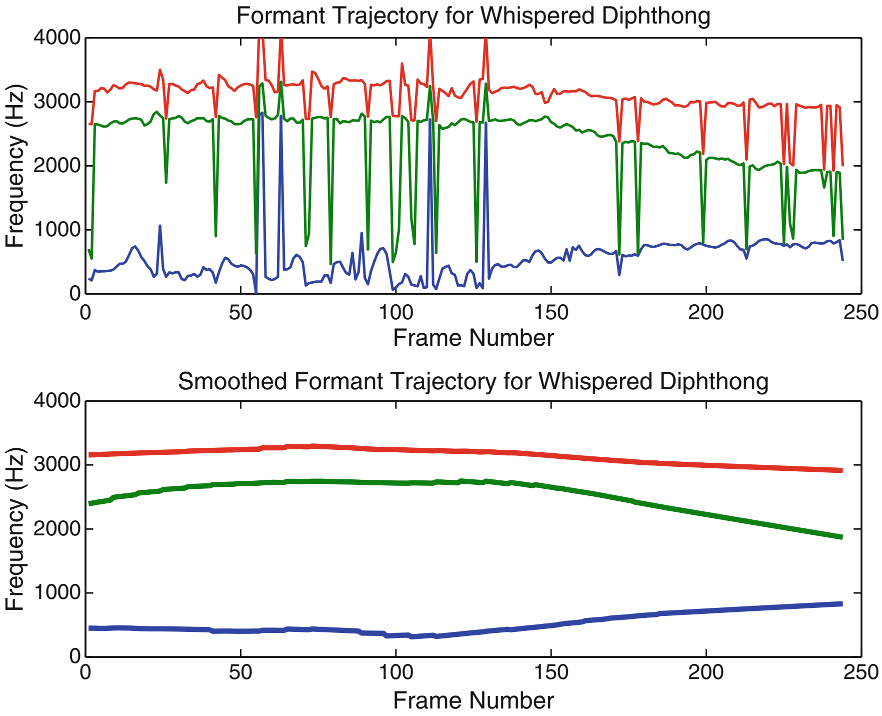


Fig. 51.3 Formant trajectory for original whispered diphthong /ie/ (*top*) and the smoothed vector (*bottom*). Note the diphthong transition toward the right hand side

Since the formant roots are complex-conjugate pairs, only those that have positive angles are applied to the algorithm, and their conjugate parts are obtained readily at the final stage. The radii modification process using (51.5) starts with the pole whose angle is the smallest and it continues until all radii are modified.

Figure 51.3 demonstrates the formant trajectory for a whispered diphthong (/ie/) before applying the spectral enhancement, and the resulting formant trajectory after the application of the smoothing technique. These show the effectiveness of the method even for transition modes of formants spoken across diphthongs.

Furthermore, listening tests have been conducted which indicate that reconstructed vowels and diphthongs, are significantly more natural than electrolarynx versions. The major deficiency in the implemented technique as it currently stands relates to the transition between phonemes: unlike the electrolarynx, the method relies upon the accurate estimation of whispered phoneme. The evaluation and listening tests conducted to date apply a manual correction to phoneme classification to ensure that mis-classification errors are not contributors to system performance loss.

51.5 Summary

This chapter has discussed the methods of returning the power of speech to post-laryngectomised patients. Three current and popular methods of prosthesis are overviewed, along with a fourth method which uses non-medical and non-invasive computational techniques to regenerate voiced speech from the whispers of post-laryngectomised patients.

By the use and analysis of whisper speech, allied with a novel method of reconstructing formant locations and reinserting pitch signals, this chapter along with our previous works [10, 11], describes a viable algorithmic approach for a system potentially able to provide such patients the capacity to re-attain a natural sounding voice.

In particular, this chapter presented an innovative method for the spectral enhancement and formant smoothing within the speech regeneration process. The smoothed formant trajectory resulting from applying the proposed enhancement method to a typical diphthong was illustrated to demonstrate the effectiveness of the method, and results from listening tests noted which indicated that the technique improves upon the quality of the most popular prosthesis: the electrolarynx.

References

1. Vary, P., Martin, R.: Digital Speech Transmission. Wiley, West Sussex (2006)
2. Corey, C.L.: Voice rehabilitation after total laryngectomy. Baylor College of Medicine, 2005. http://www.bcm.edu/oto/grand/08_25_05.htm
3. Pietruch, R., Michalska, M., Konopka, W., Grzanka, A.: 'Methods for formant extraction in speech of patients after total laryngectomy. Biomed. Signal Proc. Cont. **1**, 107–112 (2006)
4. Azzarello, M., Breteque, B.A., Garrel, R., Giovanni, A.: Determination of oesophageal speech intelligibility using an articulation assessment. Rev. Laryngol. Otol. Rhinol. (Bord) **126**, 327–334 (2005)
5. Callanan, V., Gurr, P., Baldwin, D., White-Thompson, M., Beckinsale, J., Bennet, J.: Provox valve use for post-laryngectomy voice rehabilitation. J. Laryngol. Otol. **109**, 1068–1071 (November 1995)
6. Brandenburg, J.H.: Vocal rehabilitation after laryngectomy. Arch. Otolaryngol **106**, 688–691 (November 1980)
7. Morris, H.L., Smith, A.E., Van Demark, D.R., Maves, M.D.: Communication status following laryngectomy: the Iowa experience 1984-1987. Ann. Otol. Rhinol. Laryngol. **101**, 503-510 (1992)
8. Liu, H., Zhao, Q., Wan, M., Wang, S.: Enhancement of electrolarynx speech based on auditory masking. IEEE Trans. Biomed. Eng. **53**, 865–874 (2006)
9. Goldstein, E.A., Heaton, J.T., Kobler, J.B., Stanley, G.B., Hillman, R.E.: Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. IEEE Trans. Biomed. Eng. **51**, 325–332 (2004)
10. Ahmadi, F., McLoughlin, I.V., Sharifzadeh, H.R.: Analysis-by synthesis method for whisper-speech reconstruction. In: Proceedings of IEEE APCCAS, pp. 1280–1283 2008
11. Sharifzadeh, H.R., McLoughlin, I.V., Ahmadi, F.: Regeneration of speech in voice-loss patients. In: Proc. ICBME **23**, 1065–1068 (2008)

12. Sharifzadeh, H.R., McLoughlin, I.V., Ahmadi, F.: Speech reconstruction in post-laryngectomised patients by formant manipulation and pitch profile generation. In: Proc. ICSBB **II**, 1838–1843 (2009)
13. Gates, G.A., Ryan, W., Cooper, J.C., Lawlis, G.F., Cantu, E., Hayashi, T., Lauder, E., Welch, R.W., Hearne, E.: Current status of laryngectomee rehabilitation: I. Results of therapy. Am. J. Otolaryngol. **3**, 1–7 (1982)
14. Hillman, R., Walsh, M., Wolf, G., Fisher, S.: Functional outcomes following treatment for advanced laryngeal cancer. Part 1. Voice preservation in advanced laryngeal cancer. Part II. Laryngectomy rehabilitation: the state-of-the-art in the VA system. Ann. Otol. Rhinol. Laryngol. **107**, 1–27 (1998)
15. Culton, G., Gerwin, J.: Current trends in laryngectomy rehabilitation: a survey of speech language pathologists. Otolaryngol. - Head Neck Surg. **115**, 458–463 (1998)
16. Liu, H., Ng, M.: Electrolarynx in voice rehabilitation. Auris Nasus Larynx **34**, 327–332 (2006)
17. Solomon, N.P., McCall, G.N., Trosset, M.W., et al.: Laryngeal configuration and constriction during two types of whispering. J. Speech Hear. Res. **32**, 161–174 (1989)
18. Tartter, V.C.: What‘s in whispers? J. Acoust. Soc. Am. **86**, 1678–1683 (1989)
19. Fant, G.: Acoustic Theory of Speech Production. Mouton, The Hague (1960)
20. Thomas, I.B.: Perceived pitch of whispered vowels. J. Acoust. Soc. Am. **46**, 468–470 (1969)
21. Stevens, H.E.: The representation of normally-voiced and whispered speech sounds in the temporal aspects of auditory nerve responses. Ph.D. Thesis, University of Illinois (2003)
22. Lehiste, I.: Suprasegmentals. MIT, Cambridge (1970)
23. Klatt, D.H., Klatt, L.C.: Analysis, synthesis, and perception of voice quality, variations among male and female talkers. J. Acoust. Soc. Am. **87**, 820–857 (1990)
24. Stevens, K.N.: Acoustic Phonetics. MIT, Cambridge, MA (1998)
25. Kondoz, A.M.: Digital Speech Coding for Low Bit Rate Communication Systems. Wiley (1994)
26. McLoughlin, I.V.: Line spectral pairs. Signal Process. J. 448–467 (2007)
27. Hsiao, Y.S., Childers, D.G.: A new approach to formant estimation and modification based on pole interaction. In: Proceedings of IEEE Asilomar CSSC, pp. 783–787 (1997)
28. Jovivcic, S.T.: Formant feature differences between whispered and voiced sustained vowels. Acustica-Acta Acustica **84**, 739–743 (1998)

Chapter 52
Study of the Tip Surface Morphology of Glass Micropipettes and Its Effects on Giga-Seal Formation

Majid Malboubi, Yuchun Gu, and Kyle Jiang

Abstract Reported is a study of applying nanofabrication technology to improve the surface roughness of micro glass pipettes to achieve giga ohm seal resistance in patch clamping processes. The surface roughness of pipette tips was first measured by 3D reconstruction of pipette tips using stereo imaging technique based on high resolution SEM images. Both the SEM images and the reconstructed images show that micro glass pipettes have rough and uneven tips which could be one of the causes of leakage in patch clamping. Then focused ion beam system was used to cut across the very end of the tip, producing a smooth and flat new tip. The average surface area roughness S_a of a milled pipette tip was within a few nanometres. Patch clamping experiments were carried out using the polished pipettes on human umbilical vein endothelial cells (HUVEC), which were well known for their extremely flat shape making them very difficult to patch. The results show that above 3 GΩ seals were achieved in 60% of the experiments, as opposed to 1.5–2.0 GΩ in average with the conventional pipettes. The highest seal resistance achieved with a focused ion beam polished pipette was 9 GΩ, well above the 3 GΩ resistance, the usually best result achieved with a conventional pipette. The research results demonstrate that the surface roughness of a pipette has a significant effect on the giga-seal formation of a patch clamping process.

Keywords Roughness · giga-seal formation · patch clamping · pipette · focused ion beam

M. Malboubi (✉) and K. Jiang
School of Mechanical Engineering, University of Birmingham, Birmingham, B15 2TT, UK
e-mail: mxm726@bham.ac.uk; k.c.jiang@bham.ac.uk

Y. Gu
School of Medicine, University of Birmingham, Birmingham, B15 2TT, UK
e-mail: y.gu@bham.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_52,

52.1 Introduction

Patch clamp technique has been extensively used for cellular ion channels studies. Introduced by [1], the technique is capable of detecting currents flowing in/out of the cell through a single ion channel at highest resolution. Nowadays it has been proven that many different diseases can be caused by the malfunction of ion channels [2]. The activities of various ion channels under different physical and chemical stimulations and the communications of cells can be studied with the aid of patch clamp method and these studies help us to better understand the fundamentals of cells [3]. In patch clamping a glass micropipette is used to isolate a patch of membrane from external solution to record the currents flowing into the patch. To achieve this small glass capillaries are heated and pulled to fabricate glass micropipettes with a tip diameter of 1–2 μm. The micropipettes are then backfilled with a conductive solution and pressed against the surface of a cell. To improve the sealing condition a gentle suction is applied to the backend of the pipette. As it is shown in Fig. 52.1 there are two electrodes in patch clamp set-up: a recording electrode inside the pipette and a reference electrode in the bath solution. In order to be able to detect single ion channel currents which are in the order of few Pico Amperes, there should be a high resistance seal between the glass and the patch of membrane. The high resistance seal reduces the leakage current between the two electrodes, completes the electrical isolation of the membrane patch and reduces the current noise of the recording [4]. Since the electrical resistance of the seal is in the order of giga-ohms, it is called giga-seal.

The physical and chemical mechanisms behind the giga-seal formation are not fully understood yet, which could probably be a result of high number of different factors important for giga seal formation, such as: the cleanliness of the pipette and cell surface [5, 6], surface roughness of the tip [7–9], geometry of the tip [3, 10], hydrophilicity of the patch site [11, 12], material type [11, 13, 14], glass type of pipette [15, 16], tip size [3], presence of positive ions in the solution [17], gentle approaching of the pipette to the cell membrane [6], vibration of the pipette [4], skill and patience of the operator, etc.

This research work is designed to improve the sealing resistance in patch clamping and to acquire giga-seals more frequently. In this work, the effect of the surface roughness of glass micro pipettes on seal formation was examined. The tip of

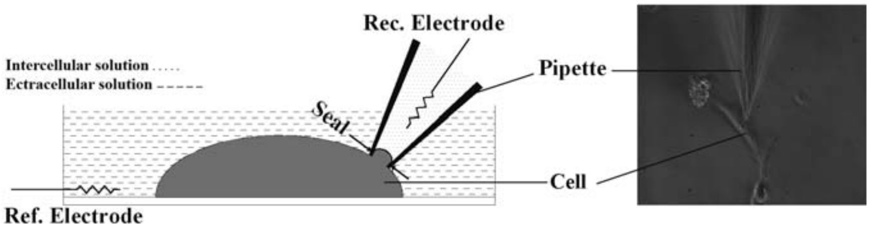


Fig. 52.1 The schematic of the patch clamping method

pipettes were imaged and reconstructed. The surface roughness was measured. The pipette tips were then milled using a focused ion beam (FIB) system resulting in a highly smooth surface. Extensive patch clamp experiments were carried out to investigate the effect of the roughness on seal formation. We call this method “FIB polishing” in comparison with the “fire polishing method”. Compared with fire polishing [18,19] the pipettes polished using FIB have much smoother tip surfaces. The nanomachined pipettes were used in patch clamp recording experiments and much improved gigaseal formation has been achieved.

52.2 3D Reconstruction of Micropipette Tip

The glass micro pipettes used in the experiments were made of borosilicate glass pipes with outer diameter of 1.5 mm and inner diameter of 0.86 mm (BF150-86-10 Sutter Instrument). They were heated and pulled with flaming/brown micro pipette puller machine (Sutter Instrument Model P-97). The machine was set to produce pipettes with approximately 1.5 μm in tip diameter. The 3D reconstruction of the pipette tip was based on high resolution scanning electron microscope (SEM) images. Stereoscopic techniques have been widely used to determine the three dimensional structure of an object in which the same area of the object is scanned from different angles by tilting the object with respect to the fixed optic axis [20]. In this technique measurements are made under two different perspectives. Surface features in different heights have different lateral displacements and depth can be calculated by measuring the parallax movement of features from their location in the first image, to the new location in the second image [21]. 3D points are computed from 2D matched points in two SEM images taken from two angles between 5° and 10° away from the norm. Figure 52.2 shows the configuration of SEM, tilting angle (α) and the projected coordination P1(X1, Y), P2(X2, Y).

The third dimension can be found from Eq. (52.1) which can be derived based on the geometry of the projection.

$$Z = \frac{x_2 - x_1 + x_2(1 - \cos\alpha)}{\sin\alpha} \tag{52.1}$$

This process is used for every point of the object to find the shape of the structure [22]. 3D surface profile of the pipette was obtained by analyzing three SEM images using a commercial software package Mex (Alicona) [23]. Figure 52.3 shows the configuration of the FIB polishing and SEM imaging used in the experiments. Also, Fig. 52.4a–c show the SEM images taken from the left, middle and right of the pipette. The tilting angle between (a–b) and (b–c) of the images is 9°. Figure 52.5 shows the 3D reconstructed surface of the pipette tip. It was found that the pipette tip was not only rough, but also wavy or inclined in its form. The surface parameters computed by considering both the roughness and shape of the tip are given in Table 52.1.

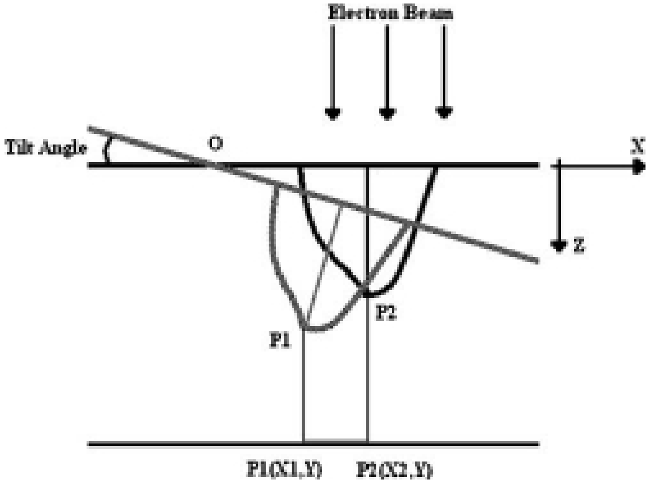


Fig. 52.2 A pair of images of a single object for reconstruction of the object. P1 is the new position of P2 after a tilt of the stage about O

Fig. 52.3 Schematic of the configuration of SEM and the tilting angle (α)

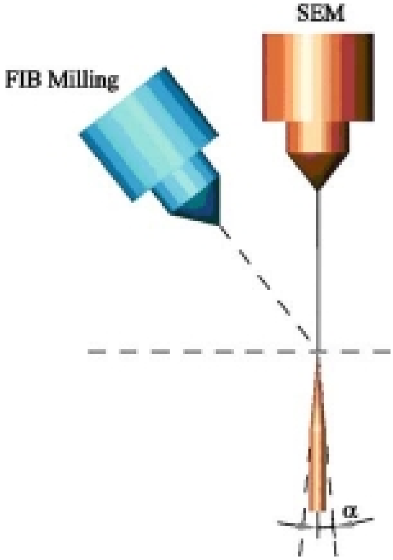


Figure 52.6 shows the shape of the tip along different profiles across the thickness of the pipette. The large variation of the surface morphology shown in Fig. 52.6 increases the chance of ion escape and compromises the formation of high resistance seals. Profile parameters of the profiles are given in Table 52.2.

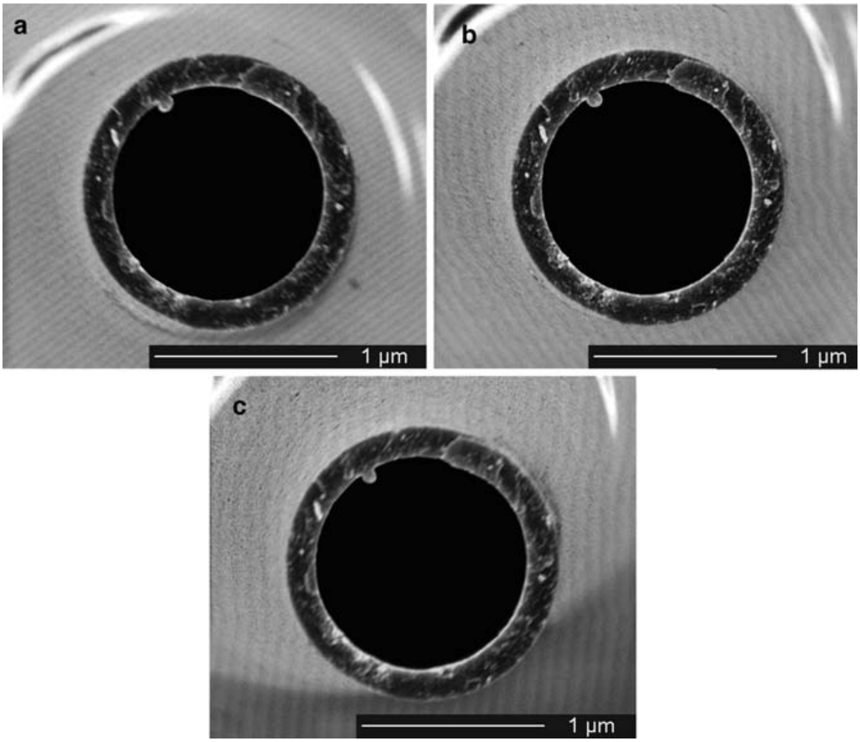


Fig. 52.4 Stereo images of the pipette tip for 3D reconstruction: (**a**) left, (**b**) middle and (**c**) right

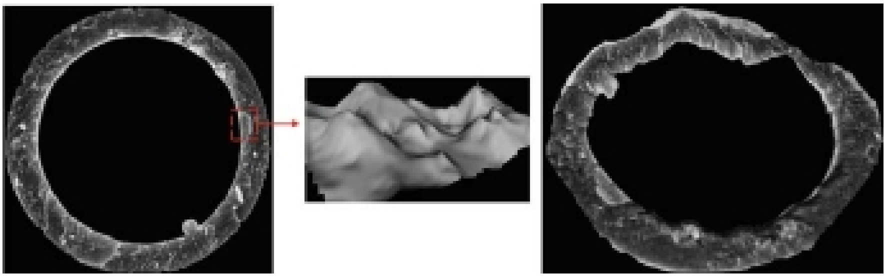


Fig. 52.5 A 3D reconstructed surface of the pipette tip shown at different viewing angles; top view (*right*), the exploded view of the area showed by dash-line (*middle*), view with an angle (*right*)

52.3 Focused Ion Beam Polishing

The uneven surface of the pipette tip was corrected by cutting the top of the pipette across using FEI dual beam focused ion beam system. Because of the conic shape of the pipette, cutting the tip changes the tip size which is an important factor in

Table 52.1 Surface parameters of the pipette tip

Name	Value	Description
Sa	27 nm	Average height of selected area
Sq	34 nm	Root-mean-square height of selected area
Sp	104 nm	Maximum peak of selected area
Sv	150 nm	Maximum valley depth of selected area
Sz	255 nm	Maximum height of selected area
Ssk	−0.225	Skewness of selected area
Sku	3.26	Kurtosis of selected area
Sdq	0.877	Root mean square gradient
Sdr	34.98%	Developed interfacial area ratio

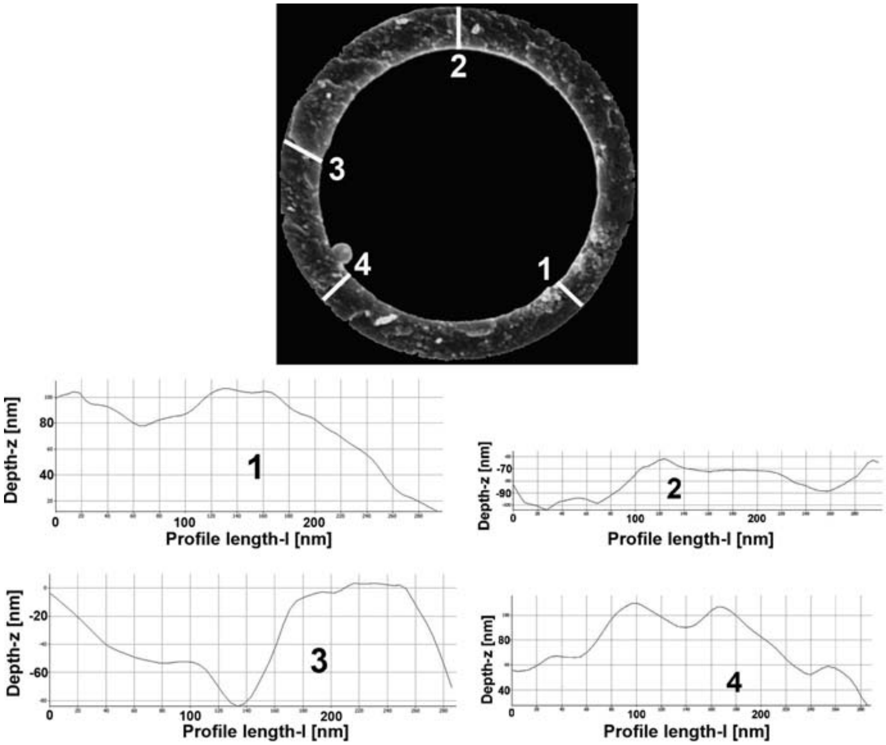


Fig. 52.6 Four different profiles of the tip surface across the thickness are shown. The large variation of surface morphology compromises the formation of a high resistance seal

patch clamping as it determines the pipette resistance. It is also well known that a giga-seal is not likely to be achieved with big tip sizes. So care was taken not to cut more than 1 μm from the top. Since the roughness of the tip of the pipette was in nanometres cutting 1 μm from the top should be sufficient to remove all rough edges

Table 52.2 Profile parameters of four different profiles

Profile	No. 1	No. 2	No. 3	No. 4	Description
Pa	16 nm	8 nm	22 nm	18 nm	Average height of profile
Pq	18 nm	9 nm	25 nm	20 nm	Root-mean-square height of profile
Pt	61 nm	36 nm	95 nm	71 nm	Maximum peak to valley height of primary profile
Pz	21 nm	21 nm	46 nm	31 nm	Mean peak to valley height of primary profile
Pp	30 nm	20 nm	43 nm	30 nm	Maximum peak height of primary profile
Pv	31 nm	16 nm	52 nm	40 nm	Maximum valley height of primary profile
Pc	0 nm	34 nm	79 nm	0 m	Mean height of profile irregularities of primary profile
Psm	0 nm	211 nm	221 nm	0 m	Mean spacing of profile irregularities of primary profile
Psk	0.1269	0.1976	−0.3735	−0.0375	Skewness of primary profile
Pku	1.7548	1.9034	2.0228	1.6307	Kurtosis of primary profile
Pdq	0.5648	0.4698	1.0879	0.7267	Root-mean-square slope of primary profile

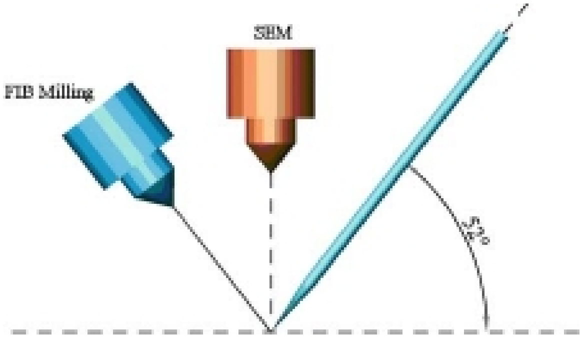


Fig. 52.7 The configuration of glass micro pipette milling in the SEM/FIB chamber. The stage was tilted by 52° so that the ion beam was perpendicular to the pipettes

without increasing the tip size significantly. In the FIB milling process, the pipettes tips were cut using Ga+ ions with 50 pA current for 100 s and dwell time of 1 μs (Fig. 52.7). The pipette before and after milling is shown in Fig. 52.8. The image of the milled pipette, shown in Fig. 52.8 (right), has a resolution of 4.5 nm. No feature could be identified on the milled surface for producing roughness parameters at this magnification. Therefore, the average surface area roughness (Sa) of the milled pipette tip should be less than 4.5 nm.

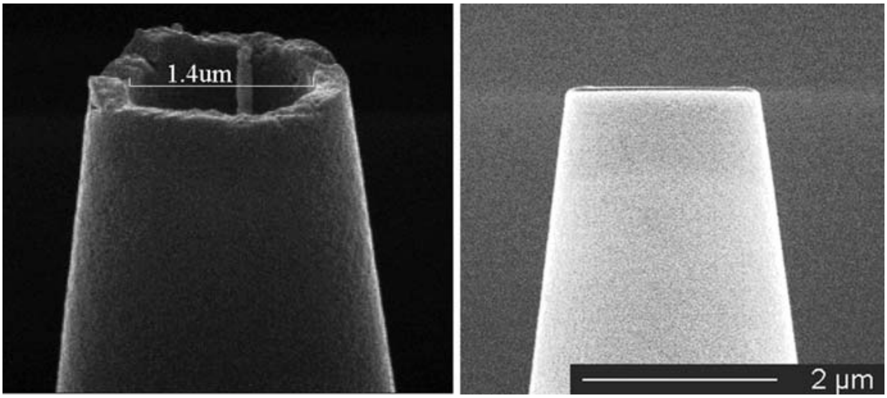


Fig. 52.8 A micro glass pipette before milling (*left*), the pipette after the milling (*right*). No surface roughness could be identified after milling, so the surface roughness should be smaller than the resolution of the SEM image, which is 4.5 nm

52.4 Patch Clamping Experiments

Human umbilical vein endothelial cells (HUVECs) were utilized to investigate the performance of the FIB polished micro pipettes in achieving giga-ohm seals. HUVECs were cultured in EBM medium (Lonza Co., CC-3121) on cover slips 2–3 days before the experiment and incubation was done at 37 °C. At the time of experiments, the confluence of the cells is over 80% and all the cells were firmly attached to the bottom of the cover slips. HUVECs are well known for their extremely flat shape; thus, it is one of the most difficult cell types for patch clamping. During experiments, individual cover slip was directly taken out from incubator and sited in the recording chamber.

Experimental equipment setup consisted of Axon 1D amplifier, Flaming/Brown micro pipette puller (Sutter Instrument Model P-97) and glass micro pipettes (BF150-86-10 Sutter Instrument). The opening of the pipette tip is about 1.4 μm in diameter. The backfill solution contained (in mM): kcl 40, K-gluconate 96, K2ATP 4, GTP 2, HEPE 10, pH 7.2, and the bath solution contained (in mM): NaCl 110, KCl 5, MgCl2 1, CaCl2 1, HEPEs 5, HEPE-Na 5 (mM), pH 7.2.

A 10 mV pulse was constantly applied on the recording electrode from the time that pipette tip was just immersed in the bath solution till it touched the cell membrane. A negative pressure was immediately applied to the pipette and then the voltage pulse was raised to 60 mV to monitor the seal resistance precisely.

To investigate the effect of the roughness of pipette tips, experiments were carried out with polished and conventional pipettes under the same conditions and the results were compared. When there was no contact between recording pipette and cell membrane, the total resistance ranged from 6.0 to 6.5 MΩ. With the FIB polished pipettes, above 3 GΩ seals were achieved in 60% of the experiments (n = 20) and the highest seal resistance reached 9 GΩ. In comparison, the seal resistance

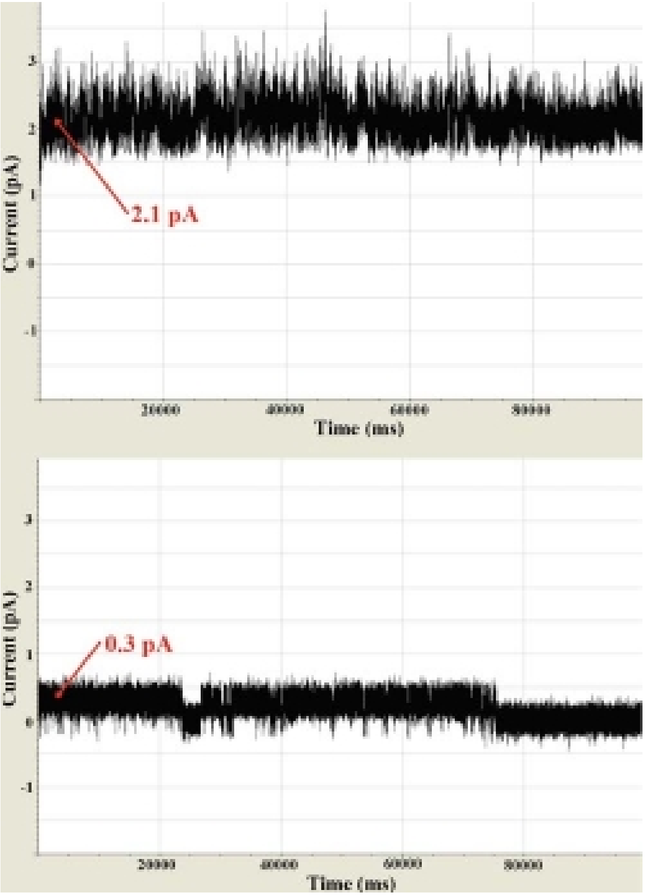


Fig. 52.9 Single channel recording from HUVECs. Conventional pipettes (*top*). FIB polished pipettes (*bottom*)

achieved using the conventional pipettes are 1.5–2.0 GΩ in average and the seal resistance could reach 3 GΩ in some excellent cases. The current in the case of polished pipettes is 0.3 pA, significantly smaller than 2.1 pA achieved using conventional pipettes, indicating higher seal resistance in this case. Single-channel currents recorded from conventional and polished pipettes are shown in Fig. 52.9.

The improved patch clamping performance with polished pipettes is obtained from better contact conditions of the smoother tip surface with membrane. This can be understood as that the smoother surface of the pipette tip leaves little concave area to hold water, opposed to the conventional pipettes, as illustrated in Fig. 52.10. This shows that the membrane can not fill the valleys of the rough surface of conventional pipettes perfectly which could be possibly the reason for reports on lower seal resistance with rough surfaces in the literature. Higher seal greatly reduces the chance of current leakage and reduces the current noise of the recording.

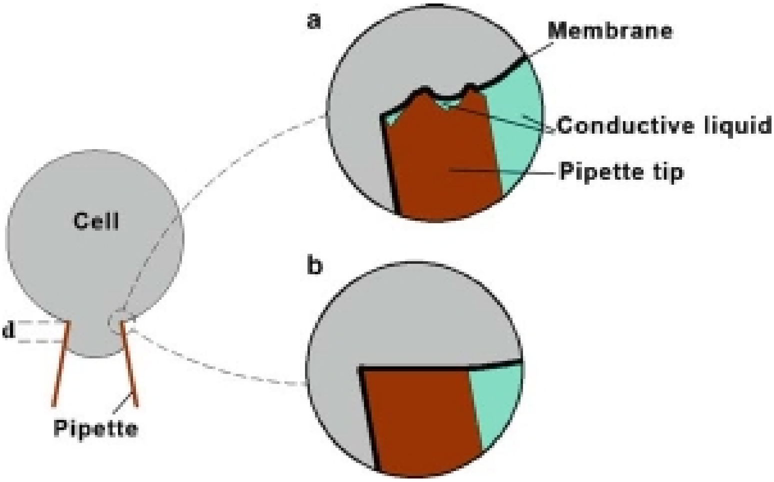


Fig. 52.10 Schematic of pipette-membrane interaction: (**a**) the original pipette tip with a bumpy surface (**b**) the tip is flat

52.5 Discussions and Conclusions

A giga-seal in patch clamping will produce improved signal-to-noise ratio and enables ion channel signal measurement to be more accurate. Currently, the formation of a giga-seal in patch clamping occurs in a sudden and all-or-nothing way. A large number of parameters affect the seal formation, making it hard to understand the physical and chemical mechanisms behind it. In this research, the SEM stereo imaging techniques were used to inspect the surface roughness of micropipettes. The high magnification images revealed the surface nature of the tips to be in contact with cells. Then the contact tips of pipettes were cut across, leaving a very smooth surface at the top of the pipettes. A large number of patch clamping experiments were conducted on HUVECs using the polished pipettes and 60% of the experiments achieved above 3 GΩ seals and the highest seal resistance reached 9 GΩ. The leakage current in single channel recording afterwards was found 0.3 pA, significantly smaller than 2–3 pA usually achieved using conventionally treated pipettes. Smaller current is the consequence of higher seal resistance. The higher seal is obtained from better contact conditions of a smoother tip surface with the cell surface. The results show that nanomachined micro glass pipettes have improved the giga-seal formation in patch clamping.

References

1. Neher, E., Sakmann, B.: Single-channel currents recorded from membrane of denervated frog muscle-fibers. Nature **260**, 799–802 (1976)
2. Dworakowska, B., Dolowy, K.: Ion channel related diseases. Acta Biochim. Polonica **47**(3), 685–703 (2000)

3. Li, S., Lin, L.: A single cell electrophysiological analysis device with embedded electrode. Sens. Actuat. A **134**, 20–26 (2007)
4. Molleman, A.: Patch Clamping: An Introductory Guide to Patch Clamp Electrophysiology. Wiley, England (2003)
5. Hamill, O.P., Marty, A., Neher, E., Sakmann, B., Sigworth, F.J.: Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. Euro. J. Physiol. **391**, 85–100 (1981)
6. Stett, A., Burkhardt, C., Weber, U., Stiphout, P., Knott, T.: CYTOCENTERING: A novel technique enabling automated cell-by-cell patch clamping with the CYTOPATCHTM chip. Recep. Channel. **9**, 59–66 (2003)
7. Malboubi, M., Ostadi, H., Wang, S., Gu, Y., Jiang, K.: The effect of pipette tip roughness on giga-seal formation. Proceedings of the World Congress on Engineering (WCE 2009), vol II, London, 1–3 July 2009
8. Matthews, B., Judy, J.W.: Design and fabrication of a micromachined planar patch-clamp substrate with integrated microfluidics. J. Microelectromech. Syst. **15**, 214–222 (2006)
9. Ong, W., Yobas, L., Ong, W.: A missing factor in chip-based patch clamp assay: gigaseal. J. Phys. Conf. Ser. **34**, 187–191 (2006)
10. Lau, A.Y., Hung, P.J., Wu, A.R., Lee, L.P.: Open-access microfluidic patch-clamp array with raised lateral cell trapping sites. Lab Chip **6**, 1510–1515 (2006)
11. Picollet-D'hahan, N., Sauter, F., Ricoul, F., et al.: Multi-Patch: A chip-based ion-channel assay system for drug screening. IEEE ICMENS 2003: International Conference MEMS, NANO & Smart Systems, pp. 251–254, Banff, Alberta, Canada (2003)
12. Ionescu-Zanetti, C., Shaw, R.M., Seo, J., Jan, Y., Jan, L.Y., Lee L.P.: Mammalian electrophysiology on a microfluidic platform. Proceeding of the national academy of science of the united states of America (PNAS) **102**, 9112–9117 (2005)
13. Klemic, K.G., Klemic, J.F., Sigworth, F.J.: An air-molding technique for fabricating PDMS planar patch-clamp electrodes. Euro. J. Physiol. **449**, 564–572 (2005)
14. Zhang, Z.L., Asano, T., Uno, H., et al.: Fabrication of Si-based planar type patch clamp biosensor using silicon on insulator substrate. Thin Solid Films **516**, 2813–2815 (2008)
15. Corey, D.P., Stevens, C.F.: Science and technology of patch-recording electrodes. In: Sakmann, B., Neher, E. (eds.) Single-channel recording, pp. 53–68. Plenum, New York (1983)
16. Levis, R.A., Rae J.L.: Technology of patch-clamp electrodes. In: Walz, W., Boulton, A.A., Baker, G.A. (eds.) Patch-Clamp Analysis Advanced Techniques, pp. 1–35. Humana Press, Totowa, NJ (2002)
17. Priel, A., Gil, Z., Moy, V.T., Magleby, K.L., Silberberg, S.D.: Ionic requirements for membrane-glass adhesion and giga seal formation in patch-clamp recording. Biophys. J. **92**:3893–3900 (2007)
18. Yaul, M., Bhatti, R., Lawrence, S.: Evaluating the process of polishing borosilicate glass capillaries used for fabrication of in-vitro fertilization (iVF) micro-pipettes. Biomed. Microdevice. **10**, 123–128 (2008)
19. Goodman, M., Lockery, S.R.: Pressure polishing: a method for re-shaping patch pipettes during fire polishing. J. Neurosci. Method. **100**, 13–15 (2000)
20. Piazzesi, G.: Photogrammetry with the scanning electron microscope. J. Phys. E: Sci. Instrum **6**, 392–396 (1973)
21. Marinello, F., Bariani, P., Savio, E., Horsewell, A., De, C.L.: Critical factors in SEM 3D stereomicroscopy. Measure. Sci. Technol. **19**, 1–12 (2008)
22. Samak, D., Fischer, A., Rittel, D.: 3D Reconstruction and visualization of microstructure surfaces from 2D images. J. Manufact. Technol. **56**(1), 149–152 (2007)
23. Alicona Imaging GmbH, Austria. http://www.alicona.com. (2010)

Chapter 53
Effect of Canned Cycles on Drilled Hole Quality

Mohammad Nazrul Islam, Noor Hakim Rafai, and Charoon Phaopahon

Abstract Several factors influence the accuracy of drilled holes. The most obvious ones are the cutting conditions (cutting speed and feed rate) and cutting configurations (tool material, diameter, and geometry). As such, most previous studies have concentrated on these factors. However, in CNC drilling operations, choosing to use canned cycles may have significant effect on drilled hole quality. The objective of this project is to explore this possibility in detail. This paper presents experimental and analytical results of an investigation into the dimensional accuracy and surface finish of drilled holes using different canned cycles. A traditional analysis, the Pareto ANOVA, and the Taguchi S/N ratio are employed to determine the effects of the three major input parameters (cutting speed, feed rate, and canned cycle) on three key accuracy characteristics of drilled holes (diameter error, circularity, and surface roughness), as well as to obtain an optimal combination of the input parameters. The work and tool materials selected are aluminum 6061 and high-speed steel (HSS), respectively. The results indicate that the canned cycle has a profound effect on drilled hole quality, and, in general, canned cycle spot drilling produces the best results.

Keywords Drilling canned cycle · hole quality · Pareto ANOVA analysis · Taguchi methods

53.1 Introduction

Drilling is one of the oldest and the most widely used of all machining processes, comprising about one third of all metal-machining operations [1]. It is used to create or to enlarge a round hole in a workpiece by the relative motion of a cutting tool,

M.N. Islam (✉), N.H. Rafai, and C. Phaopahon
Department of Mechanical Engineering, Curtin University of Technology,
GPO Box U1987, Perth, WA 6845, Australia
e-mail: m.n.islam@curtin.edu.au; n.rafai@postgrad.curtin.edu.au; z_sengki@hotmail.com

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_53,

called a *drill* or *drill bit*. Various methods of drilling are in use, such as conventional drilling, deep hole drilling, and peck drilling. The choice of a drilling method depends on the size, tolerance, and surface finish needed, as well as the production requirements and which machine is available to perform the job.

Several factors influence the quality of drilled holes. The most obvious ones are the cutting conditions (cutting speed and feed rate) and cutting configurations (tool material, diameter, and geometry). Consequently, most previous studies [2–5] have concentrated on these factors. Nonetheless, a few researchers have examined the influence of certain additional factors: Pirtini and Lazoglu [6] studied cutting force, Nouari, List, Girot and Gehin [7] studied tool wear, and Bono and Ni [8] studied thermal distortion. Islam, Jawahir and Kirby [9] included canned cycles in their input variables list; however, their treatment of the topic was brief and their findings inconclusive. Therefore, the effect of a canned cycle on the dimensional accuracy and surface finish of drilled holes needs further investigation; the quality of holes should not be compromised for the sake of higher productivity.

53.2 Drilling Canned Cycle

A *canned cycle* is a sequence of machine operations initiated by a single code. The code acts as a shortcut that simplifies the program. A number of different canned cycles are in use for computer numerical control (CNC) drilling operations, of which the chip-breaking canned cycle (G73), spot drilling canned cycle (G81), and deep hole canned cycle (G83) are the three most popular choices (Fig. 53.1). A brief description of these operations is given in the following paragraphs.

A *chip-breaking canned cycle* is used for drilling a material that has the tendency to produce stringy chips. In other words, the chips form around the tool and do not break easily. G73 can be used to break the chips out of the hole by slightly retracting the tool during a drilling operation.

A *spot drilling canned cycle* is applied for normal drilling. The tool will plunge into the bottom of the hole and then rapidly retract from the bottom of the hole. Drilling is performed from point R to point Z.

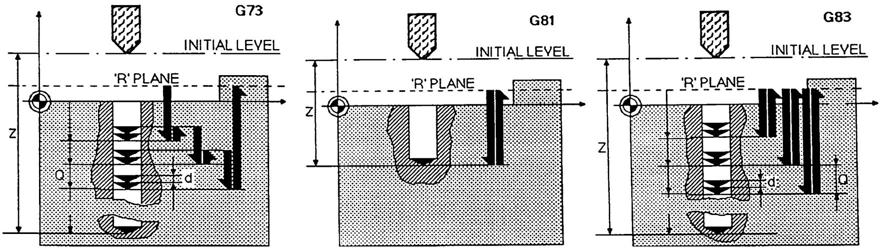


Fig. 53.1 Three drilling canned cycles

A *deep hole canned cycle* is utilized when a deep hole is being machined to allow the chips to be cleared at certain intervals. If a drill plunges too deep into the hole, the chips will pack up around the flute of the drill bit. As the drill continues to go deeper, the packing of chips will cause the drill bit to break. Drilling is performed from point R to point Z. Q is the depth of cut for each cutting feed. After clearing the chips, the drill bit rapidly traverses back to point R, and the cutting feed is performed again.

53.3 Scope

The main objective of this project is to investigate the three major quality characteristics of drilled holes produced on a CNC machining centre using three canned cycles: the chip-breaking canned cycle (G73), the spot drilling canned cycle (G81), and the deep hole canned cycle (G83). For a drilling operation, diameter error, circularity, and surface roughness (as determined by arithmetic averages or R_a values) are the three most important quality characteristics; thus, they are used here for monitoring the quality of drilled holes. The three main input parameters are cutting speed, feed rate, and canned cycles. A general-purpose coordinate measuring machine (CMM) and a surface roughness analyzer are employed for the measurement of the output parameters. The results are analyzed with three techniques: (a) a traditional analysis, (b) a Pareto analysis of variation, and (c) Taguchi's signal-to-noise ratio (S/N) analysis. The expected outcomes of this project are to find the effects of the three input parameters on the accuracy of drilled holes and, subsequently, to optimize the input parameters.

It is worth pointing out that drilling is not regarded as a precision-machining operation; as a result, additional operations are required to improve the accuracy levels of drilled holes. Even so, experience shows that while the surface roughness of drilled holes can be enhanced by applying subsequent finishing operations, improving geometric accuracies is difficult [2]. Furthermore, when different tools are used for finishing operations, which occurs in most cases, the geometric inaccuracies of drilled holes are increased. The present study examines the utility and limitations of CNC drilling operations used to achieve a better quality of drilled holes.

53.4 Experimental Work

The experiments were planned using Taguchi's orthogonal array methodology [10]. A three-level L_{27} orthogonal array was selected for our experiments. Aluminium 6061 was chosen as the work material because of its extensive use in the industry. Through holes were designed with a $\phi 12 \times 24$ mm hole size. Three blocks of aluminium 6061, each containing nine holes marked H1 to H9, were drilled on a vertical CNC machining center (Leadwell V-30 Machining Center, Taiwan). Machining

Table 53.1 Input variables

			Levels		
Input parameters	Unit	Symbol	Level 0	Level 1	Level 2
Cutting speed	rpm	A	800	2,000	3,200
Feed rate	mm/rev	B	0.1	0.2	0.3
Canned cycle	–	C	G73	G81	G83

was performed under wet conditions. Three new, 12 mm-diameter high-speed steel (HSS) drill bits, one for each component, were used to perform the drilling operation. Center drills were employed for initial positioning. The input parameters (cutting speed, feed rate, and canned cycle) were chosen on the basis of the capacity and limiting cutting conditions of the CNC machine; details are given in Table 53.1.

The precision measurement data was obtained using a general-purpose coordinate measuring machine (CMM; Model 7.10.7, Brown & Shape, USA) and the roughness parameter (R_a, the arithmetic average) was measured with a surface finish analyzer (Surftest SJ-201P, Mitutoya, Japan).

53.5 Results and Analysis

An enormous amount of data was obtained and subsequently analyzed. Due to space constraints, only a few are illustrated, although in the analysis of the work, all these relationships were considered at different stages. In the traditional analysis, the mean values of the measured variables were used. For the Taguchi method, the *signal-to-noise ratio* was calculated using the following formula [10]:

$$S/N \quad = \quad -10\log\frac{1}{n}\left(\sum\frac{1}{y^2}\right) \tag{53.1}$$

where S/N is the signal-to-noise ratio (in *dB*), n is the number of observations, and y is the observed data.

53.5.1 Diameter Error

The diameters of the holes were calculated using the standard built-in software package of the CMM. Eight points were probed to determine the diameter in the horizontal plane, and the diameter of each hole was checked at 1 mm height increments. The difference between the measured diameter and the designed diameter is the diameter error; thus, a positive error indicates over-sizing of the holes. All diameters were checked thoroughly. The results are summarized in Table 53.2, which shows, in terms of diameter error, G81 was the best, G83 was medium, and G73 was

Table 53.2 Diameter measurement data

		Canned cycles		
Input parameters	Unit	G73	G81	G83
Mean diameter	mm	12.100	12.039	12.070
Diameter error	µm	100	39	70
Range of measurement	µm	180	90	117
6 × Standard deviation	µm	134	36	118

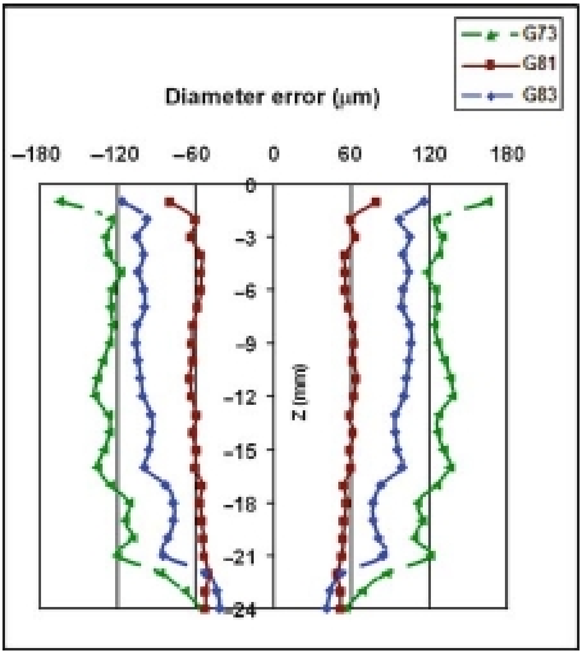


Fig. 53.2 Change of diameter error along hole axis

the worst. For a ϕ12 mm hole, the expected size limits are 12.000–12.180 mm for normal-quality drilling and 12.000–12.110 mm for high-quality drilling [11]. All holes produced were within the expected tolerance limit; however, drilling operations performed on CNC machines should produce high quality holes. In this case, only G73 qualified as high-quality, with a tolerance range of 90 mm, whereas G83 and G73 produced 118 and 134 mm ranges, respectively.

It was also noted that in all cases the holes were oversized, which is a common problem in drilling operations. Galloway [2] identified this problem in 1957, and concluded that it is caused by the variation in relative lip heights of the drill. Other possible reasons are runout of the drill when attached to the machine, thermal distortion, a non-symmetric point angle, and runout of the chisel edge.

Changes in average diameter error along the hole axis for different canned cycles are illustrated in Fig. 53.2. This type of error is commonly known as *error in* *shape*. Figure 53.2 confirms that G81 produced the most uniform size variation, whereas G73 was the worst. This suggests that a worsening of the hole profile took place due to increased vibration, caused by multiple changes of drilling direction during the pecking action of the drill. The diameter and the diameter-to-length ratio are two other major factors affecting error of shape, and are not included in this study. Figure 53.2 also shows a *bell mouth* shape for all holes; that is, enlargement at the entry of the hole, regardless of the type of canned cycle applied. The enlargement of the hole at entry could have been caused by the wobbling of drills during positioning.

Variation in the average diameter error for different holes is shown in Fig. 53.3, grouped by three levels of cutting speed. For all cutting conditions, in terms of diameter error, G81 was the best, followed by G83 and G73. Contrary to the findings of some other researchers (e.g., Kuet, Kaynak and Bagci [3]), no increase in the dimensional error was noted when the cutting speed and feed rate were increased. Kuet, Kaynak and Bagci performed their experiment under dry conditions, whereas our experiment was performed under wet conditions. As a result, in our case the effect of thermal distortion was minimized. Figure 53.3 also shows that for all three cutting speed ranges-low, medium, and high—diameter error decreased as feed rate increased. In our view, diameter error was reduced when feed rate was increased, due to the reduction of drill engagement time with the hole.

The Pareto ANOVA analysis for dimensional error given in Table 53.3 illustrates that the canned cycle (C) had the most significant effect on diameter error (P = 85.41%) when drilling aluminium 6061. Compared to the canned cycle, the other two independent parameters – the feed rate (B) and the cutting speed (A) – contributed to diameter error only by very small percentages (P = 7.35% and 2.59%, respectively). Moreover, the remaining interactions were almost negligible.

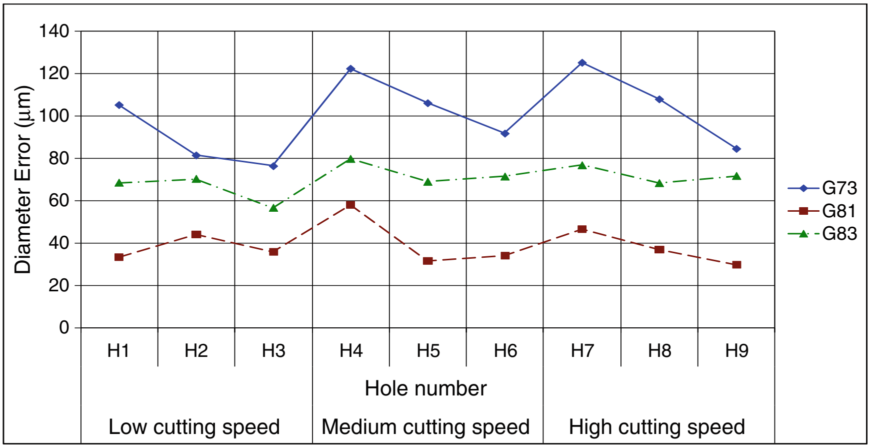
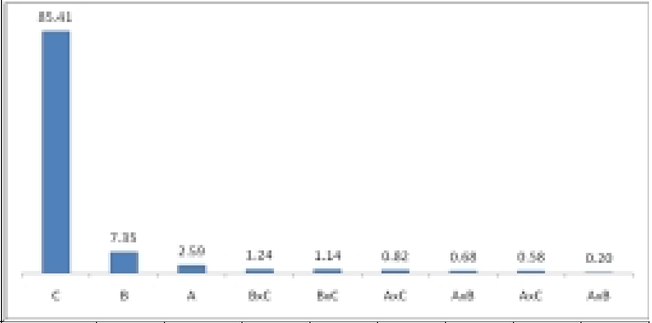


Fig. 53.3 Variation in diameter error for different cutting conditions

Table 53.3 Pareto ANOVA analysis for diameter error

Sum at factor level	Factor and interaction								
	A	B	A×B	A×B	C	A×C	A×C	B×C	B×C
0	−518.87	−520.46	−519.37	−519.16	−522.55	−519.14	−519.23	−519.82	519.88
1	−519.89	−519.35	−519.65	−519.64	−516.32	−519.73	−519.74	−519.56	519.18
2	−519.70	−518.64	−519.44	−519.65	−519.58	−519.58	−519.49	−519.08	519.40
Sum of squares of difference (S)	1.761	5.006	0.136	0.465	58.142	0.555	0.393	0.841	0.777
Contribution ratio (%)	2.59	7.35	0.20	0.68	85.41	0.82	0.58	1.24	1.14
									
Cumulative contribution	85.41	92.76	95.35	96.58	97.72	98.54	99.22	99.80	100
Check on significant interaction			B×C twoway table (Appendix E)						
Optimum combination of significant factor level			A0B2C1						

The response graph for diameter error shown in Fig. 53.4a demonstrates that the canned cycle (C) had the most significant effect on diameter error among the three cutting parameters, followed by the feed rate (B) and the cutting speed (A).

Based on the above S/N ratio and Pareto ANOVA analyses, it was found that the optimal combination to achieve a low value for diameter error was $A_0B_2C_1$; that is, a low cutting speed, a high feed rate, and using a spot drilling canned cycle (G81).

53.5.2 Circularity

For drilled holes, *circularity* (also known as roundness or out-of-roundness) is another important quality characteristic, and it is geometric in nature. A large circularity value is problematic for parts with relative motion: it induces vibration and heat. Circularity is measured from a cross-section perpendicular to the axis of a hole or a cylinder, and is defined by two concentric boundaries within which each circular element of the surface must lie. The circularity of each hole was checked at three different heights: $z = -3$ mm (i.e., near the top), $z = -12$ mm (i.e., near the middle), and $z = -21$ mm (i.e., near the bottom).

The variation in average circularity for different holes, grouped by three levels of cutting speed, is given in Fig. 53.5. From this figure, it appears that G73 produced the highest circularity error, followed by G81 and G83. There was a slight increase in circularity values with an increase in cutting speed, whereas the effect of the feed rate was minimal. The rates of change of circularity, related to cutting speed, were different for different canned cycles, suggesting an interaction between the cutting speed and canned cycle.

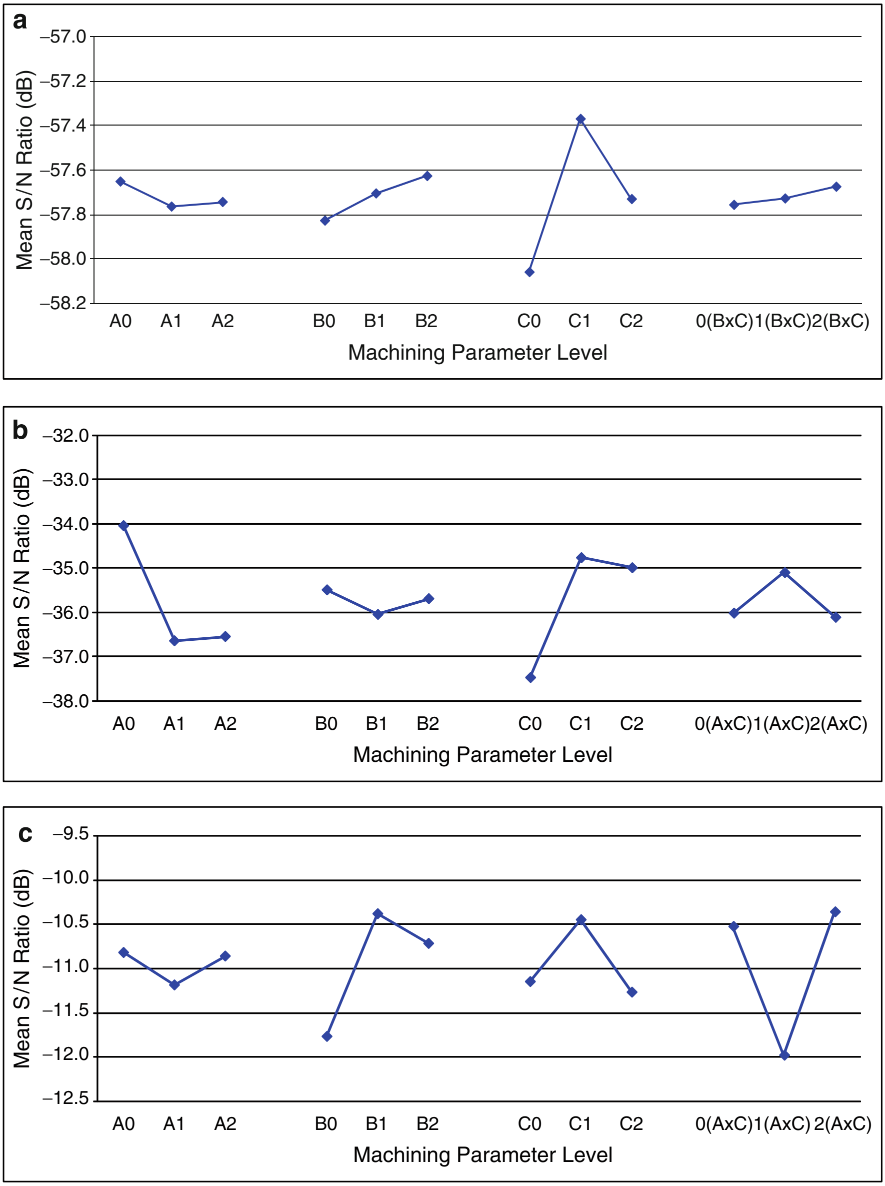


Fig. 53.4 Response graphs for (**a**) diameter error, (**b**) circularity and (**c**) surface roughness

The variation in average circularity at three hole heights is illustrated in Fig. 53.6. No clear trend was apparent. However, G73 produced the highest circularity values compared to other canned cycles, and G83 produced the highest circularity value differential at the top and at the bottom of the hole.

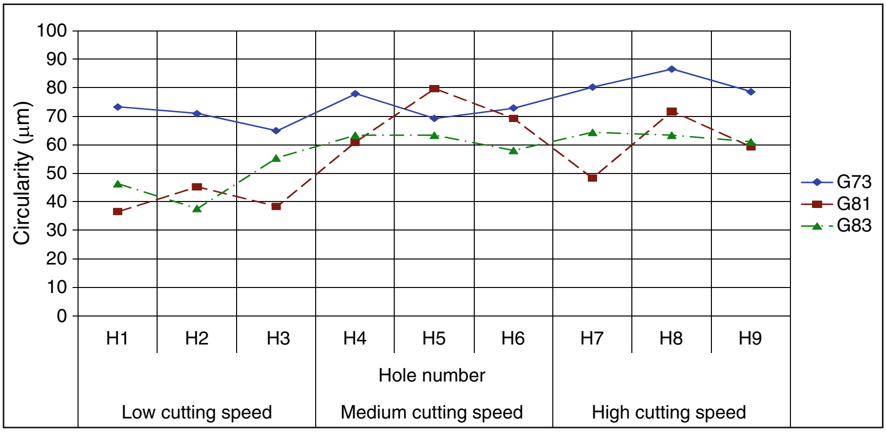


Fig. 53.5 Variation in circularity for different cutting conditions

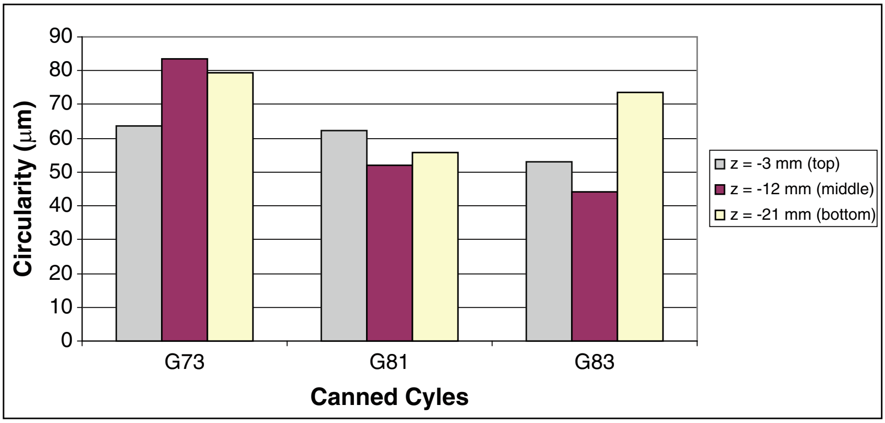
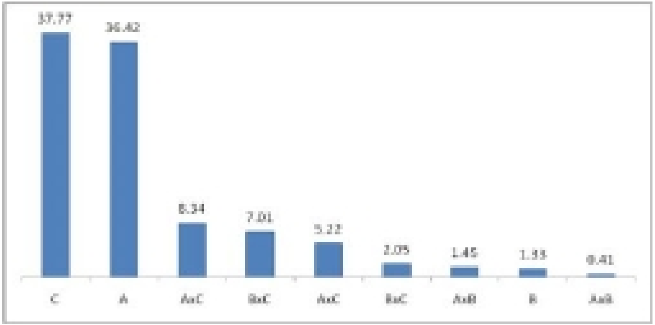


Fig. 53.6 Variation in circularity for different height of holes

The Pareto ANOVA analysis of circularity given in Table 53.4 illustrates that both canned cycle and cutting speed had a highly significant effect on circularity with C (P = 37.77%) and A (P = 36.42%). Among the interactions, A × C had the most significant effect on circularity (P = 8.34%). The main contributor to the circularity error was the runout of the drill when attached to the machine, which increased with the increase in cutting speed.

The response graph for circularity shown in Fig. 53.4b indicates that the cutting speed (A) had the most significant effect on circularity, followed by the canned cycle (C) and feed rate (B). The interaction between the cutting speed and canned cycle greatly contributed to the circularity of the holes.

Table 53.4 Pareto ANOVA analysis for circularity

Sum at factor level	A	B	A×B	A×B	C	A×C	A×C	B×C	B×C
0	–306.37	–319.44	–323.16	–321.87	–337.29	–324.18	–329.05	–320.79	–327.34
1	–329.81	–324.45	–320.34	–318.98	–312.89	–315.92	–318.19	–319.11	–315.70
2	–328.94	–321.23	–321.62	–324.27	–314.94	–325.02	–317.88	–325.22	–322.08
Sum of squares of difference (S)	1059.38	38.57	11.94	42.08	1098.72	151.71	242.71	59.75	203.82
Contribution ratio (%)	36.42	1.33	0.41	1.45	37.77	5.22	8.34	2.054	7.01
									
Cumulative contribution	37.77	74.20	82.54	89.55	94.76	96.82	98.26	99.59	100.00
Check on significant interaction		A×C two-way table (Appendix E)							
Optimum combination of significant factor level		A0B0C1							

Based on the above S/N ratio and Pareto ANOVA analyses, it was found that the optimal combination for achieving a low circularity value of was $A_0B_0C_1$; that is, a low cutting speed and a low feed rate, using a spot drilling canned cycle (G81).

53.5.3 Surface Roughness

Surface roughness, another important quality characteristic of drilled holes, needs attention. For each hole, the surface roughness is measured parallel to the hole axis at four radial positions using a surface finish analyzer. The variation in average surface roughness for different holes, grouped by three levels of cutting speed, is given in Fig. 53.7. From this figure, it can be seen that the surface roughness of the holes produced by G81 was high at a low cutting speed, and it gradually decreased as the cutting speed was increased. Conversely, the surface roughness of the holes produced by G83 was the lowest when the cutting speed was also the lowest, and it increased with the increase in cutting speed. No particular trend was noted for the holes produced by G73. From this, it can be concluded that an interaction exists between cutting speeds and canned cycles, which is strongly influenced by surface roughness. This was later confirmed by the Pareto analysis (see Table 53.5). It is interesting to note that all three graphs meet where both cutting speed and feed rates are at their respective medium levels.

The Pareto ANOVA analysis for surface roughness given in Table 53.5 shows that the feed rate (B) has the most effect on the surface roughness (P = 21.19%), followed by the canned cycle (C) (P = 7.91%) and the cutting speed (A) (P = 1.64%). However, the interaction of cutting speed and canned cycle (A × C) had the most significant effect on surface roughness (P = 32.20%).

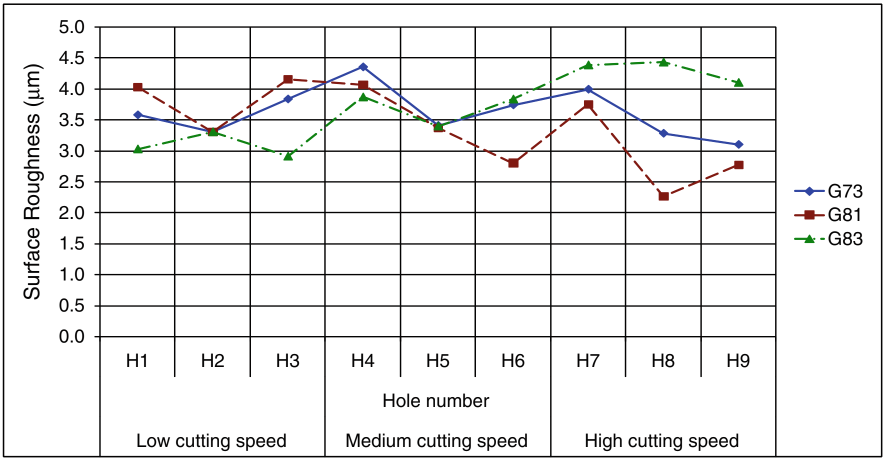
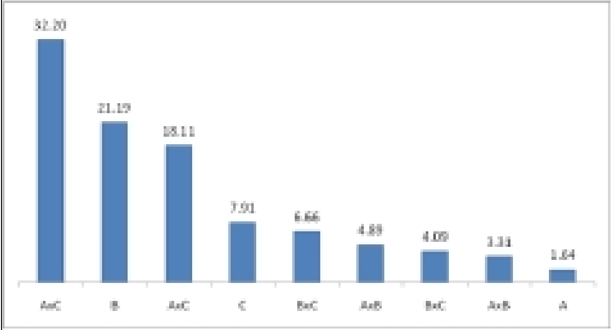


Fig. 53.7 Variation in surface roughness for different cutting conditions

Table 53.5 Pareto ANOVA analysis for surface roughness

Sum at factor level	A	B	A×B	A×B	C	A×C	A×C	B×C	B×C
0	−97.37	−105.92	−95.32	−95.66	−100.34	−94.72	−102.91	−100.07	−97.20
1	−100.68	−93.42	−98.91	−99.66	−94.04	−107.84	−101.17	−100.43	−102.74
2	−97.74	−96.44	−101.56	−100.47	−101.42	−93.23	−91.71	−95.29	−95.84
Sum of squares of difference (S)	19.76	255.30	58.88	39.85	95.33	387.86	218.17	49.28	80.22
Contribution ratio (%)	1.64	21.19	4.89	3.31	7.91	32.20	18.11	4.091	6.66
									
Cumulative contribution	32.20	53.39	71.50	79.41	86.07	90.96	95.05	98.36	100.00
Check on significant interaction			A×C two-way table (Appendix E)						
Optimum combination of significant factor level			A2B1C1						

According to the mean S/N response graph for surface roughness in Fig. 53.4c, it can be seen that the feed rate (B) had the most significant effect on surface roughness, followed by the canned cycle (C) and cutting speed (A). The influence of feed rate on surface roughness is well known, and in most cases, with an increase in feed rate, surface roughness deteriorates. The interaction between the cutting speed and canned cycle (A × C) had the most influence on the surface roughness of the holes.

Based on the above S/N ratio and Pareto ANOVA analyses, it was found that the optimal combination for achieving a low value of surface roughness was $A_2B_1C_1$; that is, a high cutting speed, a medium feed rate, and using a spot drilling canned cycle (G81).

53.6 Concluding Remarks

The research presented in this paper demonstrates that canned cycles have a profound effect on the quality of drilled holes. In general, the spot drilling canned cycle (G81) produced the best results. All three quality characteristics considered – diameter error, circularity, and surface roughness – deteriorate due to the pecking action of the chip breaking canned cycle (G73) and deep hole canned cycle (G83). Therefore, unless there are requirements compelling their use, both the chip breaking and deep hole canned cycles should be avoided.

The experimental results presented in this paper show that drilled holes are always oversized, and a bell mouth shape is present in all holes; that is, there is enlargement at the entry of the hole, regardless of the type of canned cycle applied. The enlargement of the hole at entry can be caused by the wobbling of drills during positioning.

Drilling is a complex, three-dimensional cutting process, with conditions varying along the entire cutting edge. The process is further complicated by the different pecking actions of drilling with different canned cycles. Consequently, some trends observed in this study could not be explained fully, and further research is needed for their precise understanding.

References

1. Chen, C., Tsao, C-C.: Cutting performance of different coated twist drills. J Mater. Proc. Tech. **88**, 203–207 (1999)
2. Galloway, D.F.: Some experiments on the influence of variation factors on drill performance. Trans. ASME. **57**, 191–231 (1957)
3. Kurt, M., Kaynak, Y., Bagci, E.: Evaluation of drilled quality in AL 2042 alloy. Int. J. Adv. Manuf. Technol. **37**, 1051–1060 (2008)
4. Tammimi, E., Darwish, S.M.: Geometric accuracies of NC and conventionally drilled holes. J. Mater. Process. Tech. **75**, 111–116 (1998)
5. Kurt, M., Bagci, E., Kaynak, Y.: Application of Taguchi methods in the optimisation of cutting parameters for surface finish and diameter accuracy in dry drilling process. Int. J. Adv. Manuf. Technol. **40**, 458–469 (2009)
6. Pirtini, M., and Lazoglu, I.: Forces and hole quality in drilling. Int. J. Mach. Tools Manufact. **45**, 1271–1281 (2005)
7. Nouari, M., List, G., Girot, F., Gehin, D.: Effect of machining parameters and coating on wear mechanisms in dry drilling of aluminum alloys. Int. J. Mach. Tools Manufact. **45**, 1436–1442 (2005)

8. Bono, M., Ni, J.: The effects of thermal distortion on the diameter and cylindricity of dry drilled holes. Int. J. Mach. Tools Manufact. **41**, 2261–2270 (2001)
9. Islam, M.N., Jawahir, I.S., Kirby, I.J.: A CMM-based geometric accuracy study of CNC drilling operations. ME Trans. IEAust. **16**:2261–2270 (1991)
10. Taguchi, G.: Introduction to Quality Engineering, Translated into English by Asian Productivity Organization, Tokyo (1989)
11. Gladman, C.A.: Geometric Analysis of Engineering Designs, 2nd edn. Australian Trade Publications Pvt. Ltd., Sydney, Australia (1972)

Chapter 54
Micro Machine Parts Fabricated from Aqueous Based Stainless Steel Slurry

Mohamed Imbaby, Isaac Chang, and Kyle Jiang

Abstract A fabrication process of stainless steel micro components from metallic powder is reported. The process consists of two stages. In the first stage, high quality SU-8 master moulds and their negative replicas from soft moulds are produced using photolithography and soft moulding techniques respectively. The second stage includes preparation of stainless steel slurry, filling the soft mould, obtaining the green parts, de-binding and sintering to the finial parts. A method is proposed in this research to obtain the optimum dispersant for preparation of the metallic slurry and the result was found to be 0.003 (dispersant/powder wt.). Both vacuum and forming gas atmosphere sintering conditions were investigated. The results show micro components have the similar high quality as the master moulds. The maximum sintered density was found to be 98.1% when the sample was sintered at 1,350 °C in vacuum.

Keywords Dispersant · micro parts · sintering · soft mould · stainless steel · SU-8

54.1 Introduction

MEMS components offer a wide range of various applications and their materials have expanded rapidly in recent years. Among the new MEMS materials, metals show many favourable properties. 316-L stainless steel is one of the metallic materials providing good mechanical and corrosion resistance properties. Recently, many micro fabrication techniques are emerging to cover a wide range of materials and

M. Imbaby (✉) and K. Jiang
University of Birmingham, School of Mechanical Engineering, Birmingham,
Edgbaston, B15 2TT, UK
e-mail: mohamed.imbaby@gmail.com; mfi606@bham.ac.uk; k.c.jiang@bham.ac.uk

I. Chang
University of Birmingham, School of Metallurgy and Materials,
Birmingham Edgbaston, B15 2TT, UK
e-mail: i.t.chang@bham.ac.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_54,

applications. Micro system technology (MST) is one of the leading technologies of producing micro components ranging from nanometres to millimetres [1]. X-ray Lithography is a conventional micro fabrication technique which is suitable for fabricating polymeric and silicon based materials [2]. In other fabrication techniques, micro electro discharge machining (EDM) and laser micro machining provides a wide range of materials micro components [3]. EDM is limited to conductive materials and laser micro machining produces structures with rough edges. Electroforming is used for producing micro components from metallic materials [4]. Electroforming has a difficulty in producing micro components with thickness higher than 0.5 mm because of its slow metal deposition speed.

Micro metal injection moulding (μMIM) is a technique developed from metal injection moulding (MIM) to fabricate micro components from wide range of materials [5, 6]. Soft lithography is a new micro fabrication technique emerging few years ago. It was successfully used in production of free standing micro machines parts from metallic to ceramic materials [7, 8]. The main difference between soft lithography and MIM is the type of mould inserts used. In soft lithography, the soft mould inserts used is usually PDMS moulds. On the other hand, rigid mould inserts are always used in μMIM. Due to soft mould inserts (PDMS), powder-binder slurries are a very effective way to fill the soft moulds. Consequently, the binder must be selected carefully to be removed completely before the sintering process. The unburned binder increases the carbon content during the sintering and reduces the corrosion resistance of stainless steel [9, 10]. Stainless steel parts have been successfully sintered at various sintering temperatures and atmospheres [11]. The mechanical properties of the sintered parts are enhanced with the increase of the density [12].

In this research, fabrication of free standing stainless steel micro machine parts, gears, pistons and connecting rods was studied. The fabrication process was previously [7] and modification has been made in this work. Preparation of stainless steel slurry from aqueous based binder and dispersant were investigated in details and the optimum dispersant and binder are obtained. The density of the micro components was also studied in details for different sintering atmospheres and temperatures.

54.2 SU-8 Master Moulds and Their Negative Replicas

In this study, SU-8 2075 [MicroChem, USA] was used for fabricating ultra-thick micro moulds. SU-8 is imagable with UV light in the range of 320–420 nm. UV photolithography was used for fabricating SU-8 structures with depth from few microns to over 1 mm. The detailed fabrication process was discussed in previous reports [7]. The fabrication procedure started with casting SU-8 resist onto 4 in. Silicon wafer and soft baking it at 65 °C for 2 h followed by 95 °C for 34 h. Exposure was done in Canon PLA-501 FA UV-mask aligner. Afterwards, the wafer went through post exposure bake and development in EC solvent. The steps used to replicate soft mould (PDMS) inserts are as follows: (a) PDMS raw material (DOW Sylgard Silicone)

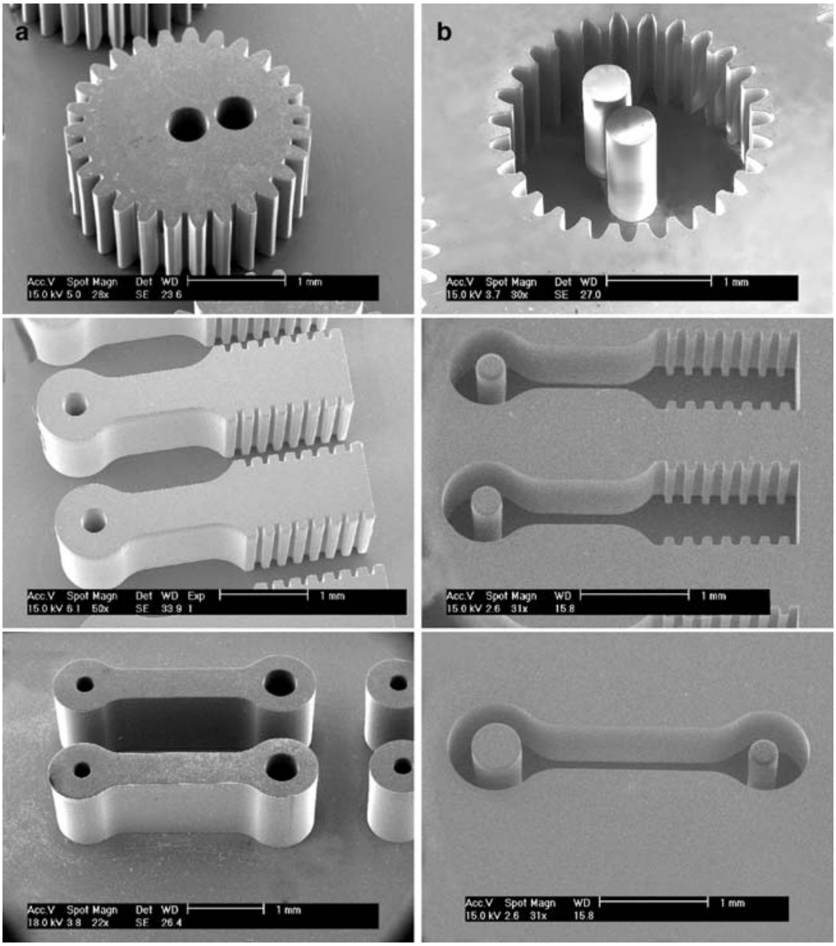


Fig. 54.1 (**a**) SU-8 master moulds and (**b**) their negative replicas soft moulds

and curing agent were added in 10:1 weight ratio; (b) the mixture was placed in a vacuum chamber to remove any trapped bubbles; (c) the degassed mixture was poured on SU-8 moulds and degassed again; and (d) the PDMS was cured in an oven at 90 °C. Figure 54.1 shows the SEM images of the SU-8 and PDMS micro moulds.

54.3 Preparing Stainless Steel Slurry

Stainless steel metallic slurry was prepared frpm 316-L stainless steel powder, dispersant, binder and distilled water. 316-L stainless steel powder (Sandvik Osprey Ltd., UK) with 81.7% powder <5 μm was used in this research. The chemical compositions delivered by supplier include 65.5% Fe, 18.5% Cr, 11.6% Ni, 2.3% Mo, 1.4% Mn and other minor elements. The particle size distribution: D10, D50 and D90 measure 1.9, 3.4 and 5.8 μm respectively (supplier). Duramax B-1000 (Rohm and Hass, USA) is an ammonium salt of acrylic homopolymer which is used as a dispersant for different ceramic and metallic powder is used in this work as a dispersant [13, 14]. A mixture of and aqueous emulsion of acrylic polymer Duramax B-1000 and B-1007 (Chesham Speciality Ingredients Limited, UK.) is used as the binder. The slurry is prepared as follows: (a) the dispersant and distilled water are mixed in specimen tube using ultrasonic bath for 5 min; (b) stainless steel powder is added and the mixture is stirred using mechanical stirrer for 20–30 min in order to disperse the powder properly; (c) binder is added and the whole mixture is stirred again for 15 min, (d) the slurry is degassed using vacuum to remove the bubbles formed during mixing and poured onto the micro gear soft mould; (e) the filled mould is put under vacuum in order to completely fill the soft moulds and the excess slurry on the top of the mould opening is removed with a blade to keep the patterns flat.

54.4 De-binding and Sintering

Two different sintering conditions were investigated in this research as follows: (a) de-binding and sintering were carried out in one continuous cycle in a tube furnace with dynamic flow of forming gas atmosphere of 90% Nitrogen and 10% Hydrogen, and (b) de-binding was carried out in a tube furnace with nitrogen atmosphere and sintering was done under vacuum atmosphere. In both cases, the de-binding rate was adjusted to 1.2°C per minute until 700°C and maintained at this temperature for 1 h. During the sintering stage, the heating rate is adjusted to be 5°C per minute until reaching the sintering temperature. Four different sintering temperatures investigated in this research were 1,200°C, 1,250°C, 1,300°C, and 1,350°C. The holding time during sintering was adjusted to be 1.5 h for all cases.

54.5 Results and Discussions

54.5.1 Optimization of the Slurry Properties

To improve the density of the green parts, a proposed method was used in this research to obtain the optimum dispersant and binder used. The process used for obtaining the optimum dispersant is presented as follows: (a) stainless steel slurries containing different dispersant/powder ratio with the same solid loading and binder was prepared to study only the effect of dispersant on the green density, (b) the PDMS moulds were filled with the slurry and the green parts obtained (c) the green density of each dispersant/powder ratio was measured, (d) The maximum

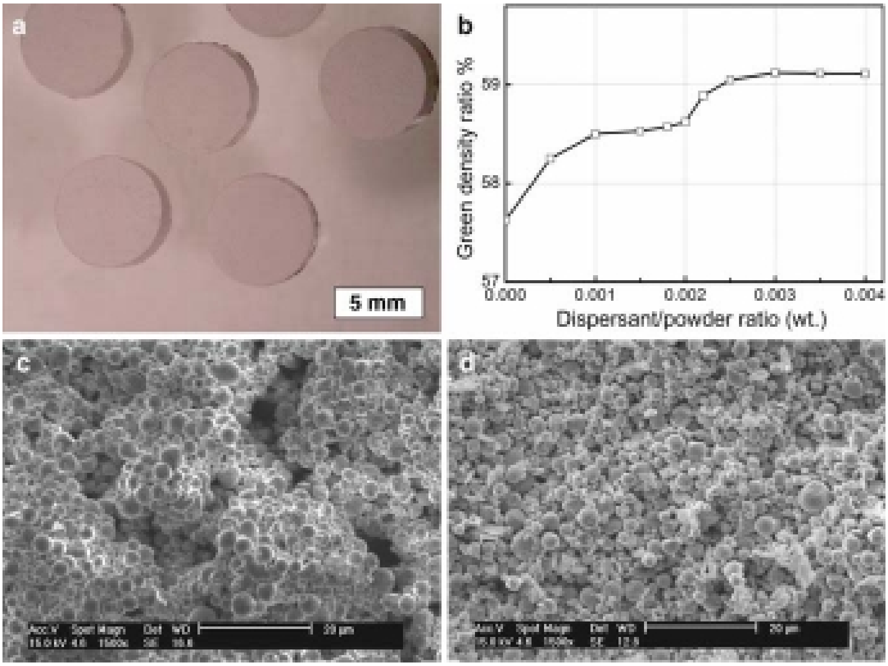


Fig. 54.2 (**a**) Cylindrical green parts, (**b**) effect of dispersant on the green density. Fractured green parts: (**c**) without dispersant and (**d**) with optimum dispersant

green density was obtained and its corresponding dispersion ratio was selected as the optimum. Due to the complexity of micro components fabricated and difficulties in measuring the green density, PDMS mould inserts with cylindrical shape 5 mm in both diameter and high were fabricated. The cylindrical green parts were obtained and their images are shown in Fig. 54.2a. The green density is measured directly by mass volume method. The relation between the density of the green parts (based on wrought material density 8 g/ml) and the dispersant/powder weight ratio is shown in Fig. 54.2b. It was found that increasing the dispersant ratio reduced the green density. Moreover, the maximum green density was obtained when the dispersant ratio was 0.003 by weight. After this ratio no improvements in the green density was found. Therefore, this ratio is selected as the optimum dispersion ratio used in this work. The fractured green parts fabricated from both optimum and without dispersant are shown in Fig. 54.2c and d, respectively. It is found that when no dispersant is used, the fractured green parts show high tendency of particles aggregation for different sizes which produce inhomogeneous distribution. The tendency of forming gaps inside the green parts is increased. On the other hand, using the optimum dispersant not only reduces the tendency of particles aggregation but also improves the particles distribution in the green parts and the density packing. That is why the density of the green parts increases as the dispersant increase.

The binder has a great effect on the preparation of stainless steel aqueous slurries and their corresponding green parts. It is cleared that using too much binder not only increases the green strength but also increases the overall de-binding time and shrinkage after sintering. However, using less binder produces green micro parts with insufficient strength which is likely to be damaged during de-moulding process. Actually, there is no rule to select the optimum amount of binder desired for obtaining damage free green micro parts because de-moulding is performed manually and the samples may be damaged either due to in-experience de-moulding or insufficient strength. After several trials, it was found that using a 0.02 binder/powder ratio (wt.) produced at least 1/2 of the de-moulded micro parts damage free when careful de-moulding was taken in account.

54.5.2 Green Micro Parts

The green micro parts fabricated from optimum dispersant and binders are investigated. Three different defects are observed in the green micro parts after de-moulding. The first defect is resulted from insufficient strength in which the micro components are retained except a piece is damaged and a crack appears as shown in Fig. 54.3a. A de-moulding defect is clearly shown at the edge of a micro part caused during the extraction from the soft mould as shown in Fig. 54.3b. This type of defects commonly happens and cannot be completely eliminated. The defect caused by incomplete filling is another type appearing on the corners of micro parts. It is usually the result of the air trapped inside the soft moulds which prevent the slurry from filling all the micro features properly, as illustrated in Fig. 54.3c. This type defects can be controlled and eliminated when the soft moulds filled with slurry is completely degassed under vacuum. The defect free green micro components are also investigated and shown in Fig. 54.3d–f in which high quality green micro parts are obtained, retaining the similar geometric quality to the SU-8 master moulds and all the micro features are duplicated.

54.5.3 Sintered Micro Components

After sintering, high quality micro components are obtained as shown in Fig. 54.4. It was observed that both vacuum and forming gases produce the same quality in excellent shape retention, while the parts sintered in using vacuum look brighter than sintered in forming gas.

54.5.4 Density of the Sintered Parts

The density of the sintered components is measured using buoyancy method. The effect of sintering temperature on the density in relation to the theoretical value of

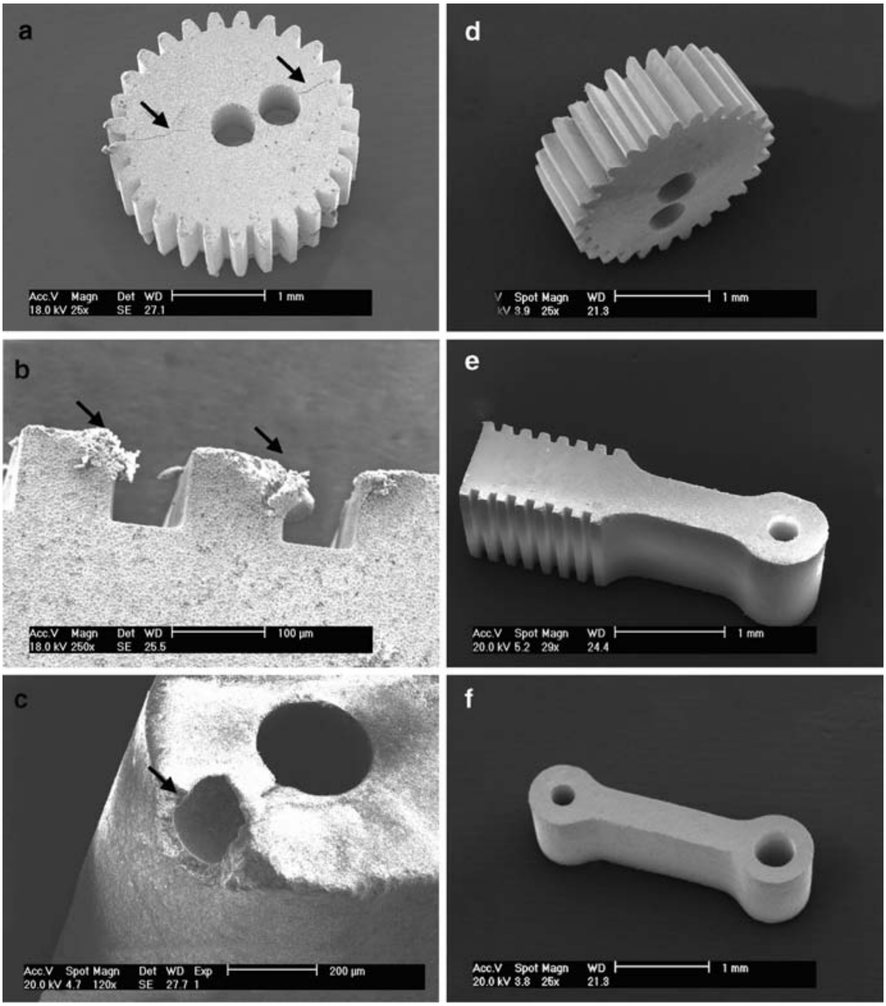


Fig. 54.3 Defected green micro parts: (**a**) insufficient strength, (**b**) de-moulding defect and (**c**) incomplete filling defect. Deffect free micro parts: (**d**) gear, (**e**) piston and (**f**) linkage rod

8g/ml for both vacuum and forming gas are investigated and shown in Fig. 54.5. It has been shown that increasing the sintering temperature increases the density. Moreover, both vacuum and forming gases produce nearly the same density at each sintering temperature, although sintering in vacuum results in a little bit higher density at 1,350°C and lower at other sintering temperatures. The maximum density is obtained when the sintering is done at 1,350°C and found to be 98.1%. On the other hand, the linear shrinkage is also investigated for different sintering conditions.

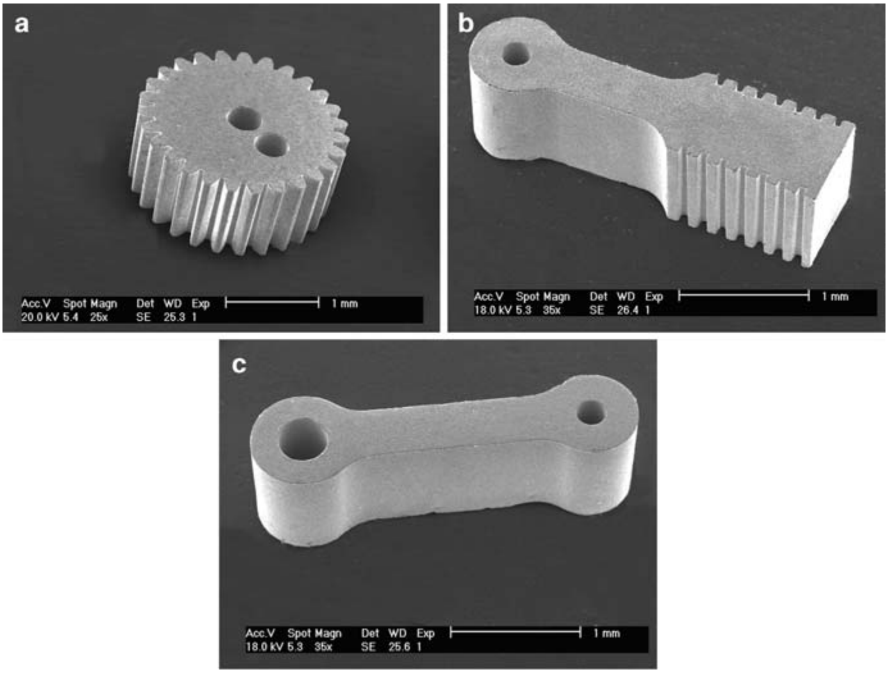
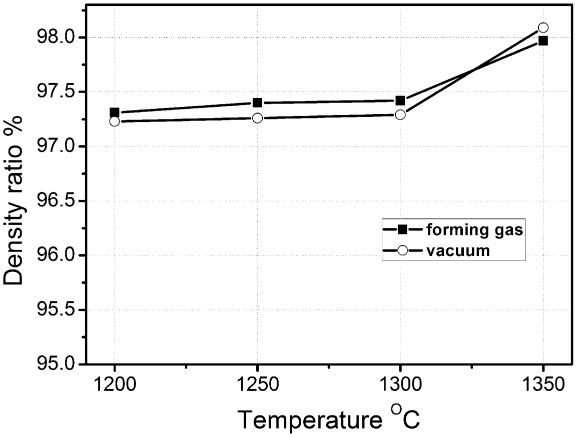

a
Acc.V Spot Magn Det WD Exp 1 mm
20.0 kV 5.4 25x SE 25.3 1
b
Acc.V Spot Magn Det WD Exp 1 mm
18.0 kV 5.3 35x SE 26.4 1
c
Acc.V Spot Magn Det WD Exp 1 mm
18.0 kV 5.3 35x SE 25.6 1

Fig. 54.4 Sintered micro components

Fig. 54.5 Effect of sintering conditions on the density of the micro parts



54.6 Conclusions

In this research, 316-L stainless steel micro machine components were successfully fabricated with high quality shape retention using dispersant-binders aqueous slurry. The optimum slurry properties are obtained experimentally. Using the optimum dispersant in the preparation of stainless steel slurry improves the green density. Using both vacuum and forming gas sintering atmospheres produce nearly the same density, while sintering in vacuum produces a little bit higher density than in forming gas at the high sintering temperature. The maximum sintering density was obtained when the sintering was carried out under vacuum at 1,350°C. The work presents a valid fabrication technique of producing various micro components with complex shapes without additional machining processes.

References

1. Beeby, S., Enell, G., Kraft, M., White, N.: MEMS Mechanical Sensors. London Artech House Inc., London (2004)
2. Mappes, T., Worgull, M., Heckele, M.: Submicron polymer structures with X-ray lithography and hot embossing. J. Mohr. Microsyst. Technol. **14**, 1721–1725 (2008)
3. Sheu, D.Y.: High-speed micro electrode tool fabrication by a twin-wire EDM system. J. Micromech. Microeng. **18**, 1–5 (2008)
4. Lee, S., Chen, Y.P., Huang, C.H.: Electroforming of metallic bipolar plates with micro-featured flow field. J. Power Sources **145**, 369–375 (2005)
5. Gietzelt, T., Jacobi, O., Piotter, V., Ruprecht, R., Hausselt, J.: Development of a micro annular gear pump by micro powder injection molding. J. Mat. Sci. **39**, 2113–2119 (2004)
6. Loh, N.H., Tor, S.B., Tay, B.Y., Murakoshi, Y., Maeda, R.: Fabrication of micro gear by micro powder injection molding. Microsyst. Technol. **14**, 43–50 (2007)
7. Imbaby, M., Jiang, K., Chang, I.: Fabrication of 316-L stainless steel micro parts by softlithography and powder metallurgy. Mater. Lett. **62**, 4213–4116 (2008)
8. Zhu, Z., Wei, X., Jiang, K.: A net-shape fabrication process of alumina micro-components using a soft lithography technique. J. Micromech. Microeng. 17, 193–198 (2007)
9. Castro, L., Merino, S., Levenfeld, B., Varez, A., Torralba, J.M.: Mechanical properties and pitting corrosion behaviour of 316L stainless steel parts obtained by a modified metal injection moulding process. J. Mater. Process. Technol. **144**, 397–402 (2003)
10. Ji, C.H., Loh, N.H., Khor, K.A., Tor, S.B.: Sintering study of 316L stainless steel metal injection molding parts using Taguchi method: final density. Mater. Sci. Eng. A **311**, 74–82 (2001)
11. Li, S., Huang, B., Li, D.Y., Liang, S., Zhou, H.: Influences of sintering atmospheres on densification process of injection moulded gas atomised 316L stainless steel. Powder Metall **46**, 241–245 (2003)
12. Shimizu, T., Mutrakoshi, Y., Wechwiayakhlug, K.: Characterization of the molding methods and the binder system in the MIM process. J. Mater. Process. Technol. **63**, 753–758 (1997)
13. Liu, Z., Loh, N., Tor, S., Khor, K., Murakoshi, Y., Maeda, R.: Binder system for micropowder injection molding. Mater. Lett. **48**, 31–38 (2001)
14. Li, S., Huang, B., Li, Y., Qu, X., Liu, S., Fan, J.: A new type of binder for metal injection molding. J. Mater. Process. Technol. **137**, 70–73 (2003)

Chapter 55
Voxel-Based Component Description for Functional Graded Parts

Juergen Gausemeier, Jan Broekelmann, and Dominic Dettmer

Abstract Graded components are a resource-conserving alternative to today's composite materials. Functional gradation means a steady gradient of the property values through the three spatial dimensions of the component. At the present time there is no methodology to specify the graded properties in the component description. Its starting point takes place in the conventional CAD-Models. The component geometry is reproduced with the aid of voxels. The property values are assigned to these voxels. The effort of the description is reduced by the use of interpolation techniques. As a result, we have an enhanced component model which contains all the necessary information for the description of a functional graded component. This model constitutes the basis of the manufacturing process planning for the component production.

Keywords Graded components · Voxel model · component description · CRC transregio 30

55.1 Introduction

The functions of mechanical components are frequently the result of innovative material combinations and a complex geometry. This applies especially for high performance components in the automobile and aviation industry. Graded components are a resource-conserving alternative to today's composite materials. *"Functional gradation is the targeted and reproducable adaptation of a material's microstructure with the intention to establish the macroscopic properties of the component. The objective is the steady progress of the microstructure's variation through at*

J. Gausemeier (✉), J. Broekelmann, and D. Dettmer
Heinz Nixdorf Institute, University of Paderborn, Product Engineering,
Fuerstenallee 11, 33098 Paderborn, Germany
e-mail: Juergen.Gausemeier@hni.uni-paderborn.de; Jan.Broekelmann@hni.uni-paderborn.de; Dominic.Dettmer@hni.uni-paderborn.de

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_55,

least one spatial dimension" [1]. In different places of the component are located apparently contradictory properties, which specifically support the posterior function of the component.

Nevertheless only one material is used. It is possible to define for example the hardness and the damping behavior through the cross-section of the component. This definition takes place within the manufacturing process. The investigation of graded structures and their manufacturing process is the goal of the collaborative research center Transregio 30 (CRC/TR TRR30).

Within the scope of the presented work a flanged steel shaft is used as demostrator of a component with graded properties. The flanged steel shaft is produced in a three-steps-process within the research of the CRC (Fig. 55.1). With the aid of an induction coil, a steel cylinder is locally heated (a) and after that, through a two-steps deformation, the cylinder is transformed into a flanged steel shaft. The reshape process consists of the steps: (b) + (c), which are tool-independent, and the step (c), which is tool-dependent.

After the deformation, the flanged shaft is cooled down. There are two types of cooling: contact-cooling, which takes place inside of the forming tool; outside of the tool, with the aid of compressed air cooling. The gradation takes place within the manufacturing process by the use of a new thermo-mechanically coupled process. Among other things it is possible to define the hardness, the tenacity and the damping behaviour.

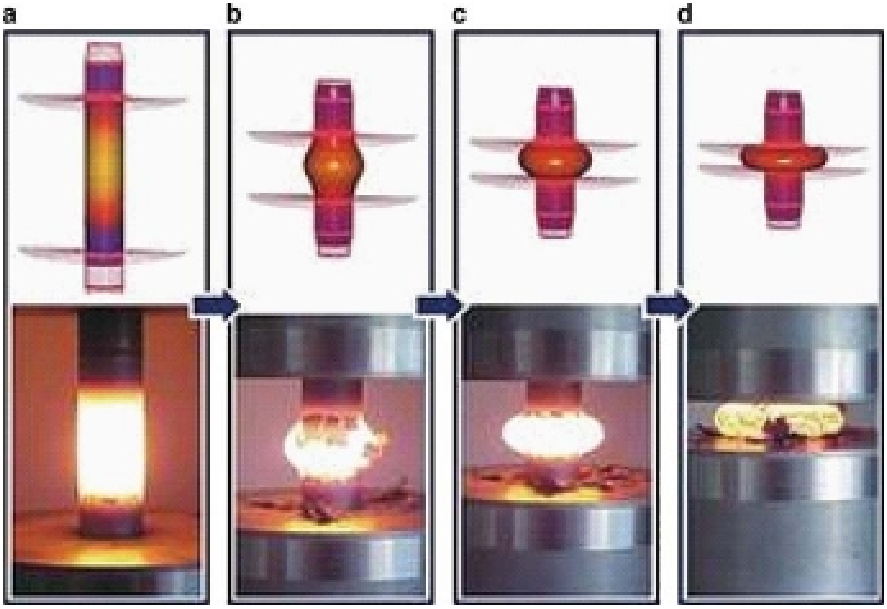


Fig. 55.1 Transforming process of the steel shaft (Professor Steinhoff, University of Kassel)

For the planning of manufacturing processes of components with graded properties a five-steps-systematic was developed:

1. Component description: For this step exists the necessity of projecting the graded properties of the component into a CAD model. From this CAD model is created a voxelmodel. Each single volume element (voxel) can be linked with information of the component properties. Interpolation algorithms are used to simplify the input of the property data. The enhanced component description is used as basis for the manufacturing process planning.
2. Determination of manufacturing functions: The production engineer determines the manufacturing functions based on the enhanced component model. In the future it is planned to assist the production engineer with heuristics, cognition and an inference engine. For each manufacturing function are determined manufacturing technologies. The selection is supported by an expert system. The expert system consists of a knowledge base and the mentioned inference engine. The knowledge base is realized by a combination of a data base and an ontology that save all the information about the manufacturing technologies and the corresponding manufacturing functions as well as the dependencies between them. Afterwards, the determined manufacturing functions and technologies are mapped into a morphological matrix.
3. Synthesis of the process chain: The manufacturing technologies are evaluated against each other in a consistency matrix. This analysis provides highly consistent combinations of manufacturing technologies. The consistent technologies' combinations are used to create process chains. A process chain is a chain of manufacturing technologies in which for each manufacturing function a manufacturing technology is assigned. The identified process chains are verified to be capable for the production of the functional graded component within the process chain optimisation.
4. Process chain optimisation: The relationships between the manufacturing technologies and the component properties are described by empirical models. The models of each technology are combined and used in the optimisation process. By a multi-objective optimisation the values of the parameters of each manufacturing technology are optimized with regard to the desired component properties. Finally a hierarchic optimisation is used to optimize the whole process chain. The optimisation is applied in all the consistent process chains. Only the suitable process chains for the production of the graded component are passed on to the following specification step.
5. Specification of the process chain: The optimisation process delivers process chains which can manufacture the graded properties in an optimal way. The process chains are specified with the optimized parameters' values of the selected manufacturing technologies. Thereby, a specification technique which enables the description of the process and resources is used. With this, a funded conception of a manufacturing system for the production of a component with graded properties is available. This conception constitutes the starting point of the manufacturing system's concretization in the domains: production resource planning, shop floor planning and production logistics [2].

55.2 Concept of the Work

In this work, we present a procedure model for the description of components with graded properties. Through the focused construction of components for the intended application there are property characteristics. These must expand into the component description and require a new methodology for the description.

To present continuous properties modifications within the component geometry it is insufficient to describe only the surface of the component. Today's methods for the component description, like for example the Boundary Representation Method or the Constructive Solid Geometry Method, describe in this way. Component models that are made with the aid of the Boundary Representation Method, describe the component only through delineating edges and areas. An enhancement of this method is the Constructive Solid Geometry Method. With this method, the component is described through the combination of different bodies. The bodies are linked to each other through Boolean operations and represent a 3D model of the component [3, 4].

None of the mentioned methods enables us neither to select any desired point inside the volume nor to describe this point through a linkage with the properties' information. Furthermore no three-dimensional property characteristics are mapped into the component models, but this is necessary for the description of components with graded properties.

Because of this, a methodology for the description of components with graded properties was developed in the sub-project D5 of the CRC TR30. This enables the integration of important properties in the component model which are relevant for the component description.

55.3 Procedure Model

It was developed a procedure model for the description of functional graded components. In the Fig. 55.2 are presented the phases of the component description. In the following paragraphs, each phase is explained with the aid of the flanged steel shaft demonstrator.

55.3.1 Component Construction

Initially, the component is made in a 3D CAD system. After that, the CAD model is exported to the next processing stage. In this phase, there is no difference between the description of traditional components and graded components [5].

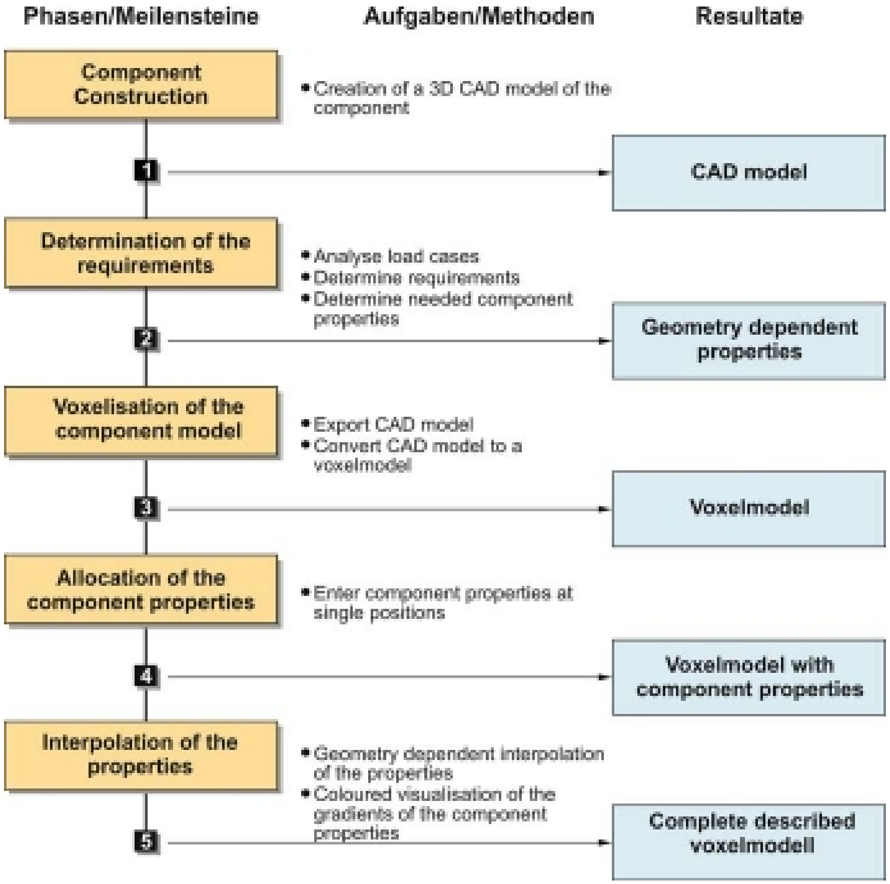

Phasen/Meilensteine
Aufgaben/Methoden
Resultate
Component Construction
• Creation of a 3D CAD model of the component
1
CAD model
Determination of the requirements
• Analyse load cases
• Determine requirements
• Determine needed component properties
2
Geometry dependent properties
Voxelisation of the component model
• Export CAD model
• Convert CAD model to a voxelmodel
3
Voxelmodel
Allocation of the component properties
• Enter component properties at single positions
4
Voxelmodel with component properties
Interpolation of the properties
• Geometry dependent interpolation of the properties
• Coloured visualisation of the gradients of the component properties
5
Complete described voxelmodell

Fig. 55.2 Procedure model for the description of functional graded components

55.3.2 Determination of the Requirements

The required graded properties are determined by the expected application of the component. For each application case are designated load cases. From these cases, the component requirements are derivated. The constructor has to translate these requirements into concrete component properties. The properties of the functional graded components are dependent of the geometry. This implies that also the position in the component is important and must be considered. Because the characteristics have a three-dimensional continuous distribution, one has to define if they are required at the surface or at a certain point within the geometry. The applied example, the flanged steel shaft, presents the following requirements:

- High toughness in the transition section of the shaft and flange to avoid a fracture of the flange during loading

- Soft changing of the characteristics in the transition section of the shaft and flange to avoid crack growth
- A thin layer of high hardness in the boundary section of the flange

The requested component properties are subsequently integrated to the component description. This is not possible for positions in the volume of traditional CAD Models. Therefore is applied a voxel based method for the volume description [4]. The conversion of the CAD Model into a voxelmodel is described below.

55.3.3 Voxelisation of the Component Model

The initial point of the voxelization is the exported CAD model of the component. The volume is divided with the aid of volume elements, the so called voxels. With these voxels, the volume of the component is reproduced. A voxel is the three-dimensional equivalent of a pixel. The geometry is approximated by the combination of the voxels. The accuracy depends on the size of the voxels. For the voxelisation a depth buffer algorithm is used (Fig. 55.3).

This algorithm takes pictures of the component from the x-, y- and z-axis and analyses the depth information. In this way the algorithm manages to decide wether a voxel is inside the component model or outside and how far away the voxel is.

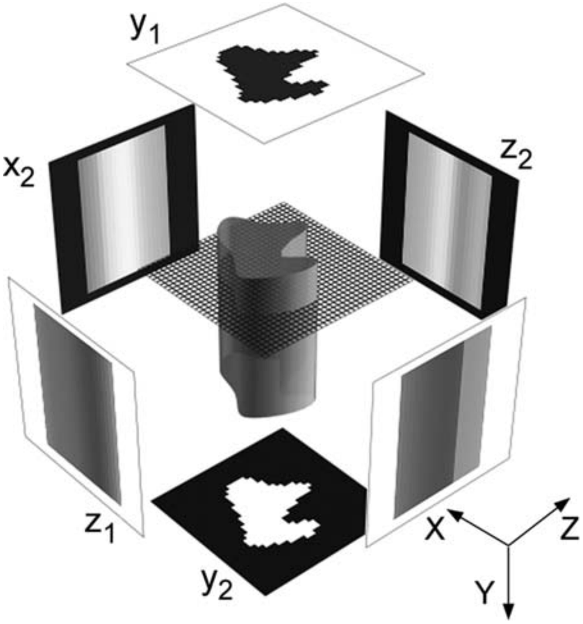


Fig. 55.3 Depth buffer algorithm for the voxelisation of the component model

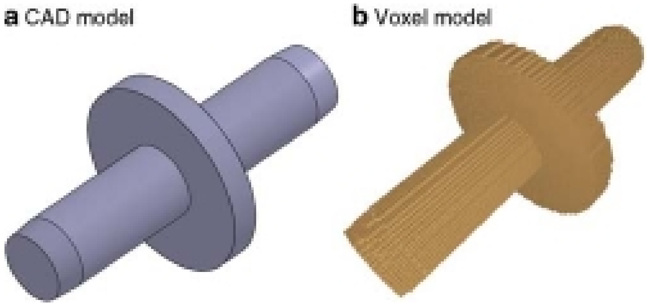


Fig. 55.4 Component models of the flanged shaft

All the voxels outside the component are deleted and a three dimensional voxelmodel is generated. The Fig. 55.4 shows the CAD model and the voxelmodel of the flanged steel shaft. Each volume element is accessible to the user. It is possible in further steps, to link them with information of the component properties.

55.3.4 Allocation of the Component Properties

With the aid of the voxel technology for the component description, it is possible to describe specific points of the component volume. A volume element is chosen and then characteristic properties' information is allocated in it. This is made for all the relevant positions. The data of the properties, including its position, is saved in a database.

Because of the huge number of voxel, depending on the size of the component model more than thousand, the input of the characteristics of each voxel is very complex. For areas with huge gradients of the characteristics more voxels have to be described than in uninteresting areas of the component. In these areas is merely an indication of the boundary values necessary. The assignment of the characteristics to the undescribed volume elements takes place in the next step of the component description.

55.3.5 Interpolation of the Properties

The interpolation of the characteristics values between the described positions follows the description of the needed positions. Because of the geometry dependency of the components' characteristics, the interpolation has to respect these. A linear interpolation of four points on a circular geometry would result from the direct connection line of the points. Therefore the characteristics are described by a rectangular, which does not fit the circular gradient expected.

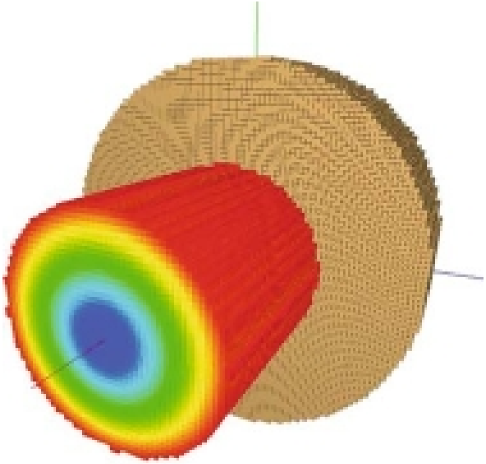


Fig. 55.5 Voxelmodel with interpolated characteristic gradient in the stub shaft

To avoid this problem an interpolation method is used, that implicates the geometry in its interpolation. The identification of the geometrical bodies the component consists of is done manually by now. The planner decides how to intersect the component model and the identified geometrical bodies are used as input for the voxelisation algorithms. By the use of these algorithms the characterization of the undescribed voxels is simplified and the description is easy to manage (Fig. 55.5).

Output of the interpolation phase is a complete described component model. It is described by voxels that are linked to the components' characteristics. The characteristics are depending on their position in the component. To simplify the recognition of the geometry, the application of a feature-recognition is planned. This should help in describing the gradients of the characteristic properties. For detailing the characteristics additional voxels can be described at any time.

The characteristic gradients of the components properties are visualized by a colour gradient in the component model to make a check intuitive and easy. The visualisation for the properties of the flanged steel shaft is shown in Fig. 55.6. From the stubs of the shaft a hardness of 200 HV proceeds until the change to the flange. From the inner region of the flange the hardness shows a gradient up to 700 HV in the boundary region.

55.4 Prototype of the Voxelisation Tool

At the moment a prototype of the software tool for the description of functional graded components is developed and programmed (Fig. 55.7).

The tool uses the exported CAD model of the component. First the dimension of the voxels is chosen. For the voxelisation an algorithm is used, that takes six two-dimensional recordings of different directions of the imported component model.

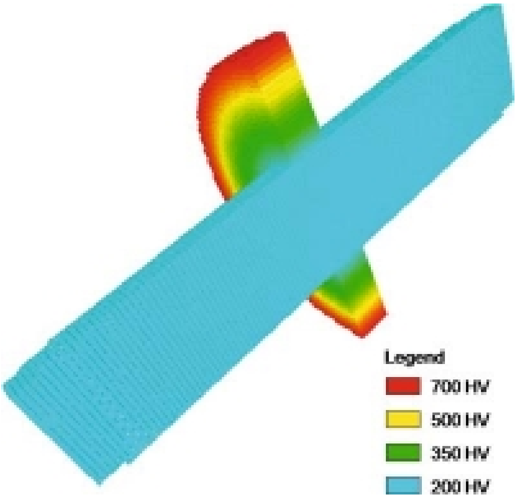
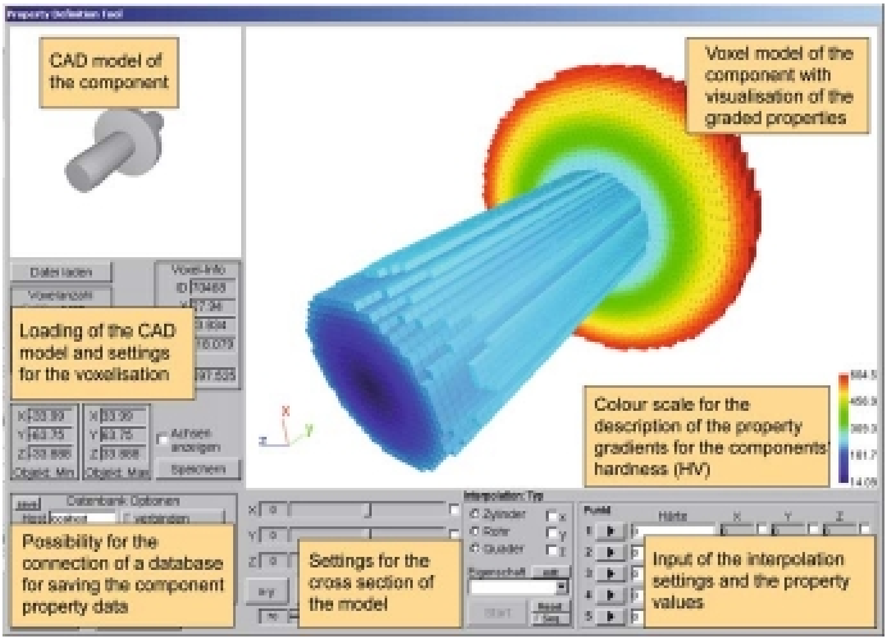


Fig. 55.6 Voxelmodel of the cut flange steel shaft with visualisation of the hardness gradient included



CAD model of the component

Voxel model of the component with visualisation of the graded properties

Loading of the CAD model and settings for the voxelisation

Colour scale for the description of the property gradients for the components' hardness (HV)

Possibility for the connection of a database for saving the component property data

Settings for the cross section of the model

Input of the interpolation settings and the property values

Fig. 55.7 Screenshot of the prototype of the voxelisation tool

By a comparison of the recordings, the boundaries of the component are defined. The components' volume is filled up by voxels afterwards. Thereby a three-dimensional model of the component is generated, where the geometry is copied by the use of cubical elements. The accuracy of the voxelmodel is determined through the combination of the voxel dimension and the dimensions of the component.

To enter the properties values, the voxels are directly selected. For assisting the interpolation at the moment the geometry of the components region to describe is manually forced. After the interpolation step, the properties gradients can be visualised by choosing the property that someone wants to see.

Single steps of the component description of functional graded components can be accomplished by the use of the prototype of the voxelisation tool at present. Further steps of the procedure have to be integrated and to be improved, to assure the description of functionally graded components to be accurate. The enhancement of the tool is one of the goals of our work.

55.5 Prospect

To get information about the geometry of the component the use of feature recognition in the process of the component description is targeted. Feature recognition means to check the CAD model for basic bodies. Features are objects, a component can be build of in a CAD system. For the component description of functional graded parts, geometrical bodies are used as features. These can be bowls, cubes or special features like a drill hole, a flute or a mating surface. The features can be manipulated in their dimensions and it is possible to decompose components into such basic bodies [6]. It is planned to use the identified geometrical bodies to decide how to interpolate the graded properties within the voxelmodel.

The description of functional graded components is used for the development of the manufacturing processes. The information about the geometry as well as the properties of the component is necessary for the planning, because the properties shall be produced integrated in the manufacturing process. This can only be reached when at the beginning of the planning process all gradients of the properties are known. By using the information about the component for the selection and the validation of the manufacturing technologies, a manufacturing process is developed, which is designed especially for manufacturing of the requested component properties. In the optimisation step of the process chains the description of the component properties is used for the evaluation of the process quality.

55.6 Conclusion

The description of functionally graded components is not possible using todays CAD systems. New description methods are needed to describe gradients of properties within the description of functionally graded components. The explained voxel-based approach for the description gives the possibility to do so. An uncomplicated way to describe components and to save Information in it is given by the voxelisation. For the further development of the description approach the integration of feature recognition is planned.

Acknowledgement This work is based on investigations of the Collaborative Research Centre Transregio 30 which is kindly supported by the German Research Foundation.

References

1. Reyes-Perez, M., Gausemeier, J., Nordsiek, D.: Ontology development for a manufacturing data base for products with graded properties. International Conference on Information, Process, and Knowledge Management, eKNOW '09, pp. 105–109 (2009)
2. Eversheim, W., Schuh, G.: Integrierte Produkt- und Prozessgestaltung. Springer (2004)
3. Hoffmann, C.M.: Geometric and Solid Modeling: An Introduction. Morgan Kaufman, San Mateo, CA (1989)
4. Karabassi, E.A., Papaioannou, G., Theoharis, T.: A fast depth-buffer-based voxelization algorithm. J. Graphic. Tools ACM **4**(4), (1999)
5. Frei, N.: Neue wissensbasierte Möglichkeiten im Konstruktionsprozess – Automatisierung des Übergangs der konzeptionellen in die gestalterische Phase im Konstruktionsprozess, CAD-FEM Forum, Knowledge-based Eingineering (KBE), Stuttgart (2003)
6. VDI-Gesellschaft Entwicklung Konstruktion Vertrieb (VDI-EKV): VDI2218 Informationsverarbeitung in der Produktentwicklung – Feature Technologie, VDI Richtlinien (2003)

Chapter 56
A Multi-Parametric Analysis of Drift Flux Models to Pipeline Applications

Joseph Ifeanyichukwu Achebo

Abstract Several interactions occur between the constituents of effluents within a pipeline (fluid, particles, and the pipeline interface). These interactions are birthed from their constant motion in one point in a pipeline relative to another point within the same pipeline. These constant motions expressed through various Drift Flux models are amenable to multi-parametric analysis. This particular exercise successfully elucidates the working parameters used in obtaining the drift flux equations. It utilizes a step by step self explanatory method for calculating the terminal velocity of effluents, being the volumetric flux or relative velocity of fluid/fluid or fluid/particle or fluid-particle/wall interfacial flow contact. Thus, forces encountered as a result of these relative motions are then specifically examined within the parameters of drift flux models. This study, in further applying a multi-parametric analysis of these drift flux models therefore acts as a template which could be used for solving pipeline problems involving these relative motions, once the necessary data has been collated and subsequently computed.

Keywords Continuum Phase Flow · Drift Flux Models · Erosion Wear Rate · Pipeline · Terminal Velocity · Volumetric Flux · Volume Fraction

56.1 Introduction

A multi-parametric analysis involves the collective study of various investigations on drift flux models as applied to both vertical and horizontal pipelines. Drift flux models were first developed by Zuber in 1965. The Drift flux models fall into the category of Computational Fluid Dynamics CFD used for particle transport prediction equations. A drift flux model is employed to represent slip between fluid

J.I. Achebo (✉)
Department of Production Engineering, Faculty of Engineering, University of Benin,
Benin City, Edo State, Nigeria
e-mail: josephachebo@yahoo.co.uk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_56,

phases [4]. Brethour and Hirt [7] were of the opinion that the concept behind the Drift Flux models is that the relative motion between these components can be described as a continuum rather than by discrete elements. A continuum considers the entire process as a whole with no distinct exclusive or conclusive attention given to the elements or parts separately. It is uniquely used for the study of sedimentation, fluidised beds and other flow processes that involve relative motion (interaction) between phases that are controlled by buoyancy and fluid drag forces. The relative flow moves in a slip pattern propagated by Kinematic Shocks or Expansion Waves mostly caused by turbulent fluid motion aided by external and internal forces such as the inward radial pressure generated by interfacial surface tension of a fluid in a stable high thermal environment as in the case of a viscoplastic slurry or paste transport situation.

Applications range from scenarios and processes that see the occurrence of elements as simple as bubbles and slug particles, to more complex and subtly devastating affects such as pipeline erosion and the attendant wear rate. Drift Flux models consider the different densities and sizes of the volume fraction of particles assumed to be continuously slipping; in other words it considers the relative motion between and within the fluids or fluid/particle or fluid/particle/pipe wall at constant velocity due to gravitational and/or centrifugal forces [9].

The aforementioned Zuber considered a one dimensional flow of a mixture of two components, A and B. The volumetric fluxes of the two components, j_A and j_B, were related to the total volumetric flux, j, the drift flux, j_{AB} and the volume fraction, $\alpha_A + \alpha_B = 1$. To determine the relative motion (Drift Flux) by applying the theory of dynamics to the forces on the individual phases, the momentum and energy equations would have to be understood although not exclusively or conclusively.

Drift flux models are not without limitations in that some multiphase flows can not be approximated especially when the relative motion is intimately connected with the pressure and velocity gradients in the two or more phases. However, since Zuber, many researchers have applied the model to a two phase flow with success [2, 3, 8].

A fluid carrying pipeline is rife with opportunities for applying drift flux models. In this research work, the application of the various drift flux models to pipeline engineering is examined.

56.2 Application to a Vertical Pipe Considering the Buoyancy Effect

Since their introduction in the 1960s the Drift Flux models have proven very adaptable to various engineering challenges encountered. This versatility is of immense value in the prediction of expected and anticipated engineering failures, it is also important for examining post failure root cause analysis. The need for a full fledged parametric analysis of various drift flux models cannot be overemphasised.

The governing equations describing one dimensional two phase drift flux transport equations in vertical pipes are considered by stating the mass conservation, momentum and internal energy conservation equations; this is a basic start:

For Mass Conservation

$$\frac{\partial p_j}{\partial t} + \frac{\partial}{\partial z}\left(P_j U_j\right) = 0 \tag{56.1}$$

For Drift Flux Momentum Conservation

$$\rho_j \frac{\partial v_j}{\partial t} + \rho_j V_j \frac{\partial V_j}{\partial z} + \frac{\partial}{\partial z}\left(\frac{\rho_v \rho_1 \alpha V_{vj}^2}{\rho_j \left(1-\alpha\right)}\right) = -\sum_{i=1}^{N} F_{gl} - \frac{\partial P_j}{\partial z} - \rho_j g \cos\theta \tag{56.2}$$

where θ is the angle of contact between the surface of the liquid and the surface of the pipe which is assumed to be 0°.

Drift flux internal Energy conservation

$$\begin{aligned} &\frac{\partial}{\partial t}\left(\rho_j u_j\right) + \frac{\partial}{\partial z}\left(\rho_j u_j v_j\right) + \frac{\partial}{\partial z}\left[\frac{\alpha \rho_l \rho_v \left(u_v - u_l\right) V_{vj}}{\rho_j}\right] \\ &\quad + P\frac{\partial V_j}{\partial z} + P_m \frac{\partial}{\partial z}\left[\frac{\alpha\left(\rho_l - \rho_v\right) V_{vj}}{\rho_j}\right] \\ &= \sum_{i=1}^{N} q_l^n \frac{P_1}{A_l} + V_j\left(\sum_{i=1}^{N} F_{wi}\right) \end{aligned} \tag{56.3}$$

where A is the surface area of the pipe; $\mathrm{u_v}$ and $\mathrm{u_l}$ are the vapor and liquid velocities respectively, V is the volumetric flow rate, $\mathrm{F_g}$ is the gravitational force, P is the operating pipe pressure, Z is the height of the pipe, ρ_l and ρ_v are the liquid and vapor density respectively, q is the heat flux whereas, $\mathrm{F_{wi}}$ is the wall shear force.

This study was done by examining the works of Holman [13] and DF Models [9]. In nucleate boiling, Holman [13] observed that bubbles are created by the expansion of entrapped gas or vapor at small cavities in the surface. The bubbles increase in size depending on the surface tension at the liquid vapor interface and the temperature and pressure.

In this scenario, a superheated liquid at its boiling point would have bubbles of vapor form on the heating element surface. These bubbles collapse as the heat increases, and the entrapped gases escape through the liquid to the surface of the vertical pipe being investigated. The volumetric drift flux of bubbles as they move through the liquid is represented by Eq. (56.4)

$$J_{VL} = \alpha\left(1-\alpha\right) U_{VL} \tag{56.4}$$

where $\mathrm{J_{VL}}$ is the drift flux, $\mathrm{U_{VL}}$ is the relative velocity and α is the volume fraction.

The relative velocity can also be represented in Eq. (56.5)

$$U_{VL} = U_{VLO}\,(1-\alpha) \tag{56.5}$$

In terms of the terminal velocity of single bubble in the dispersed vapor phase, U_{VLO}, as represented in Eq. (56.6) and the corresponding drift flux written in Eq. (56.7)

$$U_{VL} = U_{VLO}\,(1-\alpha)^{b-1} \tag{56.6}$$

$$J_{VL} = U_{VLO}\,(1-\alpha)^{b} \tag{56.7}$$

The term b is some constant of the order of 2 or 3. b takes on values from 3 for very minute bubbles to 2 for somewhat larger bubbles.

To determine the terminal velocity of individual bubbles rising, U_{VLO}, the first step here is to determine the radius of the bubble, R. The buoyancy force, F_b which propels the gas through the liquid is considered and expressed in Eq. (56.8)

$$F_b = \frac{4}{3}\pi R^3 g\,(\rho_L - \rho_V) \tag{56.8}$$

where R is the radius of the bubble, g is the acceleration due to gravity and ρ_L, ρ_V is the density of the liquid and vapor, respectively. The surface tension force F_σ, is also considered

$$F_\sigma = 2\pi R\sigma \tag{56.9}$$

where σ is the surface tension of the liquid and vapor interface

$$\sigma = \frac{1}{2}\left[(\rho_L - \rho_V)\,gR\left(H + \frac{R}{3}\right)\right]$$

where R is the radius of the bubble and H is the rise of the bubble [11].

Equating the two forces of Eqs. (56.8) and (56.9) gives a formula for R, to be

$$R = \left[\frac{3\sigma}{2g\,(\rho_L - \rho_V)}\right]^{\frac{1}{2}} \tag{56.10}$$

The second step here is to determine U_{VLO}. This is achieved by equating the drag force. F_D to the buoyancy force, F_b in Eq. (56.8)

$$F_D = \frac{C_D \pi R^2 \rho_L U_{VLO}^2}{2} \tag{56.11}$$

where C_D is the drag coefficient.

Equating Eqs. (56.8) to (56.11) generated Eq. (56.12)

$$U_{VLO} = \left[\frac{8Rg\,(\rho_L - \rho_v)}{3\rho_L C_D}\right]^{\frac{1}{2}} \tag{56.12}$$

When Eq. (56.12) is substituted into Eq. (56.6) and Eq. (56.7), the values of the volumetric drift flux and the relative velocity of the bubble and liquid interface would be obtained.

However, Holman [13] was of the opinion that when a liquid is heated above the saturation temperature, boiling occurs and the heat flux will depend on the difference in temperature between the surface and the saturation temperature. Zuber and Findlay [21] proposed an equation to determine the peak heat flux in nucleate boiling as expressed in Eq. (56.13)

$$\left(\frac{q}{A}\right)_{\max} = \frac{\pi}{24} h_{fg} \rho_v \left[\frac{\sigma g \left(\rho_L - \rho_V\right)}{\rho_v^2}\right]^{\frac{1}{4}} \left(1 + \frac{\rho_V}{\rho_L}\right)^{\frac{1}{2}} \tag{56.13}$$

where q is the heat flux, A is the surface area of the pipe, represented in Eq. (56.14)

$$A = \pi dL \tag{56.14}$$

d and L is the diameter and length of the pipe respectively. The heat transfer coefficient, h_{fg} is expressed in Eq. (56.15) as

$$h_{fg} = 2.54 \left(T_V - T_l\right)^3 e^{\frac{p}{1.551}} \left(W/m^2{}^{\circ}C\right) \tag{56.15}$$

$(5 < P < 170$ atm$)$
where P is the pressure in meganewtons per square meter. T_v, T_L are the vapor and liquid temperature, respectively.

Holman [13] stated that in saturated boiling, when the bubbles break away from the surface because of the buoyancy action, the bubbles move back into the body of the liquid. This results when the temperature of the surrounding liquid is lower than the saturated temperature in the bubble. This can be explained by deriving an expression for the pressure gradient that exists between the interface of the vapor and liquid phase.

The pressure force F_p and the surface tension force, F_σ are considered at equilibrium

$$F_\rho = \pi R^2 \left(P_V - P_l\right) \tag{56.16}$$

where, P_v is the vapor pressure inside the bubble and P_L is the liquid pressure.

Equating Eqs. (56.9) to (56.16) generated Eq. (56.17)

$$P_v - P_L = \frac{2\sigma}{R} \tag{56.17}$$

Holman [13] was of the opinion that Eq. (56.17) indicates that when the pressure inside the bubble is reduced, the corresponding vapor temperature will also reduce. This implies that the bubble will rise and move further away from the heat source to where the liquid temperature is lower. This means that heat is conducted out of the bubble and the vapor inside the bubble condenses and collapses back to the liquid especially in a forced convective condition.

In this condensed state, to determine the drift flux and relative velocity of the vapor–liquid interface. The terminal velocity, U_{VLO} should be obtained.

Here, the net gravitational force, F_g is equated to the drag force, F_D

$$F_g = \frac{4}{3}\pi R^3 g\left(\rho_L - \rho_V\right) \tag{56.18}$$

Eq. (56.18) is same as Eq. (56.8)

$$F_D = \frac{C_D \pi R^2 \rho_v U_{VLO}^2}{2} \tag{56.19}$$

Equating Eqs. (56.18) to (56.19) generated Eq. (56.20)

$$U_{VLO} = \left[\frac{8Rg\left(\rho_L - \rho_v\right)}{3\rho_v C_D}\right]^{\frac{1}{2}} \tag{56.20}$$

The value for R, as determined by Zuber et al. [20] is expressed in Eq. (56.21)

$$R \approx \lambda\alpha\left[\frac{\sigma}{g\left(\rho_L - \rho_V\right)}\right]^{\frac{1}{2}} \tag{56.21}$$

where λ is the wavelength in its unstable state related to Rayleigh–Taylor unstable surface and it is assumed to be equal to the size of water droplets at the vapor/liquid interface [9].

Sun and Lienhard [16] proposed an equation for determining q_L, to be

$$q_L = \frac{0.061}{K} \tag{56.22}$$

where

$$K = \frac{d}{\left[\frac{\sigma}{g\left(\rho_l - \rho_v\right)}\right]^{\frac{1}{2}}} \tag{56.23}$$

where d is the diameter of the tube. Equation (56.23) should be used when $K < 2.3$, however where $K < 0.24$, there is no nucleate boiling.

56.3 Drift Flux Models as Applied to Wear Rate in Horizontal Pipelines

In this case, the wear rate effect on the interface between the volume fraction of particles immersed in a transport fluid and the internal walls of a pipeline have been studied [1]. However, the Eulerian continuum flow model, the particle equation of motion and the erosion prediction equation are explained here in detail.

56.3.1 The Continuous Model

This model describes the behaviour of fluid flow patterns in a continuous phase. In this phase the conservation equations for mass and momentum in combination with transport equations for a turbulence model are applied. Tian [17] was of the opinion that in CFD model equations, governing equations are fundamentally based on fluid dynamics, which represents the mathematical statements of the conservation law of physics. These laws have been derived from the fact that certain measures must be conserved in a particular volume, known as a control volume. The governing equations for axisymmetric turbulent flow were expressed as follows [4, 18].

$$\frac{\partial}{\partial x_j}\left(\rho u_j\right) = 0 \tag{56.24}$$

where U_j is the average or mean velocity component and ρ is the fluid density.

Equation (56.24) is expanded as expressed in Eq. (56.25)

$$\frac{\partial}{\partial x_j}\left(\rho_f u_{if} u_{jf}\right) = \frac{-\partial P_t}{\partial x_i} + \frac{\partial}{\partial x_j}\left(\mu \frac{\partial u_i}{\partial x_j}\right) - \frac{\partial}{\partial x_j}\left(\rho u_i u_j\right) \tag{56.25}$$

where P is the static pressure and the stress tensor was further expanded as written in Eq. 56.26 as proposed by Hinze [12]

$$-\rho u_i u_j = \left[\mu_{f,t}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right] - \frac{2}{3}\rho k \delta_{ij} \tag{56.26}$$

where δ_{ij} is the Kronecker delta and μ_{gt} is the eddy viscosity or turbulent viscosity. The turbulent quantity, K which is the Kinetic energy of turbulence expressed in Cartesian tensor notation as

$$K = \frac{1}{2}\overline{u_i u_i} \tag{56.27}$$

Can be simply expressed as

$$K = 0.01 u_f^2 \tag{56.28}$$

The rate of production of turbulent kinetic energy, P_k is given by

$$P_k = -\rho_f \overline{u_i^1 u_j^1} \frac{\partial u_{if}}{\partial x_j} \text{ and}$$

the rate of dissipation of turbulent energy, ε is expressed as

$$\varepsilon = \frac{\mu_{ft}}{\rho_t} \overline{\left(\frac{\partial u_i^1}{\partial x_j}\right)\left(\frac{\partial u_i^1}{\partial x_j}\right)} \tag{56.29}$$

ε can simply be calculated from Eq. (56.30)

$$\mu_{ft} = \rho_f C_\rho \frac{K^2}{\varepsilon} \tag{56.30}$$

where C_ρ is the specific heat capacity of the fluid and was given by Reynolds [14] as 0.0845.

The turbulent viscosity, if not given can be determined from Eq. (56.31)

$$\mu_{f,t} = \upsilon \rho_t \tag{56.31}$$

where υ is the Kinematic viscosity of the fluid. Considering Eqs. (56.28)–(56.31), the term K can be obtained and substituting the value of K into Eq. (56.28), U_m would be obtained.

The mass rate of flow, $\dot{m}$ is calculated from Eq. (56.32)

$$\dot{m} = \rho_t u_f A \tag{56.32}$$

where A is the cross sectional area of the pipe given as πdl, where d and l have been previously defined.

The mass flow velocity, G is given in Eq. (56.33) as

$$G = \frac{\dot{m}}{A} = \rho_f u_m \tag{56.33}$$

The term U_f is the fluid terminal velocity.

56.3.2 Particle Equation of Motion

In deriving this equation, two assumptions were employed

1. The solid particles do not interact with each other.
2. The influence of particle motion on the fluid flow field is very small and could be neglected.

These assumptions were also adopted by Edwards et al. [10] and Wallace et al. [19] in their various research works.

The governing particle equation of motion is given as

$$\frac{du_p}{\partial t} = F_D(u_f - u_p) + \frac{g\ (\rho_\rho - \rho_f)}{\rho_\rho} + \sum F_x + \Delta P + F_d \tag{56.34}$$

where $F_D\ (u_f - u_p)$ is the drag force per unit particle mass and F_D is given by

$$F_D = \frac{3C_D\mu\mathrm{Re}_p}{4\rho_p d_p^2} \tag{56.35}$$

where ρ_p is the density of particle material, $\mathrm{d_p}$ is the particle diameter, $\mathrm{u_p}$ is the particle velocity and $\mathrm{Re_p}$ is the relative Reynolds number written as expressed in Eq. (56.36)

$$\mathrm{Re}_p = \frac{\alpha_f\rho_f d_p(u_p - u_f)}{\mu_{f,t}} \tag{56.36}$$

where α_f is the volume fraction of the fluid, $\mathrm{U_p}$ is the particle velocity and $\mathrm{C_d}$ is the drag coefficient, this is defined in Eq. (56.37)

$$C_D = \begin{cases} 0.44 & Re_p > 1000 \\ \dfrac{24}{Re_p}\left(1.0 + \dfrac{1}{6}\mathrm{Re}_p^{0.66}\right) & \mathrm{Re}_p \le 1000 \end{cases} \tag{56.37}$$

$\frac{g(\rho_p-\rho_f)}{\rho_p}$ represents the particle buoyancy force that keeps the particles in continuous flow suspension when it is at equilibrium with the pressure force $\mathrm{F_d}$ is the Saffman lift force proposed by Saffman [15].

$\sum F_x$ is the increase in momentum flux in the fluid around the particles. This could be represented as

$$\sum F_x = \frac{\partial(mu_T)x}{\partial\tau} \tag{56.38}$$

where $\mathrm{U_T}$ is the friction velocity and τ is the shear force due to flow.

Most of the energy loss takes place during the algebraic particle-wall collision at the interface. This causes the disintegration of particles and pipewall deformation. However, large energy loss due to molecular level forces, such as adhesion is not reversible and occurs primarily during rebound [6, 17].

The friction velocity or relative velocity at the interface can be calculated from Eq. (56.39)

$$U_T = \sqrt{\frac{\tau_w}{\rho_f}} \tag{56.39}$$

where τ_w is the wall shear stress

$$\text{The pressure gradient,}\Delta P = f\frac{L}{d}\rho_f\frac{U_T^2}{2g} \tag{56.40}$$

where f is the frictional force, L is the length of the pipe and d is the diameter of the pipe.

Here Eqs. (56.36)–(56.41) is used to determine the velocity of the particle, $\mathrm{U_p}$.

The value of $\mathrm{U_p}$ is used to compute the Erosion wear rates of the pipeline depending on the angle of contact between the particle transported by the fluid and the internal pipewall, α.

56.3.3 The Erosion Prediction Equation

The erosion prediction equations suggested by Wallace et al. [19] were used because of their simplicity and high level of accuracy for the prediction of erosion rates. These equations are given in Eqs. (56.41) and (56.42).

$$E = \left\{ \frac{\frac{1}{2}U_p^2 \cos^2 \alpha \sin 2\alpha}{\Upsilon} + \frac{\frac{1}{2}U_p^2 \sin^2 \alpha}{\sigma} \right\} \tag{56.41}$$

For $\alpha \leq 45°$

And

$$E = \left\{ \frac{\frac{1}{2}U_p^2 \cos^2 \alpha}{\Upsilon} + \frac{\frac{1}{2}U_p^2 \sin^2 \alpha}{\sigma} \right\} \tag{56.42}$$

For $\alpha > 45°$.

Where Y and σ are the cutting wear and deformation wear coefficients having the values 33,316.9 and 77,419.7, respectively.

From the study made by Bitter [5] peak erosion rates have been measured to occur at impact angles of 25–30°, indicating that cutting wear dominates. The difference between U_p and U_f gives the drift flux velocity. The term of $\left|U_f - U_p\right|$ can be used to replace the term $\left|U_p\right|$ in Eqs. (56.41) and (56.42) to obtain erosion rates based on the relative motion of fluid/particle interface.

56.4 Conclusion

The parametric analysis of these select drift flux models has introduced a wide range of applications. The application of drift flux models of fluid/fluid flow in a stagnant position and fluid/particle flow in a continuum phase to pipeline have been clearly expressed. The effect of interfacial motion between the fluid-particle and pipe wall as related to the relative motion of the fluid velocity and the particle velocity was applied to the erosion wear equation to determine the resistance of the pipe wall material to wear. These models therefore act like templates for solving pipeline engineering problems whenever they arise or when anticipated.

References

1. Achebo, J.I.: Computational analysis of erosion wear rate in a pipeline using the drift flux models based on eulerian continuum equations. In: Proceedings of the World Congress on Engineering (I) 1–3 July pp. 719–721 London, UK (2009)
2. Ahmad, K.M.: Effect of solid catalyst on bubble rise velocity and gas drift flux in three phase columns. Eng. Tech. **25**(1), 97–102 (2007)

3. Ambroso, A., Chalons, C., Coquel, F., Galie, T., Godlewski, E., Raviart, P., Segun, N.: The Drift Flux asymptotic limit of barotropic two phase two pressure models. Commun. Math. Sci. **6**(2), 521–529 (2008)
4. Bahr, H.M., Habib, M.A., Ben-Mansour, R., Said, S.A.M.: Effect of flow velocity and particle size on erosion in a pipe with sudden contraction. In: The 6th Saudi Engineering Conference, KFUPM, Dhahran held in December (5), 79–88 (2002)
5. Bitter, J.G.A.: A study of erosion phenomena parts 1 and 2. Wear **6**, 5–21 & 169–190 (1963)
6. Brach, R.M., Dunn, P.F.: Models of rebound and capture for oblique microparticle impact. Aerosol. Sci. Tech. **29**, 379–388 (1998)
7. Brethour, J.M., Hirt, C.W.: Drift flux model for two component flow. http://www.flow3d.com/pdfs/tn/flosci-TN83.pdf) (2009)
8. Dai, W., Woodward, P.R.: A high order iterative implicit – explicit hybrid scheme for magneto hydrodynamics. SIAM J. Sci. Comput. **19**, 1827–1846 (1998)
9. DF Models: Chapter 14: drift flux models. http://www.caltechbook.library.caltech.edu/51/1/chap14.pdf (2009)
10. Edwards, J.K., Mclanry, B.S., Shirazi, S.A.: Evaluation of alternative pipe bend fittings in erosive service. In: Proceedings of ASME Fluids Engineering Summer Meeting. Paper No. FEDSM 2000–11245, Boston, MA, 11–15 June (2000)
11. Experiment 446.1.: Surface tension of liquids. http://www.udel.edu/pchem/C446/Experiments/exp1.pdf (2006)
12. Hinze, J.O.: Turbulence. McGraw-Hill, New York (1975)
13. Holman, J.P.: Heat Transfer, 4th edn, pp. 364–379. McGraw-Hill, Tokyo (1976)
14. Reynolds, W.C.: Fundamentals of turbulence for turbulence modeling and simulation. In: Lecture Notes for Von Karm. Inst, Agard. Report No. 755 (1987)
15. Saffman, P.G.: The lift on a small sphere in a slow shear flow. J. Fluid Mech. **22**(2), 385–400 (1965)
16. Sun, K. H., Lienhard, J.H.: The peak boiling heat flux on horizontal cylinders. Int. J. Heat Mass Trans. **13**, 1425 (1970)
17. Tian, Z.: Numerical modelling of turbulent gas – particle flow and its applications. PhD Thesis in the School of Aerospace Manufacturing andMechanical Engineering, RMIT University (2006)
18. Versteeg, H.K., Malalasekera, W.: An Introduction to Computational Fluid Dynamics: the Finite Volume Method. Longman, London (1995)
19. Wallace, M.S., Peters, J.S., Scanlon, T.J., Dempster, W.M., McCulloch, S., Ogilvie, J.B.: CFD-based erosion modelling of multi-orifice choke valves. In: Proceedings of ASME Fluids Eng Sum. Meeting, Paper No. FEDSM2000–11244, Boston, MA, 11–15 June (2000)
20. Zuber, N. et al.: The dynamics of vapor bubbles in non uniform temperature fields. Int. J. Heat Mass Trans. **2**, 83–98 (1961)
21. Zuber, N., Findlay, J.A.: Average volumetric concentration in two phase flow systems. J. Heat Trans. **68**, 453–468 (1965)

Chapter 57
Influence of Preventive Maintenance Frequency on Manufacturing Systems Performances

Antonio C. Caputo and Paolo Salini

Abstract Preventive maintenance (PM) is often scheduled in order to minimize the maintenance cost or to comply with production planning requirements. However, maintenance interruptions affect process time variability and resource utilization, thus causing adverse effects on the operational performance of the manufacturing system such as Work in Process (WIP) and cycle time. In this paper some approximate queueing models are utilize to asses the impact of PM interval on WIP. It is shown that maintenance intervals corresponding to a minimum maintenance cost or minimum WIP can be quite different thus asking for trade-off decisions by plant managers.

Keywords Maintenance optimization · manufacturing systems performances · preventive maintenance · queueing models

57.1 Introduction

Manufacturing systems are subject to deterioration with usage and age. In case of repairable systems, maintenance can restore the operational status of manufacturing equipment after failures or can preserve it by reducing the occurrence of breakdowns. However, maintenance downtimes increase resources utilization and system variability, negatively affecting relevant performance measures, such as work in process (WIP) and cycle time. While a vast body of literature about maintenance planning and optimization exists [1–3], the interactions between maintenance planning and manufacturing systems performances has been scarcely investigated. In particular, some criteria have been proposed to dynamically determine maintenance actions based on the system status [4–7]. However, most of these approaches

A.C. Caputo (✉) and P. Salini
Department of Mechanical, Energy and Management Engineering, University of L'Aquila, Faculty of Engineering, Via Campo di Pile, Zona industriale di Pile, 67100 L'Aquila, Italy
e-mail: antoniocasimiro.caputo@univaq.it; paolo.salini@univaq.it

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_57,

are quite complex to adopt in real life conditions. In this paper, instead, some easy to use approximate queueing models are utilized to assess the impact of preventive maintenance interval on manufacturing systems performance, mainly focusing on WIP. This is made to show how the selection of preventive maintenance frequency, be it arbitrary or aimed at minimizing maintenance cost, can be detrimental to WIP and cycle time, so that trade-off decisions may be required.

57.2 Economic Optimization of Preventive Maintenance Schedule

In systems subject to wear and thus having an increasing failure rate (IFR), preventive maintenance bringing the system into its original state (as good as new) is beneficial as it reduces the average failure rate. The problem then arises of determining the preventive maintenance interval T_P. This is usually selected by maintenance planners in order to respect some external requirement (such as instructions from the equipment manufacturers or directives from the production planning department) or, too often, is decided arbitrarily. In case one wishes to optimize T_P, a maintenance cost minimization approach is usually pursued. In this respect many policies exist [1–3] which mostly fall into two classes, namely age replacement policies and periodic replacement policies. In age replacement policy a unit is always replaced at failure, or at time T_P if it has not failed up to time T_P. In both cases an "as good as new" intervention is often assumed which means that the failure rate is restored to its initial value. In periodic replacement policy, instead, a unit is replaced periodically at planned times $kT_P(k = 1, 2, \ldots)$. The preventive maintenance interval T_P is often chosen to minimize the overall maintenance cost including both corrective repairs after a breakdown and planned preventive replacements. Over an infinite time horizon one usually refers to the expected maintenance cost per unit time of a maintenance cycle $C_{AU}(T_P)$. In the periodic replacement policy, considered in this paper, $C_{AU}(T_P)$ is computed as the ratio of the average maintenance cost over a cycle to the average cycle duration.

$$C_{AU}(T_P) = \frac{C_P + C_B\, N(T_P)}{T_P}. \tag{57.1}$$

In (57.1) C_P is the cost of a preventive maintenance intervention and C_B is the cost of a corrective intervention following a breakdown, while T_P is the fixed cycle length and $N(T_P)$ the average number of corrective repairs expected over the cycle length as given by renewal theory [1]. Since with short T_P one has high preventive maintenance costs but low corrective maintenance costs, while for long T_P the opposite occurs, an optimal value of T_P which minimizes $C_{AU}(T_P)$ may exist and numerical methods can be used to determine it. However, this widely practiced approach only focuses on costs while neglects other performance measures of a manufacturing system which may be worsened by an unsuitable choice of T_P. In fact, the preventive maintenance interval besides affecting the average failure rate and maintenance cost, also influences resource availability and utilization, as well as the variability of effective processing times, thus determining congestion problems which may increase WIP accumulation and cycle time at the workstations.

57.3 Queueing Models for Unreliable Machines

Queueing theory [8] allows quick estimation of the main performance measures of manufacturing systems. Analytical queueing models provide generalizable results and explicitly show the role of the influencing parameters, while this is not possible in discrete events computer simulation models. The latters, on the other hand, are much more flexible and powerful, but are very time consuming to create and validate. However, even if from a long time a number of queueing models have been developed for unreliable servers [9–11], most of them are based on the assumption of exponentially distributed time to failure (i.e. constant failure rate) and only address preemptive interruptions, such as breakdowns, or non-preemptive such as preventive maintenance. This prevents from using simple queueing models to optimize maintenance policies, as one needs to model both kind of interruptions. Moreover, the assumption of constant failure rate makes the models unsuitable to examine preventive maintenance policies, which are only useful when the system shows an increasing failure rate caused by progressive wear and deterioration. Furthermore, due to the complex nature of interruptions in manufacturing, it is often difficult to properly select the appropriate queueing model. To this end Wu et al. [12] propose a useful classification. At first they distinguish between preemptive and non-preemptive interruptions. Preemptive interruptions are unscheduled and can occur during the processing of a job, thus inflating the average process time respect the value of the natural process time. Non-preemptive interruptions, instead, are usually scheduled and, in any case, can be postponed until the job processing is terminated. Then they distinguish between run-based and time-based interruptions. Run-based interruptions can occur only if WIP exist in the system or are indirectly caused by the presence of WIP. For instance, the breakage of a tool can occur only if the machine is processing a part. Time-based interruptions, instead, can occur even in absence of WIP. Examples of run-based and time-based preemptive interruptions are, for instance, breakdowns or out of spec inputs, and power outages respectively. Cases of run-based and time-based non-preemptive interruptions are, for instance, setups and preventive maintenance respectively. Finally, they further consider state-induced or product-induced events as sub-cases of run-based non-preemptive events (i.e. a state-induced event is an interruption deriving from a change of state of the machine such as a warm up period when the machine passes from stand-by to working conditions). According to this classification, in this work we are interested in run-based preemptive events (i.e breakdowns) and time-based non-preemptive events (i.e. preventive maintenance.). While the paper of Wu et al. [12] gives a more complete classification of M/M/1, M/G/1 and G/G/1 queueing models referring to run-based or time-based interruptions when the uptime is exponentially distributed, here we consider only models for queues with Poisson arrivals and general service processes in single servers applications (M/G/1), which better fit the scope of this paper. Considering that no simple queueing models are available which include explicitly both preemptive and non-preemptive interruptions for unreliable servers with Weibull distributed increasing failure rate, in this paper we adapt two approximate queueing models from the literature to explore the impact of maintenance policy on system performances. In particular, the models are utilized to estimate the average WIP at the workstation when the preventive maintenance interval T_P is changed, in order to compare the T_P value corresponding to a minimum WIP, if any, to the value corresponding to the minimum maintenance cost per unit time, and to assess the overall effect of changing T_P on system WIP.

57.3.1 Model I

By using the heavy traffic approximation of Whitt [13], valid for queues with general arrival and general service processes in single servers applications (G/G/1), the queueing time (QT) can be estimated as,

$$E(QT) = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho}{1-\rho}\right) t_e \qquad (57.2)$$

where $c_a = \sigma_a / t_a$ is the coefficient of variation of interarrival time, c_e is the coefficient of variation of effective process time, $\rho = \lambda / \mu A$ is the resource utilization, A the breakdown induced server availability, λ the average arrival rate and μ the average processing rate, t_e is the expected value of the Effective Process Time (EPT). EPT represents the average effective process time as modified respect the average natural process time t_0 to account for the interruptions. Please note that this definition of t_e does not include state-induced, run-based non-preemptive events. In case of Poisson arrivals obviously $c_a = 1$. Hopp and Spearman [14] provide the following equations to compute the parameters of (57.2) in case of either preemptive or non-preemptive run-based events.

In Table 57.1 t_a is the average interarrival time and σ_a its standard deviation. σ_0 is the standard deviation of natural process time, MTTR and σ_r the mean value and

Table 57.1 Effective process time computation equations [14]

Preemptive interruptions		
$t_e = \frac{t_0}{A}$	$\sigma_e^2 = \frac{\sigma_0^2}{A^2} + \frac{\left(MTTR^2 + \sigma_r^2\right)(1-A)t_0}{A\,MTTR}$	$c_e^2 = c_0^2 + \frac{\left(1 + c_r^2\right)A(1-A)\,MTTR}{t_0}$
Non-preemptive interruptions		
$t_e = t_0 + \frac{t_t}{N_t}$	$\sigma_e^2 = \sigma_0^2 + \frac{\sigma_t^2}{N_t} + \frac{N_t - 1}{N_t^2} t_t^2$	$c_e^2 = \frac{\sigma_e^2}{t_e^2}$

standard deviation of time to repair, c_0 is the coefficient of variation of natural process time, c_r is the coefficient of variation of repair time, t_t is the average duration of non-preemptive interruption (i.e. here the preventive maintenance downtime), σ_t its standard deviation, and N_t is the average number of jobs processed between non-preemptive interruptions.

In case one deals with both preemptive and non-preemptive interruptions Hopp and Spearman [14] suggest at first to compute the t_e and σ_e values including only preemptive interruptions and then to use such values as actual starting values t_0, σ_0 to compute the final t_e and σ_e values from non-preemptive interruptions equations.

Unfortunately, their expression for preemptive c_e^2 is valid only in case of constant failure rate, and an explicit analytical expression for the coefficient of variation of effective process time with increasing failure rate is difficult to obtain. Moreover, their expression for non-preemptive σ_e^2 could be used only to account for setups but not for preventive maintenance as the latter is useless in case of constant failure rate. Nevertheless, when the failure rate is increasing, the maintenance interval, indirectly represented by $\mathrm{N_t}$, influences the actual value of the MTTF so that the above described two-step procedure can not be applied or becomes iterative. Finally, Wu et al. [12] point out that this model does not account for time-based and state-induced events.

A model such as the above one could thus be used as an approximation only, for cases including both preemptive and non-preemptive interruptions provided that one includes in the effective process time the "inflation" effect of both breakdowns and preventive maintenance interruptions. However, the problem remains the estimation of c_e in case of IFR and non-preemptive interruptions. Therefore, in the following we refer to two sub-models. In all cases the expected cycle time is computed as $\mathrm{E(CT)} = \mathrm{E(QT)} + \mathrm{t_e}$.

57.3.1.1 Model Ia

This is the basic Whitt model (57.2) where c_e is simply assigned in a parametric manner. This avoids the need to explicitly compute its value based on the reliability characteristics of the machine and the timing of maintenance actions. The expected effective process time has a value which includes the natural process time, and the average downtimes occurring during the processing of the job owing to breakdowns and preventive maintenance. The computation of $\mathrm{t_e}$ is performed including the availability and considering the average number N_t of units processed during the time interval $\mathrm{T_P}$ between two consecutive preventive maintenance actions, obtaining

$$t_e = \frac{t_0}{A} + \frac{t_t}{N_t} = t_0 \left(\frac{1}{A} + \frac{t_t}{T_P\, A} \right). \tag{57.3}$$

The breakdown induced availability is $A = \frac{MTTF}{MTTF + MTTR}$ where MTTF is the average time to failure computed when a preventive maintenance with as good as new repair policy is adopted [15], which is

$$MTTF = \frac{\int_0^{T_P} R(t)dt}{1 - R(T_P)} \text{ and } R(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \text{ with } \lambda_B(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta - 1}$$

being $R(T_P)$ the reliability computed over the preventive maintenance interval assuming a Weibull distributed time to failure, and where $\lambda_B(t)$ is the failure rate function with parameters β an η [15]. Finally, the resource utilization is computed as $\rho = t_e/t_a$.

57.3.1.2 Model Ib

This is the Whitt model (57.2) where t_e is computed as shown for model Ia, while breakdown-related c_e is computed according to Table 57.1 but using the MTTF value computed as shown above with IFR and preventive maintenance. Given that systems with IFR are more predictable in their failure time, the coefficient of variation of uptime is lower than 1 [16] so that it is expected that computing the preemptive ${c_e}^2$ with a formula valid for constant failure rate overestimates rather than underestimates its value. Moreover, in practice the variability effect is often less relevant than the resource saturation effect. The final value of c_e including non-preemptive interruptions is again obtained following the above two-step procedure.

57.3.2 Model II

A model for M/G/1 non-preemptive priority queues with two priorities, by Adan and Resing [17], can be instead utilized as an approximation for time-based non preemptive interruptions cases, provided that average value of the process time is corrected to account for the effects of preemptive interruptions. Here the low priority jobs are the preventive maintenance interruptions. This model computes the expected cycle time CT as shown in Table 57.2. $\lambda_1 = 1/t_a$ and $\lambda_2 = 1/T_P$ are the

Table 57.2 Model II equations [17]

$E(CT) = E(QT) + \frac{\lambda_1}{\lambda_1 + \lambda_2}E(S_1) + \frac{\lambda_2}{\lambda_1 + \lambda_2}E(S_2)$	
$E(QT) = \frac{\lambda_1}{\lambda_1 + \lambda_2}E(QT_1) + \frac{\lambda_2}{\lambda_1 + \lambda_2}E(QT_2)$	$E(R_i) = \frac{E\left(S_i^2\right)}{2\,E(S_i)} \quad i = 1, 2$
$E(QT_1) = \frac{\rho_1\,E(R_1) + \rho_2\,E(R_2)}{1 - \rho_1}$	$E(QT_2) = \frac{\rho_1\,E(R_1) + \rho_2\,E(R_2)}{(1 - \rho_1 - \rho_2)(1 - \rho_1)}$

arrival rates of high and low priority jobs, and $\mu_1 = 1/t_e$ and $\mu_2 = 1/t_t$ their corresponding service rates with $\rho_1 = \lambda_1/\mu_1, \rho_2 = \lambda_2/\mu_2$. S_1 and S_2 are the effective process times for the two customers classes (i.e. the effective process time with preemptive interruptions only and the preventive maintenance downtime respectively), with $E(S_1) = t_e$, $E(S_2) = t_t$, where t_e, including the effects of breakdowns only is computed according to Table 57.1 preemptive case equation. Finally, from the definition of variance $E({S_i}^2) = \sigma(S_i)^2 + E(S_i)^2$. In case of the low priority jobs $\sigma(S_i)$ is assigned. For the high priority jobs, instead, it is computed by assuming a c_e value (this gives rise to Model IIa) or is computed from Table 57.1 equations for the preemptive interruptions case (Model IIb).

57.4 Analysis Results

In the following, the above approximate queueing models will be used as a means to describe the impact of preventive maintenance interval on the operational performance of a single machine, considering WIP as the performance measure of interest. WIP, in fact, apart from involving carrying costs, also implies space occupation on the shop floor (determining blocking phenomena when the buffer space is limited), and contributes to the increase of manufacturing lead time according to Little's law.

In the rest of the paper the following assumptions are made. A constant throughput is assumed, according to an imposed value of the average interarrival time of jobs to be processed which has an exponential distribution (Poisson arrivals), while the processing time area generally distributed. The adopted queueing models will then belong to the M/G/1 class ($c_a = 1$). The machine is assumed to have an increasing failure rate with time to failure modeled resorting to a Weibull distribution, while preventive maintenance is carried out at constant time intervals according to a periodic replacement policy. Maintenance is carried out under the "as good as new" assumption, i.e. after replacement the equipment failure rate is restored to the initial value. The average WIP is computed from Little's law as WIP $= \lambda$ E(CT) for Model I and WIP $= \lambda_1$ E(CT) for Model II.

It should be pointed out that this paper has not the goal of developing a new or exact queueing model for unreliable servers subject to both preemptive and non-preemptive interruptions, but rather to use some existing approximate queueing model to point out the following issues often neglected by maintenance planners:

- WIP and cycle time can be quite sensitive to the frequency of preventive maintenance actions.
- To choose a preventive maintenance interval based only on maintenance cost minimization or other criteria (production planning requirements, instructions from equipment manufacturers etc.) may have a negative impact on other operational performances of the manufacturing system such as WIP and cycle time, which can also bring an adverse effect on overall costs.
- The value of preventive maintenance interval which minimizes the maintenance cost can be quite different from the value which minimizes WIP, thus asking for a trade-off decision.

In this section, to provide an evidence of the above issues, some numerical results will be shown using the above approximate queueing models. We do not expect that any of the adopted model will provide precise numerical results, owing to the large number of approximations involved, but their combined utilization can give a measure of the effects of improperly selecting the preventive maintenance intervals, to guide maintenance planners. Improved queueing models or simulation studies will be required to obtain precise results. Only one numerical case is shown here owing to space limitations but a number of other numerical experiments were performed to confirm that it is representative of a typical system behavior. Assumed parameters values are Weibull parameters $\eta = 150$, $\beta = 2$, natural average process time $t_0 = 5$ min, standard deviation of natural process time $\sigma_0 = 1$ min, WIP holding cost $h = 10$ €/unit hour, mean preventive maintenance duration $t_t = 1$ h, standard deviation of preventive maintenance duration $\sigma_t = 0.16$ h, mean interarrival time $t_a = 6$ min, corrective maintenance MTTR $= 10$ h, standard deviation of corrective MTTR $\sigma_r = 2$ h.

Figure 57.1 shows the WIP trends computed resorting to the two adopted models when T_P changes from 20 to 500 h. Dashed curves, referring to Model Ia) in Fig. 57.1a and to Model IIa) in Fig. 57.1b are plotted for c_e values ranging from 0.6, to 3 with step 0.2. Solid curves refer to Model Ib) and IIb) respectively. While significant differences in the computed values occur, owing to the different approximations involved, all models show the same trend. In particular, all models strongly agree in the estimation of the T_P value giving the minimum WIP, which is the relevant information here. This value of T_P can be, in fact, compared to the value minimizing maintenance cost in order to determine whether trade-off situations occur. Nevertheless, the models are not able to precisely estimate the absolute value of WIP because in Models Ia and IIa a costant c_e is assumed irrespective of the T_P value, while c_e is actually dependent from T_P, and in models Ib and IIb the equation used to update c_e when T_P changes is valid for a constant failure rate instead of an IFR. Overall, as T_P is gradually increased, the WIP trend involves at first a rapid reduction until a minimum is reached, then an increase, followed by a final stabilization to an asymptotic value. In fact, apart from the variability term, WIP directly depends from the resource saturation which, in turn, is directly dependent on the value of the effective process time. The higher the resource saturation the

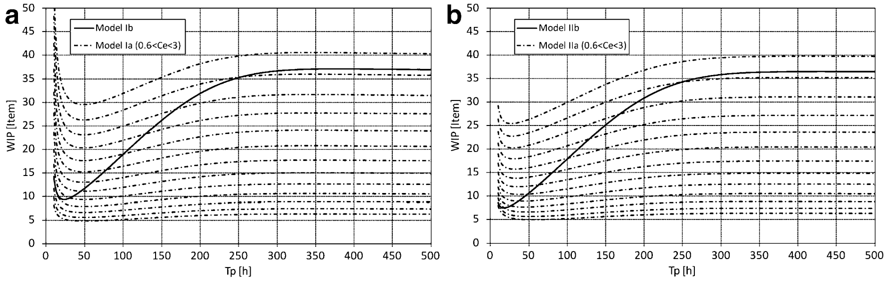


Fig. 57.1 WIP vs preventive maintenance interval (**a**) = Model I; (**b**) = Model II

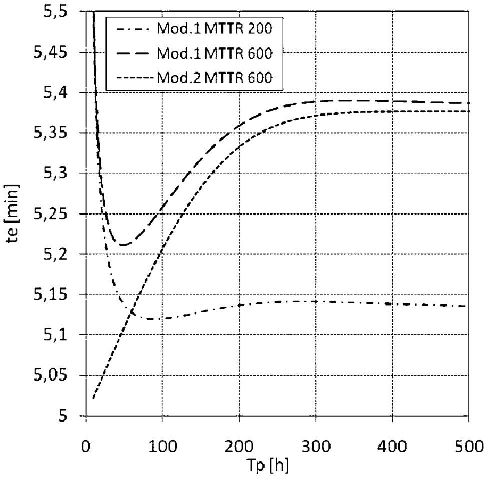


Fig. 57.2 Effective process time versus preventive maintenance interval

higher is the WIP sensitivity to changes in T_P (owing to high values of t_e respect t_a caused by high values of natural process time or preventive and corrective downtimes). As a consequence, the effect of the preventive MTTR is higher at low T_P values while that of corrective MTTR is higher at high T_P values. Simulations also demonstrated that the WIP sensitivity to T_P is enhanced the higher is the value of β and the lower is η. Figure 57.2 then shows the variation of effective process time. In the Model I plot for MTTR $= 600\,\text{min}$, we observe three distinct zones. In the first zone, for small values of T_P, the value of t_e is high, but rapidly decreasing, because preventive maintenance downtime is distributed over a comparatively small number of pieces given the high frequency of preventive maintenance. Contribution of corrective maintenance downtime is instead negligible as the frequent preventive maintenance keeps the system failure rate low and few breakdowns are expected given that MTTF $\gg T_P$.

In the third zone, for high values of T_P, the contribution of preventive downtime becomes negligible, given that a high number of pieces are processed between preventive interruptions. The corrective downtime contribution stabilizes, instead, to the value corresponding to distributing the corrective downtime over the average number of pieces processed between two consecutive breakdowns (for $T_P >$ MTTF the frequency of preventive maintenance only has a negligible effect on the failure rate and average time to failure). In the intermediate zone, a minimum of t_e, and thus WIP, may occur or not depending on the trade-off between the competing effects of reducing preventive interruptions and increasing corrective interruptions. Here T_P and MTTF become comparable and changes in T_P strongly affect MTTF variations and resource availability. In the Model II curve only the effect of preemptive interruptions is included, being the non-preemptive interruptions treated separately. In that curve the asymptotically stabilizing contribution of breakdowns is evident. Finally, the third curve shows the trend of t_e computed with Model I but when corrective downtime is decreased from 600 to 200 min, pointing out that

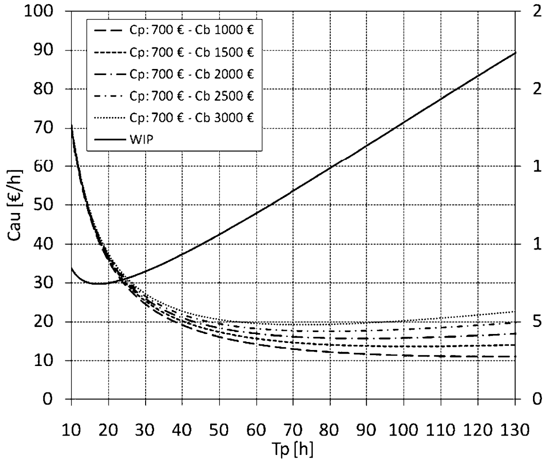


Fig. 57.3 Trends of WIP and maintenance costs

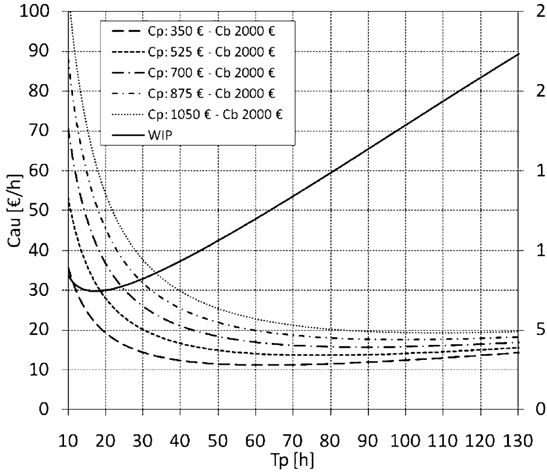


Fig. 57.4 WIP and maintenance costs

the above described trade-off between corrective and preventive downtime does not necessarily give rise to a minimum value of t_e. Figures 57.3 and 57.4, show instead the WIP curve of Model IIb, which has a minimum representative also of the minimum of WIP holding cost, superimposed to the total maintenance cost per unit time $C_{AU}(T_P)$ curves computed according to Eq. (57.1) for various C_P and C_B values. Curves are plotted for T_P values <MTBF only, because with IFR the breakdowns mostly occur in close proximity to the MTBF and it would be useless to plan a preventive maintenance after a breakdown, and the corresponding system renovation, has most likely occurred.

This shows that while the actual maintenance costs define a strongly variable optimal preventive maintenance interval, this interval is obviously not related to the interval minimizing WIP, and significant percent increases of WIP may occur

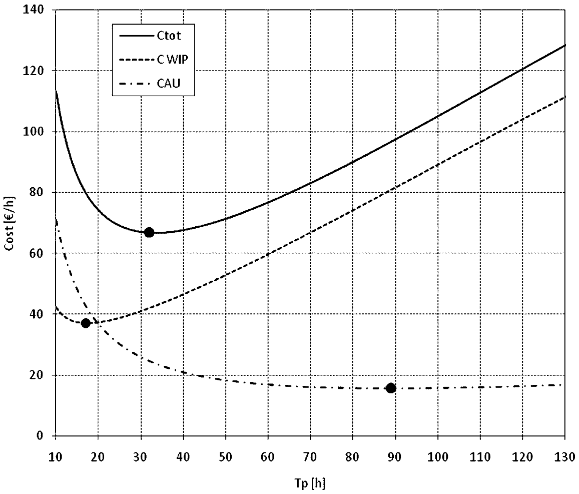


Fig. 57.5 WIP, maintenance and total cost

when T_P is set at the minimum maintenance cost respect the minimum WIP value. Therefore, if we sum the overall maintenance cost $C_{AU}(T_P)$, to the WIP holding cost, $C_{WIP}(T_P) = h$ WIP, where h (€/unit hour) is the unit WIP carrying cost, we can compute a total cost. A sample case is shown in Fig. 57.5. The corresponding T_P value minimizing this total cost can be another option for planning a maintenance policy.

57.5 Conclusion

Approximate queueing models can help in assessing the effects of preventive maintenance frequency on the performances of manufacturing systems characterized by an increased failure rate and undergoing both preemptive an non preemptive interruptions. Obtained results shows that WIP can be very sensitive to the preventive maintenance interval, and that the minimum cost maintenance interval is not related to a minimum WIP. Presented models allow, therefore, to choose the maintenance interval giving the minimum total cost or can help maintenance planners to make more informed decision related to costs-performances trade offs.

References

1. Ben-Daya, M., Duffuaa, S.O., Raouf, A.: Maintenance, Modeling and Optimization. Springer, Berlin (2000)
2. Nakagawa, T.: Maintenance Theory of Reliability. Springer, Berlin (2005)
3. Wang, H.: A survey of maintenance policies of deteriorating systems. Euro. J. Oper. Res. **139**, 469–489 (2002)

4. Gupta, D., Günalay, Y., Srinivasan, M.M.: The relationship between preventive maintenance and manufacturing systems performance. Euro. J. Oper. Res. **132**, 146–162 (2001)
5. Iravani, S.M.R., Duenyas, I.: Integrated maintenance and production control of a deteriorating production system. IIE Trans. **34**, 423–435 (2002)
6. Ke, J.C.: The optimal control of an M/G/1 queueing system with server vacations, startup and breakdowns. Comput. Indus. Eng. **44**, 567–579 (2003)
7. Perry, D., Posner, M.J.M.: A correlated M/G/1-type queue with randomized server repair and maintenance modes. Oper. Res. Lett. **26**, 137–147 (2000)
8. Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M.: Queueing Theory. Wiley, New York (2008)
9. Avi-Itzhak, B., Naor, P.: Some queueing problems with service station subject to server breakdown. Oper. Res. **10**, 303–320 (1962)
10. Federgruen, A., Green, L.: Queueing systems with service interruptions. Oper. Res. **34**(5), 752–768 (1986)
11. White, H.C., Christie, L.S.: Queueing with preemptive priorities or with breakdowns. Oper. Res. **6**, 79–95 (1958)
12. Wu, K., McGinnis, L.F., Zwart, B.: Queueing models for single machine manufacturing systems with interruptions. Proceedings of 2008 Winter Simulation Conference, pp. 2083–2092 (2008)
13. Whitt, W.: Approximations for the GI/G/m queue. Product. Oper. Manag. **2**, :114–161 (1993)
14. Hopp, W.J., Spearman, M.L.: Factory Physics. McGraw–Hill, Boston, MA (2000)
15. Ebeling, C.E.: An Introduction to Reliability and Maintainability Engineering. McGraw-Hill, Boston, MA (1997)
16. Li, J., Meerkov, S.M.: On the coefficients of variation of uptime and downtime in manufacturing equipment. Mathemat. Problem. Eng. **1**, 1–6 (2005)
17. Adan, I., Resing, J.: Queueing Theory, Lecture Notes. http://www.win.tue.nl/~iadan/queueing.pdf.

Chapter 58
On the Numerical Prediction of Stability in Thin Wall Machining

Oluwamayokun B. Adetoro, Ranjan Vepa, Wei-Ming Sim, and P.H. Wen

Abstract In this chapter, the numerical prediction of stability margin in thin wall machining is introduced. The Nyquist criterion is applied to the stability model presented by Adetoro, while a newly discovered damping prediction approach is presented, which when applied to the FEM and Fourier approach presented by Adetoro, would allow the prediction of stability margins without the need for experimentally extracted damping parameters.

Keywords Nyquist criterion · FEA · frequency response function · transfer function · damping ratio · damping matrix

58.1 Introduction

In aerospace, the manufacturing process is progressively limiting the use of joints through the manufacturing of structures as one monolithic piece. Machining is a very common operation in manufacturing, due to its versatility and its high material removal rate in producing parts of desired dimensions. Aircraft wing sections, fuselage sections, turbine blades and jet engine compressors, are all typical parts with sections produced from machined aluminium or titanium blocks. With environmental concerns and the general demand for higher efficiency, weight requirements compel the design of much thinner sections. In order to maintain the quality of the machined parts there is usually a dimensional tolerance, which the machined parts

O.B. Adetoro (✉), R. Vepa, and P.H. Wen
School of Engineering and Materials Science, Queen Mary,
University of London, Mile End Road, London, E1 4NS
e-mail: o.adetoro@qmul.ac.uk; r.vepa@qmul.ac.uk; p.h.wen@qmul.ac.uk

W.-M. Sim
Manufacturing Engineering Technology/Optimum Processes, Airbus Operations Ltd.,
New Filton House, Filton, Bristol, BS99 7AR
e-mail: WeiMing.Sim@airbus.com

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_58,

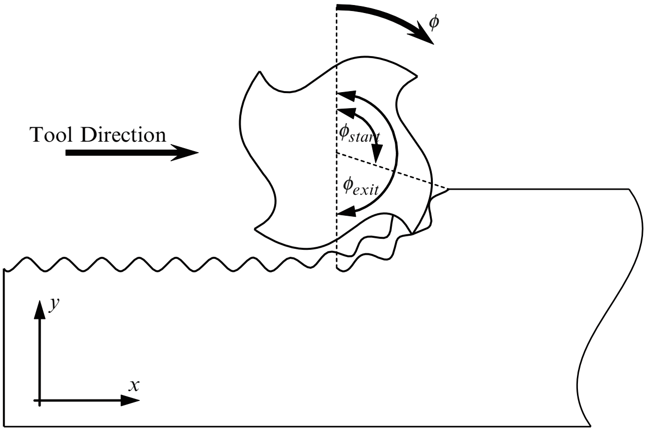


Fig. 58.1 Dynamic milling model

have to satisfy. To enforce this, it is a general practice for machined parts to undergo inspection before they are certified for use. While parts that fail this inspection are either scrapped or subjected to many hours of manual labour to remove the bad surface finish.

In milling the workpiece is fed past a rotating tool with one or more teeth (Fig. 58.1), which makes it possible to attain very high ‘Material Removal Rate’ (MRR). The tooth/teeth remove the material from the workpiece in the form of small individual ‘chips’. The study into the numerical simulations of the machining of thin-walled sections [1] is the focus of this chapter. In thin wall machining, the cutting conditions used are very important and must be chosen with care as they directly influence the cutting forces. The cutting forces cause structural vibrations in the workpiece, tool and spindle. These vibrations can be classified as free vibrations (occur after an external energy source is removed), forced vibrations (occur during the presence of an external energy source) and self-excited vibrations [2]. The self-excited vibration has its source from the inherent structural dynamics of the machine tool-workpiece and feedback responses through the undulations left on the machined surface. The optimum cutting case is when the undulations left on the machined surface are in phase with the undulations from previous tooth pass. The worst case however is when the phase angle between the two undulations is out of phase by 90°. This leads to the phenomena known as “regenerative chatter” or simply “chatter”. Chatter is usually characterised by a very bad surface finish and a drastic increase in both cutting forces and vibrations.

The prediction of stable conditions in the form of charts started when, Tobias [3] and Tlusty [4] simultaneously made the remarkable discovery that the main source of self-excited regenerative vibration/chatter was not related to the presence of negative process damping as was previously assumed. However, it is related to the structural dynamics of the machine tool-workpiece system and the feedback response between subsequent cuts. Their model is only applicable to orthogonal metal cutting where the directional dynamic milling coefficients are constant and not periodic. Other studies on the stability of orthogonal metal cutting were reported by Merritt [5].

Altintas and Budak [6–8] later proposed an analytic approach to predict stability margin. Perhaps, the first analytical approach in which the zeroth order term in the Fourier series expansion of the time varying coefficients was adopted. The analytical model was later extended to include three directions by Altintas [9], where the axial immersion angle was assumed to be constant. Except for flat end mills however, the axial immersion angle, is a function of the axial depth of cut. Campa et al. [10] later proposed an averaging approach to calculating the axial immersion angle in order to solve the stability model analytically. However, the axial immersion angle was still assumed to be a constant. This is the main analytical approach generally used in predicting stable cutting conditions in machining [11, 12]. The model has recently been improved by Adetoro et al. [13, 14] to include the nonlinearity of the cutting force coefficients, axial immersion angle and system dynamics.

The accuracy of the predicted stable region relies on the dynamic parameters identified at the cutter-workpiece contact zone. The cutter and workpiece dynamics consist of its damping, stiffness and mass parameters. Damping is the dissipative factor present in every real-life system/structure. Unlike the well developed mass/inertia and stiffness forces, the damping forces are at present extracted through experiments known as modal testing/analysis. This is because the physics behind the damping forces are not fully understood especially for a wide range of systems. It is however always desirable for an analyst to be able to predict the damping ratio (either analytically or numerically) for any given geometry without having to rely solely on experimental results.

A significant contribution at the early development of modal analysis was proportional damping model. It was first proposed by Lord Rayleigh in 1878, where he indicated that if the viscous damping matrix is proportional to mass and stiffness matrices (the damping forces are proportional to the kinetic and potential energies of the system) then it can be expressed [15] as,

$$\mathbf{C} = \alpha_0\mathbf{M} + \alpha_1\mathbf{K}, \tag{58.1}$$

where α_0 and α_1 are real positive constants. The model is termed 'Rayleigh damping' or 'classical damping'. The significance of this model is that the damped system would have the same mode shapes compared to its undamped counterpart, thus the system is said to possess 'classical normal modes'.

The equation of motion for a multi degree of freedom (MDoF) system can be expressed as,

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{F}, \tag{58.2}$$

where $\mathbf{M}$ is the mass matrix, $\mathbf{C}$ is the viscous damping matrix, $\mathbf{K}$ is the stiffness matrix, $\ddot{\mathbf{x}}$, $\dot{\mathbf{x}}$, $\mathbf{x}$ and $\mathbf{F}$ are the acceleration, velocity, displacement and excitation force vectors respectively.

In 1960, Caughey and O'Kelly [16] provided a generalization of Rayleigh's condition for discrete systems in form of the series,

$$\mathbf{C} = \mathbf{M} \sum_{u=0}^{L-1} \alpha_u \left(\mathbf{M}^{-1}\mathbf{K}\right)^u, \tag{58.3}$$

where L is the number of identified modes used in the curve fitting, α_u are real positive constants obtained through using experimentally identified damping parameters. The Rayleigh damping model is the first two series of the expansion.

Thus this chapter is structured as follows; the Nyquist Criterion is applied to the dynamic milling model; a newly discovered approach to predicting the structural damping parameters is presented.

58.2 Application of the Nyquist Criterion

The dynamic milling model is explained in the paper by Budak and Altintas [6] and summarised by Adetoro et al. [1, 17]; while a new 3-D mode was developed in the paper by Altintas [9], with improvements in the paper by Adetoro et al. [13]. For purposes of applying the Nyquist stability criterion the dynamic milling model may be re-stated in terms of the dynamic forces acting the machine tool in the form,

$$\mathbf{F}(t) = (1/2)aK_t\mathbf{AG}(D)\left(\mathbf{F}(t) - \mathbf{F}(t-T)\right) + \mathbf{F}_c(t), \tag{58.4}$$

or as,

$$\mathbf{F}(t) = (1/2)\,aK_t\mathbf{AG}(D)\left(\mathbf{I} - \exp\left(-DT\right)\right)\mathbf{F}(t) + \mathbf{F}_c(t), \tag{58.5}$$

where $\mathbf{F}(t)$ is the dynamic force vector acting on the cutting tool, a is the axial depth of cut, K_t is the tangential cutting force coefficient, $\mathbf{A}(t)$, is the immersion dependent directional cutting force coefficient matrix, which could in general be a periodic function of time satisfying the condition, $\mathbf{A}(t+T) = \mathbf{A}(t)$, $\mathbf{G}(D)$ is the direction dependent and frequency dependent transfer function relating the static and dynamic cutting force vector, T is the inter-tooth time of passage, $\mathbf{F}_c(t)$ is the control force vector acting on the cutting tool. Equation (58.5) is expressed in terms of the force vector. Further information on the formulation of the force equations for a generic case, including the application of Floquet theory to machine tool chatter may be found in the paper by Minis and Yanushevsky [18].

In many applications to machine tool chatter the immersion dependent directional cutting force coefficient matrix $\mathbf{A}$ may be approximated by a constant coefficient matrix. The problem can be reformulated as a closed loop control problem and the control law defining the control force vector acting on the cutting tool, $\mathbf{F}_c\,(t)$ may be synthesized using standard techniques of control law synthesis. Thus if the control law takes the form,

$$\mathbf{F}_c(t) = -\mathbf{K}_{cl}(D)\mathbf{F}(t), \tag{58.6}$$

and it follows that,

$$\mathbf{F}(t) = (1/2)aK_t(\mathbf{I} + \mathbf{K}_{cl}(D))^{-1}\mathbf{A}\mathbf{G}(D)\,(1 - \exp(-DT))\,\mathbf{F}(t) \tag{58.7}$$

Equation (58.7) may be expressed as,

$$\begin{aligned}&\left(\mathbf{I} - (1/2)aK_t(\mathbf{I} + \mathbf{K}_{cl}(D))^{-1}\mathbf{A}\mathbf{G}(D)\,(1 - \exp(-DT))\right)\mathbf{F}(t)\\ &\quad\equiv (\mathbf{I} + \mathbf{\Phi}(D))\mathbf{F(t)} = 0\end{aligned} \tag{58.8}$$

The conditions for stability may now be stated in terms of the Nyquist stability criterion. The return difference equation is given by,

$$[\mathbf{I} + \mathbf{GA}] = (\mathbf{I} + \mathbf{\Phi}(D)) \tag{58.9}$$

Thus the multi-input multi-output Nyquist plot may be obtained by computing the eigenvalues, λ, defined by,

$$[\mathbf{\Phi}(D) + \lambda\mathbf{I}]\,\mathbf{x} = \mathbf{0} \tag{58.10}$$

As $\mathbf{\Phi}(D)$ is a complex matrix, the above is a complex eigenvalue problem. To obtain the Nyquist plot corresponding to an eigenvalue it is plotted on the complex plane as ω traverses the Nyquist contour. In the classical Nyquist plot the gain crossover point (i.e. when the gain exceeds unity in magnitude) and phase crossover point (i.e. when phase increases from less than 180° to greater than 180°) are critical in the assessment of relative stability. The gain margin of stability is measured at the phase crossover point and the phase margin at the gain crossover point. The gain margin is the additional gain factor in dB that force the plot to pass through crossover point while the phase margin is the additional phase which when added to the phase would force the phase plot to pass through the phase crossover point. Moreover when the gain crossover and the phase crossover points are relatively close to each other, it signifies that a point of neutral stability is in the vicinity of the gain crossover point.

58.3 System's Transfer Function

The system's transfer function matrix, $\mathbf{G}(D)$ in Eq. (58.4) is required in order to predict the system's stability margin. It is therefore defined for the cutter and the workpiece as,

$$\mathbf{G}(D) = \mathbf{G}_c(D) + \mathbf{G}_w(D), \tag{58.11}$$

Where for a 2-D system,

$$\mathbf{G}_\beta(D) = \begin{bmatrix} G_{\beta xx}(D) & G_{\beta xy}(D) \\ G_{\beta yx}(D) & G_{\beta yy}(D) \end{bmatrix}, \quad (\beta = c, w), \quad (D = i\omega_c) \tag{58.12}$$

When considering only the tool's dynamics, the cutter's transfer function $\mathbf{G}_c(D)$ can be assumed to be constant and generally extracted experimentally, while the workpiece's transfer function matrix $\mathbf{G}_w(i\omega_c)$ is simply ignored. In the case of thin wall machining however, the workpiece dynamics cannot be ignored and the current adopted experimental methods would be inefficient as the dynamics are not constant along the thin wall as shown by Adetoro et al. [14]. An FEM and Fourier approach to extracting the system's transfer function was presented by Adetoro et al. [1, 17]. The main drawback of this approach is that it requires the damping parameters of the structure, therefore a new approach to predicting the damping parameters is presented in the next section.

58.3.1 Damping Ratio Prediction

An approach to predicting the damping ratio was discovered by Adetoro et al. [19, 20], which can be used to predict the damping parameters used in the FEM simulations. In many practical situations the use of the Nyquist stability criterion is unnecessary if one can extract the damping ratio of all the vibration modes relatively quickly. The approach proposed by Adetoro et al. [19] is a quick, simple and yet significantly accurate approach to predicting the damping ratio in terms of the frequency for a given wall; based on the use of the known damping ratios of a wall with same height (provided only the wall thickness is changed). From a range of extracted structural dynamics, it was discovered that there was a certain trend between the different damping ratios for different wall thicknesses. It was found that a new set of parameters, $\overline{\zeta}_p$ and $\overline{\omega}_p$, can be defined as follows,

$$\overline{\zeta}_p = \frac{\zeta_p^a}{t_a}, \tag{58.13}$$

for damping ratio and

$$\overline{\omega}_p = \frac{\omega_{np}^a}{t_a}, \tag{58.14}$$

for natural frequency, where t_a is the reference current wall thickness, ζ_p^a is the modal damping ratio and ω_p^a is the natural frequency for the reference wall respectively. These parameters ($\overline{\zeta}_p$ and $\overline{\omega}_p$) are then used to predict the damping ratio, ζ_p^b in terms of frequency, ω_p^b for any new geometry (provided only the wall thickness is changed) by simply multiplying $\overline{\zeta}_p$ and $\overline{\omega}_p$ by the new wall thickness t_b as follows,

$$\zeta_p^b = \overline{\zeta}_p \cdot t_b, \tag{58.15}$$

$$\omega_p^b = \overline{\omega}_p \cdot t_b \tag{58.16}$$

It should be noted that ζ_p^b and ω_p^b are not necessarily the precise modal damping and natural frequencies of the new wall. Studying the series in Eq. (58.3) propose by Caughey [16], the zeroth order approximation gives,

$$\mathbf{C}_0 = \alpha_0 \mathbf{M}, \tag{58.17}$$

which is not realistic as there the stiffness term has to always exist in whatever level of approximation, hence in an attempt to preserve the stiffness and mass terms a new series is proposed, which is defined as,

$$\zeta_p = \frac{1}{2} \sum_{u=1}^{L/2} \left(\alpha_{2u-1} \cdot \omega_n^{-u} + \alpha_{2u} \cdot \omega_n^{u} \right), \tag{58.18}$$

where the first term expands out to,

$$\zeta = \frac{1}{2} \left(\frac{\alpha_1}{\omega_n} + \alpha_2 \omega_n \right) \tag{58.19}$$

Therefore, Eq. (58.19) can be written for the first series as,

$$C_p = \alpha_1 M_p + \alpha_2 K_p, \tag{58.20}$$

which, shows that both mass and stiffness retained in the first series. Expanding the proposed series gives,

$$C_p = \alpha_1 M_p + \alpha_2 K_p + \alpha_3 M_p^{1.5} + \alpha_4 K_p^{1.5} + \alpha_5 M_p^2 + \alpha_6 K_p^2 + \cdots \tag{58.21}$$

Therefore by dividing Eq. (58.19) through by the wall thickness t_a, as done in Eqs. (58.13) and (58.14) we obtain the following series expansion,

$$\overline{\zeta} = \frac{1}{2} \sum_{u=1}^{L/2} (\alpha_{2u-1} \cdot \overline{\omega}^{-u} + \alpha_{2u} \cdot \overline{\omega}^{u}), \tag{58.22}$$

where constants α_{2u-1} and α_{2u} are real constants obtained using least squares method. This series expands out in the form,

$$\overline{\zeta} = \frac{1}{2} \left(\frac{\alpha_1}{\overline{\omega}} + \alpha_2 \overline{\omega} + \frac{\alpha_3}{\overline{\omega}^2} + \alpha_4 \overline{\omega}^2 + \frac{\alpha_5}{\overline{\omega}^3} + \alpha_6 \overline{\omega}^3 + \cdots \right) \tag{58.23}$$

Therefore, by dividing the different numerically extracted natural frequencies (refer to the FEM approach in the paper by Adetoro et al. [1,17]) for each identified mode for the new wall thickness by t_b to obtain $\overline{\omega}$ in Eq. (58.22) and then multiplying the calculated $\overline{\zeta}$ by t_b, the corresponding damping for that mode is obtained. This significance of this new damping modelling approach is its application in thin wall machining, as the workpiece thickness reduces and its damping parameters change. Several case studies are presented in the paper by Adetoro et al. [19,20].

58.3.2 Damping Matrix

The damping ratio, ζ_p^b in terms of frequency can be readily used directly by most commercial Finite Element (FE) packages, however the damping matrix **C** in Eq. (58.2) is sometimes required. To obtain the damping matrix, the numerically extracted natural frequency of the new structure for each mode is divided by the wall's thickness t_b, to obtain $\overline{\omega}$ and used in Eq. (58.22), to calculate $\overline{\zeta}$, which is then multiplied back by the wall's thickness t_b, to obtain the new modal damping ratio for the corresponding mode.

The modal damping C_p is simply calculated in a similar fashion as in SDoF. The damping matrix **C** is finally obtained by pre-multiplying by the modal matrix and then post-multiplying by the transpose of the modal matrix or eigenvectors obtained in FEM simulations. This orthogonal property only applies to systems that possess classical normal modes or proportional damping.

58.3.3 Examples

To demonstrate the new damping prediction approach, the FEM approach presented by Adetoro [1] was used. The workpiece material used in the FEM model is "Aluminium Alloy 7010 T7651" and the properties are: Density, $\rho = 2.823 \times 10^3$ (kg m^{-3}), Young's Modulus, $E = 69.809$ (GPa) and Poisson Ratio, $\nu = 0.337$. Two different examples taken from the paper by Adetoro [20] are shown here. The dimensions are given in the paper by Adetoro [20] and the corresponding experimentally identified damping parameters. The experimentally extracted damping parameters for the reference wall are used to predict the damping parameters, ζ_p^b for other structures using the approach presented in previous section. The damping parameters predicted, ζ_p^b and the force data, $f(t)$ measured by the instrumented hammer (in time domain) during experimental impact tests were used in each corresponding FE analysis.

During the experimental impact test, the workpiece was bolted at the back surface to a milling machine, hence in the FEM simulations it was assumed to be perfectly clamped (characterised by stiffness values of 1×10^{36} for the corresponding degrees of freedom) and that the resonant frequency of the machine is much higher than the excited frequencies during impact tests.

Two FEM simulations were carried out for each structure; one using the experimentally extracted damping parameters and the second using the predicted damping parameters. The comparison between the two (Figs. 58.2 and 58.3) shows the accuracy of the new approach to predicting damping parameters. The FEM simulations are also compared with experimentally measured accelerations to depict the accuracy of the FEM simulations.

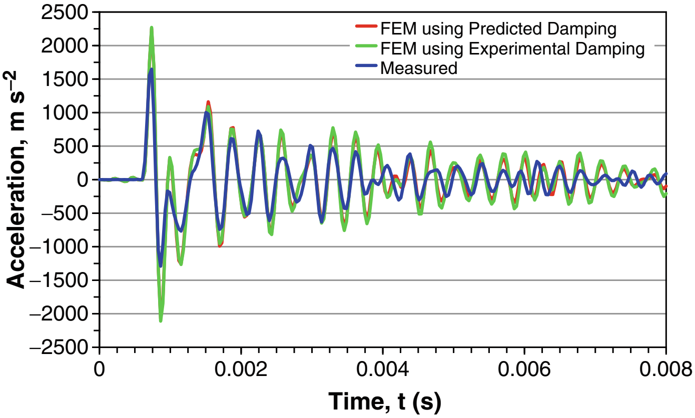


Fig. 58.2 Example 1: Wall 30 height, 3.0 thickness

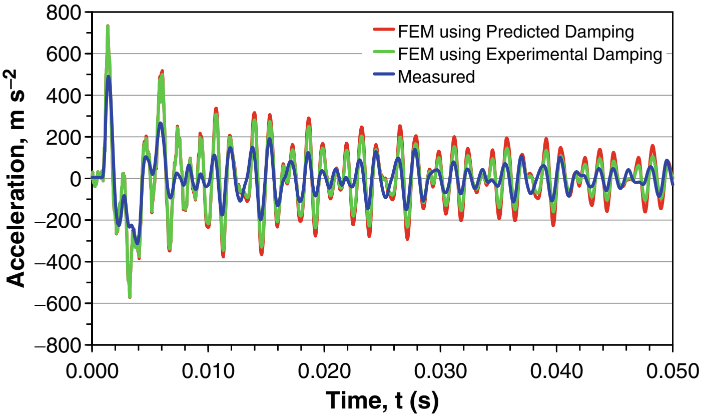


Fig. 58.3 Example 2: Wall 70 height, 3.5 thickness

References

1. Adetoro, O.B., Wen, P.H., Sim, W.M., Vepa, R.: Stability lobes prediction in thin wall machining. Proceedings of the International Multi-Conference of Engineers and Computer Scientist **1**, 520–525 (2009). www.iaeng.org/publication/WCE2009_pp20–525.pdf
2. Huo, D., Cheng, K.: Fundamentals, Applications and Practices, Springer Series in Advanced Manufacturing, pp. 7–20 (2009)
3. Tobias, S.A., Fishwick, W.: Theory of regenerative machine tool chatter. Eng. London **205**, 109–203 (1958)
4. Tlusty, J., Polacek, M.: The stability of machine tools against self excited vibrations in machining. Int. Res. Product. Eng. **41**, 465–474 (1963)
5. Merritt, H.E.: Theory of self-excited machine tool chatter. Trans. ASME J. Eng. Indus. **87**, 447–454 (1965)
6. Budak, E., Altintas, Y.: Analytical prediction of stability lobes in milling. CIRP Annal. Manufact. Technol. **44**(1), 357–362 (1995)
7. Budak, E., Altintas, Y.: Analytical prediction of chatter stability in milling – Part I: General formulation. Trans. ASME – J. Dynam. Syst. Measure. Contr. **120**, 22–30 (1998)
8. Budak, E., Altintas, Y.: Analytical prediction of chatter stability in milling – Part II: Application of the general formulation to common milling systems. Trans. ASME J. Dynam. Syst. Measure. Contr. **120**, 31–36 (1998)
9. Altintas, Y.: Analytical prediction of three dimensional chatter stability in milling. JSME Int. J. Ser. C: Mechan. Syst. Mach. Element. Manufact. **44**(3), 717–723 (2001)
10. Campa, F.J., Lopez de Lacalle, L.N., Lamikiz, A., Sanchez, J.A.: Selection of cutting conditions for a stable milling of flexible parts with bull-nose end mills. J. Mater. Process. Technol. **191**(1–3), 279–282 (2007)
11. Bravo, U., Altuzarra, O., Lopez de Lacalle, L.N., et al.: Stability limits of milling considering the flexibility of the workpiece and the machine. Int. J. Machine Tools Manufact. **45**(15), 1669–1680 (2005)
12. Lacerda, H.B., Lima, V.T.: Evaluation of cutting forces and prediction of chatter vibrations in milling. J. Brazil. Soc. Mechan. Sci. Eng. **26**(1), 74–81 (2004)
13. Adetoro, O.B., Sim, W.M., Wen, P.H.: Stability lobes prediction for corner radius end mill using nonlinear cutting force coefficients. Machin. Sci. Technol. (2009, submitted for publication)
14. Adetoro, O.B., Sim, W.M., Wen, P.H.: Accurate prediction of stability lobes using nonlinear thin wall dynamics. J. Mater. Process. Technol. (2009, submitted for publication)
15. Rayleigh, L.: Theory of Sound, vol. 2. Macmillan, New York (1878) (Reprinted 1945 by Dover Publications, New York)
16. Caughey, T.K., O'Kelly, M.E.J.: Classical normal modes in damped linear systems. Trans. ASME J. Appl. Mechan. **27**:269–271 (1960)
17. Adetoro, O.B., Wen, P.H., Sim, W.M., Vepa, R.: numerical and experimental investigation for stability lobes prediction in thin wall machining. Engineering Letters, **17**(4) (2009)
18. Minis, I., Yanushevsky, T.: A new theoretical approach for the prediction of machine tool chatter in milling. Trans. ASME J. Eng. Indus. **115**, 1–8 (1993)
19. Adetoro, O.B., Sim, W.M., Vepa, R., Wen, P.H.: Numerical and experimental investigation of damping parameters for thin wall structures. J. Advance. Manufact. Syst. (2009, submitted for publication)
20. Adetoro, O.B., Sim, W.M., Vepa, R., Wen, P.H.: A new damping modelling approach and its application in thin wall machining. Int. J. Advan. Manufact. Technol. (2009, submitted for publication)

Chapter 59
Risk Analysis of ERP Projects in the Manufacturing SMES: Case Study

Päivi Iskanius

Abstract This chapter introduces two risk management tools targeted for SMEs in their Enterprise Resource Planning (ERP) adoption projects. The purpose is to identify and assess the main risks in the ERP projects through the case study of two manufacturing SMEs. By using company-specific risk analysis method (RAM), the critical risks of the ERP projects are identified and assessed. Then, by using characteristics analysis method (CAM), the recommendations of how to divide the ERP projects into manageable sub projects are given.

Keywords Enterprise resource planning · ERP · SME · risk analysis method · characteristics analysis method

59.1 Introduction

The business environment is dramatically changing. Enterprises today face the challenges of globalization, international competition, technological complexity, and increasing customer orientation. To fulfil market demands, companies have to increase product portfolio, reduce time-to-market, shorten product-life cycles, and produce high quality products with quick response, lower costs, and greater customization [1]. Collaboration has become a common trend and success factor in today's business and industry practice and companies focus on their core competences and collaborate with other organisations with complementary knowledge and resources [2, 3]. Companies that move closer to a completely collaborative model must improve their own business practices and procedures [4]. Companies must also share with their suppliers, distributors, and customers the critical in-house information they previous aggressively protected [5]. Also, functions within the company

P. Iskanius (✉)
University of Oulu, Thule Institute, P.O. Box. 7300, 90014 Oulu, Finland
e-mail: paivi.iskanius@oulu.fi

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_59,

must upgrade their capability to generate and communicate timely and accurate information [4]. To accomplish these objectives, companies are increasingly turning to Enterprise Resource planning (ERP) systems.

ERP systems, when successfully implemented, links all functions of an enterprise including order management, manufacturing, human resources, financial systems, and distribution with external suppliers and customers into a tightly integrated system with shared data and visibility [6]. ERP systems promise seamless integration of information flowing through an organization [7, 8]. They fulfill this promise by integrating information and information-based processes within and across the functional areas in an organization, and further, by enabling the integration of information beyond the organizational boundaries.

The effective implementation of such a system can bring about many benefits, beginning with the most general, such as cost reduction, productivity improvement, and quality improvement, but also customer service improvement, better resource management, improved decision-making and planning, and organizational empowerment [7]. Consequently, improvement of economic indicators is achievable, which finally leads to an increase in enterprise profitability [9].

Despite the significant benefits of ERP systems, the Statistics show that only 30% of previous ERP implementations have been successful [10]. Many ERP implementations are difficult, lengthy and over budget, are terminated before completion, or failed to achieve their business objectives even a year after implementation [7, 11]. To achieve the desired benefits, the ERP project must be carefully managed [12–14] and the risks involved the projects must be properly assessed [15, 16]. Management of an ERP project is a challenging task for any company [7], but especially challenging it is for SMEs, which have sufficient resources, capabilities and ERP project experience. Even with significant investments in time and resources, there is no guarantee of a successful outcome [17].

Several standardized tools, methods and techniques are developed to help enterprises to better manage their IT projects, though they are often too general for ERP applications [18]. Also, consulting, project management, change management and risk management methods are normally specified for large enterprises [19]. The needs, operating requirements, logistics fulfillment and financial capabilities of SMEs are vastly different from that of large enterprises. In order to support SMEs in their ERP project, targeted risk management processes are needed in this context.

This chapter introduces two risk management tools targeted for SMEs in their ERP adoption projects. The purpose is to identify and assess the main risks in the ERP projects through the case study of two manufacturing SMEs. By using company-specific risk analysis method (RAM), the critical risks of the ERP projects have been identified and assessed. Then, by using characteristics analysis method (CAM), the recommendations of how to divide the ERP projects into manageable sub projects have been given.

59.2 Risks in ERP Systems

59.2.1 Main Characteristics of ERP Projects

There is a substantial difference between an ERP project and a simple software project [7, 18]. Most software projects focus on developing a software system. But an ERP project consists of tightly linked interdependencies of software systems, business processes, and process reengineering [16]. A major difference between ERP projects and traditional IT projects comes from the integrated nature of ERP software applications. The implementation of an ERP software package involves a mix of business process change and software configuration to align the software with the business processes. In an ERP implementation, the key focus has shifted from a heavy emphasis on technical analysis and programming towards business process design, business-focused software configuration, and legacy data clean-up [20].

ERP project can also be viewed as an organizational change project, due to the large number of changes it brings to an organization [12, 21]. Many enterprises install their ERP systems without fully understanding the implications for their business or the need for compatibility with overall organizational goals and strategies [22]. The result of this hasty approach is failed projects or weak systems whose logic conflicts with organizational goals. Usually enterprises also do not realise the full benefits that the ERP system offers because they are not organised in the correct fashion to achieve the benefits. Many companies that attempt to implement ERP system run into difficulty because the organisation is not ready for integration and the various departments within it have their own agendas and objectives that conflict with each other [23].

An ERP system as such seldom totally fits the existing business processes of an enterprise. In order to have efficient business processes with the new ERP system, an enterprise has either to change its business processes to fit the ERP system or modify the ERP system to fit its business processes [24]. For SMEs, a good fit between company business processes and the ERP system functionality is the most important selection criteria [25]. Further, SMEs, with their sufficient resources, have to focus on only the most critical business needs.

To implement an ERP system successfully, the way organizations do business will need to change and the ways people do their jobs will need to change too [13]. Almost half of ERP projects fail to achieve expected benefits because companies underestimate the efforts involved change management [20]. Thus, change management is essential for preparing an organization to the introduction of an ERP system, and its successful implementation.

59.2.2 Risk Factors

The enterprise-wide ERP projects represent a new type of management challenge. The management approaches for these projects may be altogether different from the managerial approaches for traditional IT projects [26]. An ERP project is a major and risky exercise for any size of enterprise, however, risks are higher for SMEs as the cost overruns during implementation may put financial strain on the firm and thus substantially impact firm performance [27]. Further, SMEs have less of a chance of recovering from a failed ERP implementation attempt than large enterprises [28].

The main reason for any IT project failure is that managers do not properly assess and manage the risks involved their projects [15]. Also, most project managers perceive risk management processes as extra work and expense, thus, risk management processes are often expunged if a project schedule slips. The main risk effects for SMEs can be summarized [18]: budged exceed, time exceed, project stop, poor business performances, inadequate system reliability and stability, low organisational process fitting, low user friendliness, low degree of integration and flexibility, low strategic goals fitting and bad financial/economic performances.

ERP risks can be classified in various ways (e.g. [16, 26, 29]). Popa-Nzaou et al. [30] identifies six main dimensions of risks in ERP implementation: organisational; business-related; technological; entrepreneurial; contractual; and financial risks. Organisationalrisk derives from the environment in which the system is adopted. Business-related risk derives from the enterprise's post-implementation models, artefacts, and processes with respect to their internal and external consistency. Technological risk is related to the information processing technologies required to operate the ERP system – for example the operating system, database management system, client/server technology and network. Entrepreneurial or managerial risk is related to the attitude of the owner-manager or management team, while contractual risk derives from relations with partners and financial risk from cashflow difficulties, resulting in an inability to pay license fees or upgrading costs, for example [30].

According to Sumner [26], ERP project-specific risks, in contrast to IT project risks are failure to redesign business projects, failure to follow enterprise-wide design that supports data integration, insufficient training and reskilling, insufficient internal expertise, lack of business analysts with business and technology knowledge, failure to mix internal and external expertise effectively, failure to adhere to standardized specifications which the software supports, lack of integration, and attempting to build bridges to legacy applications. Somers and Nelson [11] summarizes the critical success factors for ERP implementations, in which eight of the top ten are related to human factors: top management support, project team competence, interdepartmental cooperation, clear goals and objectives, project management, interdepartmental communication, management of expectations, and careful system selection. Finally, based on the previous research, Finally, Aloini et al. [18] summarizes the ERP risk factors: inadequate ERP selection, poor project team skills, low top management involvement, ineffective communication system, low key user involvement, inadequate training and instruction, complex architecture and high numbers of modules, inadequate business processes, bad managerial conduction, ineffective project management techniques, inadequate change management, inadequate legacy system management, ineffective consulting services experiences, poor leadership, inadequate IT system issues, inadequate IT system maintainability, inadequate IT supplier stability and performances, ineffective strategic thinking and planning, and inadequate financial management.

To minimize the risk of the ERP project, Markus and Tanis [8] recommend the application of a risk management plan at different ERP implementation project stages: selection, implementation, and usage. A planned and systematically adopted risk management procedure throughout the ERP project reduces the possibility to risks occurring. Consequently, according to Soja [9], major mistakes are made in the early stages of the ERP project, even prior to the implementation process. However, Kliem [31] emphasizes the efficiency of risk management when it is introduced at the earliest possible opportunity in the life cycle of the system in question, when planning issues are most important and the criteria for system selection are determined. Instead of using abovementioned ready-made risk lists, a company might consider identifying their own, company-specific ERP implementation risk list. These risks could be complemented by common risk lists, such as Sumner [26].

59.3 Risk Management Tools

59.3.1 Risk Analysis Method

Risk analysis method (RAM) identifies the most essential risks and their probability in the company context. In this study, the risk list has been formed based on the risk list of Vilpola [32]. The risk list is formed out of 63 questions or statements dealing with the ERP selection, implementation, and usage. The basic aim is to identify the ERP risks arising from the company's reality and therefore the employees from various levels of organisation have been interviewed and observed. The company-specific risk list has been filled in close interaction with company personnel. Risk assessment is done by evaluating each risk's probability and effect in a scale from one to five. The number one means very small probability and effect, and number five means high probability and catastrophic effect. Then, the risk multiplication as an indicator of risk significance has been used. It is calculated as multiplying the value of the probability by the value of the effect. Range of this value is from 1 to 21 [32].

59.3.2 Characteristics Analysis Method

Characteristics analysis method (CAM) ensures that an IT project is manageable and consistent by its different goals content and development approaches. The result of the CAM is a recommendation of how to divide a large and complex IT project, such as an ERP project, into manageable sub projects. Further, the CAM

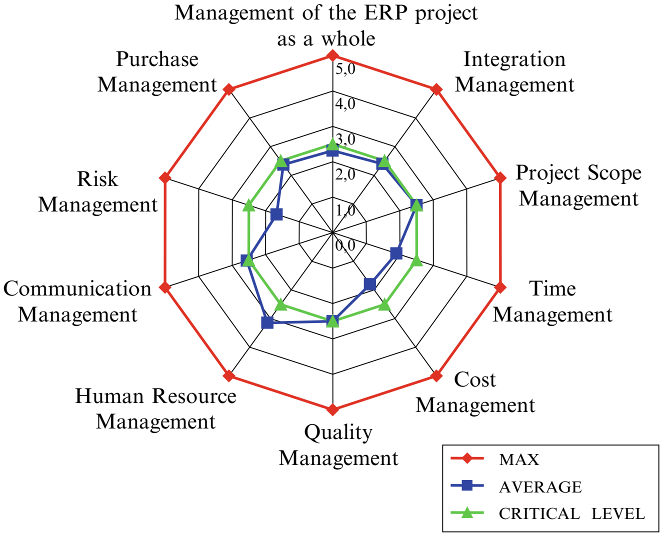


Fig. 59.1 CAM diagram of the company A

provides the project proposition documents, the knowledge and experience from prior development projects, and the requirements of the of the project portfolio [33].

In this paper, the CAM analysis is formed out of 90 questions dealing with the ERP project management factors. The basic aim is to find out the manageable size of the ERP project of the case companies. Also, CAM provides recommendations what management aspects should be put more attention to successfully manage the ERP projects (management of a project as a whole, management of integration, project scope management, time management, cost management, quality management, human resource management, management of communication, risk management, management of purchase). The questions are either positive or negative statements for which their applicability to the project will be evaluated (0 = fault, not true, 5 = exactly right; N/A = don't know). The tool has been implemented as an MS Excel worksheet with automatic tabulation based on decision rule sets. The result is can also be illustrated graphically (see Figs. 59.1 and 59.2) [33].

59.4 Case Study

This study has been carried out through the case study of two manufacturing SMEs. The case SMEs are in different phases of the ERP project. Company A is still contemplating the ERP implementation, Company B is in the selection phase, and Company C is already in the usage phase. In practice, this study has been carried out during January 1 to December 12, 2008.

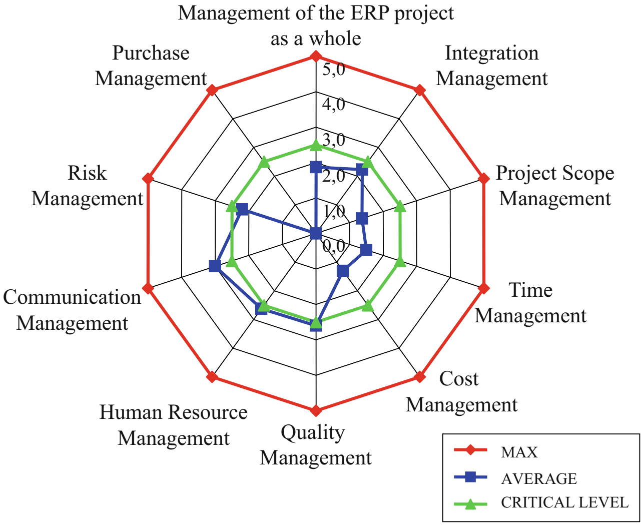


Fig. 59.2 CAM diagram of the company B

59.4.1 Company A

Company A develops blast cleaning technology and manufactures automated blast cleaning machines and robots (turnover about 1 2 M€ and number of personnel about 20). Company A has not an ERP system, but is contemplating the ERP implementation in near future. The need for the new ERP system has grown internally because of the problems in the current IT system. Today, the company is using an Excel-based IT system, which includes e.g. customer relationship management (CRM), product data management (PDM), purchase and order management, and product lifespan management. The problem of the current system is how to manage hundreds of different versions and variations of Excel, Word, and AutoCAD documents. The purpose of company A is to adopt an ERP system, which helps production capacity planning and control so that the scheduling and resource allocation for different projects can be planned in detail before the project is started. Furthermore, the new system should include warehouse and stock management functions and it have to support purchase process.

The risk list has been filled with the company key persons, and the effects and probability of risks have been assessed. In the *ERP selection phase*, the most critical risks are: misunderstanding between an ERP supplier and a company (12), an ERP system is not flexible enough (12), and the special company-specific ERP needs are not defined (10). In the *ERP installation phase*, the most critical risks are: a company's project manager is not a full time PM (20), data transfer from old to new ERP system is difficult (16), integration an ERP system with other IT systems creates problems (16), and an ERP supplier is not committed enough to the company's ERP system implementation (15). In the *ERP usage phase*, the most critical risks are: An ERP system does not support the company's business (12) and the ERP supplier does not develop the system in the future (10).

In the RAM results, in every phase (selection, implementation, usage), the crucial factors are depended on the decision of what ERP system and ERP supplier the company is choosing. The technical and functional factors related to an ERP system itself, and the factors related the system supplier, are considered the most critical. Even the company A has very few employees, under 20, the lack of resource, skills and expertise, and other factors related personnel have - surprisingly – not aroused as potential risks in this analysis.

According to CAM, 'Human resource management (HRM)' is the management/leadership field that clearly exceeds the critical level. Company A should direct a special attention to this factor in its ERP project management. In addition, several other management/leadership fields, such as 'Communications management', 'Purchase management', 'The project as a whole', 'Integration management', 'Project scope management' and 'Quality management' are right at the critical level. Only 'Cost management' and 'Time management' and 'Risk management' are clearly under critical level. According to CAM, risk factors related to personnel training and increasing personnel skills and knowledge require more from management, although they are not considered to be amongst the most potential risk factors according to RAM. On the basis of the CAM, it can be deduct that company A has a clear understanding of the costs caused by the ERP project, the time spent for it, as well as the technical and operational risks involved. The results analysed by CAM is presented in Fig. 59.1.

59.4.2 Company B

Company B provides demanding sheet metal work, welding, and heavy metal works, specialising in steel, paper, chemistry, and ship manufacturing related machinery and equipment. In addition, the company manufactures offshore equipment and ship propellers. The company B employs about 150 employees. Company B is in the selection phase of ERP project. The current IT systems are already in the end of their life cycle, and the company has to invest in a new ERP system. The company has interviewed several ERP suppliers. The company has made a preliminary requirement specification, a type of demand list, through which they have been able to limit the potential ERP suppliers into two options. Also, some IT consultants have worked for the company.

The risk list has been filled with company key persons, and the effects and probability of risks have been assessed. In the *ERP selection phase*, the most critical risks are: An ERP system will be a poor compromise for all stakeholders (12), selecting an improper project manager or project team, and misunderstandings between an ERP supplier and a company (10), and selecting an improper ERP system (10). In the *ERP installation phase*, the most critical risks are: normal business disturbs ERP project activities (20), ERP project disturbs normal business (16), an ERP project will be late (16), software configuration and testing don't function swiftly (16), and an ERP system not used in a disciplined manner (16). In the *ERP usage phase*, the most critical risks are: An ERP system not used in a disciplined manner (12), and only part of the ERP system is used (12).

In the RAM results, the crucial factors are mostly depended on the personnel (including project manager/team and top management level) behaviour, skills, and IT experience. Company B is also worried of the changes what the new ERP system will affect to the company's normal business, and in opposite, how the normal business hinders the ERP project progress.

According to CAM, 'Communications management' is the management/ leadership field that clearly exceeds the critical level. Company B should direct special attention to the factor considered people skills, knowledge and expertise. In addition, 'Human resource management' and 'Quality management' are right at the critical level. To manage ERP project successfully, the company should pay attention to these three management/leadership factors. The results analysed by CAM; is presented in the Fig. 59.2.

59.5 Conclusion

This study presents a case study of two SMEs which have analyzed their ERP project risks by the risk analysis method (RAM) and characteristics analysis method (CAM). The RAM presents crucial risks in a form and language that is understandable, because the analysis have been done in the company context. As negative aspect of RAM is that it requires a significant amount of resources and support of external experts. The CAM helps the case companies in dividing their ERP project into manageable entities and provides them recommendations on what leadership or management aspects they should devote special attention to. The CAM also shows inadequacies in the fields of management and leadership that the implementation of ERP system causes in companies.

This study has been done in deep cooperation with researchers and company key persons. Cooperation with the research group has provided companies extra skills and support in their ERP project endeavours. Company A is taking the next step in their ERP project and is in the extensive requirement specification process with the aim of selecting the suitable ERP solution for the company in 2010. Company B have made their decision on which ERP system they will select at the end stage of this study in 2008. Implementation will commence in 2009.

References

1. Canavesio, M.M., Martinez, E.: Enterprise modelling of a project-oriented fractal company for SME networking. Comput. Indus. **58**, 794–813 (2007)
2. Bititci, C.S., Martinex, V., Albores, P., Parung, J.: Creating and managing value in collaborative networks. Int. J. Phys. Distribut. Logistic. Manag. **34**(3/4), 251–268 (2004)

3. Zhou, Q., Ristic, M., Besant, C.B.: An information management architecture for production planning in a virtual enterprise. Int. J. Advan. Manufact. Technol. **16**, 909–916 (2000)
4. Umble, E.J., Haft, R.R., Umble, M.M.: Enterprise resource planning: Implementation procedures and critical success factors. Euro. J. Oper. Res. **146**, 241–257 (2003)
5. Loizos, C.: ERP: Is it the ultimate software solution. Industry Week **7**, 33 (1998)
6. Chen, I.J.: Planning for ERP systems: analysis and future trend. Business Process Manag. J. **7**(5), 1463–7154 (2001)
7. Davenport, T.H.: Putting the enterprise into the enterprise system. Harv. Business Rev. (July–August):121–131 (1998)
8. Markus, M.L., Tanis, C.: The enterprise systems experience – From adoption to success. In: Zmud, R.W. (ed.) Framing the Domains of it Management: Glimpsing the Future Through the Past, pp. 173–207, Pinnaflex, Cincinnati, OH (2001)
9. Soja, P.: Success factors in ERP systems implementations. Lessons from practice. J. Enterprise Inform. Manag. **19**(4), 418–433 (2006)
10. Standish Group International, Inc.. The Quarter Research Report (2004)
11. Somers, T.M., Nelson, K.G.: A taxonomy of players and activities across the ERP project life cycle. Inform. Manag. **41**, 257–278 (2004)
12. Bingi, P., Sharma, M.K., Godla, J.K.: Critical issues affecting an ERP implementation. Inform. Syst. Manag. **16**(3), 7–14 (1999)
13. Davenport, T.H.: Mission Critical: Realizing the Promise of Enterprise Systems. Harvard Business School Press, Boston, MA (2000)
14. Motwani, J., Mirchandani, D., Mandal, M., Gunasekaran, A.: Successful implementation of ERP projects: evidence from two case studies. Int. J. Product. Economic **75**, 83–96 (2002)
15. Markus, M.L., Tanis, C.: The enterprise systems experience – From adoption to success. In: Zmud, R.W. (ed.) Framing the Domains of it Management: Glimpsing the Future Through the Past, pp. 173–207, Pinnaflex, Cincinnati, OH (1998)
16. Wright, S., Wright, A.M.: Information system assurance for enterprise resource planning systems: implementation and unique risk considerations. J. Inform. Syst. **16**, 5–15 (2001)
17. Mabert, V.A., Soni, A., Venkataramanan, M.A.: Enterprise resource planning: Managing the implementation process. Euro. J. Oper. Res. **146**, 302–314 (2003)
18. Aloini, D., Dulmin, R., Mininno, V.: Risk management in ERP project introduction: Review of a literature. Inform. Manag **44**, 547–567 (2007)
19. Koh, S.C.L., Maguire, S.: Identifying the adoption of e-business and knowledge management within SMEs. J. Small Business Enterprise Dev. **11**(3), 338–348 (2004)
20. Al-Mudimight, A., Zairi, M., Al-Mashari, M.: ERP software implementation: an integrative framework. Euro. J. Inform. Syst. **10**, 216–226 (2001)
21. Hammer, M., Stanton, S.: How process enterprises really work. Harv. Business Rev. (November–December):108–118 (1999)
22. Wei, C.-C., Chien, C.-F., Wang, M.-J. J.: An AHP-based approach to ERP system selection. Int. J. Product. Economic. **96**, 47–62 (2005)
23. Langenwalter, G.A.: Enterprise Resource Planning and Beyond—Integrating Your Entire Organization. St. Lucie Press, Boca Raton, FL (2000)
24. Buonanno, G., Faverio, P., Pigni, F., Ravarini, A., Sciuto, D., Tagliavini, M.: Factors affecting ERP system adoption. A comparative analysis between SMEs and large companies. J. Enterprise Informat. Manag. **18**(4), 384–426 (2005)
25. Everdingen, Y., Hillegersberg, J., Waarts, E.: ERP adoption by European midsize companies. Commun. ACM **43**(4), 27–31 (2000)
26. Sumner, M.: Risk factors in enterprise-wide/ERP Projects. J. Inform. Technol. **15**, 317–327 (2000)
27. Cereola, S.J.: The Performance Effects of Latent Factors on Assimilation of Commercial Open-Source ERP Software on Small-Medium Enterprises. Virginia Commonwealth University, Richmond, VA (2006)
28. Muscatello, J., Small, M., Chen, I.: Implementing enterprise resource planning (ERP) systems in small and midsize manufacturing firms. Int. J. Oper. Product. Manag. **23**(7/8): 850–871 (2003)

29. Huang, S.M., Chang, I.C., Li, S.H., Lin, M.T.: Assessing risk in ERP projects: identify and prioritize the factors. Indus. Manag. Data Syst. **108**(8):681–688 (2004)
30. Poba-Nzaou, P., Raymond, L., Fabi, B.: Adoption and risk of ERP systems in manufacturing SMEs. A positivist case study. Business Process Manag. J. **14**(4):530–550 (2008)
31. Kliem, R.L.: Risk management for business process reengineering project. Inform. Syst. Manag. **17**(4):71–73 (2000)
32. Vilpola, I.: Applying use-centered design in ERP implementation requirements analysis, Tampere University of Technology, Publication 739, Tampere (2008)
33. Forselius, P.: Software development program characteristics http://citeseerx.ist.psu.edu/viewdoc/summary?doi = 10.1.1.131.1516. Accessed December 2008

Chapter 60
Sleeping in Sitting Posture Analysis of Economy Class Aircraft Passenger

CheeFai Tan, Wei Chen, Floris Kimman, and Matthias Rauterberg

Abstract With the rapid development of technology, the comfort of service has become an important issue. Air travels, especially long distance, may cause both physiological and psychological discomfort to passenger. Passenger comfort is clearly a main factor in user's acceptance of transportation systems. Sleeping is one of the common activities during the long haul flight. In this paper, subjective and objective measurement method was described to evaluate the sleeping in sitting posture of economy class aircraft seat passenger.

Keywords Aircraft seat · economy class · sleeping in sitting posture · subjective method · objective method

60.1 Introduction

Air travel is becoming increasingly more accessible to people both through the availability of cheap flights and because the airlines are now able to cater for individuals of all ages and disabilities. Health problems may arise due to anxiety and unfamiliarity with airport departure procedures prior to flying, whilst during the flight, problems may arise as a result of the food served on board, differences in the environmental conditions inside the cabin (pressure, ventilation, relative humidity, noise and vibration), the risk of cross-infection from fellow passengers, seat position, posture adopted and duration of the flight. These can be further compounded by changes in time zones and meal times, which may continue to affect an individual's health

C. Tan (✉)
Technical University Malaysia Melaka, Durian Tunggal, Melaka, Malaysia
e-mail: c.f.tan@tue.nl

W. Chen, F. Kimman, and M. Rauterberg
Department of Industrial Design, Technical University Eindhoven, Den Dolech 2, 5612AZ Eindhoven, The Netherlands
e-mail: w.chen@tue.nl; f..p.f.kimman@tue.nl; g.w.m.rauterberg@tue.nl

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*, Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_60,

long after arrival at the final destination [1]. Travel by air, especially long distance, is not a natural activity for human. Many people experience some degree of physiological and psychological discomfort and even stress during flying. Excessive stress may cause passenger to become aggressive, over-reaction, and even endanger the passenger's health [2, 3].

Comfort is an attribute that is highly demanded by today's passenger. The aircraft passenger comfort depends on different features and the environment during air travel. Seat comfort is a subjective issue because it is the customer who makes the final determination and customer evaluations are based on their opinions having experienced the seat [4]. The aircraft passenger seat has an important role to play in fulfilling the passenger comfort expectations. The seat is one of the important features of the vehicle and is the place where the passenger spends most of time during air travel. The aviation industry is highly competitive and therefore airlines try to maximize the number of seats [5]. Often this results in a very limited amount of seating space for passengers, especially in economy class [6]. In this paper, we described the subjective and objective measurement to analyze the sleeping in sitting posture of economy class aircraft seat passenger.

60.2 Aircraft Seat

Seat is one of the important elements for the passenger comfort. Different seat aspects have to be seen and taken into account in the comfort model. In charter and economy class the two least satisfactory characteristics are 'seat comfort' and 'leg room' [1].

The Civil Aviation Authority (CAA) is the regulatory body for the safety guidelines for aircraft seat spacing. The guidelines are set with safety, not comfort, in mind and relate to robustness of aircraft seats at the time of a crash and the ease of passenger evacuation in the event of an emergency [1]. There are three kinds of seat position in the aircrafts, such as window, aisle and isolated. For passengers seated in the central position of three or more seat row, the feeling of being surrounded is one of the worst aspects of economy air travel.

InNova [7] created a seat design called the bubble. The innovation of the design is to relocate the hand baggage to underneath the seat, therefore eliminating the need for overhead bins; this in turns increase the passenger's perception of space by reducing the tunnel effect. B/E Aerospace developed the moving set called ICON seating [8]. The moving seat surface allows the passenger to adopt multiple postures, including back and side sleep. Side support wings on the seat bottom can be adjusted to provide leg support in a side sleep posture. ICON seating allows passenger in full control of comfort and personal space.

A Swiss company developed the pneumatic cushions comfort system for aircraft seat. The new system is replaced conventional foams with air-filled chambers. Passenger can adjust the pneumatic pressure of the seat to suit their personal preferences, from firm when seated upright and medium when relaxing to soft in the fully flat position [9].

60.3 Relationship of Subjective Method to Comfort and Discomfort

Due to the lack of proven analytical metrics, seat manufacturers have opted to rely on subjective evaluations as the main indicator of seat comfort. The seat manufacturers developed elaborative subjective evaluation protocols that involved highly structure questionnaires [10]. The questionnaires direct occupants to assign feelings of discomfort to a specific region of seat. The questionnaires, which typically contain numeric scales (e.g. 1 = very uncomfortable to 10 = very comfortable), produce subjective ratings that are translated into performance requirements/specifications [11]. A properly designed questionnaire is paramount because it affords researchers an instrument from which to establish theories [12].

In the study by Mehta and Tewari [13], ten point scale local discomfort is used to measure the tractor seat comfort. The work is to project the most appropriate method of assessment and selection of tractor seats from engineering and biomechanical view point. Eklund and Corlett [14] used local discomfort with visual analogue scale to study the correlation between trunk and back discomfort. Kyung et al. [15] used a visual body mapping analogue scale as shown in Fig. 60.1 to obtain overall ratings of comfort and discomfort for the whole body.

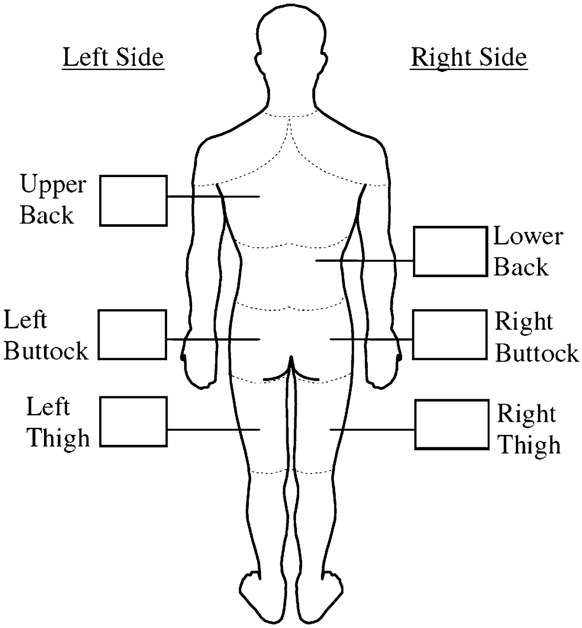


Fig. 60.1 The body mapping for comfort and discomfort rating [15]

60.4 Relationship of Objective Method to Comfort and Discomfort

Comfort measurement of seat is difficult because of such factors as user subjectivity, occupant anthropometry, seat geometry, and amount of time spent sitting [16]. A great deal of research has been performed in recent years to find objective measures for predicting seat comfort perception. Some of the proposed objective measures include vibration, interface pressure, posture and muscle activity. These objective measures are correlated with subjective data to determine the relative effects of each measure related to comfort [17].

The seat industry strongly encourages research in the field of objective comfort assessment, especially dedicated to the seat and the related postures [18, 19]. The posture is one of the important issues to be considered in the seat design process [20] regarding not only the car and the user [21, 22] but also the experimental conditions. The instruments that used in the posture measurement are camera, optoelectronic system, driving posture monitoring system, digital signal processing, ultrasonic device, 3D motion analysis, and motion measurement system.

A vast majority of objective measures used for evaluating comfort and discomfort. From the literature search, the objective measurement methods for seat such as pressure distribution, posture, computer-aided design (CAD), computer-aided engineering (CAE), temperature, humidity, vibration, electromyography (EMG), and adrenaline. Figure 60.2 shows an overview of different objective measurement methods for seat comfort and discomfort.

60.5 Sleeping Posture Analysis

Two analyses were conducted to study the sleeping in sitting posture of economy class aircraft passenger.

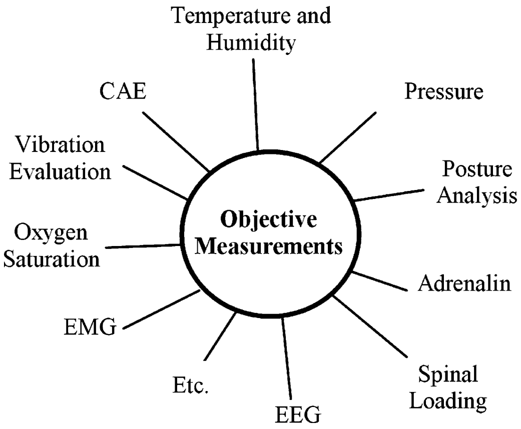


Fig. 60.2 Overview of different objective measurement methods for seat comfort and discomfort

60.5.1 Observation on Sleeping Posture

The main purpose of the observation is to find out the sleeping in sitting posture and sleeping behavior of seated economy class aircraft passenger during long distance travel. The observation was conducted in a long haul flight from Amsterdam, the Netherlands to Kuala Lumpur, Malaysia. The duration of the trip was 12 h. The researcher documented the activity of the passengers in his visual range. There were 15 subjects, eight female and seven male selected in the observation. The age of subjects was between 19 and 62 years old. The average age was 28 years old.

Based on the observation results, seven different sleeping positions identified. Observation in a long haul flight established a ground protocol on sleeping behavior of economy class passenger in a sitting position.

The protocol of sitting position while sleeping is based on four general sitting positions and one open sitting position. The sitting position while sleeping protocol as follows (Fig. 60.3):

1. Neutral position (a)
2. Slid down on seat in neutral position
   - With pillow (b)
   - Without pillow (c)
3. Head in tilted position (left and right)
   - With pillow (between shoulder and head) (f, g)
   - Without pillow (d, e)
   - Supported with hand (between shoulder and head) (h, i)
4. Torso sitting position
   - With pillow (head in diagonal with backrest) (l, m)
   - Without pillow (head in diagonal with backrest) (j, k)
   - Head resting on head rest (head perpendicular with backrest) (n, o)

60.5.2 Objective Analysis on Sitting Posture While Sleeping

The purpose of the objective method is to measure and validate the sleeping in sitting protocol that based on observation method. The objective analysis was conducted in a low cost aircraft cabin simulator (Fig. 60.4).

The low cost aircraft cabin simulator is a testbed that is developed for European project, namely, SEAT (Smart tEchnologies for Stress free Air Travel). The SEAT project aims to develop a new radical approach through integration of cabin systems with multimedia features. The aircraft cabin simulator is fully designed and built by us. The simulator consists of a small scale cabin-like testing platform, an inventory section, a simulation section and a control section. The interior of the aircraft cabin consists of an economy class section, a business class section, a galley and a lavatory.

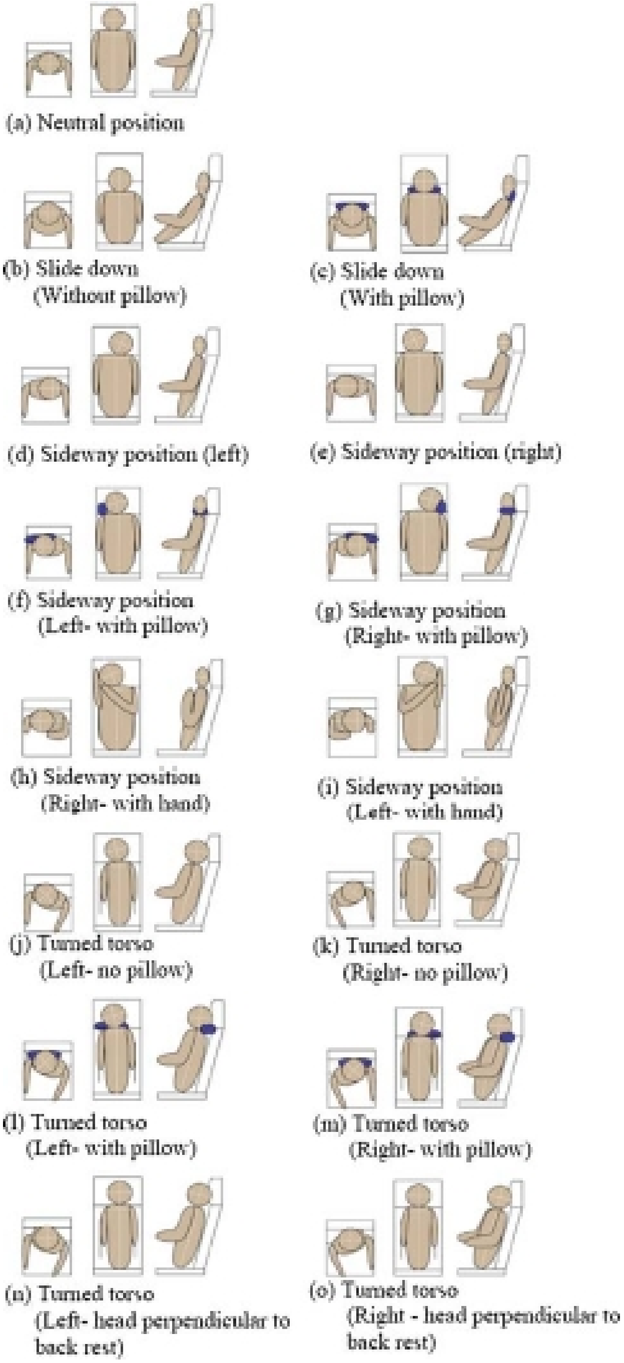

(a) Neutral position
(b) Slide down (Without pillow)
(c) Slide down (With pillow)
(d) Sideway position (left)
(e) Sideway position (right)
(f) Sideway position (Left- with pillow)
(g) Sideway position (Right- with pillow)
(h) Sideway position (Right- with hand)
(i) Sideway position (Left- with hand)
(j) Turned torso (Left- no pillow)
(k) Turned torso (Right- no pillow)
(l) Turned torso (Left- with pillow)
(m) Turned torso (Right- with pillow)
(n) Turned torso (Left- head perpendicular to back rest)
(o) Turned torso (Right - head perpendicular to back rest)

Fig. 60.3 The observed sleeping in sitting protocol

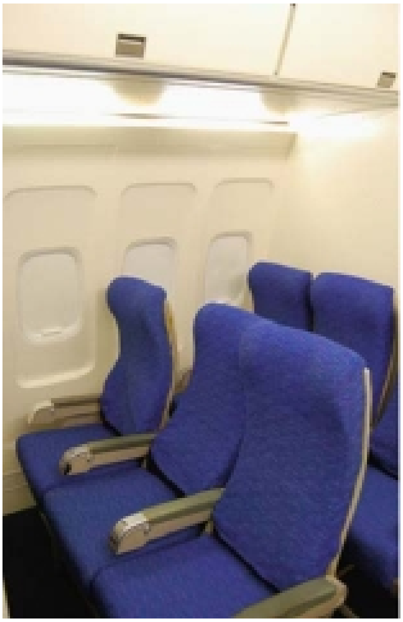
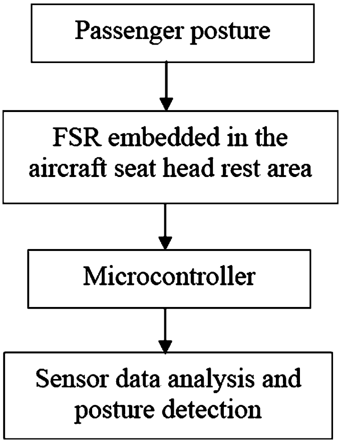


Fig. 60.4 Aircraft cabin simulator



Passenger posture
FSR embedded in the aircraft seat head rest area
Microcontroller
Sensor data analysis and posture detection

Fig. 60.5 Flowchart of the experimental setup

The experiment was conducted for each individual separately. Before the experiment, the participant was briefed with the experiment procedure and regulation. The participant sat in the prepared seat, interpreted the ten sitting positions from the protocol as shown in Fig. 60.3 for 30 s. The measurement was started when the participant confirmed in the correct sitting position. Each position was measured with microcontroller. Force sensitive resistor (FSR) was used for the posture measurement. The experimental setup to detect seated person posture is summarized in Fig. 60.5. Figure 60.6 shows the preparation of the sensors on seat head rest area.

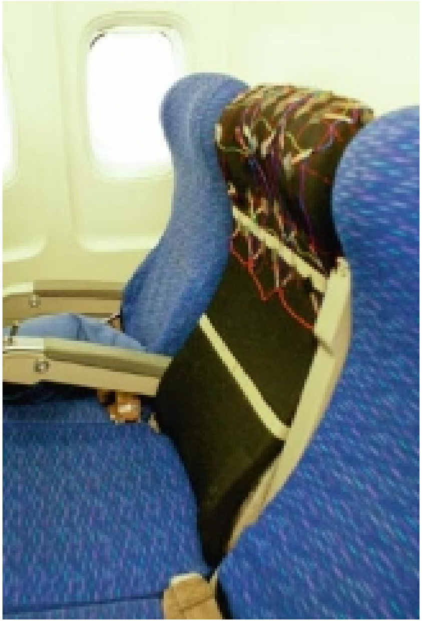


Fig. 60.6 FSR sensors on head rest area

Twelve participants, four female and eight male, participated in the experiment. The age range of participant was between 22 and 25 years old, with an average of 24 years old. Their average height is 1.82 m.

From the experiment, the sensor outputs corresponded with the sitting posture protocol (Figs. 60.7 and 60.8).

The sitting position with P4C-turned torso, with head facing the seat in front is the most comfortable sitting position for participants. The sitting position with P3C-head tilted with hand supporting between neck and head are criticized by many participants. For position 5 – freedom for personal sleeping preferences, it is the most preferable sleeping position among the participants. During the experiment, most of participants turned their torso slightly up to perpendicular towards the back-rest as well as leaning to their side of their face (with or without pillow) against the headrest.

60.6 Conclusion

In this chapter, we have described the subjective and objective measurement techniques used to measure sleeping in sitting posture of the economy class aircraft passenger. For subjective measurements, observation method is used to determine the sleeping in sitting posture. The protocol enables the ongoing research to quantify

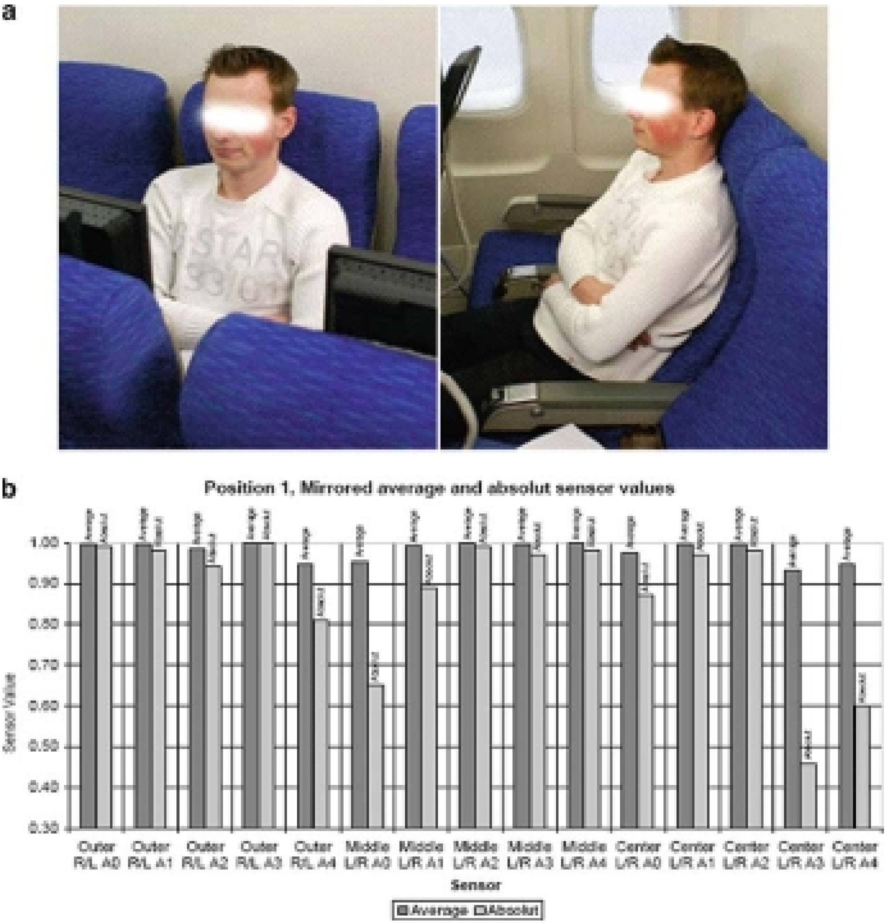


Fig. 60.7 Position 1, neutral sitting position (**a**) a participant in position 1 (**b**) sensor output

the sleeping in sitting posture, predict the sitting pattern and make the comparison. For objective measurement, the developed posture measurement method is used to detect the posture change of aircraft passenger. Twelve participants involved in the experiment to validate the protocol with the developed sensor platform. It is recommend that objective and subjective measurement should be correlated together for better understanding of comfort and discomfort in order to design comfortable economy class aircraft seat.

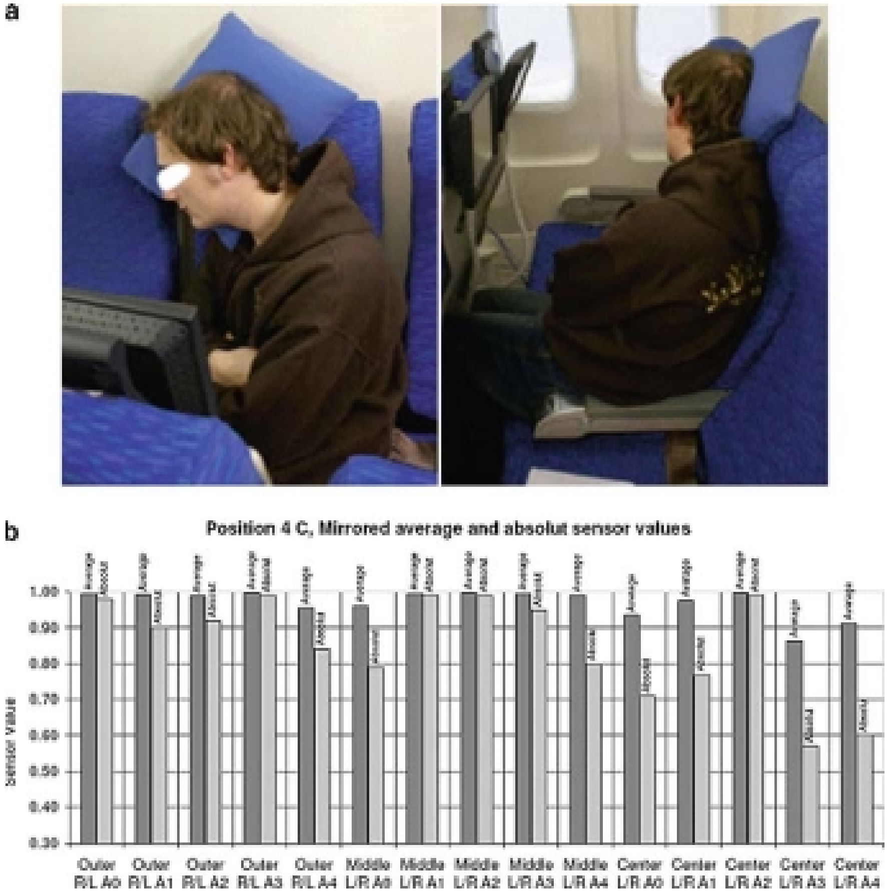


Fig. 60.8 Position 4C- sitting position in turned torso position with a pillow. (**a**) A participant in position P4C; (**b**) sensor output

References

1. Brundrett, G.: Comfort and health in commercial aircraft: a literature review. J. Roy. Soc. Promot. Health **121**(1), 29–37 (2001)
2. Kalogeropoulos, S.: Sky rage. In: Flight Safety Australia, pp. 36–37 (1998)
3. World Health Organization. Travel by air: health considerations. http://whqlibdoc.who.int/publications/2005/9241580364_chap2.pdf. Accessed March 2007
4. Runkle, V.A.: Benchmarking seat comfort. SAE Technical Paper, No. 940217 (1994)
5. Quigley, C., Southall, D., Freer, M., Moody, A., Porter, M.: Anthropometric study to update minimum aircraft seating standards. In: EC1270, ICE Ergonomics Ltd (2001)
6. Hinninghofen, H., Enck, P.: Passenger well-being in airplanes. Auton Neurosci **129**(1–2), 80–85 (2006)
7. Sutter, C.B., Acuna, M.: Tall order. In: Aircraft Interiors International. UK International Press, pp. 58–64 (2003)

8. Elliott, C.: One day, that economy ticket may buy you a place to stand. *The New York Times* (2006)
9. Lantal Textiles. Status May 2008: Lantal's Pneumatic Cushion Comfort System (2006)
10. Ahmadian, M., Seigler, M., Clapper, D., Sprouse, A.: A comparative analysis of air-inflated and foam seat cushions for truck seats. SAE Technical Paper, no. 2002-01-3108 (2001)
11. Kolich, M., Pielemeier, W.J., Szott, M. L.: A comparison of occupied seat vibration transmissibility from two independent facilities. J. Vibrat. Contr. **12**(2), 189–196 (2006)
12. Scarlett, A.J., Price, J.S., Stayner, R.M.: Whole body vibration: evaluation of emission and exposure levels arising from agricultural tractors. J. Terramech. **44**, 65–73 (2007)
13. Mehta, C.R., Tewari, V.K.: Seating discomfort for tractor operators- a critical review. Int. J. Indus. Ergonomic. **25**(6), 661–674 (2000)
14. Eklund, J.A.E., Corlett, E.N.: Evaluation of spinal loads and chair design in seated work tasks. Clin. Biomech. **2**, 27–33 (1987)
15. Kyung, G., Nussbaum, M.A., Banski-Reeves, K.: Driver sitting comfort and discomfort (part 1): use of subjective ratings in discriminating car seats and correspondence among ratings. Int. J. Indus. Ergonom. **38**(5–6), 512–525 (2008)
16. Hertzberg, H.T.E.: The human buttocks in sitting: pressures, patterns, and palliatives. SAE Technical Paper, no. 72005 (1972)
17. Kolich, M.: Review: A conceptual framework proposed to formalize the scientific investigation of automobile seat comfort. Appl. Ergonom. **39**(1), 15–27 (2008)
18. De Looze, M.P., Kuijt_Evers L.F.M., Dieen, J.V.: Sitting comfort and discomfort and the relationship with objective measures. Ergonomics **46**(10), 985–997 (2003)
19. Shen, W., Vertiz, A.: Redefining Seat Comfort. SAE Technical Paper, no. 970597 (1997)
20. Gruber, G.J.: Relationship between whole body vibration and morbidity patterns among interstate truck drivers, Southwest Research Institute, pp. 77–167, San Antonio, TX, Center for Disease Control Publication (1976)
21. Transafety Reporter. Truck Survey Highlights Causes of Drowsy Driving and Suggests Preventative Measures, Transafety Incorporated (1998)
22. Gillespie, T.D.: Heavy Truck Ride. SAE Technical Paper, no. 850001 (1985)

Chapter 61
A Scheduling Method for Cranes in a Container Yard with Inter-Crane Interference

Mak Kai Ling and Sun Di

Abstract Effective and efficient scheduling of yard crane operations is essential to guarantee a smooth and fast container flow in a container terminal, thus leading to a high terminal throughput. This paper studies the problem of scheduling yard cranes to perform a given set of loading and unloading jobs with different ready times in a yard zone. In particular, the inter-crane interference between adjacent yard cranes which results in the movement of a yard crane being blocked by adjacent yard cranes is studied. The objective is to minimize the sum of yard crane completing times. Since the scheduling problem is NP-complete, a new hybrid optimization algorithm combining the techniques of genetic algorithm and tabu search method (GA-TS) is proposed to solve the challenging problem. Two new operators, namely the Tabu Search Crossover (TSC) and the Tabu Search Mutation (TSM), are introduced into the proposed algorithm to ensure efficient computation. A set of test problems generated randomly based on real life data is used to evaluate the performance of the proposed algorithm. Computational results clearly indicate that GA-TS can successfully locate cost-effective solutions which are on average 20% better than that located by GA. Indeed, the proposed hybrid algorithm is an effective and efficient means for scheduling yard cranes in computer terminals.

Keywords Yard crane · inter-crane interference · hybrid algorithm · genetic algorithm · Tabu search

61.1 Introduction

With the rapid trade globalization, the marine transportation is getting more and more popular. Large numbers of cargos are moved in containers through ports. Therefore, effective and efficient management of port container terminals is quite

M.K. Ling (✉) and S. Di
Department of Industrial and Manufacturing Systems Engineering,
The University of Hong Kong (HKU), Hong Kong
e-mail: makkl@hkucc.hku.hk; h0795497@hku.hk

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_61,

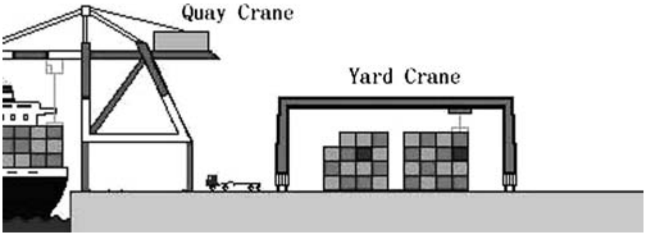


Fig. 61.1 Typical container flow in terminal operations

important in marine transportation development. In addition, container ports compete with each other for better customer service. Of all the service performance measures, vessel turnaround time, which is the average time that a vessel stays in a terminal, is the key one. The most effective method to reduce the terminal turnaround time is to improve the productivity of handling activities.

It is essential to study the operational processes of a port container terminal. Figure 61.1. shows the typical container flow between major handling equipment in a port container terminal. When a vessel arrives at the terminal, containers are normally discharged from the vessel onto trucks by quay cranes. Unloaded import containers are transported to the yard and off-loaded by yard cranes for storage. For export containers, containers are mounted onto trucks by yard cranes and transported to the quayside for loading onto the vessel by quay cranes. Steenken et al. [1] described and classified the main logistics processes and operations in container terminals and presented a survey of methods for their optimization. Stahlbock and Voβ [2] extended the study of Steenken et al. [1].

In a container terminal, yard cranes are important equipment for loading containers onto and unloading containers from mobile trucks, and for stacking containers in storage locations in the container yard according to the sequence predetermined by terminal planners. However, the low physical operation rate of such equipment and their frequent movements when performing handling tasks in the yard very often cause bottlenecks in the container flow in the terminal. Therefore, yard crane scheduling plays a significant role in port management. An effective yard crane scheduling can reduce truck waiting time, speed up the container flow to and from vessels. A number of researchers studied scheduling of yard cranes in a container yard.

Kim and Kim [3] considered the problem of routing a single straddle carrier, which is transportation equipment with container handling capability, to transport export containers to a loading vessel. To minimize the total container handling time of a straddle carrier, a mixed integer program was formulated. Zhang et al. [4] investigated the Rubber Tyred Gantry Cranes (RTGCs) deployment problem. The objective was to find the times and routes of cranes movement among blocks. So that the total delayed workload in the yard was minimized. A mixed integer program (MIP) model was formulated and solved by Lagrangean relaxation. Linn and Zhang [5] also studied the RTGCs deployment problem. The objective was to minimize the

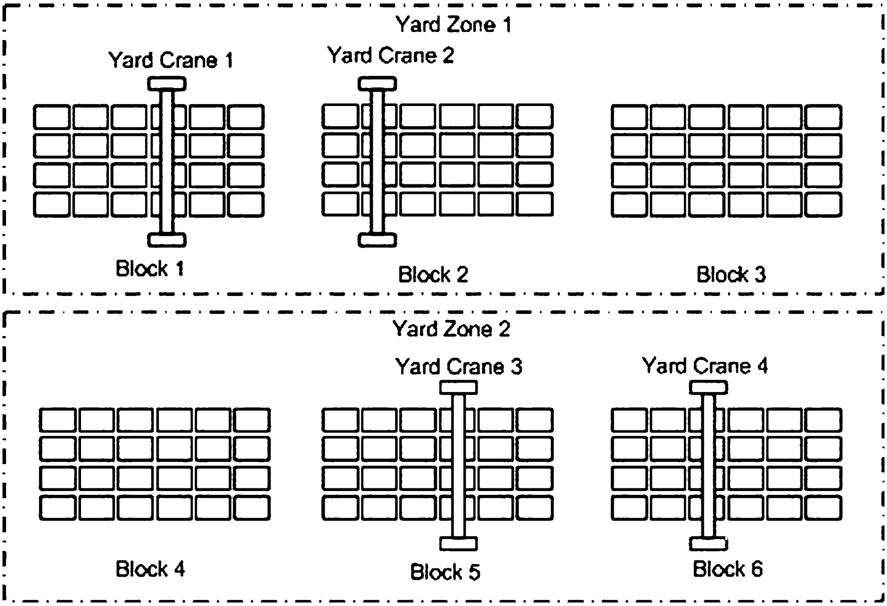


Fig. 61.2 Layout of container yard

total workload overflow through determining the crane deployment frequency and routes over a planning horizon. A heuristic algorithm was also developed to provide a near optimal solution for crane deployment.

Rail Mounted Gantry Cranes (RMGs) are also used frequently in practice. These cranes are particularly effective for rail/road transshipments of large quantities of containers. For yard cranes (RMGs) running on rails, movement is restricted to a predetermined zone. Due to sharing of the traveling lane among two or more yard cranes in a yard zone, inter-crane interference, a planned move of a yard crane blocked by the other yard cranes, may happen. For example, in Fig. 61.2, if yard crane 2 is handling its job, yard crane 1 cannot move across yard crane 2 to handle jobs on its right hand side. Therefore, if yard crane 1 is going to handle jobs in block 3, it has to wait until yard crane 2 has completed its job.

Due to the complexity of considering inter-crane interference constraints in scheduling yard cranes, Ng and Mak [6] neglected the crossing movement restriction by assuming only a yard crane in a yard zone, where the yard crane can move freely to perform all handling jobs generated by different vessels. Only a few papers addressed the routing problem regarding yard cranes with inter-crane interference constraints. Lim et al. [7] studied a model that took into account of interference between yard cranes as "non-crossing" constraint and used a tabu search heuristic for solutions. However, the paper did not consider the handling time of yard cranes, the travel time between two jobs and the waiting time due to yard crane interference. Ng [8] addressed the scheduling problem for yard cranes considering interference among adjacent yard cranes. The paper divided the yard to several zones and used a dynamic programming model to determine the sequence of jobs for each yard crane. In Ng [8], the time was discretized beforehand. However, in real world, the time is continuous. Jung and Kim [9] scheduled loading operations when multiple yard cranes are operating in the same block. They considered interferences between adjacent yard cranes. The objective was the minimization of the make-span of the yard crane operation. The paper used a genetic algorithm and a simulated annealing method to solve the model. This paper studies the problem of scheduling multiple yard cranes in a yard zone to minimize the sum of the completion times of yard cranes. When calculating the completion time of each yard crane, we consider the container ready time, the handling time, the yard crane travelling time and the waiting time of yard cranes due to inter-crane interference. The remainder of the paper is organized as follows. Section 61.2 proposes a mixed integer program model for the scheduling problem. A new hybrid genetic algorithm and tabu search method for solving the scheduling problem is given in Section 61.3. Section 61.4 presents the results of computational experiments. Conclusions are in Section 61.5.

61.2 Model Development

A mixed integer program mathematical model describing the characteristics of the yard crane scheduling problem is developed. Some important assumptions for the formulation are listed below:

1. A yard truck can transport an export/import container from the yard storage place/quayside to quayside/yard storage place.
2. A job is defined as a yard crane loading/unloading an export/import container onto/from a yard truck from/to its storage yard place.
3. The handling time of a job is fixed.
4. Multiple yard cranes serve simultaneously in a yard zone which can move freely as long as they do not cross over each other. Inter-crane interference between yard cranes is taken into consideration.

There are n jobs to be handled by m identical yard cranes in a yard zone of θ slots. The yard cranes are ordered in an increasing order of their relative locations in the yard zone. The meanings of variables in the model are listed below:

r_i = the ready time of job i
l_i = the location of job i
d_{ij} = the time required for yard cranes to travel from l_i to l_j
h = the time required by a yard crane to handle one job

Decision variables:

$W_i = (S_i, D_i)$ the handling time window for job i
D_i = the completion time of job i
S_i = the time at which the yard crane assigned starts to handle job i
t_i = the arrival time of the yard crane assigned to job i

C_k = the completion time of yard crane k

$$X_{ij}^k = \begin{cases} 1 & \text{if yard crane } k \text{ performs job } i \text{ before job } j \\ 0 & \text{otherwise} \end{cases}$$

Y_i = the yard crane assigned to handle job i

The mathematical model describing the characteristics of the scheduling problem is shown below:

$$\text{Minimize } \sum_{k=1}^{m} C_k$$

Subject to

$$\sum_{j=1}^{n} X_{0j}^k = 1, \quad k = 1, \ldots, m, \tag{61.1}$$

$$\sum_{i=1}^{n} X_{iT}^k = 1, \quad k = 1, \ldots, m \tag{61.2}$$

$$\sum_{k=1}^{m} \sum_{i=0}^{n} X_{ij}^k = 1, \quad j = 1, \ldots, n \tag{61.3}$$

$$\sum_{j=1}^{n} X_{ij}^k - \sum_{j=1}^{n} X_{ji}^k = 0, \quad k = 1, \ldots, m \tag{61.4}$$

$$D_i = S_i + h, \quad i = 1, \ldots, n, \tag{61.5}$$

$$S_i = \max\{r_i, t_i\}, \quad i = 1, \ldots, n \tag{61.6}$$

$$D_j - D_i \geq d_{ij} + h - \left(1 - X_{ij}^k\right) M, \; i, j = 1, \ldots, n, \quad \text{and } i \neq j \tag{61.7}$$

$$\begin{array}{ll} (Y_i - Y_j)(l_i - l_j) > 0 & \text{if } \bigcap W_i \cap W_j \neq \phi \\ i, j = 1, \ldots, n & \text{and} \quad i \neq j \end{array} \tag{61.8}$$

$$D_j + d_{jT} - C_k \leq M\left(1 - X_{jT}^k\right), \quad j = 1, \ldots, n, \; k = 1, \ldots, m \tag{61.9}$$

$$X_{ij}^k \in \{0, 1\}, \quad i, j = 1, \ldots, n, \; k = 1, \ldots, m \tag{61.10}$$

$$D_i, \; S_i, \; t_i, \; C_k \geq 0, \quad i = 1, \ldots, n, \; k = 1, \ldots, m \tag{61.11}$$

$$Y_i \in \{1, 2, \ldots\} \quad i = 1, \ldots, n \tag{61.12}$$

In Eq. (61.7), M is a big positive number.

The objective of the scheduling problem is to minimize the sum of the completion times of the yard cranes. Constraints (61.1) and (61.2), select the first and last tasks for each yard crane, respectively. Constraint (61.3) indicates that every job must be completed by one yard crane. Constraint (61.4) is a flow balance constraint for yard crane travels. Constraint (61.5) computes a job's completion time. Constraint (61.6) shows that the yard crane assigned for a job would start to handle the job after the job ready time and its arrival time to the job. Constraint (61.7) implies the relationship between the completion time of a job and that of its successors. By constraint (61.8), interference among yard cranes can be avoided. The completion time of each yard crane is defined by constraint (61.9). X_{ij}^k can be 0 or 1 by constraint (61.10). Variables in Constraint (61.11) have nonnegative value, while Y_i is a nonnegative integer variable by constraint (61.12).

61.3 Hybrid Genetic Algorithm and Tabu Search

In order to find an efficient and effective solution, this paper develops a new hybrid genetic algorithm and tabu search method. Two new TS-based operators of genetic algorithm are designed. The first one is adding tabu search to the mutation operation of genetic algorithm called (TSM). Genetic algorithm and tabu search depend on disparate search principles. In general, TS is good at performing deeper exploitation since it is based on local search heuristics. GA is short of the power of deeper exploitation within the promising regions but performs well at exploring different regions due to the implicit parallelism feature. It would be desirable to hybridize them to achieve a better balance between exploration and exploitation. The TSM fulfills the objective.

The other one is the new tabu search crossover operator called (TSC). In canonical GA, all chromosomes are replaced by their offspring after the crossover procedures as the population evolves. Due to the extremely short life span of the individuals, the search algorithms therefore do not have sufficient time to sample out the useful schemata from individuals. To avoid the shortage, the parents would stay in the population to join the selection process with their offspring. However, this method would cause new problem. Some good individuals would stay permanent and give birth to new generations which would lead to premature and stalling of the search process. In order to ameliorate this situation, this paper designs the TSC. It is a memory-based strategy like tabu search. A TSC list is designed to store the individuals who have been selected as parents. The TSC list records the number of times (*np*) of each chromosome selected as parent. The probability to become parent would be inversely proportional to the *np* of each individual in the TSC list. Using this method, this paper effectively avoids premature caused by crossover operation.

The general outline of each step of the proposed algorithm is presented below:

Step 1: Set $k = 0$, and randomly generate the initial population P_0. The size of P_0 is n_0.

Step 2: Select n_0 individuals from population P_k to form population S_k by using the selection operator of conventional GA.

1. Select parent individuals to give birth to new individuals according to the individuals' crossover probabilities calculated by using the value of *np* of each individual in the TSC list. Population S_k and the new individuals together form population C_k. The size of C_k is n_1 $(n_1 > n_0)$
2. Apply the TSM mutation operator to population C_k to obtain population M_k
3. Set $k = k + 1$, $P_k = M_k$

Step 3: If $k <$ Total Generation, go to Step 2

61.3.1 Representation

This paper uses the same two-part chromosome structure as in Ng et al. [10]. (2 4 8 9 5 3 9 6 1 | 4 3 2) illustrates an example chromosome for representing a crane schedule for assigning three yard cranes ($m = 3$) to process nine jobs ($n = 9$) using the two-part chromosome structure. There are two distinct parts in the chromosome with a total length of $m + n$. In the first part, the n jobs are represented by a permutation of the integers from 1 to n. The second part of the chromosome, which is of length m, represents the number of jobs assigned to each of the m yard cranes. The values assigned to the second part of the chromosome are constrained to be m non-negative integers whose sum must equal the number n of jobs. In the example shown above, the first yard crane would sequentially process jobs 2, 4, 8 and 9, the second yard crane should process jobs 5, 3 and 7, and the last yard crane would process jobs 6 and 1.

61.3.2 Fitness Evaluation and Selection Operation

Inter crane interference occurs when the next job for yard crane A is located on the other side of yard crane B, and yard crane B is working on another job. In this case, yard crane A has to wait until the yard crane B has completed its current job. Consequently, the procedure to evaluate the objective function value of an individual is as follows:

Step 1: Calculate the objective function value of the individual without considering constraint (61.8). During this process, we can obtain the value of $W_i = (S_i, D_i)$ of job i $(i = 1, 2, \ldots, n)$.

Step 2: (Check possible inter crane interference)

1. Sort the n jobs in ascending order of the value of its S_i to determine the sequence $(1', 2', \ldots, n')$

2. Find the first pair of jobs (i, j) such that $\bigcap_{i \neq j} W_i \cap W_j \neq \phi$ and constraint (61.8) is not satisfied

Let $i = 1'$ to $n' - 1$. For each i, check $S_j < D_i$ for $j = i + 1$ to n' until the condition is satisfied and jobs (i, j) does not satisfy constraint (61.8). If such a pair of (i, j) exits, inter crane interference occurs. Otherwise, inter crane interference does not exit and the current result obtained is the final objective function value of the individual.

Step 3: (Conflict resolution).Yard crane Y_j must wait until yard crane Y_i has completed job i. Then, Y_j moves from l_i to l_j and begins to handle job j. Update all the W_k affected by this operation.

Step 4: Repeat Steps 2 and 3 until no inter crane interference could be found.

61.3.3 Tabu Search Crossover (TSC) Operation

In crossover procedure, there are two questions should be answer. The first one is how to choose individuals to be parents. The other one is how to use information of two parents to generate good offspring. Since this paper uses the method suggested by [5] to generate offspring, the following of this section focuses on the first question. The individuals are chosen as parents depending on their crossover probabilities which is computed as following:

$$Pb(np(b_i)) = \exp\left(-\frac{(np(b_i) - a)^2}{b^2}\right)$$

where $np(b_i)$ is the value of solution b_i in TSC list. TSC list records the number of times that an individual has been selected as parent. a and b are given parameters which control the shape of probability function $Pb(np(b_i))$. If $np(b_i) = a$, then $Pb(np(b_i)) = 1$. $Pb(np(b_i))$ gradually converges to 0 with $np(b_i)$ growing up. The procedure to choose the individuals as parents is as following:

Step 1: Calculate crossover probability for each individual, $Pb(np(b_i))$.

Step 2: Generate a random number RN between 0 and 1.

Step 3: If $RN < Pb(np(b_i))$, the individual b_i is chosen as parents. $np(b_i) = np(b_i) + 1$.

61.3.4 Tabu Search Mutation (TSM) operation

Mutation forces GA searching new areas. Adding Tabu Search in mutation operation would improve the deeper exploitation for GA. By using Tabu List, it could avoid premature and finally get global optimal solution. The procedure of tabu search mutation is following:

x_0 the initial solution. The tabu list is TL. Neighborhood of x_0 is $N(x_0)$.

```
Input $x_0$
While $(t < T)$
$TL = \phi$; Set the best solution $x = x_0$; $t = 0$
Get set $CN(x_0) \subseteq N(x_0)$ and $CN(x_0) \cap TL = \phi$
Get $x' \in CN(x_0)$ and $x'$ has the best fitness value in
$CN(x_0)$
Move $x$ to $x'$; $t = t + 1$
Update $(x; x_0; TL)$
Output $x$
```

61.4 Computation Results

The test problems are randomly generated using the method proposed in Ng [8]. The parameters of GA-TS need to be determined. This paper assumes that $a = 3$, $b = 2$ in the TSC operator. In the mutation operation, the length of the tabu list is four, the number of neighbors searched is five and the termination condition is $T = 6$. In both GA-TS and GA, the crossover and mutation probabilities are fixed to 0.6 and 0.4, respectively. The initial population size is set to 50 and the stop condition is 100 iterations. The computer used is equipped with Inter Pentium 2.4 GHz CPU and 512 MB RAM. The calculation time is shown in seconds.

61.4.1 Comparison with Branch and Bound Algorithm

This paper uses the technique of Least Cost Branch and Bound to get the optimal solution. In such an approach, the solution space is often organized as a tree. Details of the method can be found in Sahni [11]. A small-size test problem is used in the experiment for performance evaluation. Let TC_{GT} and TC_{IP} be the best objective function values found by using the hybrid GA-TS method and the Branch and Bound algorithm, respectively. $Time_{GT}$ and $Time_{IP}$ are the computational times of the algorithms. A set of 20 test problems is randomly generated for each combination of parameters. The results are shown in Table 61.1.

Table 61.1 Performance of proposed algorithm on small-scale test problem

			$(TC_{GT} - TC_{IP})/TC_{IP}$			$Time_{GT}$			$Time_{IP}$		
n	m	θ	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean
5	1	10	0.0	0.0	0.0	5.03	3.09	0.61	151	45	74
5	2	10	0.0	0.0	0.0	5.79	3.84	4.56	172	50	88
10	2	20	2.2	0.0	1.3	6.12	4.43	5.61	924	472	650
10	2	30	3.3	0.0	2.1	6.83	4.74	6.10	1057	561	732

Comparing the results between branch and bound algorithm and GA-TS for small-scale problems, for $n = 5$, GA-TS can always get the global optimum. Although in some cases GA-TS may not always reach the optimum, its solutions are quite near optimum with only average 2% above optimal solutions. Thus, the quality of solutions found by GA-TS is acceptable. In addition, by using GA-TS, it would averagely cost less than 5 s to solve problems when $n = 5$ and 7 s when $n = 10$. In contrast, the runtime used by Branch-and-Bound is much longer. Since the Branch and Bound algorithm requires a n-element permutation, the solution space tree is a permutation tree with $n!$ leaves. Searching through all nodes of the tree would cost $\Omega(n!)$ time. Thus, the runtime of Branch-and-Bound algorithm increases by geometric rate as the number of jobs growing. In this experiment, using Branch-and-Bound algorithm, the average CPU time required is 74 s to solve the problems when $n = 5$. As to the problems when $n = 10$, it increases to 650 s. Therefore, when the number of jobs is large, Branch-and-Bound algorithm is not suitable as a solution approach.

61.4.2 Comparison with Genetic Algorithm

Results given by hybrid Genetic Algorithm and Tabu Search method and canonical Genetic Algorithm are presented in Table 61.2. For each combination of parameters, 20 random test problems are generated. TC_{GA} is the objective function value found by Genetic Algorithm. The results show the performance of GA-TS clearly

Table 61.2 Comparison between GA-TS and GA

			$(TC_{GT} - TC_{IP})/TC_{IP}$		
n	m	θ	Max	Min	Mean
10	2	40	11.8	0.0	5.7
20	2	40	33.3	6.7	22.6
30	3	40	39.6	5.5	21.9
40	3	40	41.4	11.7	23.8
50	4	40	24.2	6.0	10.6
60	4	40	23.4	2.0	11.6
10	2	80	17.0	0.0	8.9
20	2	80	41.9	4.7	28.0
30	3	80	44.9	4.1	23.1
40	3	80	26.1	−0.1	14.6
50	4	80	36.5	4.3	27.1
60	4	80	14.1	1.9	7.0
10	2	120	33.9	4.0	20.6
20	2	120	42.6	18.0	22.9
30	3	120	53.8	3.1	35.7
40	3	120	44.6	10.7	35.3
50	4	120	58.6	9.0	40.2
60	4	120	25.6	7.8	10.2

outperforms GA. Almost all solutions obtained by GA-TS are better than the corresponding solutions obtained by GA except one. The GA-TS solutions are on average 20% better than solutions found by GA.

61.5 Conclusion

This paper studies the yard crane scheduling problem with inter-crane interference. The objective is to minimize the sum of the completion time of the yard cranes. When calculating the completion time of each yard crane, we consider the container ready time, handling time, the yard crane travelling time and the waiting time of yard cranes due to inter-crane interference. Since the problem is NP-complete, a new hybrid Genetic Algorithm and Tabu Search method has been developed to resolve the challenging problem. The computation results show the new algorithm is superior to GA with its solutions on average 20% better than solutions found by GA. Tabu Search Crossover (TSC) and Tabu Search Mutation (TSM) are effective to avoid premature and speed up convergence. Comparing with branch and bound algorithm, GA-TS can always give a reasonable solution within limited time.

References

1. Steenken, D., Voß, S., Stahbock, R.: Container terminal operation and operations research – a classification and literature review. OR Spectrum. **26**, 282–292 (2004)
2. Stahlbock, R., Voβ, S.: Operations research at container terminals: a literature update. OR Spectrum, **30**(1), 1–52 (2008)
3. Kim, K.H., Kim, K.Y.: An optimal routing algorithm for a transfer crane in port container terminals. Transport. Sci. **33**, 17–33 (1999)
4. Zhang, C., Wan, Y., Liu, J., Linn, R.J.: Dynamic crane deployment in container storage yards. Transport. Res. Part B **36**, 537–555 (2002)
5. Linn, R.J., Zhang, C.: A heuristic for dynamic yard crane deployment in a container terminal. IIE Trans. **35**, 161–174 (2003)
6. Ng, W.C., Mak, K.L.: Yard crane scheduling in port container terminals. Appl. Math. Model. **29**, 263–276 (2005)
7. Lim, A., Xiao, F., Rodrigues, B., Zhu, Y.: Crane scheduling using tabu search. In 14th IEEE International Conference on Tools with Artificial Intelligence, Washington DC, USA (2002)
8. Ng, W.C.: Crane scheduling in container yards with inter-crane interference. European J. Oper. Res. **164**, 64–78 (2005)
9. Jung, S.H., Kim, K.H.: Load scheduling for multiple quay cranes in port container terminals. J. Intell. Manufact. **17**, 479–492 (2006)
10. Ng, W.C., Mak, K.L, Zhang, Y.X.: Scheduling trucks in container terminals using a genetic algorithm. Eng. Optimizat. **39**:33–47 (2007)
11. Sahni, S.: Data Structures, Algorithms, and Applications in C++, pp.788–810. McGraw-Hill, New York, (1998)