Gi-Chul Yang · Sio-Iong Ao
Len Gelman *Editors*

Transactions on Engineering Technologies

World Congress on Engineering 2014

Springer

Transactions on Engineering Technologies

Gi-Chul Yang • Sio-Iong Ao • Len Gelman
Editors

Transactions on Engineering Technologies

World Congress on Engineering 2014

Springer

Editors
Gi-Chul Yang
Multimedia Engineering
College of Engineering
Mokpo National University
Chonnam, Korea
Republic of (South Korea)

Len Gelman
Department of Applied
Mathematics & Computing
School of Engineering
Cranfield University
Cranfield, Bedfordshire, UK

Sio-Iong Ao
IAENG Secretariat
International Association of Engineers
Hong Kong, Hong Kong SAR

ISBN 978-94-017-9803-7 ISBN 978-94-017-9804-4 (eBook)
DOI 10.1007/978-94-017-9804-4

Library of Congress Control Number: 2015936284

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

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

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in London, U.K., 2–4 July, 2014, under the World Congress on Engineering 2014 (WCE 2014). The WCE 2014 is organized by the International Association of Engineers (IAENG); the Congress details are available at: http://www.iaeng.org/WCE2014. 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 to exchange ideas. The conferences have also struck a balance between theoretical and application development. The conference committees have been formed with over 300 committee members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries. The congress is truly global international event with a high level of participation from many countries. The response to the Congress has been excellent. There have been more than 900 manuscript submissions for the WCE 2014. All submitted papers have gone through the peer review process, and the overall acceptance rate is 51 %.

This volume contains 51 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include mechanical engineering, bioengineering, internet engineering, wireless networks, signal and image engineering, manufacturing engineering, and industrial applications. The book offers the state of art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent reference work for researchers and graduate students working on engineering technologies and physical science and applications.

Chonnam, Korea, Republic of (South Korea) — Gi-Chul Yang
Hong Kong, Hong Kong SAR — Sio-Iong Ao
Cranfield, Bedfordshire, UK — Len Gelman

Contents

Chapter 1
Numerical Study of Conjugate Natural Convection from Discrete Heat Sources

Farouq A. Gdhaidh, Khalid Hussain, and Hong-Sheng Qi

Abstract The coupling between natural convection and conduction within rectangular enclosure was investigated numerically. Three separate heat sources were flush mounted on a vertical wall and an isoflux condition was applied at the back of heat sources. The governing equations were solved using control volume formulation. A modified Rayleigh number and a substrate/fluid thermal conductivity ratio were used in the range $10^4 - 10^7$ and $10 - 10^3$ respectively. The investigation was extended to examine high thermal conductivity ratio values. The results illustrated that, when Rayleigh number increased the dimensionless heat flux and local Nusselt number increased and the boundary layers along hot, cold and horizontal walls were reduced significantly. An opposite behaviour for the thermal spreading in the substrate and the dimensionless temperature, were decreased for higher Rayleigh number. Moreover, the thermal spreading in the substrate increased for higher substrate conductivity, which affected the temperature level. However the effect of the substrate is negligible when the thermal conductivity ratio higher than 1,500.

Keywords CFD study • Conjugate heat transfer • Control volume • Natural convection • Rayleigh number • Thermal conductivity

1 Introduction

Coupling between natural convection and conduction heat transfer is evident in many practical applications. In natural convection the fluid movement is created by the buoyancy force due to the temperature difference. Therefore no external force such as fans or coolers are needed which could be the main cause of noise and vibration. The conducting solid wall could also give additional heat transfer from the heat source to the fluid. The main drawback of natural convection is the rate of heat transfer is small compared to forced convection. Previous researches showed, this problem could be overcome by immersing the system in dielectric liquid [1].

F.A. Gdhaidh (✉) • K. Hussain • H.-S. Qi
Faculty of Engineering and Informatics, University of Bradford, Bradford BD7 1DP, UK
e-mail: F.A.Gdhaidh@bradford.ac.uk; K.Hussain1@bradford.ac.uk; H.Qi@bradford.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_1

Zinnes [2] investigated the conjugate effects in natural convection. He observed that, the coupling between conduction in a substrate and convection in a fluid is hugely affected by the substrate/fluid thermal conductivity ratio. Most of the published works in the field of natural convection from rectangular enclosures [3–5] whether these enclosures considered vertical or horizontal revealed that, the isothermal condition was applied to create temperature gradient. Several numerical investigations [6, 7] considered the conjugate natural convection from a heater mounted on a substrate immersed in liquid within an enclosure. It was concluded that most of the generated power was dissipated by the substrate for high value of substrate/fluid thermal conductivity ratio and the maximum temperature of the heater was decreased.

Heindel et al. [8] carried out a 3-D numerical and experimental study. Water and FC-77 were used within the enclosure and an isothermal cold wall at room temperature was assumed. They concluded that when the applied heat flux went up the convection coefficients and vertical velocity along heater faces increased. In addition, the flow inside the cavity became very complex when the Rayleigh number increased. Moreover, the 2-D numerical predictions exceeded those of 3-D model by 10–15 %. Further work by Heindel et al. [9] showed that the heat transfer was enhanced as much as 15 and 24 times for horizontal and vertical cavity orientations respectively by installing fins on the discrete heaters. The cold wall kept isothermally at 15 °C by using parallel plate fin arrays.

Numerical predictions of steady state natural convection in a square cavity was given by Banerjee et al. [10] to determine the sizes of heaters and the value of applied heat fluxes to ensure the operation within the specified thermal limit. Wroblewski and Joshi [11, 12] noticed that, at high substrate/fluid thermal conductivity ratio $R_s > 10$, the effects of the substrate conductivity especially on the maximum temperature were found to be very important due to the low thermal conductivity of the coolant liquid (FC-75). Experimental and numerical studies of conjugate heat transfer on a heated vertical wall were studied by Bilgen [13]. It was noticed that the Nusselt number depended on the Rayleigh number and wall thickness but it was a weak function of conductivity ratio. The thermal performance of the enclosure was improved with an increase in the Rayleigh number as reported in [14, 15]. Hyun and Kim [16] noticed that, in the case when the fluid confined between two plates, the temperature difference had to exceed a certain value between the plates before fluid motion could be detected.

Many efforts from previous studies using both numerical and experimental methods are focused on investigation of the effects of Rayleigh number R_a, substrate/fluid thermal conductivity ratio R_s and different fluids on the flow and temperature fields. It is shown that the maximum temperature is decreased with increasing R_s especially when $R_s > 10$. In the present study, numerical investigation was carried out, which shows that although the maximum temperature is reduced by increasing R_s, but there is a limit on R_s where the maximum temperature remains constant with increasing R_s over that limit.

2 Problem Description

A 2-D numerical investigation of the coupling between conduction and natural convection heat transfer from three discrete heat sources mounted on vertical wall of a cavity is presented. The opposite vertical wall and the horizontal walls are assumed to be at constant temperature (isothermal) and adiabatic respectively. Figure 1.1 shows a schematic of the two-dimensional rectangular cavity filled with different fluids of FC-77 (a dielectric fluorocarbon liquid) and air used in this investigation. The isoflux condition is applied at the back of each heat source whereas the back of the substrate was assumed to be adiabatic.

The height and length of the cavity where assumed as H and $(L_s + L_f)$ respectively. Aspect ratio of the cavity is fixed at $(A_z = H/L_f = 8)$. Table 1.1 illustrates the dimension of the parameters in mm.

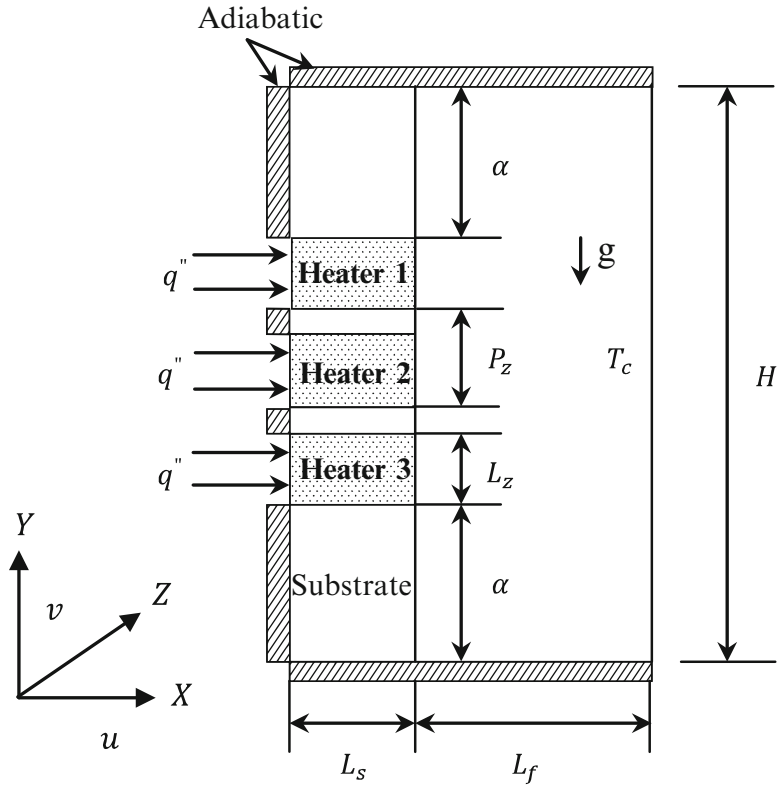


Fig. 1.1 A schematic diagram of the physical model

Table 1.1 Physical model parameters

H	L_z	P_z	L_f	L_s	α
96	12	16.8	12	6	25.2

3 Mathematical Model

The mathematical model is constrained by the following assumptions: 2-D steady state heat transfer, laminar natural convection flow without viscous dissipation, contact resistance between the heater/substrate interfaces and the radiation effects are neglected, the Boussinesq approximation is applied and the gravity acts in the vertical downward direction.

By using the above assumptions in the differential equations of continuity, momentum and energy, the governing equations are written in dimensionless form as:

- *In fluid region*:

  Continuity:

$$\frac{\partial U}{\partial X} + \frac{\partial U}{\partial Y} = 0 \tag{1.1}$$

  X-momentum:

$$U\ \frac{\partial U}{\partial X} + V\ \frac{\partial U}{\partial Y} = -\frac{\partial P}{\partial X} + P_r\ \left[\frac{\partial^2 U}{\partial X^2} + \frac{\partial^2 U}{\partial Y^2}\right] \tag{1.2}$$

  Y-momentum:

$$U\ \left[\frac{\partial V}{\partial X}\right] + V\ \left[\frac{\partial V}{\partial Y}\right] = -\left[\frac{\partial P}{\partial Y}\right] + P_r\ \left[\frac{\partial^2 V}{\partial X^2} + \frac{\partial^2 V}{\partial Y^2}\right] + R^*_{aL_z} P_r \theta \tag{1.3}$$

  Energy:

$$U\frac{\partial \theta}{\partial X} + V\frac{\partial \theta}{\partial Y} = \frac{\partial^2 \theta}{\partial X^2} + \frac{\partial^2 \theta}{\partial Y^2} \tag{1.4}$$

- *In solid region*:

  In this region there is only energy equation because the velocity components are zero, therefore the energy equation is written as:

$$K_1\frac{\partial^2 \theta}{\partial X^2} + K_2\frac{\partial^2 \theta}{\partial Y^2} = 0 \tag{1.5}$$

where K_1 and K_2 can be either the values of R_h or R_s depends on the position of the calculation whether in the heater or in the substrate region and:

$$R_s = \frac{k_s}{k_f},\ R_h = \frac{k_h}{k_f}$$

The above equations are obtained using the following dimensionless parameters:

$$X = \frac{x}{L_z},\ Y = \frac{y}{L_z},\ U = \frac{uL_z}{\alpha_f},\ V = \frac{vL_z}{\alpha_f} \tag{1.6}$$

$$P = \frac{p}{\rho\left(\alpha_f \Big/ L_z\right)^2},\ \theta = \frac{T - T_c}{\left(q'' L_z / k_f\right)} \tag{1.7}$$

$$R^*_{al_z} = \frac{g\beta q'' L_z^4}{k_f \alpha_f \nu}, \qquad P_r = \frac{\nu}{\alpha_f} \tag{1.8}$$

In order to complete the mathematical model, the following dimensionless boundary conditions are used:

- *At* $X = 0$:

$$U = V = 0,\ U = V = 0, \partial\theta/\partial X = \begin{cases} -1/R_h & at\ heater \\ 0 & at\ substrate \end{cases} \tag{1.9}$$

- *At* $X = \left(L_s + L_f\right)/L_z$:

$$U = V = 0,\ \theta = 0 \tag{1.10}$$

- *At* $Y = 0$:

$$U = V = 0,\ \frac{\partial\theta}{\partial Y} = 0 \tag{1.11}$$

- *At* $Y = H/L_z$:

$$U = V = 0,\ \frac{\partial\theta}{\partial Y} = 0 \tag{1.12}$$

The thermal conductivity for dissimilar materials is obtained by the harmonic mean formulation [17].

The local heat transfer coefficient at solid/fluid interface is defined as $h_s = q''/\left(T_{s(x)} - T_c\right)$ where $T_{s(x)}$ is the local temperature on the surface [10]. Therefore, the rate of convection heat transfer at any point in the solid/fluid interface wall is introduced by the dimensionless number (local Nusselt number $Nu(X)$) based on length of heat source, which is written as:

$$N_u(X) = -\frac{\Gamma}{\theta_w}\left(\frac{\partial\theta}{\partial X}\right)_w \tag{1.13}$$

4 Solution Procedure and Validation

In present study, the control volume technique was used to discretise the governing Eqs. (1.1, 1.2, 1.3, 1.4, and 1.5). The resulting algebraic equations were solved sequentially by "TDMA" (Tri-Diagonal Matrix Algorithm). The "SIMPLE" algorithm (semi-implicit method for pressure linked equations) was used to handle the coupling between pressure and velocity as described by Patankar [18]. To avoid the divergence in the iteration process, Under-relaxation techniques were used to slow down the change between iterations. It is noted that, although the solid and fluid regions have different equations, the numerical solutions within the computational domain for continuity, momentum and energy equations are obtained simultaneously in both regions.

The effect of the number of grid size ($Nx * Ny$) on the numerical analysis was carried out under the condition of the enclosure filled with FC-77 and $R^*_{al_z} = 10^4$. Three different uniform grid sizes were tested in both x and y directions. The results of dimensionless temperature at solid/fluid interface show that, the mesh size of $36 * 80$ gave an optimum computational time in comparison with other mesh sizes and also gave accurate results.

The code is validated against the benchmark results of Heindel et al. [18]. The validation results show that, dimensionless temperatures at solid/fluid interface have the same trends and the deviation between them decreases with increase of $R^*_{al_z}$. The biggest deviation was at the base of the enclosure with 11 % when $R^*_{al_z} = 10^4$ and the percentage is reduced to 6 % at the top of the enclosure.

5 Results and Discussion

The geometry used in this study is represented in Fig. 1.1. Firstly the numerical study examined the effects of modified Rayleigh number which is based on the

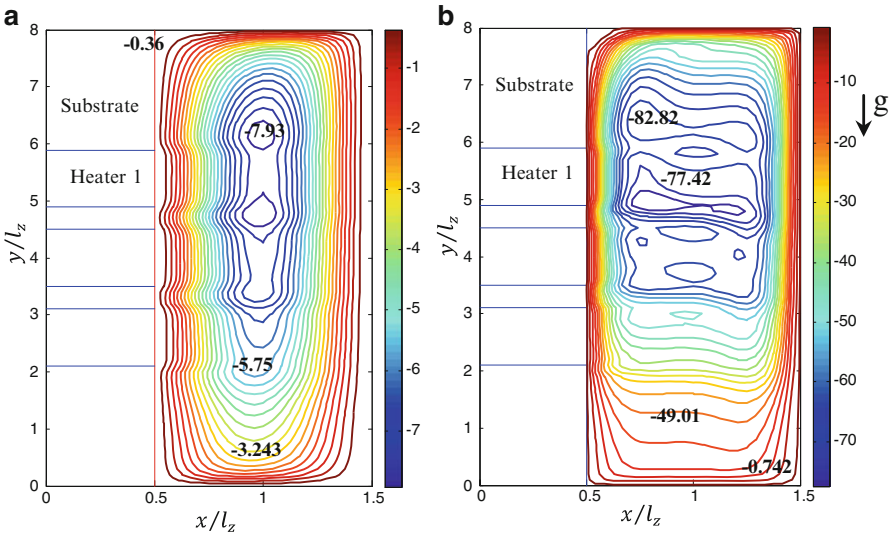


Fig. 1.2 Dimensionless streamlines ψ at (**a**) $R^*_{al_z} = 10^4$ and (**b**) $R^*_{al_z} = 10^7$ for FC-77 $P_r = 25$, $R_h = 2350$ and $R_s = 10$

applied isoflux condition in the range $R^*_{al_z} = 10^4 - 10^7$ for both fluids (FC-77 and air) and then the effect of thermal conductivity ratio R_s was examined in the range $10 - 1000$ for FC-77 only when the modified Rayleigh number fixed at 10^6.

5.1 Effects of Modified Rayleigh Number

Modified Rayleigh number was varied by changing the applied power to each heat source. The heaters' material used for the study was silicon with thermal conductivity ($k_h = 148\ W/mK$) producing constant heat flux and the heater/fluid thermal conductivity ratio is $R_h = 2350$ for FC-77 and $R_h = 5627$ for air. The fluid Prandtl number is assumed to be 25 corresponding to FC-77 and 0.7 for air, respectively. The flow field inside the enclosure is presented by dimensionless stream function as follow:

$$U = \partial\psi/\partial Y, \quad V = -\partial\psi/\partial X \tag{1.14}$$

Figures 1.2 and 1.3 illustrate the dimensionless streamlines for FC-77 and air respectively. Figure 1.2a represents the results for $R^*_{al_z} = 10^4$, where the flow is weak with $|\psi|_{max} = 7.93$, where ψ is an absolute value. There is a small core of nearly stagnant fluid located slightly above the centre of the enclosure, and the heaters regions are very clear.

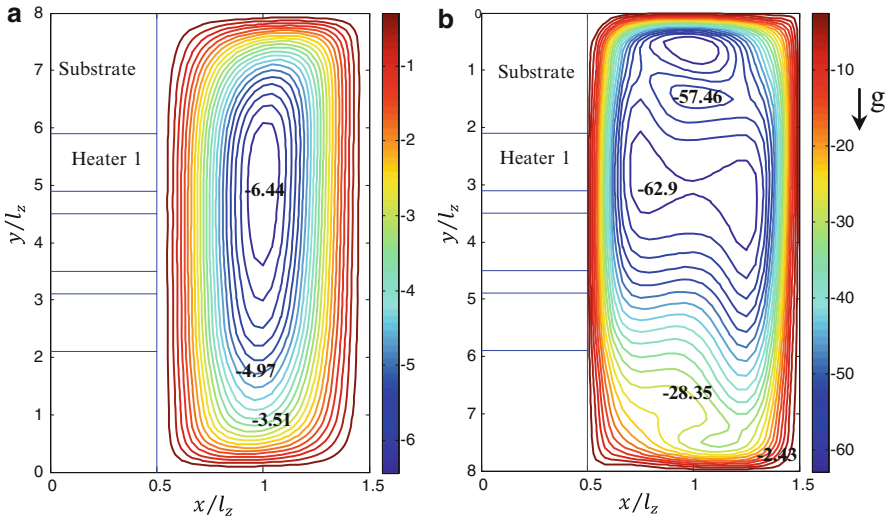


Fig. 1.3 Dimensionless streamlines ψ at (**a**) $R^*_{al_z} = 10^4$ and (**b**) $R^*_{al_z} = 10^7$ for air $P_r = 0.7$, $R_h = 5627$ and $R_s = 24$

Figure 1.3a shows that, the flow inside the enclosure for air is also weak with $|\psi|_{max} = 6.436$ but the flow is smooth. Additionally there is a thick thermal boundary layer along solid/fluid interface wall for both fluids.

It is noticed that, with increasing $R^*_{al_z}$ the boundary layers along hot, cold and horizontal walls become thinner. Also the flow developed in the central region of the cavity and becomes complex. Additionally as a result of the fluid circulation in clockwise direction, the cold fluid swept the hot fluid near the leading edge of each heater row and that is more noticeable for FC-77.

The dimensionless isotherms θ are shown in Figs. 1.4 and 1.5. For both fluids when $R^*_{al_z} = 10^4$, the contour lines of θ is nearly vertical in fluid region because the heat transfer is controlled by conduction. The lines for air are smoother than FC-77 where the effects of heater edges are clear.

With increasing $R^*_{al_z} = 10^7$ the central of fluid region is completely stratified and the thermal boundary layers of hot and cold walls are extremely thin. Because of the large thermal conductivity of heaters, each heater face is isothermal but the temperatures are different from one heater to another.

The dimensionless temperatures (θ) at solid/fluid interface for both fluids are studied and presented in Gdhaidh et al. [19]. The results indicate that, the dimensionless temperature decreases $R^*_{al_z}$ increase due to the increase in $(T - T_c)$ and hence not equal to the increase in q''.

Also the dimensionless local heat flux $\widehat{q}''$ and the local Nusselt number $Nu(X)$ are investigated. In general, both of $\widehat{q}''$ and $Nu(X)$ increase for higher Rayleigh number. For more details, see Gdhaidh et al. [19].

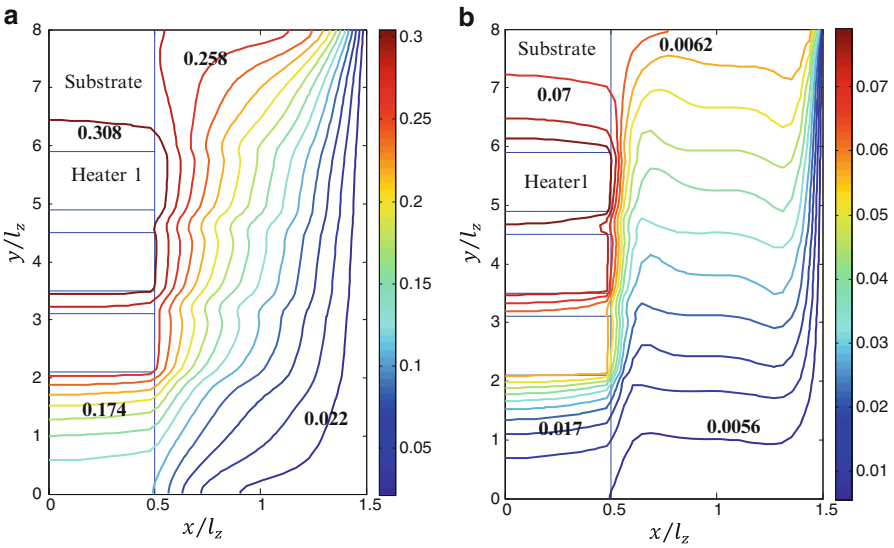


Fig. 1.4 Isotherms θ at (**a**) $R^*_{al_z} = 10^4$ and (**b**) $R^*_{al_z} = 10^7$ for FC-77 $P_r = 25$, $R_h = 2350$ and $R_s = 10$

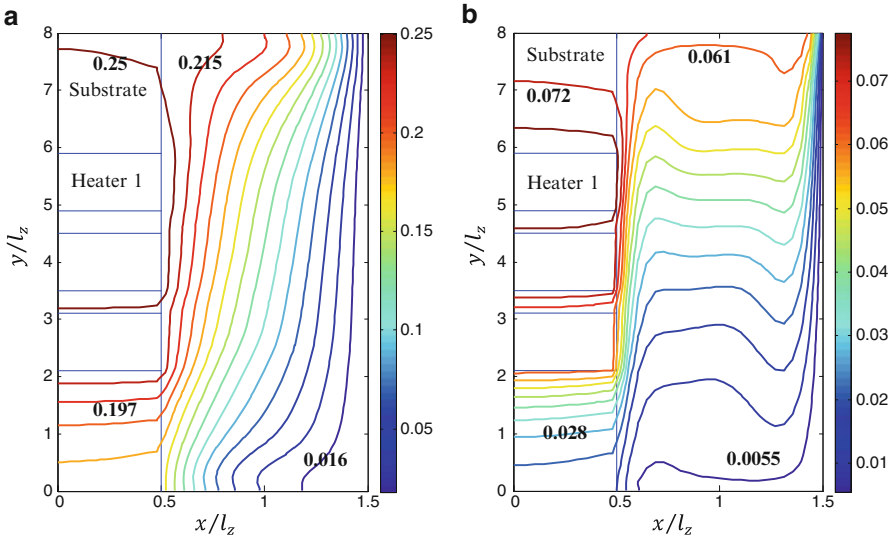


Fig. 1.5 Isotherms θ at: (**a**) $R^*_{al_z} = 10^4$ and (**b**) $R^*_{al_z} = 10^7$ for air $P_r = 0.7$, $R_h = 5627$ and $R_s = 24$

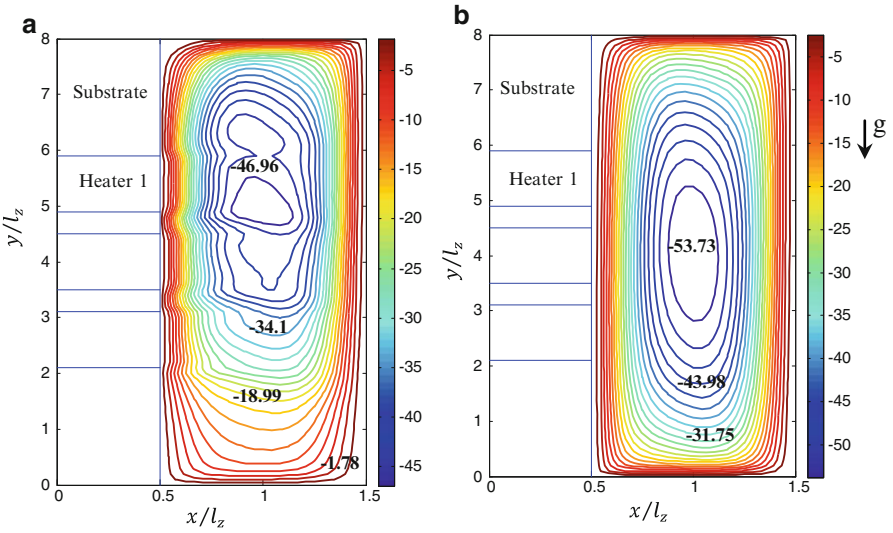


Fig. 1.6 Dimensionless streamline ψ at (**a**) $R_s = 10$ and (**b**) $R_s = 1000$ for FC-77 ($P_r = 25$, $R_h = 2350$ and $R^*_{al_z} = 10^6$)

5.2 *Effects of Substrate/Fluid Thermal Conductivity Ratio*

Further investigation was carried using FC-77 to study the effects of the substrate/fluid thermal conductivity ratio with a fixed modified Rayleigh number $R^*_{al_z} = 10^6$.

Figure 1.6 displays the dimensionless streamlines ψ for different values of R_s. It can be noticed that, the value of ψ increases from $|\psi|_{max} = 46.9$ when $R_s = 10$ to $|\psi|_{max} = 53.7$ when $R_s = 10^3$. As R_s increases, the flow pattern becomes smoother, and the stagnant core is located at the centre of the fluid region. The difference between the two streamlines in Fig. 1.6a, b is due to a higher thermal conductivity ratio for the case presented in Fig. 1.6b. Also the flow regions related to the discrete heaters become indistinguishable at high values of R_s, where the fluid below Heater 3 is preheated causing fluid movement. Additionally the flow is characterised by a clockwise circulation arising from the heaters due to the buoyancy effects.

Figure 1.7 shows the isotherms θ for different values of R_s. From the figure, the heaters' faces are isothermal but there are significant differences between the heaters' temperatures when $R_s = 10$. Those differences are due to the large thermal conductivity of the heaters and low thermal conductivity of the substrate. The heaters' temperature increases from Heater 3 to Heater 1 as a result of the increase in the local fluid temperature along the solid/fluid interface. For high value of $R_s = 10^3$, the maximum dimensionless temperature is decreased as more energy is dissipated by the substrate before being transported to the fluid. Moreover, the

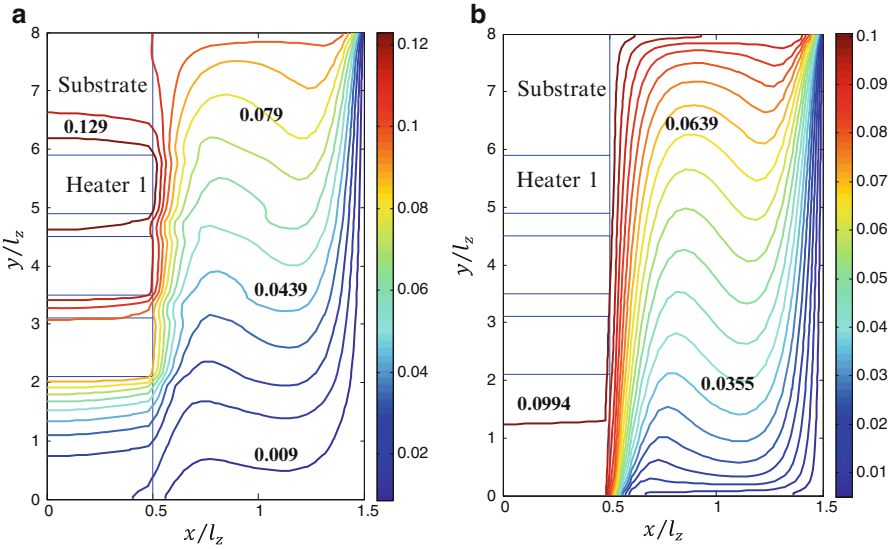


Fig. 1.7 Isotherms θ at (**a**) $R_s = 10$ and (**b**) $R_s = 1000$ for FC-77 ($\mathrm{P_r} = 25$, $R_h = 2350$ and $R^*_{al_z} = 10^6$)

heaters and substrate along the solid/fluid interface are approximately at the same temperature.

The dimensionless temperature distributions at the solid/fluid interface for different values of R_s are shown in Fig. 1.8. Increase in R_s gives another path to dissipate the heat from the heaters where more energy passes through the substrate. When the results from using two values of $R_s = 10$ and $R_s = 10^3$ are compared, there is a noticeable difference between them at the substrate region under Heater 3 where the dimensionless temperature of $R_s = 10^3$ is about 6 times higher than that of $R_s = 10$ at the base of the cavity. This difference decreases as y/l_z increase until $y/L_z = 3.5$, where the values of θ for $R_s = 10$ passes those of $R_s = 10^3$ and the maximum temperature when $R_s = 10^3$ is reduced by 23 %. Furthermore, the dimensionless temperature over the entire cavity height becomes nearly isothermal for large values of R_s.

In most applications, copper and aluminium are the preferred materials for the substrate. These two materials have a large thermal conductivity where the thermal conductivity ratios are 6,350 and 3,970 for copper and aluminium respectively. The results of θ along the solid/fluid interface are converged for both materials, also the heaters and substrate regions cannot be distinguished as shown in Fig. 1.8. Therefore, with very high values of R_s, the effect of thermal conductivity ratio on the dimensionless temperature is disappeared. As a result, the aluminium can be used instead of copper.

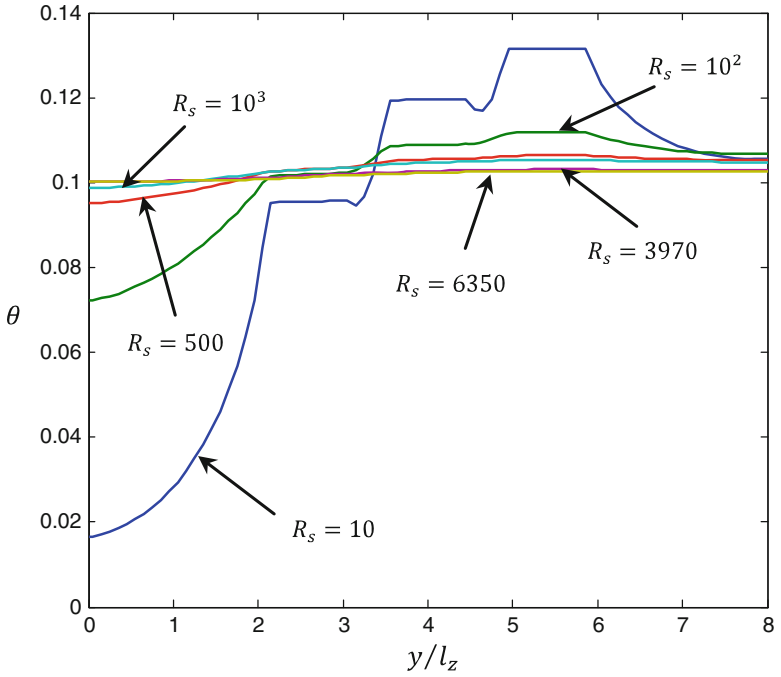


Fig. 1.8 Local dimensionless temperature distribution at the solid/fluid interface for FC-77 ($P_r = 25$, $R_h = 2350$ and $R^*_{al_z} = 10^6$)

5.3 *Temperature Results in Dimensional Form*

The previous results are presented in dimensionless form but it is important to know the value of the temperature in dimensional form. To present the temperature distribution at the solid/fluid interface in dimensional form (°C), as shown in Fig. 1.9 two values of modified Rayleigh number are selected, $R^*_{al_z} = 10^4$ and $R^*_{al_z} = 10^5$, for the given values of $R_s = 10$ and $R_s = 24$ for FC-77 and air respectively.

The value of q″ could be found from modified Rayleigh number equation $R^*_{al_z} = g\beta q'' L_z^4 / k_f \alpha_f \nu$ where the properties of FC-77 and air are known constant and the cold wall has a constant temperature of 20 °C. So the temperature in dimension form can be found from the following relation:

$$T = \frac{\theta q'' L_z}{K_f} + T_c \tag{1.15}$$

Table 1.2 shows the values of heat flux q'' for both working fluids at two selected modified Rayleigh number 10^4 and 10^5.

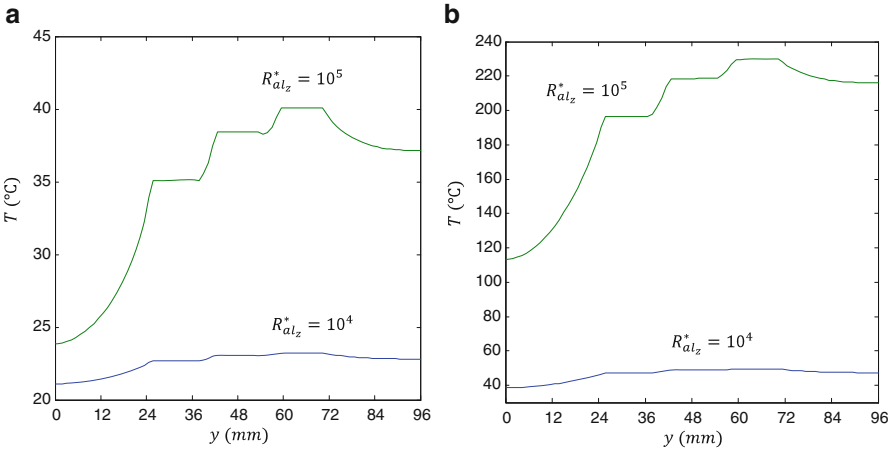


Fig. 1.9 Local dimension temperature distribution at the solid/fluid with different $R^*_{al_z}$: (**a**) for FC-77 and (**b**) for air

Table 1.2 The values of q″ for different modified Rayleigh number for both working fluids

$R^*_{al_z}$	$q''(W/m^2)$ for FC-77	$q''(W/m^2)$ for air
10^4	52.1	240
10^5	521	2400

6 Conclusions and Future Work

Steady state natural conjugate convection analysis for rectangular cavity with discrete heat sources flush mounted on one vertical wall has been conducted numerically. The control volume technique with the "SIMPLE" algorithm is used to simulate the problem. The results from this research show that:

1. At small values of modified Rayleigh number $R^*_{al_z} = 10^4$ the heat transfer is controlled by conduction. With increase $R^*_{al_z}$ the thermal boundary layers of the hot and cold walls become extremely thin and also the flow becomes more complex.
2. The dimensionless temperature θ at the solid/fluid interface for both working fluids (FC-77 and air) decreases as $R^*_{al_z}$ increases, due to the increment in the temperature differences is not equal to the increase in q''.
3. There is a strong effect of the substrate thermal conductivity on the temperature distribution as well as the maximum temperature. With increase R_s, more energy is dissipated by the substrate, which results in a decrease of the maximum temperature level.
4. For high values $R_s > 10^3$, the solid/fluid interface temperature becomes isothermal and the discrete heater locations become almost indistinguishable. Moreover when the $R_s > 1500$ the substrate has no effect on the maximum temperature.

In this study, an isothermal boundary conditions were is used. Such conditions are unrealistic for a normal desktop computer or other electronic devices. To keep one wall or more of the enclosure at a constant temperature, an adjustable water flow with pump and a heat exchanger is required. Moreover, when the ambient temperature is higher than the temperature of the walls, a refrigeration system is needed. Therefore, a large space is necessary and the levels of noise and vibration induced could be increased. In future study the above problem could be overcome by assuming the cold wall as a heat sink that is cooled by the air flow generated from the exhaust fan.

References

1. Bar-Cohen, A.: Physical design of electronic systems–methodology, technical trends, and future challenges. In: Advances in Thermal Modeling of Electronic Components and Systems, pp. 1–60. SME/IEEE Press, New York (1993)
2. Zinnes, A.E.: The coupling of conduction with laminar natural convection from a vertical fiat plate with arbitrary surface heating. J. Heat Transf. **92**, 528–535 (1970)
3. Hasnaoui, M., Bilgen, E., Vasseur, P.: Natural convection heat transfer in rectangular cavities heated from below. J. Thermophys. Heat Transf. **6**, 255–264 (1992)
4. Valencia, A., Frederick, R.L.: Heat transfer in square cavities with partially active vertical walls. Int. J. Heat Mass Transf. **32**, 1567–1574 (1989)
5. Selamet, E.E., Arpaci, V.S., Borgnakke, C.: Simulation of laminar buoyancy driven flows in an enclosure. Numer. Heat Transf. **22**, 401–420 (1992)
6. Sathe, S.B., Joshi, Y.: Natural convection arising from a heat generating substrate-mounted protrusion in a liquid-filled two-dimensional enclosure. Int. J. Heat Mass Transf. **34**, 2149–2163 (1991)
7. Du, Z.G., Bilgen, E.: Coupling of wall conduction with natural convection in a rectangular enclosure. Int. J. Heat Mass Transf. **35**, 1969–1975 (1992)
8. Heindel, T.J., Ramadhyani, S., Incropera, F.P.: Conjugate natural convection from an array of discrete heat sources: part 1 two- and three-dimensional model validation. Int. J. Heat Fluid Flow **16**, 501–510 (1995)
9. Heindel, T.J., Incropera, F.P., Ramadhyani, S.: Enhancement of natural convection heat transfer from an array of discrete heat sources. Int. J. Heat Mass Transf. **39**, 479–490 (1996)
10. Banerjee, S., Mukhopadhyay, A., Sen, S., Ganguly, R.: Natural convection in a bi-heater configuration of passive electronic cooling. Int. J. Therm. Sci. **47**, 1516–1527 (2008)
11. Wroblewski, D.E., Joshi, Y.: Computations of liquid immersion cooling for a protruding heat source in a cubical enclosure. Int. J. Heat Mass Transf. **36**, 1201–1218 (1993)
12. Wroblewski, D.E., Joshi, Y.: Liquid immersion cooling of a substrate-mounted protrusion in a three-dimensional enclosure: the effects of geometry and boundary conditions. J. Heat Transf. **116**, 112–119 (1994)
13. Bilgen, E.: Conjugate heat transfer by conduction and natural convection on a heated vertical wall. Appl. Therm. Eng. **29**, 334–339 (2009)
14. Aminossadati, S.M., Ghasemi, B.: Enhanced natural convection in an isosceles triangular enclosure filled with a nanofluid. Comput. Math. Appl. **61**, 1739–1753 (2011)
15. Sathe, S.B., Joshi, Y.: Natural convection liquid cooling of a substrate-mounted protrusion in a square enclosure: parametric study. Int. J. Heat Transf. **114**, 401–409 (1992)
16. Hyun, M.T., Kim, M.C.: Onset of buoyancy-driven convection in the horizontal fluid layer subjected to time-dependent heating from below. Int. Comm. Heat Mass Transf. **30**, 965–974 (2003)

17. Patankar, S.V.: Numerical Heat Transfer and Fluid Flow. McGraw-Hill, Washington (1980)
18. Heindel, T.J., Incropera, F.P., Ramadhyani, S.: Conjugate natural convection from an array of discrete heat sources: part 2 – a numerical parametric study. Int. J. Heat Fluid Flow **16**, 511–518 (1995)
19. Gdhaidh, F.A., Hussain, K., Qi, H.S.: Numerical investigation of conjugate natural convection heat transfer from discrete heat sources in rectangular enclosure. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 1304–1309. London, 2–4 July (2014)

Chapter 2
Study of Soil-Structure Interaction Problems Using Mixed FEM-BEM Formulations

Dimas Betioli Ribeiro and João Batista de Paiva

Abstract The objective of this paper is to present formulations developed for soil-building interaction analysis, including foundations. The soil is modeled with the boundary element method (BEM) as a layered solid which may be finite for the vertical direction, but is always infinite for radial directions. Infinite boundary elements are employed for the far field simulation, allowing computational cost reduction without compromising the result accuracy. Beams, columns and piles are modeled with the finite element method (FEM) using one dimensional elements. Slabs and rafts are also modeled with the FEM, but with two dimensional elements. The analysis is static and all materials are considered homogeneous, isotropic, elastic and with linear behavior.

Keywords Boundary elements • Finite elements • Soil-structure interaction • Pile • Raft • Slab

1 Introduction

The construction of buildings involve complex soil-structure interaction effects that require previous studies to be correctly considered in the project. The basis of these studies has to be chosen among many options available and each one of them implies on advantages and disadvantages, as described below.

When possible, a good choice is to employ analytical methods. When correctly programmed they give trustful results in little processing time. In Ref. [2], for example, a solution is presented for an axially loaded pile with a rectangular cross section and immersed in a layered isotropic domain. The main disadvantage of

D.B. Ribeiro (✉)
Federal Institute of Education, Science and Technology of São Paulo, Rua Antonio Fogaça de Almeida, s/n, Jardim Elza Maria, Jacareí, SP, Brazil
e-mail: dimasbetioli@gmail.com

J.B. de Paiva
Department of Structural Engineering, School of Engineering of São Carlos, University of São Paulo, Av. Trabalhador Sãocarlense, 400, São Carlos, SP, Brazil
e-mail: paiva@sc.usp.br

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_2

these solutions is that they suit only specific situations, so many researches keep developing new ones to include new problems. Another reference that may be cited is [9].

If analytical solutions cannot be used, one alternative would be a numerical approach. The developments [6] of the numerical methods in the latter years and its versatility made them attractive to many researchers. The finite element method (FEM) is still popular [5, 16], however has some disadvantages when compared to other options such as the boundary element method (BEM). The FEM require the discretization of the domain, which has to be simulated as infinite in most soil-structure interaction problems. This implies on a high number of elements, leading to a large and sometimes impracticable processing time.

It becomes more viable solving these problems with the BEM, once only the boundary of the domains involved is discretized. This allows reducing the problem dimension, implying on less processing time. This advantage is explored in several works [1, 7, 12] and new developments are making the BEM even more attractive to future applications. One is simulating non-homogeneous domains using an alternative multi-domain BEM technique [13], another is using mapping functions to make boundary elements infinite [14].

The objective of this paper is to present a formulation for building-soil interaction analysis that uses recent developments accomplished by the authors in Refs. [13–15]. The proposed formulation is applied into two examples. In the first, a squared raft resting on an infinite layered domain is considered. Results are compared with other formulations available in the literature including an analytical approach and good agreement is observed. The objective of the second example is to show all functionalities of the proposed formulation, considering a complete building interacting with a layered soil. No comparison with other authors is presented, nevertheless the results obtained may be considered coherent. Finally, it is concluded that the presented formulation may be considered a practical and attractive alternative in the field of soil-structure interaction simulation.

2 Boundary Element Formulation

The equilibrium of a solid body can be represented by a boundary integral equation called the Somigliana Identity, which for homogeneous, isotropic and linear-elastic domains is

$$c_{ij}(y)\,u_j(y) + \int_{\Gamma} p^*_{ij}(x,y)\,u_j(x)\,d\Gamma(x) = \int_{\Gamma} u^*_{ij}(x,y)\,p_j(x)\,d\Gamma(x) \quad (2.1)$$

Equation 2.1 is written for a source point y at the boundary, where the displacement is $u_j(y)$. The constant c_{ij} depends on the Poisson ratio and the boundary geometry at y, as pointed out in Ref. [11]. The field point x goes through the whole

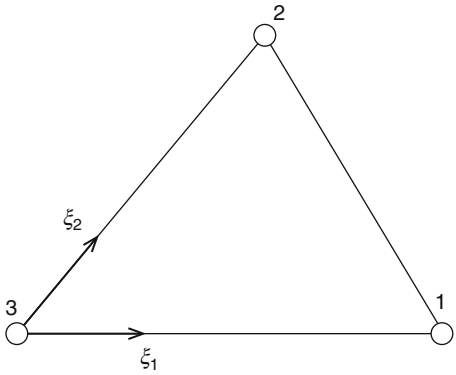


Fig. 2.1 Triangular boundary element

boundary Γ, where displacements are $u_j(x)$ and tractions are $p_j(x)$. The integral kernels $u^*_{ij}(x, y)$ and $p^*_{ij}(x, y)$ are Kelvin three-dimensional fundamental solutions for displacements and tractions, respectively. Kernel $u^*_{ij}(x, y)$ has order $1/r$ and kernel $p^*_{ij}(x, y)$ order $1/r^2$, where $r = |x - y|$, so the integrals have singularity problems when x approaches y. Therefore the stronger singular integral, over the traction kernel, has to be defined in terms of a Cauchy Principal Value (CPV).

To solve Eq. 2.1 numerically, the boundary is divided into regions within which displacements and tractions are approximated by known shape functions. Here these regions are of two types, finite boundary elements (BEs) and infinite boundary elements (IBEs). The BEs employed are triangular, as shown in Fig. 2.1 with the local system of coordinates, $\xi_1\xi_2$, and the local node numbering. The following approximations are used for this BE:

$$u_j = \sum_{k=1}^{3} N^k u_j^k, \quad p_j = \sum_{k=1}^{3} N^k p_j^k \tag{2.2}$$

Equation 2.2 relates the boundary values u_j and p_j to the nodal values of the BE. The BEs have 3 nodes and for each node there are three components of displacement u_j^k and traction p_j^k. The shape functions N^k used for these approximations are

$$N^1 = \xi_1, \quad N^2 = \xi_2, \quad N^3 = 1 - \xi_1 - \xi_2 \tag{2.3}$$

The same shape functions are used to approximate the boundary geometry and to interpolate displacements and tractions for the IBEs. The IBE geometry, on the other hand, is approximated by special mapping functions, as discussed in more detail in Sect. 3. By substituting Eq. 2.2 in 2.1, Eq. 2.4 is obtained:

$$c_{ij}u_j + \sum_{e=1}^{N_{BE}} \left\{ \sum_{k=1}^{3} \left[\Delta p_{ij}^{ek} u_j^k \right] \right\} + \sum_{e=1}^{N_{IBE}} \left\{ \sum_{k=1}^{Np} \left[\Delta^{\infty} p_{ij}^{ek} u_j^k \right] \right\}$$

$$= \sum_{e=1}^{N_{BE}} \left\{ \sum_{k=1}^{3} \left[\Delta u_{ij}^{ek} \, p_j^k \right] \right\} + \sum_{e=1}^{N_{IBE}} \left\{ \sum_{k=1}^{Np} \left[\Delta^{\infty} u_{ij}^{ek} \, p_j^k \right] \right\} \tag{2.4}$$

N_{BE} is the number of BEs and N_{IBE} is the number of IBEs. For BEs:

$$\Delta p_{ij}^{ek} = \int_{\gamma_e} |J| \, N^k \, p_{ij}^* \, (x, y) \, d\gamma_e, \quad \Delta u_{ij}^{ek} = \int_{\gamma_e} |J| \, N^k u_{ij}^* \, (x, y) \, d\gamma_e \tag{2.5}$$

In Eq. 2.5, γ_e represents the domain of element e in the local coordinate system and the global system of coordinates is transformed to the local one by the Jacobian $|J| = 2A$, where A is the element area in the global system. On the other hand, for IBEs:

$$\Delta^{\infty} p_{ij}^{ek} = \int_{\gamma_e} |^{\infty} J| \, N^k \, p_{ij}^* \, (x, y) \, d\gamma_e, \quad \Delta^{\infty} u_{ij}^{ek} = \int_{\gamma_e} |^{\infty} J| \, N^k u_{ij}^* \, (x, y) \, d\gamma_e \tag{2.6}$$

Equation 2.6 is analogous to 2.5. Integrals of Eqs. 2.5 and 2.6 are calculated by standard BEM techniques. Non-singular integrals are evaluated numerically by using integration points. The singular ones, on the other hand, are evaluated by the technique presented in Ref. [10]. Finally, the free term c_{ij} may be obtained by rigid body motions. Writing Eq. 2.4 for all boundary nodes leads to

$$\Delta p \cdot u = \Delta u \cdot p \tag{2.7}$$

3 Infinite Boundary Elements

Three types of mapping are considered, as illustrated in Fig. 2.2.

In the first type, represented in Fig. 2.2a, only direction ξ_1 is mapped to infinity and node 1 is placed at infinity. The IBE is represented in the local coordinate system on the left and in the global coordinate system on the right. The global coordinates x_i are related to the local ones by special mapping functions, M^k, and the nodal global coordinates, x_i^k. Node 4 is created only to replace node 1 in the mapping and does not contribute to the integrals.

Figure 2.2b is analogous to Fig. 2.2a, but in this case only direction ξ_2 is mapped to infinity and node 2 is placed at infinity. Therefore, node 5 is created to facilitate the mapping. Finally, in Fig. 2.2c both local directions are mapped to infinity and nodes 1 and 2 are placed at infinity. As a result, the auxiliary nodes 4 and 5 must be created to replace them in the mapping.

In Ref. [14], auxiliary coordinates $\bar{\xi}_1$ and $\bar{\xi}_2$ are used to obtain the mapping functions for each case. When only direction ξ_1 is mapped to infinity, the result is:

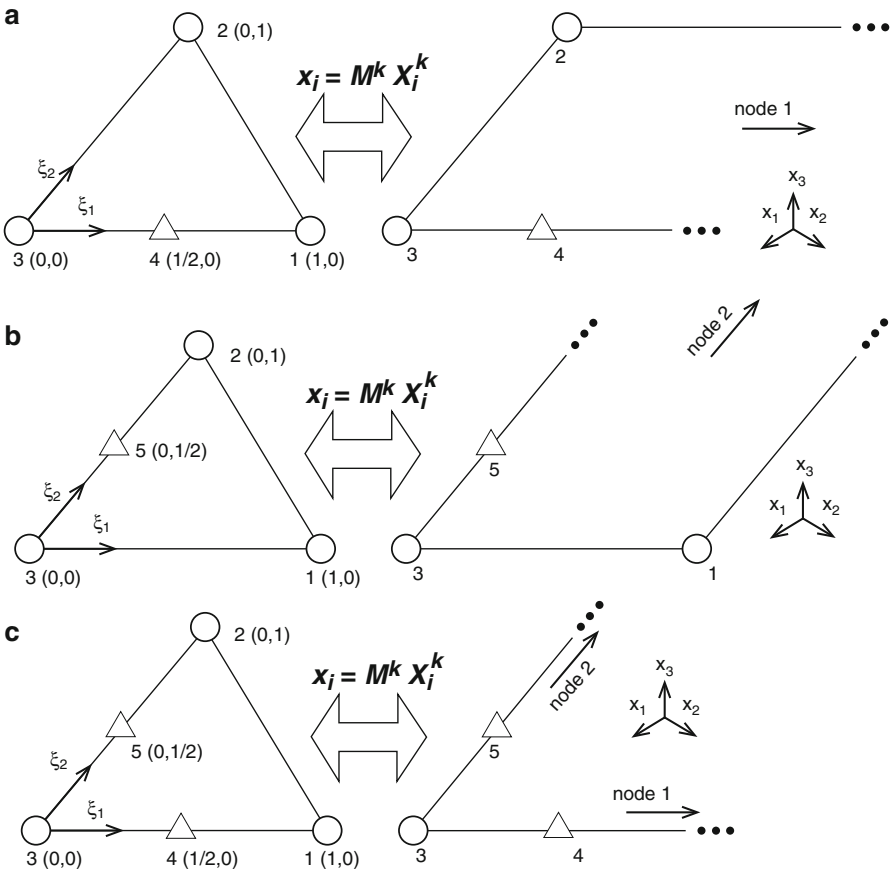


Fig. 2.2 Types of mapping: (**a**) node 1 to infinity, (**b**) node 2 to infinity, (**c**) nodes 1 and 2 to infinity

$$M^4_{1\infty} = \bar{\xi}_1(\xi_1) = \frac{\xi_1}{1-\xi_1},\quad M^2_{1\infty} = \xi_2,\quad M^3_{1\infty} = 1 - \bar{\xi}_1(\xi_1) - \xi_2 = 1 - \frac{\xi_1}{1-\xi_1} - \xi_2 \tag{2.8}$$

The symbol "1∞" is used to indicate that these expressions are valid if only direction ξ_1 is mapped to infinity. These functions are then employed to relate the local system of coordinates to the global one. In other words:

$$x_i = M^4_{1\infty} x_i^4 + M^2_{1\infty} x_i^2 + M^3_{1\infty} x_i^3 \tag{2.9}$$

After obtaining Eq. 2.9, the Jacobian used when only direction ξ_1 is mapped to infinity may be calculated as follows:

$$|^{\infty}J_1| = \frac{\partial x_1}{\partial \xi_1}\frac{\partial x_2}{\partial \xi_2} - \frac{\partial x_2}{\partial \xi_1}\frac{\partial x_1}{\partial \xi_2} = \frac{2A_1}{(1-\xi_1)^2} \tag{2.10}$$

where A_1 is the area of the triangle drawn between nodes 2, 3 and 4 in the global system of coordinates.

For mapping only in direction ξ_2 to infinity, the functions obtained are:

$$M_{2\infty}^1 = \xi_1, \;\; M_{2\infty}^5 = \bar{\xi}_2(\xi_2) = \frac{\xi_2}{1-\xi_2}, \;\; M_{2\infty}^3 = 1-\xi_1-\bar{\xi}_2(\xi_2) = 1-\xi_1-\frac{\xi_2}{1-\xi_2} \tag{2.11}$$

The symbol "2∞" is used to indicate that only direction ξ_2 is mapped to infinity. Therefore, the global system is related to the local one as follows:

$$x_i = M_{2\infty}^1 x_i^1 + M_{2\infty}^5 x_i^5 + M_{2\infty}^3 x_i^3 \tag{2.12}$$

and the Jacobian is

$$|^{\infty}J_2| = \frac{2A_2}{(1-\xi_2)^2} \tag{2.13}$$

where A_2 refers to the area of the triangle drawn between nodes 1, 3 and 5 in the global system of coordinates.

Finally, for mapping in both directions ξ_1 and ξ_2 to infinity, the mapping functions are

$$M_{\infty}^4 = \frac{\xi_1}{1-\xi_1}, \;\; M_{\infty}^5 = \frac{\xi_2}{1-\xi_2}, \;\; M_{\infty}^3 = 1-\frac{\xi_1}{1-\xi_1}-\frac{\xi_2}{1-\xi_2} \tag{2.14}$$

The symbol "∞" is used to indicate that both directions are mapped to infinity. The local system of coordinates is related to the global one as follows:

$$x_i = M_{\infty}^4 x_i^4 + M_{\infty}^5 x_i^5 + M_{\infty}^3 x_i^3 \tag{2.15}$$

and the Jacobian is now

$$|^{\infty}J_3| = \frac{2A_3}{(1-\xi_1)^2(1-\xi_2)^2} \tag{2.16}$$

where A_3 is the area of the triangle drawn between nodes 3, 4 and 5 in the global system.

4 Load Lines in the Soil

In this work, the reactive tractions from the piles are applied in the soil as load lines. Figure 2.3 presents the model adopted, with four nodes equally spaced along the pile. The load lines influence may be computed in Eq. 2.1 with an additional term:

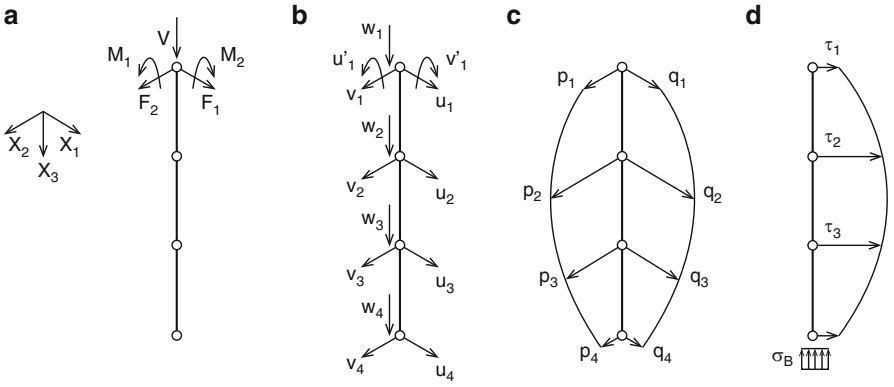


Fig. 2.3 Model for load lines: (**a**) possible loads at the top node, (**b**) degrees of freedom, (**c**) horizontal tractions, (**d**) vertical tractions

$$c_{ij}(y)\,u_j(y) + \int_{\Gamma} p^*_{ij}(x,y)\,u_j(x)\,d\Gamma(x) = \int_{\Gamma} u^*_{ij}(x,y)\,p_j(x)\,d\Gamma(x) + \sum_{e=1}^{nl}\left[\int_{\Gamma^e} u^*_{ij}(x,y)\,s^e_j(x)\,d\Gamma^e(x)\right] \tag{2.17}$$

where nl is the number of load lines, Γ^e are their external surface and s^e_j are the tractions presented in Fig. 2.3c, d. The tractions are approximated from the nodal values s^{ek}_j using nf polynomial shape functions ϕ:

$$s^e_j = \sum_{k=1}^{nf} \phi^k s^{ek}_j \tag{2.18}$$

Shape functions are written with a dimensionless coordinate $\xi = {}^{2x_3}/_{L} - 1$, where L is the load line length and x_3 is the vertical global coordinate. For the horizontal tractions, illustrated in Fig. 2.3c, $nf = 4$ and the shape functions are:

$$\begin{aligned} \phi^1 &= \tfrac{1}{16}\left(-9\xi^3 + 9\xi^2 + \xi - 1\right), \quad \phi^2 = \tfrac{1}{16}\left(27\xi^3 - 9\xi^2 - 27\xi + 9\right), \\ \phi^3 &= \tfrac{1}{16}\left(-27\xi^3 - 9\xi^2 + 27\xi + 9\right), \quad \phi^4 = \tfrac{1}{16}\left(9\xi^3 + 9\xi^2 - \xi - 1\right) \end{aligned} \tag{2.19}$$

For shear tractions in direction x_3, $nf = 3$ and the shape functions are

$$\phi^1 = \frac{1}{8}\left(9\xi^2 - 1\right), \quad \phi^2 = \frac{1}{4}\left(-9\xi^2 - 6\xi + 3\right), \quad \phi^3 = \frac{1}{8}\left(9\xi^2 + 12\xi + 3\right) \tag{2.20}$$

Finally, for the base reaction $nf = 1$ and a constant approximation is used. The integrals that are not singular may be numerically calculated using Gauss points, while singular ones are analytically evaluated. Writing Eq. 2.20 for all boundary points plus the points defined on each load line, it is obtained:

$$[H]\{u\} = [G]\{p\} + [M]\{s\} \tag{2.21}$$

Matrix $[M]$ is obtained from the integrals calculated for all load lines, and vector $\{s\}$ contains the tractions prescribed for them. As the number of equations is equal to the number of unknowns, the system may be solved obtaining all unknowns.

5 FEM-BEM Coupling

Each pile is modeled using a single finite element with polynomial shape functions. Lateral displacements are approximated using fourth degree polynomials $\{\varphi\}$. Vertical displacements and lateral tractions are approximated using third degree polynomials $\{\phi\}$. Vertical tractions are approximated using second degree polynomials $\{\omega\}$ and the tractions at the pile base are considered constant. Using a dimensionless coordinate $\xi = \frac{x_3}{L}$, where x_3 is the global vertical coordinate and L is the pile length, $\{\varphi\}$, $\{\phi\}$ and $\{\omega\}$ may be written as:

$$\{\varphi\} = \begin{Bmatrix} -\frac{99}{4}\xi^4 + 45\xi^3 - \frac{85}{4}\xi^2 + 1 \\ -\frac{9}{2}\xi^4 L + 9\xi^3 L - \frac{11}{2}\xi^2 L + \xi L \\ \frac{81}{2}\xi^4 - \frac{135}{2}\xi^3 + 27\xi^2 \\ -\frac{81}{4}\xi^4 + 27\xi^3 - \frac{27}{4}\xi^2 \\ \frac{9}{2}\xi^4 - \frac{9}{2}\xi^3 + \xi^2 \end{Bmatrix}, \quad \{\phi\} = \begin{Bmatrix} -\frac{9}{2}\xi^3 + 9\xi^2 - \frac{11}{2}\xi + 1 \\ \frac{27}{2}\xi^3 - \frac{45}{2}\xi^2 + 9\xi \\ -\frac{27}{2}\xi^3 + 18\xi^2 - \frac{9}{2}\xi \\ \frac{9}{2}\xi^3 - \frac{9}{2}\xi^2 + \xi \end{Bmatrix},$$

$$\{\omega\} = \begin{Bmatrix} \frac{9}{2}\xi^2 - \frac{9}{2}\xi + 1 \\ -9\xi^2 + 6\xi \\ \frac{9}{2}\xi^2 - \frac{3}{2}\xi \end{Bmatrix} \tag{2.22}$$

The next step is obtaining the total potential energy function, considering internal and external contributions. To obtain the final system of equations, such function must be minimized with respect to the nodal parameters. The result is:

$$[K]\{u\} = \{f\} - [Q]\{y\} \rightarrow [K]\{u\} = \{f\} - \{r\} \tag{2.23}$$

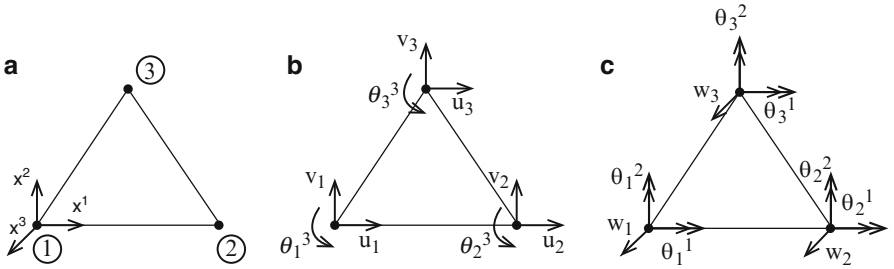


Fig. 2.4 Triangular finite element: (**a**) local node numbering, (**b**) in plane degrees of freedom, (**c**) out of plane degrees of freedom

where $[K]$ is the stiffness matrix of the pile, $\{u\}$ contains nodal displacements, $\{f\}$ contains nodal loads, $\{y\}$ contains distributed tractions and $[Q]$ is a matrix that transforms distributed tractions into nodal loads. Therefore, $\{r\}$ contains nodal loads that represent the distributed loads.

Now a brief description of the triangular finite element used for the raft and slabs will be presented. The element has three nodes at its vertices as presented in Fig. 2.4a with the local node numbering and a local rectangular system of coordinates x^i, where the superscript i indicates the direction. Each node, indicated with the subscript j, has six degrees of freedom (DOFs). Three of them, u_j, v_j and θ_j^3, may be visualized in Fig. 2.4b which refers to the membrane effects. The other three, w_j, θ_j^1 and θ_j^2, are presented in Fig. 2.4c which refers to the plate effects. In Fig. 2.4c, rotational DOFs are indicated with a double arrow for better visualization. All DOFs of the finite element may be arranged into three vectors, as shown below:

$$\{\mathbf{u}_1\}^T = \left\{u_1\ v_1\ {\theta_1}^3\ w_1\ {\theta_1}^1\ {\theta_1}^2\right\},\ \ \{\mathbf{u}_2\}^T = \left\{u_2\ v_2\ {\theta_2}^3\ w_2\ {\theta_2}^1\ {\theta_2}^2\right\},$$
$$\{\mathbf{u}_3\}^T = \left\{u_3\ v_3\ {\theta_3}^3\ w_3\ {\theta_3}^1\ {\theta_3}^2\right\} \tag{2.24}$$

Displacements at any point P of the finite element, with coordinates x_1, x_2 and x_3, may be written as

$$\{\mathbf{u}\} = \begin{Bmatrix} u \\ v \\ w \end{Bmatrix} = \begin{Bmatrix} u_0 - x_3 \frac{\partial w_0}{\partial x_1} \\ v_0 - x_3 \frac{\partial w_0}{\partial x_2} \\ w_0 \end{Bmatrix} \tag{2.25}$$

where u_0, v_0, and w_0 are the displacements for the projection of P at the mid plane of the finite element. The strain field may be obtained from the displacements:

$$\{\varepsilon\} = \{\varepsilon_m\} + \{\varepsilon_p\} = \begin{Bmatrix} \frac{\partial u_0}{\partial x_1} \\ \frac{\partial v_0}{\partial x_2} \\ \frac{\partial u_0}{\partial x_2} + \frac{\partial v_0}{\partial x_1} \end{Bmatrix} - x_3 \begin{Bmatrix} \frac{\partial^2 w_0}{\partial {x_1}^2} \\ \frac{\partial^2 w_0}{\partial {x_2}^2} \\ 2\frac{\partial^2 w_0}{\partial x_1 \partial x_2} \end{Bmatrix} \tag{2.26}$$

where index m corresponds to the membrane effect and the index p indicates the plate effect. Equation 2.26 relates the strain field to the displacement field, which may be related to the nodal displacements using the element shape functions. Using these functions and Eq. 2.26, it is possible to relate strains with the DOFs of the finite element as follows:

$$\{\varepsilon\} = [B] \begin{Bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{Bmatrix} \tag{2.27}$$

It is also necessary to relate strains with stresses. For linear elasticity this may be done using a matrix $[D]$ which is obtained from Hooke's law:

$$\{\sigma\} = [D]\{\varepsilon\} \tag{2.28}$$

In the end, the stiffness matrix of the element is obtained by integrating the domain Ω of the element:

$$[K] = \int_{\Omega} [B]^T [D][B]\, d\Omega \tag{2.29}$$

More detail about the membrane and plate effects of this element may be consulted in Refs. [3, 4], respectively.

All finite element contributions, including piles and the raft, are assembled to the same system of equations. This system has the form of Eq. 2.23, which is later used to demonstrate how the FEM/BEM coupling is made. The starting point is Eq. 2.21, which may be rewritten as:

$$[H]\{u\} = [T]\{y\} \tag{2.30}$$

Matrix $[T]$ contains the terms of matrices $[G]$ and $[M]$ and $\{y\}$ contains the distributed loads of vectors $\{p\}$ and $\{s\}$. Next step is isolating the distributed loads, which are transformed in nodal loads using a matrix $[Q]$.

$$[T]^{-1}[H]\{u\} = \{y\} \rightarrow [B]\{u\} = \{y\} \tag{2.31}$$

$$[Q][B]\{u\} = [Q]\{y\} \rightarrow [D]\{u\} = \{r\} \tag{2.32}$$

Before relating Eqs. 2.23 and 2.32, they must be expanded as to contain all degrees of freedom defined in the coupled FEM-BEM model. The result is

$$\left[\bar{K}\right]\{\bar{u}_{FEM}\} = \left\{\bar{f}\right\} - \{\bar{r}_{FEM}\}, \quad \left[\bar{D}\right]\{\bar{u}_{BEM}\} = \{\bar{r}_{BEM}\} \tag{2.33}$$

These equations are related by imposing compatibility and equilibrium conditions, which are $\{\bar{u}_{FEM}\} = \{\bar{u}_{BEM}\} = \{\bar{u}\}$ and $\{\bar{r}_{FEM}\} = \{\bar{r}_{BEM}\} = \{\bar{r}\}$. The following expression is then obtained:

$$\left[\bar{K}\right]\{\bar{u}\} = \left\{\bar{f}\right\} - \left[\bar{D}\right]\{\bar{u}\} \rightarrow \left(\left[\bar{K}\right] - \left[\bar{D}\right]\right)\{\bar{u}\} = \left\{\bar{f}\right\} \rightarrow \left[\bar{A}\right]\{\bar{u}\} = \left\{\bar{f}\right\} \quad (2.34)$$

where $\{\bar{u}\}$ contain all unknown displacements of the FEM-BEM model. Once the number of equations is equal to the number of unknowns, the system may be solved obtaining all unknowns.

6 Examples

6.1 Raft on a Layered Domain

Here a squared raft over a domain with four layers is considered, as presented in Fig. 2.5 with all geometrical and material parameters. Young's module and Poisson ratio are named E and ν, respectively, while thickness is indicated as t. The subscript R is used for the raft and numbers are used for the layers. Point A is at the center of the raft, point B is at the side midpoint and an uniform 0,1 MPa load is applied over it.

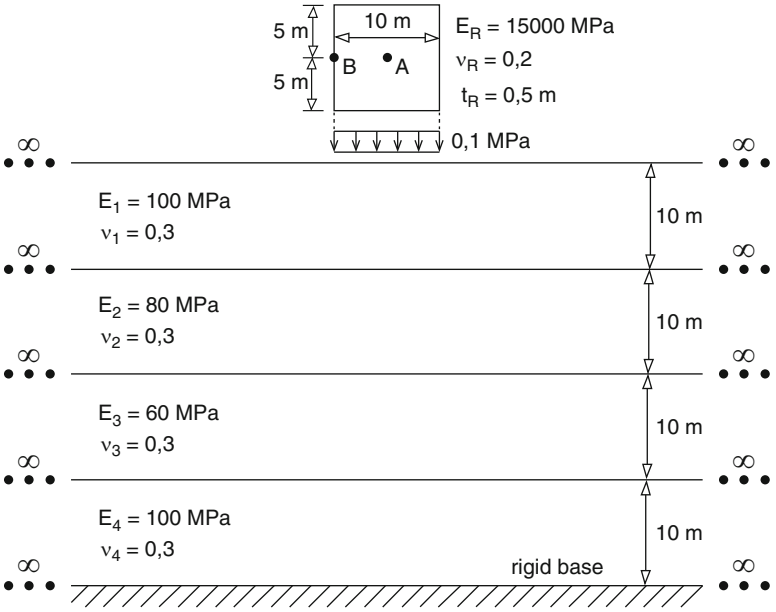


Fig. 2.5 First example

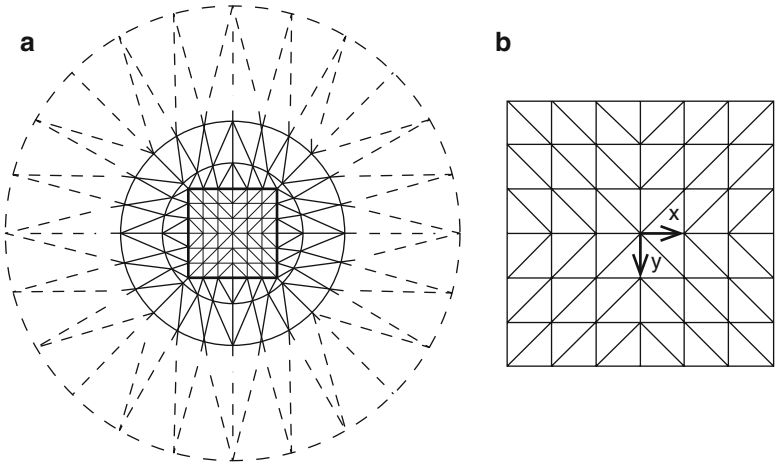


Fig. 2.6 Mesh employed: (**a**) soil surface, (**b**) raft

Table 2.1 Vertical displacements

	d_A (cm)	d_B (cm)
This work	0,97	0,74
[18]	1,07	0,78
[8]	1,14	0,87
[17]	1,20	0,89

Figure 2.6 presents the mesh. The mesh for the surface and contacts between layers is presented in Fig. 2.6a, where dashed lines represent 48 IBEs and other ones represent 168 BEs. Figure 2.6b presents the 72 FE mesh employed for the raft with an xy system of coordinates.

Displacements obtained for points A and B are presented in Table 2.1 with results obtained by other authors. Good agreement may be observed. To complement this example, the bending moment for axis x was calculated. The values obtained for points A and B were, respectively, $3,83 \times 10^{-2}$ kNm/m and $2,20 \times 10^{-2}$ kNm/m.

6.2 Building Resting on a Layered Domain

The objective of this example is to demonstrate the generality of the presented formulation. The problem to be analyzed is presented in Fig. 2.7 and considers a building with its foundations, resting on a layered media. In Fig. 2.7a the lateral view is illustrated, Fig. 2.7b contains the standard floor considered and in Fig. 2.7c is presented the top view of the structural foundations included.

The Poisson Ratio is zero for all soil layers. The elasticity modulus of the layers is 60 MPa for the top one, 80 MPa for the second and 90,000 MPa for the base layer.

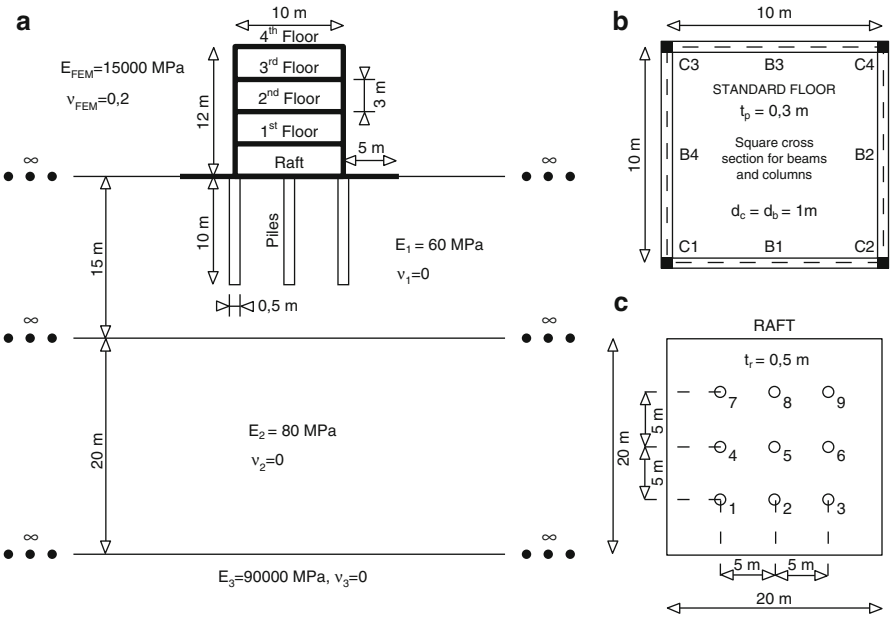


Fig. 2.7 Soil-building interaction: (**a**) vertical cut, (**b**) standard floor, (**c**) raft

The thickness is 15 m for the top layer, 20 m for the second and the base layer is considered infinite. The diameter of all piles is 0, 5 m, their length is 10 m and they are spaced of 5 m. The square raft has size 20 m and thickness 0, 5 m. The elasticity modulus of all materials modeled with the FEM is 15,000 MPa and their Poisson ratio is 0, 2. This includes all piles, beams, columns, slabs and the raft.

The building has four floors, as shown in Fig. 2.7a. All floors have the same standard geometry, as presented in Fig. 2.7b, with a slab with thickness 0, 3 m, four beams supporting this slab and four columns supporting the beams. A cross section $1 \times 1\,m^2$ is used for all beams and columns. The base of each column is connected to the raft at the same node where a corner pile is connected. Corner piles are numerated in Fig. 2.7c as 1, 3, 7 and 9.

The loads considered are vertical and presented in Fig. 2.8. They are uniformly distributed over the slabs, with an intensity of 0, 04 MPa.

Figure 2.9 presents the FE-BE-IBE mesh employed in the example. Figure 2.9a contains the top view of the mesh used for the soil surface and contacts between layers, totalizing 480 BEs and 96 IBEs. The square detached at the center indicates the position of the raft at the surface. In Fig. 2.9b is illustrated the mesh with 32 FEs used for the raft, together with the position of the piles. Finally, Fig. 2.9c contains the 32 FE mesh used for the slabs. Lines detached at the boundary indicate the FEs used for beams, totalizing 16 FEs for each floor. Furthermore, each part of the columns between floors is divided into 4 FEs. Considering all floors plus the raft,

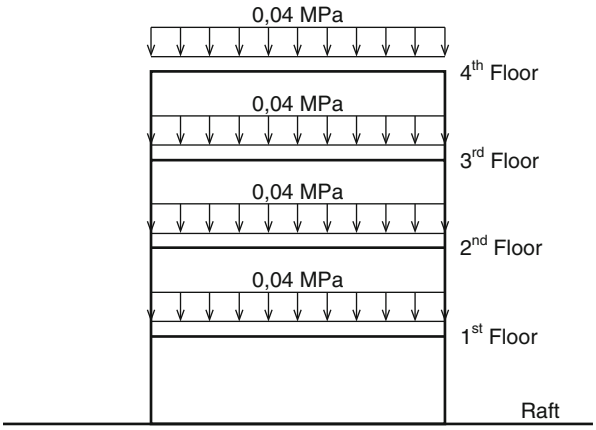


Fig. 2.8 Vertical loads applied

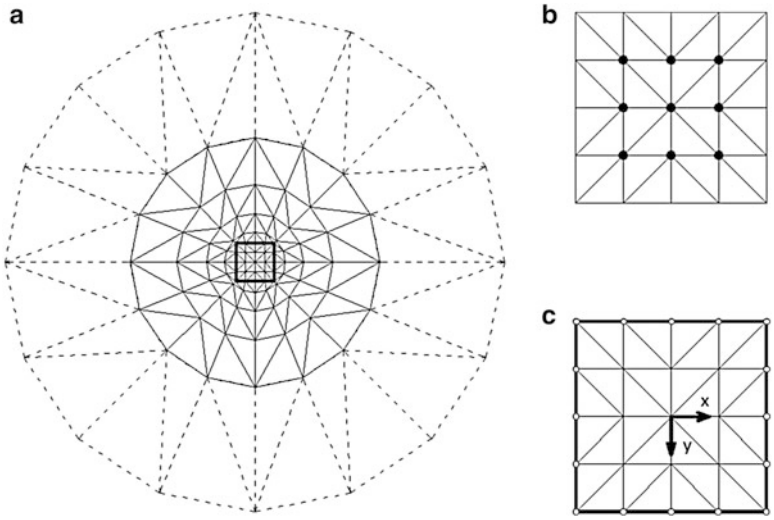


Fig. 2.9 FE/BE/IBE mesh employed: (**a**) soil surfasse, (**b**) raft, (**c**) standard floor

the total number of two-dimensional FEs is 160 and the total of one-dimensional FEs is 128.

Piles are also simulated with the FEM, employing the FE with 14 parameters presented previously. The axis of any pile is orthogonal to the surface of the soil.

The vertical displacements along the axis of the piles are shown in Fig. 2.10. Only piles number 1, 2 and 5 are presented (see Fig. 2.7c) because the results are symmetric for the other ones. Piles placed at corners presented higher displacements, with the value of 12, 2 mm at the top. This result may be considered coherent because the base of the columns of the building is placed exactly over the corner piles. The

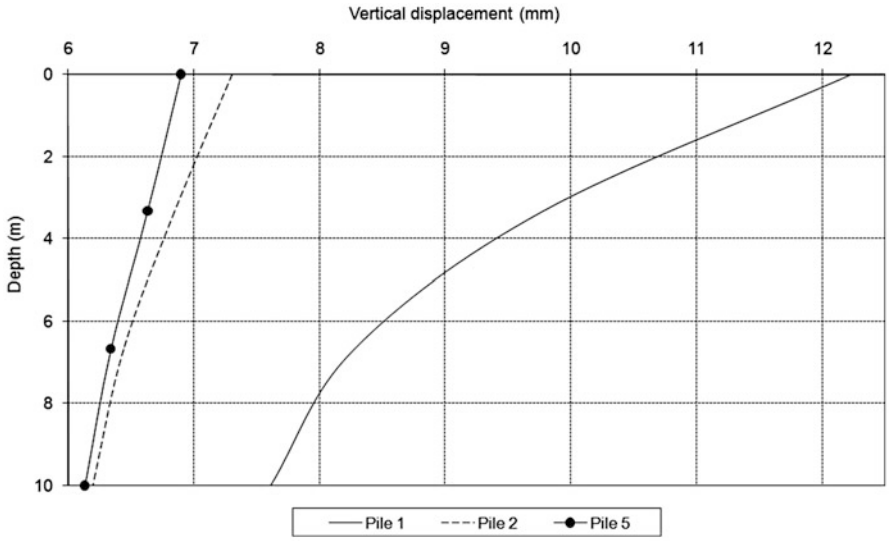


Fig. 2.10 Vertical displacement at slabs

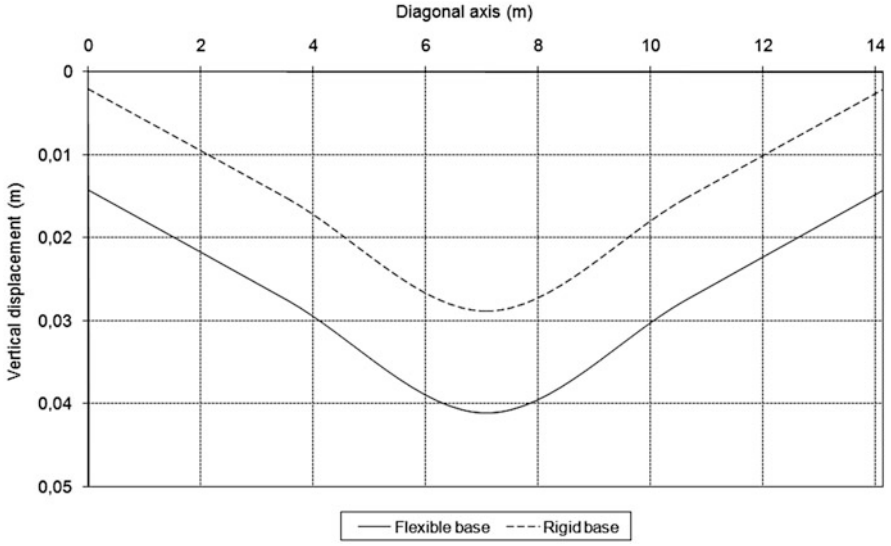


Fig. 2.11 Vertical displacement at the fourth floor

vertical displacement obtained at the top of pile 2 was 7, 3 mm and for pile 5 it was 6, 9 mm.

Figure 2.11 contains results calculated for the fourth floor, as numbered in Fig. 2.7a. Values are evaluated along a diagonal line on the floor, which extremes are placed at columns $C1$ and $C4$, as shown in Fig. 2.7b. Two results are presented.

One is considering the elastic foundation presented in Fig. 2.7a and the other is considering a rigid base. For the rigid base, displacements at the base of the columns are simply restrained. It may be observed that the vertical displacements of the slab are significantly higher when an elastic base is considered, what may be considered a predictable result. The maximum displacement at the slab for a rigid base is 28, 8 mm and for an elastic foundation it is 41, 0 mm.

7 Conclusions

In this paper a formulation for building-soil interaction analysis was presented. The FEM/BEM equations together with the techniques from Refs. [13, 14] contributed with reducing the total number of degrees of freedom. Piles are modeled using one-dimensional FEs, whose influence in the soil is computed by integrating load lines. Two examples were presented. On the first one the values obtained were compared with other publications and good agreement was observed. On the second one no comparison was presented, nevertheless the results obtained were considered coherent. In the end, it may be concluded that the presented formulation is a powerful and attractive alternative for soil-structure interaction analysis.

Acknowledgements Thanks are due to the research council FAPESP, the University of São Paulo and the Federal Institute of São Paulo.

References

1. Ai, Z.Y., Cheng, Y.C.: Analysis of vertically loaded piles in multi-layered transversely isotropic soils by BEM. Eng. Anal. Bound Elem. **37**, 327–335 (2013)
2. Basu, D., Prezzi, M., Salgado, R., Chakraborty, T.: Settlement analysis of piles with rectangular cross sections in multi-layered soils. Comput. Geotech. **35**, 563–575 (2008)
3. Batoz, J.L.: A study of tree-node triangular plate bending elements. Int. J. Numer. Methods Eng. **15**, 1771–1812 (1980)
4. Bergan, P.G., Felippa, C.A.: A triangular membrane element with rotacional degrees of freedom. Comput. Methods Appl. Mech. Eng. **50**, 25–69 (1985)
5. Bourgeois, E., de Buhan, P., Hassen, G.: Settlement analysis of piled-raft foundations by means of a multiphase model accounting for soil-pile interactions. Comput. Geotech. **46**, 26–38 (2012)
6. Clouteau, D., Cottereau, R., Lombaert, G.: Dynamics of structures coupled with elastic media – a review of numerical models and methods. J. Sound Vib. **332**, 2415–2436 (2013)
7. Elahi, H., Moradi, M., Poulos, H.G., Ghalandarzadeh, A.: Pseudostatic approach for seismic analysis of pile group. Comput. Geotech. **37**, 25–39 (2010)
8. Fraser, R.A., Wardle, L.J.: Numerical analysis of rectangular rafts on layered foundations. Geotechnique **26**, 613–630 (1976)
9. Georgiadis, K., Georgiadis, M., Anagnostopoulos, C.: Lateral bearing capacity of rigid piles near clay slopes. Soils Found. **53**, 144–154 (2013)

10. Guiggiani, M., Gigante, A.: A general algorithm for multidimensional cauchy principal value integrals in the boundary element method. J. Appl. Mech. **57**, 906–915 (1990)
11. Moser, W., Duenser, C., Beer, G.: Mapped infinite elements for three-dimensional multi-region boundary element analysis. Int. J. Numer. Methods Eng. **61**, 317–328 (2004)
12. Padron, L.A., Aznarez, J.J., Maeso, O.: 3-D boundary element-finite element method for the dynamic analysis of piled buildings. Eng. Anal. Bound. Elem. **35**, 465–477 (2011)
13. Ribeiro, D.B., Paiva, J.B.: An alternative multi-region BEM technique for three-dimensional elastic problems. Eng. Anal. Bound. Elem. **33**, 499–507 (2009)
14. Ribeiro, D.B., Paiva, J.B.: Analyzing static three-dimensional elastic domains with a new infinite boundary element formulation. Eng. Anal. Bound. Elem. **34**, 707–713 (2010)
15. Ribeiro, D.B., Paiva, J.B.: Mixed FEM-BEM formulations applied to soil-structure interaction problems. In: Proceedings of the World Congress on Engineering 2014 (WCE 2014), London, 2–4 July 2014. Lecture Notes in Engineering and Computer Science, pp. 1178–1183 (2014)
16. Su, D., Li, J.H.: Three-dimensional finite element study of a single pile response to multidirectional lateral loadings incorporating the simplified state-dependent dilatancy model. Comput. Geotech. **50**, 129–142 (2013)
17. Wang, Y.H., Tham, L.G., Tsui, Y., Yue, Z.Q.: Plate on layered foundation analyzed by a semi-analytical and semi-numerical method. Comput. Geotech. **30**, 409–418 (2003)
18. Wardle, L.J., Fraser, R.A.: Finite element analysis of a plate on a layered cross-anisotropic foundation. In: Proceedings of the First International Conference of Finite Element Methods in Engineering, University of New South Wales, pp. 565–578 (1974)

Chapter 3
A Study of the Reliability of Electronic Telecommunication Systems Working at Subsea Level

Sabuj Mallik and Franziska Kaiser

Abstract Reliability is of increasing importance for electronics systems operating at harsh environments, such as the electronic telecommunication systems used at subsea level. The aim of this research was to investigate the reliability of such electronic systems through a simulated accelerated thermal cycle test. The paper presents a step-by-step process of designing accelerated thermal cycle test using field operating conditions. The Coffin-Mansion equation was used to calculate the accelerated factor for the thermal cycle test. In order to simulate the expected life time of 25 years, the solder assembly samples were subjected to 400 temperature cycles, with every cycle lasting for 40 min. Reliability was determined by measuring shear strengths of solder joints of different electronic components at set intervals. Although some of the components showed an initial decrease in shear strength, it was generally concluded that the electronic assemblies are able to maintain their shear strength for up to 25 years. The fracture surfaces of the solder joints, after shear testing, were also analyzed for brittle and ductile fractures, with the use of scanning electron microscopy (SEM).

Keywords Failures • Fractures • Reliability • Shear strength • Solder joint • Thermal cycling

1 Introduction

In the current global competitive market it is crucial to make highly reliable products. This in turn will reduce product cost by having less warranty claims and low repair costs. Furthermore, in some application areas such as under water locations it is impossible or rather unaffordable to repair or change a faulty component.

S. Mallik (✉) • F. Kaiser
Mechanical, Manufacturing and Design Engineering, Faculty of Engineering and Science, University of Greenwich, Central Avenue, Chatham, Kent ME4 4TB, UK
e-mail: S.Mallik@gre.ac.uk; KaiserFranziska@web.de

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_3

The solder joints are the weakest part of an electronics manufacturing assembly. Due to this it is important to ensure that the solder joints are reliable for the expected lifetime of the assembly. In this research study accelerated ageing of the product by thermal cycling was used to investigate solder joint reliability. Accelerated life testing is still a relatively new subject in reliability engineering, but it is starting to gain greater acceptance in the industry. Forcing the product to fail quickly reduces test time and still allows understanding of the life characteristics of products. The main focus of this research was to study the reliability of electronic telecommunication products that operates at the ground of the Atlantic Ocean. The objectives of the study are threefold: to design accelerated thermal cycling tests by taking account of operating conditions and expected product life time; to evaluate the shear strength of solder joints for different surface mount components, at different stages of thermal cycling; and to examine the shear-fractured board surface for brittle and ductile fractures using scanning electron microscope.

2 Solder Joint Failure

Solder joints refer to the solder connections between a semiconductor package and the substrate board on which it is mounted. Solder joint failures occur for various reasons, such as – (a) poor solder joint design, (b) poor solder joint processing, (c) solder material issues, (d) excessive stresses applied to the solder joints.

Solder joint failures can be generally classified in terms of the nature of the stresses that induce the failures and the manner in which the solder joints fail. As such, solder joint failures fall under three major categories: (a) tensile fracture due to stress overloading, which is short-term, (b) creep failure due to the application of a long-term, permanent load, (c) fatigue failure due to the application of cyclical stresses. Of course, more than one of these stresses can act on a solder joint in a given situation. Added to this is the fact that solder joint degradation due to other factors such as corrosion can also occur [1].

2.1 Solder Joint Fracture due to Stress Overloading

Solder joint fractures attributed to short-term stress overloading are mainly those experienced by units subjected to gross mishandling or misprocessing, especially after these units have been mounted on the application board. This is often a result of an accident or harsh treatment. These cases bring the parts to thermo-mechanical stress levels that exceed the fracture strength of the solder joints, resulting in solder joint failures [1]. Real-world examples of events that lead to solder joint fracture due to mechanical overloading include: (a) dropping of the assembled electronic board (or final product) to the floor, (b) ‘force-fitting’ of an improperly loaded electronic

board into its module or enclosure and (c) high-impact collisions involving the module containing the electronic board. These incidents subject the solder joints of the device to very high shear stresses that tend to rip them away from their board.

2.2 Solder Joint Failure due to Creep

Solder joints that are subjected to permanent mechanical loading degrade over time and eventually fail. This failure phenomenon is known as creep. Creep is more pronounced at higher temperatures, although solder joint failures due to creep at room temperature can occur. Indeed, solder joints do operate at quite high temperature. For example, the typical value of solder joint homologous temperature (ratio of service temperature and melting temperature) is 0.64, which is above the critical creep value of 0.6 [2]. One potential instance of fracture caused mainly by creep occurs when an assembly has been insufficiently supported during soldering, particularly during the reflow stage. After soldering, the board may have a rather large permanent warp, but the joints are in an almost stress free condition. However, if this board is then screwed into an enclosure, whilst being firmly forced flat, very large forces are exerted on the joints. This may cause cracks during the mounting operation (overloading) or soon thereafter [3].

2.3 Solder Joint Failure Due to Fatigue

Fatigue, or failure resulting from the application of cyclical stresses, is the third category of solder joint failures. It is often considered to be the largest and most critical failure category. Solder joint fatigue failure is attributed primarily to stresses brought about by temperature swings and mismatches between the coefficients of thermal expansion (CTEs) of the mounted devices' solder joints and the application board. Under these circumstances, it is possible for failure to occur at a stress level considerably lower than the tensile or yield strength for a static load. During the fatigue process, successive metallurgical phenomena occur. As the strain in the joint exceeds the plastic limit, the solder will start to creep [1]. Examples of real-world events that can lead to fatigue failures include: (a) powering up of an equipment in the day and turning it off at night, (b) the frequently repeated cycle of driving a car and parking it, with the application board under the hood, (c) the orbiting of a satellite that exposes it to the alternating direct heat of the sun and cold vacuum of space.

The progress of the fatigue damage may be seen as [3]: (a) start of the crack, generally under the component at the edge of the metallization, (b) progression of the crack to the outer surface of the fillet, generally first visible at the corners of the

metallization, (c) growth of the visible cracks from the corners of the component to the middle of the joint, (d) sometimes, depending on the configuration, the cracks follow the interface between component and solder.

3 Materials and Experimentation

3.1 Materials

The electronic assemblies used for this investigation are designed and developed for operation at the ground of the Atlantic Ocean, where these are used as hubs for joining different cables for telephone connections.

The solder joints between the electronic components and printed circuit boards (PCBs) were produced from a lead-free solder paste with solder alloy composition of 95.5 % Tin 3.8 % Silver 0.7 % Copper and melting point of 217 °C.

3.2 Field Conditions

The temperature at the ground of the Atlantic Ocean is nearly constant at around 2 °C. However, the electronic systems laid on the ocean bed are continuously producing heat while they are running, heating up the PCB board and also the solder joints. Cooling system is therefore, used around the electronic units to avoid any drastic increase in temperature. Within the cooling system the electronic components experience very small variation in temperature (between 25 °C and 30 °C). However, considering the worst-case scenario, it is estimated that electronic components (and hence the solder joints) undergo a cyclic change in temperature from 20 °C to 35 °C with a cycle time of 1 h. The electronic systems are working nearly the whole year and switched off only once or twice a year. For cycle time calculations only the working period was used, the case of switching off the unit is neglected.

3.3 Thermal Cycling

In order to simulate temperature cycles in the field, the electronic systems were subjected to accelerated thermal cycling. Temperature cycling test is one of the most important tests used to assess the reliability of solder joint interconnections. The objective of temperature cycling test is to assess the resistance and robustness of the package structure to exposures at extremes of high and low temperatures and to the effect of alternate exposures to these extremes [4].

In order to define the thermal cycling parameters the following standards were looked at: IPC-9701 [5], IPC-SM-785 [6] and JESD22-A104-B [7]. The standards suggest low and high thermal cycle temperatures of 0–100 °C respectively to induce fatigue damage to the solder joints. It is also suggested that the changes in temperature should be at rates less than 20 °C/min to avoid thermal shock [6]. Dwell times of 5, 10 or 15 min are generally used for solder fatigue and creep testing [7]. Based on the guidelines from the standards and literatures, thermal cycle parameter values are identified and these are presented in Table 3.1. The duration for one thermal cycle results from the elected parameters is 40 min. A visual presentation of the designed thermal cycle profile is shown in Fig. 3.1.

The useful life of the electronic systems was estimated to be a minimum of 25 years. In order to calculate the thermal cycle test time, acceleration factor (AF) was determined using the Coffin-Mansion equation (3.1) [8]. The necessary parameters to calculate the AF are summarized in Table 3.2, based on Ref. [5].

$$AF = \left(\frac{f_{field}}{f_{test}}\right)^{-m} \cdot \left(\frac{\Delta T_{field}}{\Delta T_{test}}\right)^{-n} \cdot \left[e^{\frac{Ea}{k}\cdot\left(\frac{1}{T_{max,field}} - \frac{1}{T_{max,test}}\right)}\right]. \tag{3.1}$$

Where, $AF =$ acceleration factor, $E_a =$ activation energy in electron-volts (eV), $k =$ Bolzman constant ($=8.617 \times 10^{-5}\, eV/K$), $e = 2.71828$ (base of the natural logarithms), $f_{field} =$ cycle frequency in the field (cycles/24 h), $f_{test} =$ cycle frequency in

Table 3.1 Thermal cycling test parameters

Low temperature (°C)	0
High temperature (°C)	100
Ramp rate (°C/min)	10 °C/min

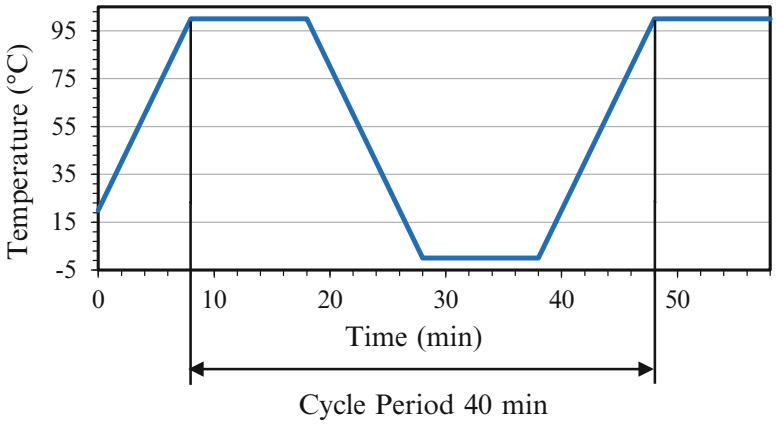


Fig. 3.1 Thermal cycling profile

Table 3.2 Values of Coffin-Manson equation parameters

Parameter	Value
m	0.136
n	2.65
Ea/k	2,185
f_{field}	24 cycles/24 h
f_{test}	36 cycles/24 h
ΔT_{field}	15 K
ΔT_{test}	100 K
$T_{max,\ field}$	308 K
$T_{max,\ test}$	373 K

the lab, ΔT_{field} = field temperature difference, ΔT_{test} = lab temperature difference, $T_{max,field}$ = maximum field temperature, $T_{max,test}$ = maximum test temperature.

Upon substituting the parameter values in (3.1), AF equates to 545.88. Again, AF also equates to the proportion of the number of field temperature cycles to the number of test temperature cycles (see (3.2)).

$$AF = \frac{N_{field}}{N_{test}} \tag{3.2}$$

Where, N_{field} = number of field temperature cycles and N_{test} = number of test temperature cycles. The cycles in the field (N_{field}) can be calculated using (3.3).

$$N_{field} = t_{field} \cdot f_{field} \tag{3.3}$$

Where, f_{field} = cycle frequency in the field, t_{field} = time in the field. $N_{field} = t_{field} \cdot f_{field}$. Equations (3.2) and (3.3) were used to calculate the analogous number of test temperature cycles (N_{test}). Finally, the total test times (t_{test}) needed to simulate the field time (t_{field}) was found out using (3.4), where t_{cycle} is the time for one cycle.

$$t_{test} = \frac{N_{field}}{t_{cycle}} \tag{3.4}$$

The calculated values of different parameters are summarized in Table 3.3 which shows the test times (in hours and days) for different field times (in years). For example, a field time 25 years is equivalent to 10.96 days of test time. Eight electronic PCB samples were used for the investigations, each representing different field times (from 0 to 25 years). The first sample board was used as reference sample and was not subjected to any thermal cycle ageing. The other seven boards were subjected to thermal cycles and taken out of the thermal chamber after different periods of test time as provided in Table 3.3. The temperature cycling tests were conducted in an environmental test chamber from Design Environmental (UK) (model FS800-70SV).

Table 3.3 Predicted test time

t_{field} (year)	N_{field}	N_{test}	t_{test} (h)	t_{test} (days)
0	0	0.00	0.00	0.00
2	17,520	31.57	21.05	0.88
4	35,040	63.15	42.10	1.75
6	52,560	94.72	63.15	2.63
10	87,600	157.87	105.25	4.39
15	131,400	236.81	157.87	6.58
20	175,200	315.74	210.49	8.77
25	219,000	394.68	263.12	10.96

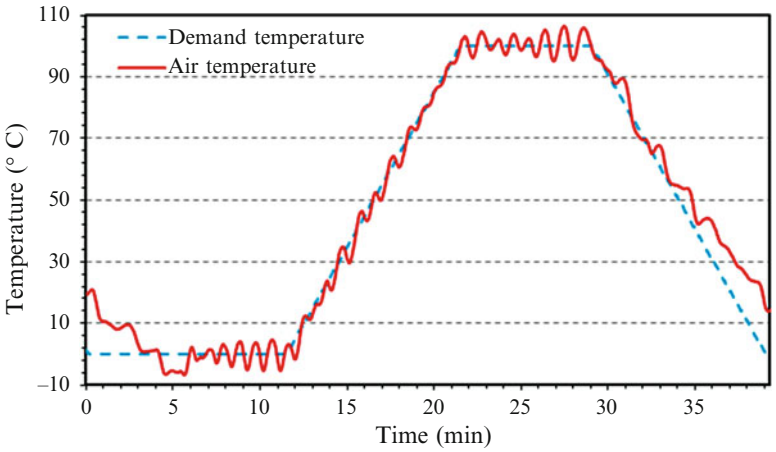


Fig. 3.2 Deviation of air temperature from demand temperature inside the thermal chamber

The thermal cycling profile (as in Table 3.1 and Fig. 3.1) was programmed with a software tool "HanseView" for up to 10.96 days. Figures 3.2 and 3.3 show the variation in Air and Product temperatures as compared to the programmed temperature. The variation was within an acceptable tolerance of ±5 °C. The solder joints were not able to follow these quick temperature variations because of their thermal time lag. Hence the small variations had a negligible impact on the overall accelerated aging process.

All eight PCBs were of same size and each of them were populated with many different types of electronic. After a careful screening, five different electronics components, which were common in all PCBs, were chosen for shear testing and SEM analysis. The standard sizes of these components are provided in Table 3.4.

Reliability of solder joints was tested by measuring the shear strength of the joints. The ball-shear tests were carried out using a 4,000 series Dage Bond Tester. The shear speed and shear height (shear tool offset) were kept at 0.7 mm/s and 0.1 mm respectively for all the electronic components. The fractured surfaces were

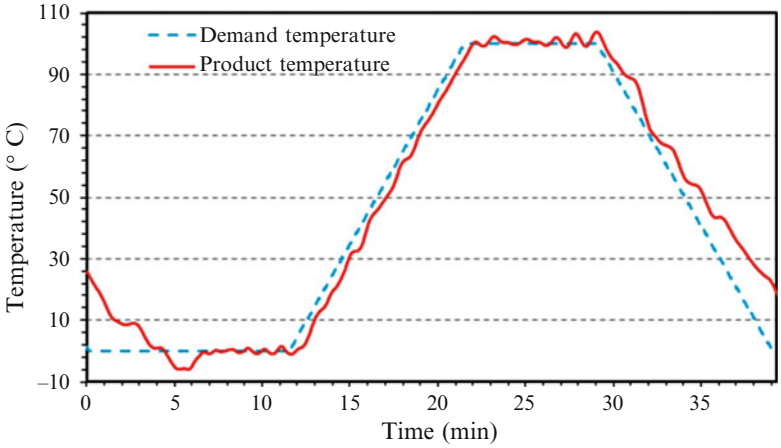


Fig. 3.3 Deviation of product temperature from demand temperature inside the thermal chamber

Table 3.4 Test materials

Size	Length (mm)	Width (mm)	Area (mm^2)
0603 (Res.)	1.55	0.8	0.24
0603 (Cap.)	1.6	0.8	0.28
0805 (Cap.)	2.01	1.25	0.625
1206 (Res.)	3.2	1.6	0.8
1210 (Cap.)	3.2	2.5	1.25
1812 (Cap.)	4.5	3.2	1.6

Res. resistor, *Cap*. capacitor

investigated for brittle and ductile fractures under a scanning electron microscope (SEM). NeoScope JEM-5000 a product of JEOL was used for SEM analysis.

4 Results and Discussions

4.1 Study of the Effect of Accelerated Thermal Cycle Ageing on Solder Joints

In order to prepare the PCBs for the shear test a lot of devices which were not chosen for shear tests had to be removed from the boards to make space for shear testing. Also the PCBs were cut into small pieces to provide the opportunity to fix them at the bench vice. During these preparations some of the selected components were damaged. It was also not possible to make enough space to investigate all of the components as planned.

From the shear tests, the ultimate shear force required to rupture the solder joints were recorded and then plotted as a function of field time (in years). For each of the five component types a number of components were sheared for any particular board and the average shear force value was taken. Figure 3.4 shows the average shear forces as a function of field time and the deviations for 0603 resistors. The average shear force values were calculated for components in the same way and plotted in Fig. 3.5. The variations in shear force (as in Fig. 3.4) could primarily be due to inconsistency in the amount of solder paste used during reflow soldering. In deed previous research found that stencil printing (used for depositing solder paste) accounts for more than 60 % of solder joint assembly defects [9].

From Fig. 3.5 it is quite clear that the different components produced different level of shear force values. This was expected, as the shear force values are depended on solder shear area and hence smaller components will produce lower shear force and vice versa. The aim of this investigation was to find out if the solder joints

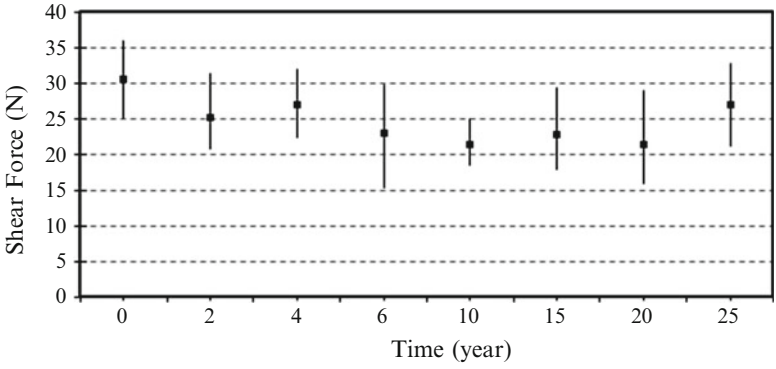


Fig. 3.4 Mean and deviations of measured shear force values for 0603 capacitors

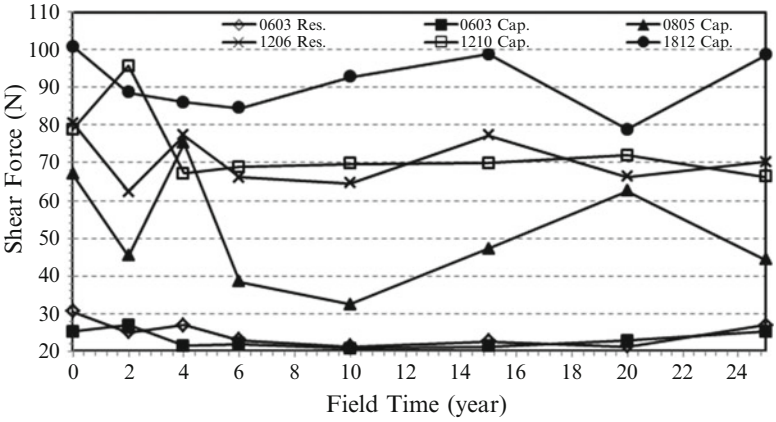


Fig. 3.5 Shear force as a function of field time

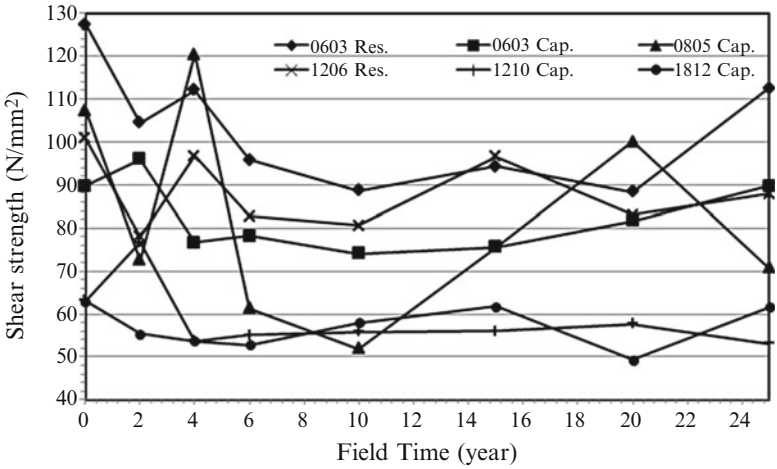


Fig. 3.6 Shear strength as a function of field time

degrade over time in the field. A careful observation of Fig. 3.5 reveals that some of the components did show an initial decrease in shear force values. In some instances the shear force values were found to increase as well. However, the decrease or increase in shear values were minimal and hence it can be generally stated that the shear force values remained same within a close limit. Therefore, from the shear test results it can be concluded that the solder joints and the components are mechanically reliable for the expected field time of 25 years.

Shear strengths of the solder joints were calculated by dividing the shear forces with their respective areas. These are plotted in Fig. 3.6. From the plot it is worth noting that shear strength of the components are not similar. So, depending on the components, some of them are definitely stronger than the others. A similar trend like the shear forces (shown in Fig. 3.5) was observed. Some components did show a decrease in shear strength with time. For example, shear strength of 0603 resistors decreased up until 20 year of field time and showed a sudden increase at 25 years. However, considering the other components, once again it can be said that the components will maintain their mechanical integrity for the expected life of 25 years. The measured shear values are comparable with similar research results reported in literature [10].

4.2 Study of Solder Fracture Surfaces

The microstructure of the fractured solder joints were analyzed using SEM. Before looking into the SEM images it is worth mentioning that the shear fractures could have been affected by the shear heights of the shear tool. The shear tester does not

allow 'on contact' shear and as mentioned earlier, during shear test the shear tool was set at a height of 0.1 mm above the PCB substrate surface. However, as the shear heights were kept constant for all measurements, the shear fractured surfaces are definitely comparable.

The fracture behaviors of solder joints are very complex in nature. For example, depending on the intensity and speed of applied load solder balls could fail through pad lift, interfacial fracture (solder /intermetallic or intermetallic/pad) and bulk solder failure [11]. Among these failures interfacial fractures are predominantly brittle and bulk solder fractures are tend to be ductile in nature. However, solder ball failure through mixed fractures are also frequently observed by various researchers [11, 12].

In general the fractured surfaces showed a mixed mode of fractures including brittle and ductile fractures. However, in most cases the ductile fracture dominates and there was no specific trend of increasing the brittle fracture according to the stages of thermal cycling. Figures 3.7, 3.8, and 3.9 show examples of SEM images of fractured surfaces of different components at various stages of thermal cycle ageing. All the fractured surface images were of the PCB side of the joints.

Figure 3.7 shows an example of a fracture surface for 1812 capacitor representing 15 years in filed. Ductile fracture is dominated in this case and is evidenced through the presence of dimpled, cup and cone features on the surface. Typical ductile fracture may also have a grey, fibrous appearance and occurs when metal and their alloys are tear off after significant plastic deformation.

Brittle fracture on the other hand, displays either cleavage (trans-granular) or inter-granular fracture. This depends upon whether the grain boundaries are stronger or weaker than the grains. Brittle fracture is characterized by rapid crack propagation with low energy release and without significant plastic deformation. The fracture may have a bright granular appearance. The fractures are generally of the flat type

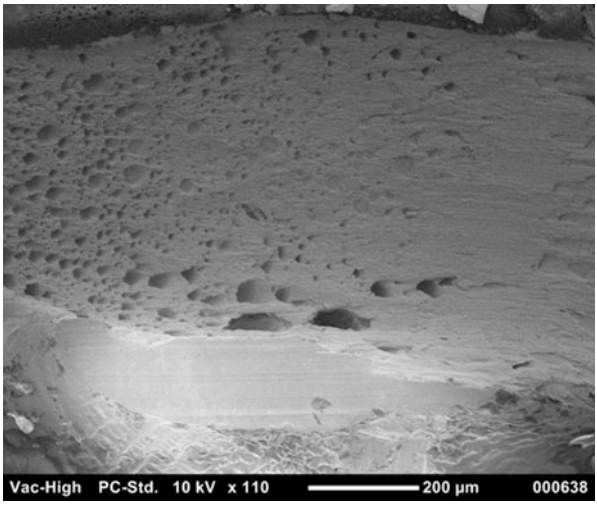


Fig. 3.7 Fracture surface of 1812 capacitor after 15 years of simulated field time

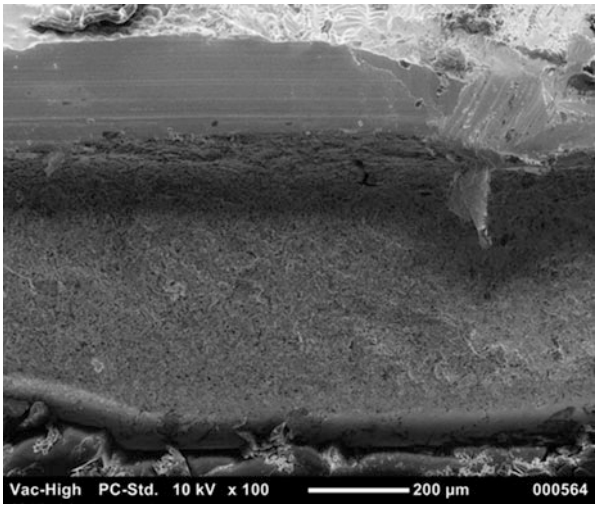


Fig. 3.8 Fracture surface of 1210 capacitor after 6 years of simulated field time

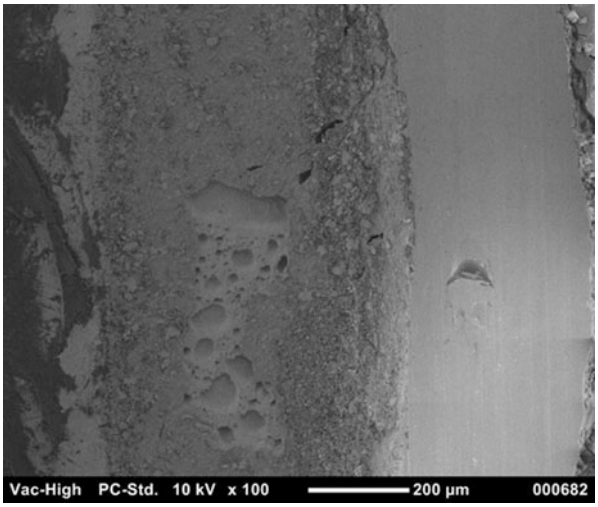


Fig. 3.9 Fracture surface of 1812 capacitor after 25 years of simulated field time

and chevron patterns may be present [13, 14]. Figure 3.8 shows the fractured surface of 1210 capacitor representing 6 years in service. In this instance brittle fracture is dominated with the presence of inter-granular fracture features.

Figure 3.9 presents the fractured surface for 1812 capacitor after 25 years in service. The presence of both brittle and ductile fractures is evidenced in this case. The dimpled and cups, typical for the ductile fracture can be seen and the inter-granular cleavage, typical for the brittle fracture can be seen as well. As stated earlier observation of fracture surfaces revealed no significant change of the microstructure of the solder joints with ageing time. Just as the shear test results showed, the SEM analysis shows that the solder joints are reliable for the expected field time of 25 years.

5 Conclusion

The study investigated the quality and reliability of the solder joints on specific PCBs used for electrical connection at the ground of the Atlantic Ocean. Repair and replacement of these subsea electronic systems is very difficult and would be very expensive. Based on an expected useful life of 25 years an accelerated thermal cycle test was designed to measure the reliability of these devices. Using an accelerated factor, the 25 year of field time was reduced down to 10.96 days in lab test time. Shear testing and SEM analysis of different components were used to ascertain their reliability at different stages of thermal cycling. The results of the shear test show a slight decrease in shear strength but there is no significant decrease in the values observed. During the SEM examination of the shear-fractured board surface, both brittle and ductile fracture modes were observed with ductile fracture being dominated. However, no significant change at the fracture microstructure was observed with thermal cycle ageing. Both the shear test and the SEM analysis showed that the solder joints and the electronic systems are mechanically reliable for the expected lifetime of 25 years.

References

1. Siliconfareast: Solder joint failures. Available: http://www.siliconfareast.com (2012)
2. Tarr, M.: Stress and its effect on materials. Online postgraduate course for the electronics industry, The University of Bolton. Available: http://www.ami.ac.uk/courses/topics/0124_seom/index.html (2003)
3. Tarr, M.: Failure in solder joints. Online postgraduate course for the electronics industry, The University of Bolton. Available: http://www.ami.ac.uk/courses/topics/0154_fsj/index.html (2003)
4. Liu, X.: Processing and reliability assessment of solder joint interconnection for power chips. Ph.D. thesis. Available at: http://scholar.lib.vt.edu/theses/available/etd-04082001-204805/ (2001)
5. IPC: IPC-9701, Performance Test Methods and Qualification Requirements for Surface Mount Solder Attachments. Available at: http://www.ipc.org/TOC/IPC-9701A.pdf (2006)
6. IPC: IPC-SM-785, Guidelines for Accelerated Reliability Testing of Surface Mount Solder Attachments. Available at: http://www.ipc.org/TOC/IPC-SM-785.pdf (1992)
7. JEDEC Solid State Technology Association: Temperature Cycling JESD22-A104-B. Available at: http://www.akrometrix.com/industrydocs/JESD22-A104-B_Temp_Cycling_2000.pdf (2000)
8. Vasudevan, V., Fan, X.: An acceleration model for lead-free (SAC) solder joint reliability under thermal cycling. In: Proceedings of electronic components and technology conference, Lake Buena Vista, Florida, USA, pp. 139–145 (2008)
9. Okuru, T., Kanai, M., Ogata, S., Takei, T., Takakusagi, H.: Optimisation of solder pate printability with laser inspection technique. In: Proceedings of IEEE/CPMT international electronics manufacturing symposium, Santa Clara, USA, pp. 57–161 (1993)
10. Dusek, M., Hunt, C.: Crack detection methods for lead-free solder joints. NPL report MATC(A) 164, March. Available: http://publications.npl.co.uk/npl_web/pdf/matc164.pdf (2004)

11. Newman, K.: BGA brittle fracture – alternative solder joint integrity test methods. In: Proceedings of electronic components and technology conference, Lake Buena Vista, Florida, USA, pp. 1194–1201 (2005)
12. Koo, J.M., Jung, S.B.: Effect of displacement rate on ball shear properties for Sn–37Pb and Sn–3.5Ag BGA solder joints during isothermal aging. Microelectron. Reliab. **47**, 2169–2178 (2007)
13. Metallurgical Consultants: Ductile and brittle failures. Available: http://www.materialsengineer.com/CA-ductbrit.htm (2012)
14. Mallik, S., Kaiser, F.: Reliability study of subsea electronic systems subjected to accelerated thermal cycle ageing. Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE 2014, pp. 1436–1440. London, 2–4 July 2014

Chapter 4
The Navier-Stokes Problem in Infinite Space

Kulyash Kaliyeva and Asset Kaliyev

Abstract This paper is devoted to the mathematical theory of the existence and uniqueness of the three dimensional Navier-Stokes solution for convergent-divergent flows. Using rotor operator and a well-known formula of vector analysis was obtained the nonlinear Volterra-Fredholm integral equation in a matrix form containing only three components of velocity vector which was solved by using the successive approximation method. Considering the pressure gradient as a potential field was determined the balance equation for defining the distribution pressure. Due to the obtained balance equation for the scalar pressure distribution were defined significant properties of the transient convergent-divergent flows with which provided a description of the constitutive relationships between three physical quantities: the velocity vector, the external and internal forces, the pressure distribution. According to the defined estimations of the velocity vector were proved the uniqueness theorems for the convergent-divergent Navier-Stokes problem in the general case.

Keywords Clay Institute's Millennium problem • Convergent-divergent flow • Navier-Stokes equations • Nonlinear Volterra-Fredholm integral equation • Potential field • Pressure distribution • Turbulent fluid motion

1 Introduction

The Navier-Stokes equations describe interactions between fluctuations and their directions for different wavelengths which have a great interest in mathematical modeling of turbulent process. Mathematical solution for this practical complex

K. Kaliyeva (✉)
Professor at Kazakh National Pedagogical Univerity named after Abai, 13 Dostyk ave, Almaty 050010, Kazakhstan
e-mail: kklara_09@mail.ru

A. Kaliyev
Specialist at University of International Business and Economics, 10 Huixin Dongjie, Chaoyang District, Beijing, China
e-mail: kka_2014@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_4

problem requires a perspective using some alternative approach which different from that is needed for studying the fundamental issues. Theoretical prediction and analysis of turbulence has been the fundamental problem of fluid dynamics. It is worth stressing that turbulence is fundamentally interesting and practical importance for engineering models of convergent-divergent turbulent effects.

Turbulence is a continuous phenomenon that exists on a large range of length and time scales. As there exists different scales which energy is transferred from the larger scales to the smaller scales where energy is dissipated into heat by molecular viscosity. The Navier-Stokes equations as nonlinear partial differential equations in real natural situation were formulated in 1821 and appeared to give an accurate description of fluid flow including laminar and turbulent features. Concerning the large literature on the Navier-Stoke problem we mention only some papers which consider particularly relevant for our purpose. We have focused on the global existence, uniqueness and smoothness. Examples of weak solution were given by L. Caffarelli [1], V. Sheffer [2]. A critical analysis for many analytic and numerical solutions of Navier-Stokes equations was given by C. Fefferman [3]. We will follow this unique idea of existence of weak solution given in [3] by using the energy conservation law for the external and internal forces with the gradient of pressure. The key idea of our approach is to exclude the pressure function from the Navier-Stokes equation by using rotor operator. According to this transformation we can give the integral representative for the velocity vector and the energy conservation condition for the determining pressure distribution. We involve this method to show that the velocity vector with respect to the pressure function exists and satisfies the energy conservation law. We split the construction of solution for the Navier-Stokes problem into two steps. In the first step we claim that we may assume

$$rot\ \overrightarrow{f} = 0,\ rot\ \overrightarrow{u}_0 = 0$$

Then we will get

$$grad\ \left(\frac{u^2}{2} + \frac{p}{\rho} - \Phi\right) = 0$$

where

$$grad\ \Phi(x,t) = -\overrightarrow{f}(x,t)$$

It is proved that under the energy conservation law there exists a unique velocity vector given by the integral representation. There we get a stable solution of the Navier-Stokes problem according to the energy conservation law. In the second step we assume that

$$rot\ \overrightarrow{f} \neq 0$$

Due to this assertion we obtain the second kind nonlinear matrix Volterra-Fredholm integral equation which is solved by using the method of the successive approximation. Under above assumption there exists a unique unstable solution with the appropriate properties in a Hilbert space.

2 Mathematical Formulation of the Problem

Denote a point in the spatial coordinate R^3 by $x = (x_1, x_2, x_3)$ let us consider the velocity vector in a point $M(x,t)$ of and in a given time t by the formula

$$\overrightarrow{u}(x,t) = \begin{cases} \overrightarrow{u}_v(x,t) & if \ \ div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{u}_\alpha(x,t) & if \ \ div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

Suppose that infinite spaces are $\Omega = R^3$, $\Omega_T = R^3 \times (0 < t < \infty)$,

$$\overrightarrow{u}(x,t) = u_1(x,t)\overrightarrow{i} + u_2(x,t)\overrightarrow{j} + u_2(x,t)\overrightarrow{k}$$

is the velocity vector and $p(x,t)$ is the fluid pressure field.

Statement of the problem: a mixture of perfect gases with simplified thermodynamics, kinetics and transport moves under some external influence. Possible modes are:

- the point of space is a stock in this field (compressible flow);
- the point of space is a source in this field (compressible flow);
- isolated point when stock and source compensate each other (incompressible flow).

We consider a convergent-divergent flow given by the Navier -Stokes equations in the general form

$$\frac{\partial \overrightarrow{u}}{\partial t} + \left(\overrightarrow{u} \cdot \nabla\right)\overrightarrow{u} = -\frac{1}{\rho}\nabla p + \nu\Delta\overrightarrow{u} + \left(\frac{\nu}{3} + \eta\right)\nabla div\ \overrightarrow{u} + \overrightarrow{f}(x,t) \text{ in } \Omega_T \tag{4.1}$$

with the initial condition

$$\overrightarrow{u}\big|_{t=0} = \overrightarrow{u}_0(x) \text{ on } \Omega \tag{4.2}$$

Here, the vector function $\overrightarrow{f}(x,t) = f_1(x,t)\overrightarrow{i} + f_2(x,t)\overrightarrow{j} + f_2(x,t)\overrightarrow{k}$ denotes an external and internal forces, ν is a kinematic viscosity, ρ is a fluid density, the symbol ∇ denotes the gradient with respect to the function, the symbol Δ denotes the three dimensional Laplace operator, η is a dynamic viscosity which is related to the kinematic viscosity by the formula $\eta = \rho\nu$. For simplication our

vision there is considered a specific case when coefficients ρ, ν, η are fixed given parameters which can be defined at any points. We will construct a solution for the Navier-Stokes initial value problem (4.1)–(4.2). The problem (4.1)–(4.2) in case when $div\ \overrightarrow{u}\ (x,t) = 0$ is Clay Institute's Millennium problem which was formulated by Fefferman [3]. The formulation of the problem (4.1)–(4.2) is based on the introduced technique for incompressible potential and swirling turbulent flows. There we assume that

$$\left|\overrightarrow{u}\right| = \sqrt{u_1^2 + u_2^2 + u_3^2} \to 0 \ for\ |x| = \sqrt{x_1^2 + x_2^2 + x_3^2} \to \infty$$

Note, that many problems formally exist for any Reynolds numbers and it can have an exact solution, but not all partial differential equation can describe real-nature phenomenon, therefore we will consider the basic equations that correctly can be solved (existence, uniqueness and stability). The requirement of stability is caused by the fact that physical evidence is usually determined from experiment and approximately, therefore we must be sure that the determined solution is the stability solution. For point $M(x,t)$ in term of physic and mechanic points of view divergence operator is an indication that the point of space are stock or source in this field (Fig. 4.1) which are characterized physical properties and electric interaction effects. In real nature turbulent process represents an average departure from the different points of the space and we have different combinations of the conditions $div\ \overrightarrow{u}\ (x,t) < 0$, $div\ \overrightarrow{u}\ (x,t) > 0$ and $div\ \overrightarrow{u}\ (x,t) = 0$.

Due to these combinations of the turbulent processes which associate with the precise organization of flux yielding a complete picture of the distribution parameters in point $M(x,t)$ and we can explain transfer mechanisms of divergent-convergent flows [5, 6]. Problem (4.1)–(4.2) deals with flows having the vector velocity for which

$$div\ \overrightarrow{u}\ (x,t) = D^*\ (x,t) \tag{4.3}$$

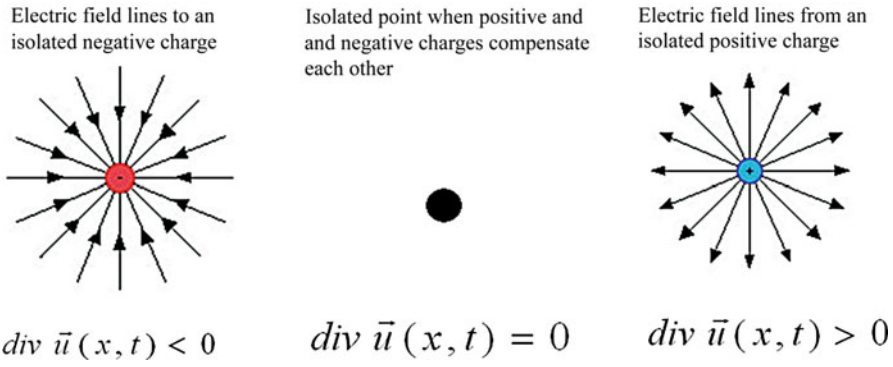


Fig. 4.1 Divergence and turbulent flux in a point $M(x,t)$

Mathematical description of the condition (4.3) can be characterized by the physical properties and electric interaction effects (Fig. 4.1). So, development of our idea is the most challenging task due to behavior of the expression

$$\left(\frac{v}{3}+\eta\right)\nabla div\ \overrightarrow{u}$$

in the Navier-Stokes problem (4.1)–(4.2).

Using the formula

$$\frac{1}{2}grad\ \overrightarrow{u}^2=\left[\overrightarrow{u}\times rot\ \overrightarrow{u}\right]+\left(\overrightarrow{u}\ \nabla\right)\overrightarrow{u} \tag{4.4}$$

from the vector analysis, we have got the expression

$$\frac{\partial\overrightarrow{u}}{\partial t}-\left[\overrightarrow{u}\times rot\ \overrightarrow{u}\right]=\nabla\left(\frac{1}{\rho}p+\frac{u^2}{2}\right)+\nu\Delta\overrightarrow{u}+\left(\frac{v}{3}+\eta\right)\nabla div\ \overrightarrow{u}+\overrightarrow{f}\ (x,t) \tag{4.5}$$

then (4.5) can be written as

$$\begin{aligned}&\frac{\partial\overrightarrow{u}}{\partial t}-\nu\Delta\overrightarrow{u}-\left[\overrightarrow{u}\times rot\ \overrightarrow{u}\right]-\left(\frac{v}{3}+\eta\right)\nabla div\ \overrightarrow{u}-2\overrightarrow{f}\ (x,t)\\&=\nabla\left(\frac{1}{\rho}p+\frac{u^2}{2}-\Phi\right)\end{aligned} \tag{4.6}$$

Using the rotor operator for (4.6) and assume that

$$rot\ \overrightarrow{u}_0=0\ rot\ \overrightarrow{f}=0$$

we have got the following problem

$$\frac{\partial\overrightarrow{u}}{\partial t}=\nu\Delta\overrightarrow{u}+\left(\frac{v}{3}+\eta\right)\nabla div\ \overrightarrow{u}+2\overrightarrow{f}\ (x,t)\ \text{in}\ \Omega_T \tag{4.7}$$

with the initial condition

$$\overrightarrow{u}\big|_{t=0}=\overrightarrow{u}_0(x)\ \text{on}\ \Omega \tag{4.8}$$

Using the divergence operator for the problem (4.7)–(4.8) we have obtained the following equation

$$\frac{\partial div\ \overrightarrow{u}}{\partial t}=\nu\Delta div\ \overrightarrow{u}+\left(\frac{v}{3}+\eta\right)div\left(\nabla div\ \overrightarrow{u}\right)+2div\ \overrightarrow{f}\ (x,t) \tag{4.9}$$

Suppose that

$$\overrightarrow{U} = div\ \overrightarrow{u}$$

we have obtained analog of the heat problem

$$\frac{\partial \overrightarrow{U}}{\partial t} = \left(\frac{4v}{3} + \eta\right) \Delta \overrightarrow{U} + 2div\ \overrightarrow{f}\ (x,t) \tag{4.10}$$

with the initial condition

$$\overrightarrow{U}\Big|_{t=0} = div\ \overrightarrow{u}_0(x) \tag{4.11}$$

For the problem (4.10)–(4.11) exists a unique conditionally stable periodic solution

$$\overrightarrow{U} = \int\limits_{R^3} div\ \overrightarrow{u}_0\,(\xi)\, G_\alpha\,(x-\xi,t)\,d\xi + 2\int\limits_0^t d\tau \int\limits_{R^3} div\overrightarrow{f}\,(\xi,\tau)\, G_\alpha\,(x-\xi,t-\tau)\,d\xi \tag{4.12}$$

Using the convolution form periodic solution for the problem (4.10)–(4.11) has been written as

$$\overrightarrow{U} = div\ \overrightarrow{u}_0 * G_\alpha^* + 2div\ \overrightarrow{f} * G_\alpha^* \tag{4.13}$$

where

$$G_\alpha\,(x,\xi,t) = \frac{e^{-\frac{(x-\xi)^2}{4\alpha t}}}{\left(2\sqrt{\pi\alpha t}\right)^3},\ \alpha = \frac{4v}{3} + \eta$$

is the three dimensional Green's function $G_\alpha(x,\xi,t)$.

Using solution (4.13) and properties of the Green's function $G_\alpha(x,\xi,t)$ we have got a solution for the problem (4.1)–(4.2) in the following form

$$\overrightarrow{u}_\alpha = grad\left\{div\ \overrightarrow{u}_0 * G_\alpha^* + 2div\ \overrightarrow{f} * G_\alpha^*\right\} \tag{4.14}$$

where $G_\alpha^*\,(x,\xi,t)$ is the Green's function

$$G_\alpha^*\,(x,\xi,t) = G_\alpha\,(\xi,\zeta,t) * \frac{1}{4\pi\,|x-\zeta|}$$

Our next concern is to investigate in which sense the hidden expression

$$div\ \overrightarrow{u}\ (x,t) = D^*\ (x,t)$$

can be associated with electric interaction effects.

Recalling that

$$\overrightarrow{u}_\alpha = \left(\frac{\partial}{\partial t} - \nu\Delta\right)\left[grad\left\{div\ \overrightarrow{u}_0 * G_\alpha^* + 2div\ \overrightarrow{f} * G_\alpha^*\right\}\right] - 2\overrightarrow{f} * G_\nu\Big)$$
$$+ \overrightarrow{u}_0 * G_\nu + 2\overrightarrow{f} * G_\nu \tag{4.15}$$

and substitute

$$\left(\frac{\partial}{\partial t} - \nu\Delta\right)\left[grad\left\{div\ \overrightarrow{u}_0 * G_\alpha^* + 2div\ \overrightarrow{f} * G_\alpha^*\right\}\right] - 2\overrightarrow{f} = \left(\frac{\nu}{3} + \eta\right)\nabla div\ \overrightarrow{u}$$

we have got simplified solution of the considered problem (4.1)–(4.2) in the another form

$$\overrightarrow{u}_\alpha = \left(\frac{\nu}{3} + \eta\right)\nabla div\ \overrightarrow{u} * G_\nu + \overrightarrow{u}_0 * G_\nu + 2f * G_\nu \tag{4.16}$$

where

$$G_\nu\ (x,\xi,t) = \frac{e^{-\frac{(x-\xi)^2}{4\nu t}}}{\left(2\sqrt{\pi\nu t}\right)^3}$$

is the Green's function.

Solution (4.14) we can rewrite as

$$\overrightarrow{u}_\alpha = grad\ div\left\{\left(\overrightarrow{u}_0 * G_\alpha\right) * \frac{1}{4\pi\ |x-\zeta|} + 2\left(\overrightarrow{f} * G_\alpha\right) * \frac{1}{4\pi\ |x-\zeta|}\right\} \tag{4.17}$$

Using (4.16) and (4.17) we have got

$$\left(\frac{\nu}{3} + \eta\right)\nabla div\ \overrightarrow{u} = \frac{\partial}{\partial t} - \nu\Delta\Big)\left[grad\left\{div\overrightarrow{u}_0 * G_\alpha^* + 2div\overrightarrow{f} * G_\alpha^*\right\}\right] - 2f$$

where

$$G_\alpha^*\ (x,\xi,t) = G_\alpha * \frac{1}{4\pi\ |x-\zeta|} \tag{4.18}$$

and expression for divergence of the velocity vector

$$D^*(x,t) = div\ \overrightarrow{u}_0 * G_\alpha + 2div\ \overrightarrow{f} * G_\alpha \tag{4.19}$$

Formula (4.19) expresses divergence operator $div\ \overrightarrow{u}(x,t)$ and three different cases of the problem (4.1)–(4.2). Value of this function can be negative, positive or equal zero: $D^*(x,t) < 0$, $D^*(x,t) > 0$ or $D^*(x,t) = 0$.

3 Unstable Solution for the Divergent-Convergent Flow of the Navier-Stokes Problem

A fundamental interest is in the study of unsteady features for an eddy divergent motion which characterizes by the higher Reynolds numbers.

Denote that

$$\overrightarrow{g} = \frac{\partial}{\partial t}\overrightarrow{u} - \left[\overrightarrow{u} \times rot\ \overrightarrow{u}\right] - \nu\Delta\overrightarrow{u}$$
$$\overrightarrow{z} = rot\left(\left(\frac{\nu}{3} + \eta\right) grad\ D^* + \overrightarrow{f}^*\right)$$

and using the system of the differential equations

$$\frac{\partial g_3}{\partial x_2} - \frac{\partial g_2}{\partial x_3} = z_1$$
$$\frac{\partial g_1}{\partial x_3} - \frac{\partial g_3}{\partial x_1} = z_2$$
$$\frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \frac{\partial g_3}{\partial x_3} = D^*$$

For problem (4.1)–(4.2) in the general case we have got second order nonlinear Volterra-Fredholm integral equation in a matrix form satisfying

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \frac{(u_1-u_2-u_3)^2}{2} & 0 & 0 \\ 0 & \frac{(u_2-u_1-u_3)^2}{2} & 0 \\ 0 & 0 & \frac{(u_3-u_1-u_2)^2}{2} \end{pmatrix} * \begin{pmatrix} \frac{\partial G_\nu}{\partial x_1} \\ \frac{\partial G_\nu}{\partial x_2} \\ \frac{\partial G_\nu}{\partial x_3} \end{pmatrix}$$
$$+ \begin{pmatrix} u_1 \cdot D^* & 0 & 0 \\ 0 & u_2 \cdot D^* & 0 \\ 0 & 0 & u_3 \cdot D^* \end{pmatrix} * \begin{pmatrix} G_\nu \\ G_\nu \\ G_\nu \end{pmatrix} + \begin{pmatrix} F_1^* \\ F_2^* \\ F_3^* \end{pmatrix} \tag{4.20}$$

where

$$\overrightarrow{F}^* = \overrightarrow{b}^* * G_\nu + u_0 * G_\nu$$

Successive approximation method can be successfully applied to solve the nonlinear Volterra-Fredholm matrix integral equation (4.20). Considering (4.20) in the operator form we have equation

$$\overrightarrow{u} = K_2\left[\overrightarrow{u}\right] + K_1\left[\overrightarrow{u}\right] + \overrightarrow{F} \tag{4.21}$$

here

$$K_2\left[\overrightarrow{u}\right] = \begin{pmatrix} \frac{(u_1-u_2-u_3)^2}{2} & 0 & 0 \\ 0 & \frac{(u_2-u_1-u_3)^2}{2} & 0 \\ 0 & 0 & \frac{(u_3-u_1-u_2)^2}{2} \end{pmatrix} * \begin{pmatrix} \frac{\partial G_\nu}{\partial x_1} \\ \frac{\partial G_\nu}{\partial x_2} \\ \frac{\partial G_\nu}{\partial x_3} \end{pmatrix}$$

$$K_1\left[\overrightarrow{u}\right] = \begin{pmatrix} u_1 \cdot D^* & 0 & 0 \\ 0 & u_2 \cdot D^* & 0 \\ 0 & 0 & u_3 \cdot D^* \end{pmatrix} * \begin{pmatrix} G_\nu \\ G_\nu \\ G_\nu \end{pmatrix}$$

Using inverse operators R_i for the operators K_i $(i = 1, 2)$ we have got

$$\overrightarrow{u} = R_2\left(K_1\left[\overrightarrow{u}\right]\right) + K_1\left[\overrightarrow{u}\right] + R_2\left(\overrightarrow{F}\right) + \overrightarrow{F}$$

or

$$\overrightarrow{u} = R_1\left(R_2\left[R_2\left(\overrightarrow{F}^*\right) + \overrightarrow{F}^*\right]\right) + R_1\left[R_2\left(\overrightarrow{F}^*\right) + \overrightarrow{F}^*\right] + R_2\left(\overrightarrow{F}^*\right) + \overrightarrow{F}^*$$

here $\overrightarrow{F}^* = \overrightarrow{d} * G_\nu + \overrightarrow{b} * G_\nu + u_0 * G_\nu$, $\overrightarrow{d} = \left(\frac{\nu}{3} + \eta\right) grad\ D^*(x,t)$ vector

$$\overrightarrow{b}(x,t) = \frac{1}{4\pi}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{4.22}$$

has components

$$b_1 = \left(\frac{\partial^3}{\partial x_3 \partial^2 x_2} + \frac{\partial^3}{\partial^3 x_3}\right) \int\limits_{R^3} \frac{1}{|x-\xi|} \int\limits_{-\infty}^{\infty} \theta(\xi_3 - \varsigma_3)\, f_1^*(\xi_1, \xi_2, \varsigma_3, t)\, d\varsigma_3 d\xi$$

$$- \left(\frac{\partial^3}{\partial x_3 \partial x_2 \partial x_1} + \frac{\partial^3}{\partial x_1 \partial^2 x_2} + \frac{\partial^3}{\partial x_1 \partial^2 x_3}\right) \int\limits_{R^3} \frac{1}{|x-\xi|} \int\limits_{-\infty}^{\infty} \theta(\xi_3 - \varsigma_3)\, f_2^*(\xi_1, \xi_2, \varsigma_3, t)\, d\varsigma_3 d\xi$$

$$+ \frac{\partial^3}{\partial x_3 \partial x_2 \partial x_1} \int\limits_{R^3} \frac{1}{|x-\xi|} \int\limits_{-\infty}^{\infty} \theta(\xi_3 - \varsigma_3)\, f_3^*(\xi_1, \xi_2, \varsigma_3, t)\, d\varsigma_3 d\xi$$

$$b_2 = -\left[\frac{\partial^3}{\partial x_3 \partial x_2 \partial x_1}\int\limits_{R^3}\frac{1}{|x-\xi|}\int\limits_{-\infty}^{\infty}\theta\left(\xi_3-\varsigma_3\right) f_1^*\left(\xi_1,\xi_2,\varsigma_3,t\right)d\varsigma_3 d\xi\right.$$
$$+\left(\frac{\partial^3}{\partial x_3\partial^2 x_1}+\frac{\partial^3}{\partial x_2\partial^2 x_3}+\frac{\partial^3}{\partial x_2\partial^2 x_1}\right)\int\limits_{R^3}\frac{1}{|x-\xi|}\int\limits_{-\infty}^{\infty}\theta\left(\xi_3-\varsigma_3\right) f_2^*\left(\xi_1,\xi_2,\varsigma_3,t\right)d\varsigma_3 d\xi$$
$$\left.+\left(\frac{\partial^3}{\partial x_2\partial^2 x_1}+\frac{\partial^3}{\partial x_3\partial^2 x_2}\right)\int\limits_{R^3}\frac{1}{|x-\xi|}\int\limits_{-\infty}^{\infty}\theta\left(\xi_3-\varsigma_3\right) f_3^*\left(\xi_1,\xi_2,\varsigma_3,t\right)d\varsigma_3 d\xi\right.$$

$$b_3 = -\left[\frac{\partial^3}{\partial x_3 \partial x_2 \partial x_1}\int\limits_{R^3}\frac{1}{|x-\xi|}\int\limits_{-\infty}^{\infty}\theta\left(\xi_3-\varsigma_3\right) f_1^*\left(\xi_1,\xi_2,\varsigma_3,t\right)d\varsigma_3\right)d\xi$$
$$-\left(\frac{\partial^3}{\partial x_2\partial^2 x_3}-\frac{\partial^3}{\partial x_3\partial^2 x_1}\right)\int\limits_{R^3}\frac{1}{|x-\xi|}\int\limits_{-\infty}^{\infty}\theta\left(\xi_3-\varsigma_3\right) f_2^*\left(\xi_1,\xi_2,\varsigma_3,t\right)d\varsigma_3 d\xi$$
$$\left.-\frac{\partial^3}{\partial x_3\partial^2 x_2}\int\limits_{R^3}\frac{1}{|x-\xi|}\int\limits_{-\infty}^{\infty}\theta\left(\xi_3-\varsigma_3\right) f_3^*\left(\xi_1,\xi_2,\varsigma_3,t\right)d\varsigma_3 d\xi\right]$$

Solution of the Navier -Stokes problem (4.1)–(4.2) can be written in following form

$$\overrightarrow{u}_\alpha(x,t) = \overrightarrow{u}_\alpha{}^*(x,t) + \overrightarrow{u}_\nu(x,t) \tag{4.23}$$

where

$$\overrightarrow{u}_\alpha{}^*(x,t) = R_1\left[R_2\left[R_2\left(\overrightarrow{F}^*\right)+2\overrightarrow{F}^*\right]\right]+R_1\left[\overrightarrow{F}^*\right]+R_2\left(\overrightarrow{d}*G_\nu\right)+\overrightarrow{d}*G_\nu$$

Using the Navier-Stokes equation (4.1) we have a unique scalar function of pressure $p(x,t)$ which satisfies

$$\frac{u^2}{2}+\frac{p}{\rho}-div\ \overrightarrow{f}*\frac{1}{4\pi\,|x-\xi|}-div\ \overrightarrow{f}^{**}*\frac{1}{4\pi\,|x-\xi|}=0$$

where

$$\overrightarrow{f}^{**} = \overrightarrow{b}^* - \overrightarrow{f}^*,\ \overrightarrow{b}^* = \overrightarrow{d}+\overrightarrow{b}$$

$$\overrightarrow{f}^8 = \begin{pmatrix} -\frac{\partial^2 f_1}{\partial x_2^2}-\frac{\partial^2 f_1}{\partial x_3^2}+\frac{\partial^2 f_2}{\partial x_1\partial x_2}+\frac{\partial^2 f_3}{\partial x_1\partial x_3} \\ \frac{\partial^2 f_1}{\partial x_1\partial x_2}-\frac{\partial^2 f_2}{\partial x_1^2}-\frac{\partial^2 f_2}{\partial x_3^2}+\frac{\partial^2 f_3}{\partial x_2\partial x_3} \\ \frac{\partial^2 f_1}{\partial x_1\partial x_3}+\frac{\partial^2 f_3}{\partial x_1\partial x_3}-\frac{\partial^2 f_3}{\partial x_1^2}-\frac{\partial^2 f_3}{\partial x_2^2} \end{pmatrix} * \begin{pmatrix} \frac{1}{4\pi|x-\xi|} & 0 & 0 \\ 0 & \frac{1}{4\pi|x-\xi|} & 0 \\ 0 & 0 & \frac{1}{4\pi|x-\xi|} \end{pmatrix}$$

Obtained condition admits solution that can be predicted specific process in terms of unstable flows by increasing the rotation velocity. There we have some analogies of bifurcating instabilities for compressible unstable motions where the Navier-Stokes equations represent the evolution for the governing distribution functions which depends on the velocity vector in the position of particles as a result of thermal excitation at any finite turbulent energy.

4 Results and Discussion

Let us gather and formulate some main exciting results about properties of the vector velocity and the scalar function of pressure. Recall the notations $\Omega = R^3$ and $\Omega_T = R^3 \times (0 < t < \infty)$ we looked for periodic solution for the problem (4.1)–(4.3).We assumed that functions $f_i(x,t)$ and $u_{0i}(x,t)$ satisfy

$$u_{0i}(x) = u_{0i}\left(x + k_j\right),\ f_i(x,t) = f_i\left(x + k_j, t\right)$$

for $1 \leq j \leq 3$, where $k_j = j^{th}$ is unit vector in R^3.

Theorem 1 Assume that $\overline{\Omega}_T = R^3 \times (0 \leq t < \infty)$ let functions

$$\overrightarrow{u}_0(x) = \begin{cases} \overrightarrow{u}_0(x) \in C(\Omega) & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{u}_0(x) \in C^{(1)}(\Omega) & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

$$\overrightarrow{f}(x,t) = \begin{cases} \overrightarrow{f}(x,t) \in C_{\overline{\Omega}_T} & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{f}(x,t) \in C^{(1,0)}_{\overline{\Omega}_T} & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

be periodic functions with conditions $rot\ \overrightarrow{f} = 0$ and $rot\ \overrightarrow{u}_0 = 0$. Then there for the Navier-Stokes problem (4.1)–(4.2) exists a unique stable periodic solution $\overrightarrow{u}(x,t) \in C^{(2,1)}(\Omega_T) \cap C^{(1,0)}\left(\overline{\Omega}_T\right)$ which is given by the formula

$$\overrightarrow{u}(x,t) = \begin{cases} \overrightarrow{u}_\nu = \overrightarrow{u}_0 * G_\nu + 2\overrightarrow{f} * G_\nu & \text{if } div\ \overrightarrow{u} = 0 \\ \overrightarrow{u}_\alpha = grad\left\{div\ \overrightarrow{u}_0 * G^*_\alpha + 2div\ \overrightarrow{f} * G^*_\alpha\right\} & \text{if } div\ \overrightarrow{u} \neq 0 \end{cases} \tag{4.24}$$

Moreover, there exists a unique scalar function of pressure $p(x,t) \in C^{(1,0)}\left(\overline{\Omega}_T\right)$ which satisfies the energy conversation law

$$\frac{p}{\rho} + \frac{u^2}{2} - div\ \overrightarrow{f} * \frac{1}{4\pi\,|x-\xi|} = 0 \tag{4.25}$$

In addition, there exists a positive constant for which is satisfied the following estimates

$$\left\|\overrightarrow{u}\right\|_{C(\overline{\Omega}_T)} \leq \begin{cases} M_0\left(\left\|\overrightarrow{u}_0\right\|_{C(\Omega)} + 2\left\|\overrightarrow{f}\right\|_{C(\overline{\Omega}_T)}\right) & \text{if } div\ \overrightarrow{u}\,(x,t) = 0 \\ M_0\left(\left\|\overrightarrow{u}_0\right\|_{C^1(\Omega)} + 2\left\|\overrightarrow{f}\right\|_{C^{(1,0)}_{\Omega_T}}\right) & \text{if } div\,\overrightarrow{u}\,(x,t) \neq 0 \end{cases} \tag{4.26}$$

where

$$\left\|\overrightarrow{u}\right\| = \sqrt{u_1^2 + u_2^2 + u_3^2}$$

$$\|p\|_{C(\overline{\Omega}_T)} \leq \begin{cases} M_0\left(\left\|\overrightarrow{u}_0\right\|_{C(\Omega)} + 2\left\|\overrightarrow{f}\right\|_{C(\overline{\Omega}_T)}\right) \text{if } div\ \overrightarrow{u}\,(x,t) = 0 \\ M_0\left(\left\|\overrightarrow{u}_0\right\|_{C^1(\Omega)} + 2\left\|\overrightarrow{f}\right\|_{C^{(1,0)}_{\Omega_T}}\right) \text{if } div\ \overrightarrow{u}\,(x,t) \neq 0 \end{cases} \tag{4.27}$$

Remark 1 From Theorem 1 we can get a new useful result which is introduced the strong solution of the Navier-Stokes problem (4.1)–(4.2) with the energy conservation law (4.25). Conversely the solution of the problem (4.1)–(4.2) is sufficiently regular in case when the pressure function $p\,(x,t) \in C^{(1,0)}\left(\overline{\Omega}_T\right)$ satisfies the energy conversation law (4.25).

Assume that $rot\ \overrightarrow{f} = 0$, $rot\ \overrightarrow{u} = 0$ are satisfied. If $\overrightarrow{f} = C\overrightarrow{x} + \overrightarrow{d}$, where C matrix

$$C = \begin{pmatrix} \frac{c_1}{m} & 0 & 0 \\ 0 & \frac{c_2}{m} & 0 \\ 0 & 0 & -gh \end{pmatrix}$$

$\overrightarrow{d}$ - a numerical vector, m- a body's mass, c_1, c_2 are independent constants which satisfy the condition $c_1 + c_2 \geq 0$, g is the acceleration of gravity, h is a height. Then fluid flow can be considered to be an incompressible flow which satisfies Bernoulli's equation

$$\frac{mp}{\rho} + \frac{mu^2}{2} + mgh = const \tag{4.28}$$

Here $\frac{mp}{\rho}$ is a binding energy of the mass elements, $\frac{mu^2}{2}$ is a kinetic energy, mgh is a potential energy

Theorem 2 Let functions

$$\overrightarrow{u}_0(x) = \begin{cases} \overrightarrow{u}_0(x) \in C\,(\Omega) \ \text{ if } div\ \overrightarrow{u}\,(x,t) = 0 \\ \\ \overrightarrow{u}_0(x) \in C^{(1)}\,(\Omega) \ \text{ if } div\ \overrightarrow{u}\,(x,t) \neq 0 \end{cases}$$

$$\overrightarrow{f}(x,t) = \begin{cases} \overrightarrow{f}(x,t) \in C_{\overline{\Omega}_T} \text{ if } div\ \overrightarrow{u}(x,t) = 0 \\ \\ \overrightarrow{f}(x,t) \in C^{(1,0)}_{\overline{\Omega}_T} \text{ if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

be periodic functions with conditions $rot\ \overrightarrow{f} = 0$ and $rot\ \overrightarrow{u}_0(x) \neq 0$. Then there for the Navier-Stokes problem (4.1)–(4.2) exists a unique stable periodic solution $\overrightarrow{u}(x,t) \in C^{(2,1)}(\Omega_T) \cap C^{(1,0)}\left(\overline{\Omega}_T\right)$ which is given by the formula

$$\overrightarrow{u}(x,t) = \begin{cases} \overrightarrow{u}_v(x,t) = R_v\left[\overrightarrow{F}\right] + \overrightarrow{F} \text{ if } div\ \overrightarrow{u}(x,t) = 0 \\ \\ \overrightarrow{u}_\alpha(x,t) = R_\alpha\left[\overrightarrow{F}\right] + \overrightarrow{F}^* \text{ if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases} \tag{4.29}$$

where $R_v[\cdot]$ and $R_\alpha[\cdot]$ are inverse operators for the nonlinear equation (4.21). Moreover, there exists a unique scalar function of pressure $p(x,t) \in C^{(1,0)}\left(\overline{\Omega}_T\right)$ which satisfies the energy conversation law (4.25).

$$\frac{p}{\rho} + \frac{u^2}{2} - div\ \overrightarrow{f} * \frac{1}{4\pi|x-\xi|} = 0 \tag{4.25}$$

In addition, there exist positive constants C_0, C_1, C_2, C_4 for which are satisfied the following estimates

$$\left\|\overrightarrow{u}_\alpha\right\|_{C(\Omega_T)} \leq \begin{cases} \left(C_0\left\|\overrightarrow{y}^*\right\| + C_1\left\|\overrightarrow{y}^*\right\|^2 e^{C_2 t\left\|\overrightarrow{y}^*\right\|^2}\right) \text{if } div\ \overrightarrow{u} = 0 \\ \left(C_0\left\|\overrightarrow{y}^*\right\| + C_1\left\|\overrightarrow{y}^*\right\|^2 e^{C_2 t\left\|\overrightarrow{y}^*\right\|^2}\right)\left(1 + C_3\left\|\overrightarrow{y}^*\right\|^2 e^{C_4 t\left\|\overrightarrow{y}^*\right\|^2}\right) \\ \quad \text{if } div\ \overrightarrow{u} \neq 0 \end{cases} \tag{4.30}$$

$$\|p\|_{C(\Omega_T)} \leq \begin{cases} \left(C_0\left\|\overrightarrow{y}^*\right\| + C_1\left\|\overrightarrow{y}^*\right\|^2 e^{C_2 t\left\|\overrightarrow{y}^*\right\|^2}\right) \text{if } div\ \overrightarrow{u} = 0 \\ \left(C_0\left\|\overrightarrow{y}^*\right\| + C_1\left\|\overrightarrow{y}^*\right\|^2 e^{C_2 t\left\|\overrightarrow{y}^*\right\|^2}\right)\left(1 + C_3\left\|\overrightarrow{y}^*\right\|^2 e^{C_4 t\left\|\overrightarrow{y}^*\right\|^2}\right) \\ \quad \text{if } div\ \overrightarrow{u} \neq 0 \end{cases} \tag{4.31}$$

where

$$\left\|\overrightarrow{y}^{*}\right\| = \begin{cases} \left\|\overrightarrow{u}_0\right\|_{C(\Omega)} + 2\left\|\overrightarrow{f}\right\|_{C(\Omega_T)} & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \left\|\overrightarrow{u}_0\right\|_{C^1(\Omega)} + 2\left\|\overrightarrow{f}\right\|_{C^1(\Omega_T)} & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

Theorem 3 Let functions

$$\overrightarrow{u}_0(x) = \begin{cases} \overrightarrow{u}_0(x) \in L_2(\Omega) & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{u}_0(x) \in H^{(1)}(\Omega) & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$
$$\overrightarrow{f}(x,t) = \begin{cases} \overrightarrow{f}(x,t) \in L_{2\Omega_T} & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{f}(x,t) \in H^{(1,0)}_{\Omega_T} & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

be periodic functions with the conditions $rot\ \overrightarrow{u}_0(x) \neq 0$ and $rot\ \overrightarrow{f} \neq 0$. Then there for the Navier-Stokes problem (4.1) – (4.2) exists a unique unstable periodic solution $\overrightarrow{u}(x,t) \in H^{(2,1)}(\Omega_T)$ which is given by the formula

$$\overrightarrow{u}(x,t) = \begin{cases} \overrightarrow{u}_\nu(x,t) = R_\nu\left[\overrightarrow{F}\right] + \overrightarrow{F} & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{u}_\alpha(x,t) = R_\alpha\left[\overrightarrow{F}\right] + \overrightarrow{F}^{*} & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases} \tag{4.32}$$

where $R_\nu[\cdot]$ and $R_\alpha[\cdot]$ are inverse operators for the nonlinear equation (4.4). Moreover there exists a unique scalar function of pressure $p(x,t) \in H^{(1,0)}(\Omega_T)$ which satisfies

$$\frac{u^2}{2} + \frac{p}{\rho} - div\ \overrightarrow{f} * \frac{1}{4\pi|x-\xi|} - div\ \overrightarrow{f}^{**} * \frac{1}{4\pi|x-\xi|} = 0 \tag{4.33}$$

where

$$\overrightarrow{f}^{**} = \begin{cases} \overrightarrow{b} - \overrightarrow{f}^{*} & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \overrightarrow{b}^{*} - \overrightarrow{f}^{*} & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases}$$

In addition, for all functions $\overrightarrow{u}(x,t) \in H^{(2,1)}(\Omega_T)$ and $p(x,t) \in H^{(1,0)}(\Omega_T)$ there exist positive constants C_0, C_1, C_2, C_4 for which are satisfied the following estimates

$$\left\|\overrightarrow{u}_\alpha\right\|_{H^{(2,1)}(\Omega_T)} \leq \begin{cases} \left(C_0\left\|\overrightarrow{y}^{*}\right\| + C_1\left\|\overrightarrow{y}^{*}\right\|^2 e^{C_2 t\left\|\overrightarrow{y}^{*}\right\|^2}\right) & \text{if } div\ \overrightarrow{u}(x,t) = 0 \\ \left(C_0\left\|\overrightarrow{y}^{*}\right\| + C_1\left\|\overrightarrow{y}^{*}\right\|^2 e^{C_2 t\left\|\overrightarrow{y}^{*}\right\|^2}\right)\left(1 + C_3\left\|\overrightarrow{y}^{*}\right\|^2 e^{C_4 t\left\|\overrightarrow{y}^{*}\right\|^2}\right) & \text{if } div\ \overrightarrow{u}(x,t) \neq 0 \end{cases} \tag{4.34}$$

$$\|p\|_{H^{(1,0)}(\Omega_T)} \leq \begin{cases} \left(C_0 \left\|\overrightarrow{y}^*\right\| + C_1 \left\|\overrightarrow{y}^*\right\|^2 e^{C_2 t \left\|\overrightarrow{y}^*\right\|^2}\right) \text{if } div\ \overrightarrow{u}\,(x,t) = 0 \\ \left(C_0 \left\|\overrightarrow{y}^*\right\| + C_1 \left\|\overrightarrow{y}^*\right\|^2 e^{C_2 t \left\|\overrightarrow{y}^*\right\|^2}\right)\left(1 + C_3 \left\|\overrightarrow{y}^*\right\|^2 e^{C_4 t \left\|\overrightarrow{y}^*\right\|^2}\right) \\ \text{if } div\ \overrightarrow{u}\,(x,t) \neq 0 \end{cases} \tag{4.35}$$

where

$$\left\|\overrightarrow{y}^*\right\| = \left\|\overrightarrow{u}_0\right\|^2_{H^{(1,0)}(\Omega)} + \left\|\overrightarrow{\Psi}\right\|^2_{H^{(2,1)}(\Omega_T)}$$

$$\overrightarrow{\Psi}\,(x,t) = \begin{pmatrix} \int\limits_{-\infty}^{\infty} \theta\,(x_3 - \varsigma_3)\, f_1\,(x_1, x_2, \varsigma_3, t)\, d\varsigma_3 \\ \int\limits_{-\infty}^{\infty} \theta\,(x_3 - \varsigma_3)\, f_2\,(x_1, x_2, \varsigma_3, t)\, d\varsigma_3 \\ \int\limits_{-\infty} \theta\,(x_3 - \varsigma_3)\, f_3\,(x_1, x_2, \varsigma_3, t)\, d\varsigma_3 \end{pmatrix},$$

$\theta(z)$ is Heaviside step function.

5 Conclusion and Future Work

In this research work we concentrate on those aspects of partial differential equations that can be represented in the terms of operators on continuous differentiable spaces. The Navier-Stokes equations have been the basis for description and analysis of all turbulent phenomena. Experimental selection of the regime turbulent fluctuation is costly and not always realizable process, therefore important argument for analytic research of the Navier-Stokes equation is to investigate an mathematical formulation which is based on the Green's function and required a good deals with the parabolic and elliptic potential theory. In processes of dealing with governing equations the main point stressed that the velocity vector and the function of pressure satisfy their criteria of stability motion which is the energy conservation law. The main mathematical difficulty lies in the determining the velocity vector which associates with the nonlinear of Volterra-Fredholm matrix integral equation. Here is found convenient procedure for the Navier-Stokes equations which allows to use 'a priori' estimates and to prove existence and regularity their weak or strong solutions. We have presented the analytic method for the divergent-convergent turbulent problem which is expected to exist for all infinite domains. There are

two unknown independent thermodynamic parameters (the velocity vector and the scalar function of pressure) which play a prominent role in the obtained integral representation of the velocity distribution for the description of the turbulent behavior of fluid motion. With respect to our method there we have shown steady motion which depends on the energy conversation law. The weak formulation of the Navier-Stokes problem is based on the extension of idea to the case where the energy falls in the critical domain, due to the pressure transition. Moreover, basic concepts of the Navier-Stokes equations have been investigated in Hilbert space and weak formulation is based on the introduced technique for the turbulent flow by using analytic investigation of the model initial-boundary problems which need in constriction the Green's function and the energy conservation law. In case when the weak solution of the Navier-Stokes problem (4.1)–(4.3) is sufficiently regular which means $\overrightarrow{u}(x,t) \in C^{(2,1)}(\Omega_T) \cap C^{(1,0)}\left(\overline{\Omega}_T\right)$ in this considered case and problem (4.1)–(4.3) also satisfies classical formulation. The obtained balance equation in terms on the energy conversation law provides equivalence of the strong and weak solution.

Conversely, unsteady behavior represents a departure from the average energy of the fluid known as eddy energy. Involving fundamental properties of the turbulent flows which demonstrate technology and principal importance at the forefront of classical approach where the expression

$$div \overrightarrow{f}^{**} * \frac{1}{4\pi |x-\xi|}$$

of the turbulent fluid energy is not regular and does not allow to use classical formulation. In this case some difficulties arise in solving of the Navier-Stokes problem which demonstrate the unstable motion for which we use weak formulation in studying turbulent behavior. These introduced mathematical endeavors can serve to enlarge our intuitive experiences respect to the nonlinear theories of partial differential equations. The fundamentals of our method have shown steady and unsteady behavior involving properties of the turbulent flows which demonstrate technological and principal importance at the forefront of classical numerical approach. Authors hope that this submitted analytic solution would be understood and would have been used for visualization of the turbulent processes and behavior of the pressure distribution in the considered areas.

In the future research we will deal with the general case when all coefficients ρ, ν, η of the Navier-Stokes equations depend on the spatial coordinate x and time t.

Acknowledgment This work was supported by Springer-ITET. The authors gratefully appreciate and acknowledge the Publishing Editor and staff of the International Association of Engineers for reading earlier draft of this paper, offering comments and encouragement.

References

1. Caffarelli, L., Kohn, R., Nirenberg, L.: Partial regularity of suitable weak solution. Pure Appl. Math. **35**, 771–831 (1982)
2. Sheffer, V.: An inviscid low with compact support in space time. J. Geom. Anal. **3**, 33–41 (1993)
3. Fefferman, C.L.: Existence and smoothness of the N-S equation, pp. 1–5. www.claymath.org (2000)
4. Kaliyeva, K.: The two-phase Stefan problem for the heat equation. Lecture notes in engineering and computer SCIENCE: proceedings of the world congress on engineering and computer science 2013, WCECS 2013, pp. 868–873. San Francisco, 23–25 Oct 2013
5. Kaliyeva, K., Kaliyev, A.: Existence and uniqueness of the Navier-Stokes problem in infinite space. In: Proceedings of the world congress on engineering 2014, WCE 2014, pp. 1288–1293. London, 2–4 July 2014
6. Kaliyeva, K.: Energy conversation law in the free atmosphere. International Journal of Engineering and Innovative Technology (IJEIT) **3**(11), 50–61 (2014). ISSN: 2277-3754

Chapter 5
Transient Problem for a Accreted Thermoelastic Block

Alexander L. Levitin and Sergei A. Lychev

Abstract The thermomechanics of growing bodies studies the distributions of mechanical and thermal fields in quasistatic and dynamic processes that occur in the bodies whose composition varies in the process of deformation and heating. These types of accretion are realized in various technological processes such as laser surfacing, gas-dynamic deposition, and vapor phase deposition. Mathematical modeling of the deformations and temperature fields arising in these processes allows one to optimize the technological processes and is a topical problem of mechanics of deformable rigid body. The present work is concerned with the initial boundary-valued problem for the thermoelastic growing block. Full coupling of mechanical and thermal fields as well as relaxing of the heat flux are taken into account. A closed form solution for a body under "smoothly rigid" heat-insulated conditions for the stationary faces and the load-free conditions on the growing face is obtained. The temperature field on the growing face is analyzed numerically for various accretion scenarios.

Keywords Coupled thermoelasticity • Discrete accretion • Growing bodies • Incompatibility • Micromechanics • Thermal inertia

1 Statement of the Problem

Mechanics of growing bodies studies the stress-strain state in bodies whose composition varies in the process of deformation [1–8]. A discretely accreted body is represented as a finite family of bodies [6, 7, 9]:

$$\mathfrak{B}_0 \subset \mathfrak{B}_1 \subset \mathfrak{B}_2 \subset \ldots \subset \mathfrak{B}_N. \tag{5.1}$$

A.L. Levitin (✉) • S.A. Lychev
Institute for Problems in Mechanics, pr-t Vernadskogo 101, str. 1, Moscow 119526, Russia
e-mail: alex_lev@ipmnet.ru; lychevsa@mail.ru

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_5

The sequence of sets (5.1) is associated with the sequence of numbers

$$0 < \tau_1 < \tau_2 < \ldots < \tau_N \tag{5.2}$$

determining the attachment times, i.e., the times at which the parts $\mathfrak{B}_{k+1} \setminus \mathfrak{B}_k$, $k = 0, \ldots, N-1$ are added to the body. The sequences (5.1) and (5.2) together determine the body growth scenario. The strain, temperature, and velocity fields of the part $\mathfrak{B}_{k+1} \setminus \mathfrak{B}_k$ added at time τ_k are generally inconsistent with the fields of the body $\mathfrak{B}_k$. Therefore, the dynamic processes in the growing body vary by jump at the times of attachments.

2 General Procedure

The process of dynamic discrete accretion can be modeled by successively solving the boundary value problems for the bodies $\mathfrak{B}_k$. The initial data for the step k ($k \geq 1$) are determined by the values of the corresponding fields at the final time of the step $k-1$ and by the values associated with the attached elements. Formally, the recursive sequence of problems can be stated as follows:

$$\forall \mathbf{x} \in \mathfrak{B}_0 \quad \mathscr{L}_0 \mathbf{y}_0 + \mathbf{f}^0 = \mathbf{0}, \quad \forall \mathbf{x} \in \partial\mathfrak{B}_0 \quad \mathscr{B}_0 \mathbf{y}_0 = \mathbf{0}, \quad \mathbf{y}_0\big|_{t=0} = \mathbf{y}_0^0, \quad \dot{\mathbf{y}}_0\big|_{t=0} = \mathbf{v}_0^0,$$

$$\ldots$$

$$\forall \mathbf{x} \in \mathfrak{B}_n \quad \mathscr{L}_n \mathbf{y}_n + \mathbf{f}_n^0 = \mathbf{0}, \quad \forall \mathbf{x} \in \partial\mathfrak{B}_n \quad \mathscr{B}_n \mathbf{y}_n = \mathbf{0}, \quad \mathbf{y}_n\big|_{t=\tau_n} = \mathbf{0}, \quad \dot{\mathbf{y}}_n\big|_{t=\tau_n} = \mathbf{v}_n^0,$$

$$\ldots$$

$$\mathbf{f}_n^0 = \mathbf{f}^0 + \begin{cases} \mathscr{L}_{n-1}\mathbf{y}_{n-1}\big|_{t=\tau_n}, & \mathbf{x} \in \mathfrak{B}_{n-1}, \\ 0, & \mathbf{x} \in \mathfrak{B}_n \setminus \mathfrak{B}_{n-1}, \end{cases} \qquad \mathbf{v}_n^0 = \begin{cases} \dot{\mathbf{y}}_{n-1}\big|_{t=\tau_n}, & \mathbf{x} \in \mathfrak{B}_{n-1}, \\ \mathbf{v}^0, & \mathbf{x} \in \mathfrak{B}_n \setminus \mathfrak{B}_{n-1}. \end{cases}$$

Here $\mathscr{L}_0, \ldots, \mathscr{L}_N$ are differential operators determined by the same differential operation (the field equations) but in different domains, $\mathscr{B}_0, \ldots, \mathscr{B}_N$ are operators of boundary conditions, $\mathbf{f}^0$ are external force and thermal fields, $\mathbf{v}^0$ are the velocities associated with the attached elements, and $\mathbf{y}_0, \ldots, \mathbf{y}_N$ are increments of the displacement and temperature fields with respect to the beginning of the step. The dot denotes the derivative with respect to time, and $\mathbf{y}_0^0$, $\mathbf{v}_0^0$ are the initial data for the first step.

The efficiency of such an algorithm depends on the solution representation for each step. For bodies of relatively simple shape and some classes of boundary conditions, one can find analytic solutions that leads to efficient computational algorithms.

3 Analytic Solution at a Step

The solution for a single step is constructed under the assumption that the displacements and excess temperatures, as well as their gradients, are small. This allows one to consider the problem for a step in the linear statement and to use the methods of expansion in biorthogonal systems of functions, which were developed in [10, 11].

The process of deformation is considered in the affine (point) space $\mathscr{E}$. Up to negligibly small variables, the body configuration images $\mathfrak{B}_k$ can be identified with a parallelepiped V_k embedded in the space $\mathscr{E}$. The quantities $\hat{a}$, $\hat{b}$, $\hat{h}$ are the parallelepiped linear dimensions, which are assumed to be constant in a single step but can vary in the course of the accretion process.

Assume that a mass measure ρ is introduced on the body $\mathfrak{B}_k$, and the body is under the action of the external field of mass forces $\hat{\mathbf{F}}(\hat{\mathbf{r}})$ and of distributed heat sources whose specific capacity is determined by the field $\hat{\omega}(\hat{\mathbf{r}})$. The displacements $\hat{\mathbf{u}} = \hat{u}_x\mathbf{i} + \hat{u}_y\mathbf{j} + \hat{u}_z\mathbf{k}$ caused by these fields and their gradients are assumed to be negligibly small with respect to the coordinates of the points $\hat{\mathbf{r}}$.

The body response is determined by the linear Duhamel–Neumann functional [12] for the stresses $\hat{\boldsymbol{\sigma}}(\boldsymbol{\varepsilon}, \hat{\theta})$

$$\hat{\boldsymbol{\sigma}}(\boldsymbol{\varepsilon}, \hat{\theta}) = 2\mu\boldsymbol{\varepsilon} + \lambda\mathbf{I} \otimes \mathbf{I}{:}\boldsymbol{\varepsilon} - (3\lambda + 2\mu)\alpha\hat{\theta}\mathbf{I}$$

and by the linear Fourier functional determining the thermal flow $\hat{\mathbf{h}}(\hat{\theta})$ depending on the gradient of the excess temperature $\hat{\theta}$

$$\hat{\mathbf{h}} = -\Lambda\hat{\boldsymbol{\nabla}}\hat{\theta}. \tag{5.3}$$

Here $\boldsymbol{\varepsilon} = \frac{1}{2}\left(\hat{\boldsymbol{\nabla}}\hat{\mathbf{u}} + (\hat{\boldsymbol{\nabla}}\hat{\mathbf{u}})^{\mathrm{T}}\right)$, $\hat{\boldsymbol{\nabla}} = \mathbf{i}\frac{\partial}{\partial\hat{x}} + \mathbf{j}\frac{\partial}{\partial\hat{y}} + \mathbf{k}\frac{\partial}{\partial\hat{z}}$ is the dimensional Hamiltonian operator, and $(\ldots)^{\mathrm{T}}$ is the transposition symbol, μ and λ are the Lamé moduli for the adiabatic state, α is the coefficient of linear thermal expansion, Λ is the thermal conductivity coefficient, $\hat{\theta} = T - T_0$ is the excess temperature, T is the absolute temperature, and $T_0 = \mathrm{const}$ is the reference temperature.

Under the above assumptions, the coupled system of equations of motion and heat conduction has the form [12]:

$$\begin{cases} \mu\hat{\boldsymbol{\nabla}}^2\hat{\mathbf{u}} + (\lambda + \mu)\hat{\boldsymbol{\nabla}}\hat{\boldsymbol{\nabla}}\cdot\hat{\mathbf{u}} - (3\lambda + 2\mu)\alpha\hat{\boldsymbol{\nabla}}\hat{\theta} - \rho\frac{\partial^2}{\partial\hat{t}^2}\hat{\mathbf{u}} + \rho\hat{\mathbf{F}} = \mathbf{0}, \\ \Lambda\hat{\boldsymbol{\nabla}}^2\hat{\theta} - \rho c\frac{\partial}{\partial\hat{t}}\hat{\theta} - (3\lambda + 2\mu)T_0\alpha\,\hat{\boldsymbol{\nabla}}\cdot\left(\frac{\partial}{\partial\hat{t}}\hat{\mathbf{u}}\right) + \rho\hat{\omega} = 0, \end{cases} \tag{5.4}$$

where c is the specific heat at constant deformation referred to unit mass and $\hat{\boldsymbol{\nabla}}^2 = \hat{\boldsymbol{\nabla}}\cdot\hat{\boldsymbol{\nabla}} = \frac{\partial^2}{\partial\hat{x}^2} + \frac{\partial^2}{\partial\hat{y}^2} + \frac{\partial^2}{\partial\hat{z}^2}$ is the dimensional Laplace operator. The fact that system (5.4) is coupled can be explained by the presence of the temperature gradient in the equations of motion and by the presence of the dilatation rate in the

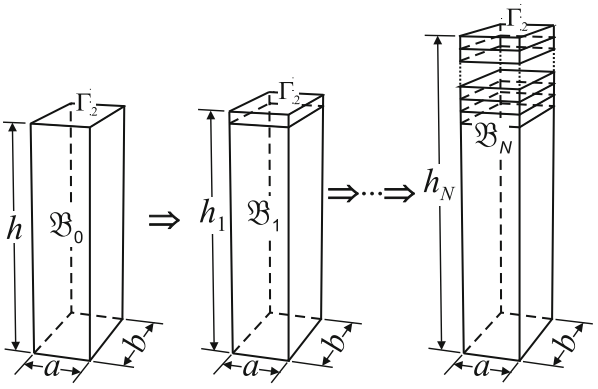


Fig. 5.1 The discrete process growing

heat equation. In the literature, the coupled thermoelasticity of such a type is said to be "completely coupled." The equations of the so-called theory of temperatures stresses [12] do not take into account the dilatation rate influence on the process of heat conduction and represent "partially" coupled thermoelasticity. It was shown in [13] that taking into account the dilatation rate in heat equations leads to a significant correction of the solution for bodies of micron dimensions.

Consider the boundary conditions corresponding to the "smoothly rigid" heat-insulated fixation (in [14], exact solutions were investigated for an elastic parallelepiped with "sliding fixation") on five of its faces Γ_1 and the free heat-insulated face Γ_2 (Fig. 5.1):

$$\hat{\mathbf{u}}\cdot\mathbf{n}\big|_{\Gamma_1}=0,\ \ \mathbf{n}\cdot\hat{\boldsymbol{\sigma}}\cdot(\mathbf{I}-\mathbf{n}\otimes\mathbf{n})\big|_{\Gamma_1}=\mathbf{0},\ \ \hat{\mathbf{h}}\cdot\mathbf{n}\big|_{\Gamma_1}=0,\ \ \mathbf{n}\cdot\hat{\boldsymbol{\sigma}}\big|_{\Gamma_2}=\mathbf{0},\ \ \hat{\mathbf{h}}\cdot\mathbf{n}\big|_{\Gamma_2}=0. \tag{5.5}$$

The initial conditions determine the distributions of the initial displacements $\hat{\mathbf{u}}_0$, velocities $\hat{\mathbf{v}}_0$ and excess temperature $\hat{\theta}_0$:

$$\hat{\mathbf{u}}\big|_{\hat{t}=0}=\hat{\mathbf{u}}_0,\quad \frac{\partial}{\partial\hat{t}}\hat{\mathbf{u}}_0\big|_{\hat{t}=0}=\hat{\mathbf{v}}_0,\quad \hat{\theta}\big|_{\hat{t}=0}=\hat{\theta}_0. \tag{5.6}$$

For our purpose it is more convenient to pass to dimensionless spatial coordinates $\mathbf{r}$ relative to the characteristics dimension R_0 and to the dimensionless time t which are related to the dimensional variables as follows:

$$\mathbf{r}=\frac{\hat{\mathbf{r}}}{R_0}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k},\quad x=\frac{\hat{x}}{R_0},\quad y=\frac{\hat{y}}{R_0},\quad z=\frac{\hat{z}}{R_0},\quad t=\frac{\hat{t}}{R_0}\sqrt{\frac{\mu}{\rho}}.$$

$$\mathbf{u}=\frac{\hat{\mathbf{u}}}{R_0}=u_x\mathbf{i}+u_y\mathbf{j}+u_z\mathbf{k},\quad u_x=\frac{\hat{u}_x}{R_0},\quad u_y=\frac{\hat{u}_y}{R_0},\quad u_z=\frac{\hat{u}_z}{R_0},\quad \theta=\frac{\hat{\theta}}{T_0}.$$

In terms of dimensionless variables, Eqs. (5.4) are reduced to the form

$$\begin{cases} \nabla^2\mathbf{u} + A\nabla\nabla\cdot\mathbf{u} - B\nabla\theta - \ddot{\mathbf{u}} + \mathbf{F} = \mathbf{0}, \\ \nabla^2\theta - C\dot{\theta} - D\nabla\cdot\dot{\mathbf{u}} + \omega = 0, \end{cases} \tag{5.7}$$

where A, B, C, D are dimensionless coefficients

$$A = \frac{\lambda+\mu}{\mu}, \quad B = \frac{(3\lambda+2\mu)\alpha T_0}{\mu}, \quad C = \frac{R_0 c\sqrt{\rho\mu}}{\Lambda}, \quad D = \frac{(3\lambda+2\mu)R_0\alpha\sqrt{\mu}}{\Lambda\sqrt{\rho}}.$$

In Eqs. (5.7) and later, the dot denotes the derivative $\frac{\partial}{\partial t}$ on dimensionless time t, and the fields $\mathbf{F}$ and ω are the dimensionless densities of the force and thermal actions $\mathbf{F} = \frac{R_0\rho\hat{\mathbf{F}}}{\mu}$ and $\omega = \frac{R_0^2\rho\hat{\omega}}{\Lambda T_0}$.

The boundary and initial conditions (5.5)–(5.6) in terms of dimensionless variables are stated as

$$\begin{aligned} \mathbf{u}\cdot\mathbf{n}\otimes\mathbf{n} + \mathbf{n}\cdot\big(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathrm{T}}\big)\cdot(\mathbf{I} - \mathbf{n}\otimes\mathbf{n})\big|_{\Gamma_1} = \mathbf{0}, \quad \mathbf{n}\cdot\nabla\theta\big|_{\Gamma_1} = 0, \\ \mathbf{n}\cdot\big(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathrm{T}}\big) + \mathbf{n}\left((A-1)\nabla\cdot\mathbf{u} - B\theta\right)\big|_{\Gamma_2} = \mathbf{0}, \quad \mathbf{n}\cdot\nabla\theta\big|_{\Gamma_2} = 0, \\ \mathbf{u}\big|_{t=0} = \mathbf{u}_0, \quad \dot{\mathbf{u}}\big|_{t=0} = \mathbf{v}_0, \quad \theta\big|_{t=0} = \theta_0, \end{aligned} \tag{5.8}$$

where $\mathbf{u}_0 = \frac{\hat{\mathbf{u}}_0}{R_0}$, $\mathbf{v}_0 = \hat{\mathbf{v}}_0\sqrt{\frac{\rho}{\mu}}$ and $\theta_0 = \frac{\hat{\theta}_0}{T_0}$ are dimensionless initial data.

Equations (5.7) along with conditions (5.8) present the statement of the initial–boundary value problem in dimensionless form. The solution of this problem is sought as the form of spectral expansion according to [10, 11, 15, 16]. Since the differential operators generated by the system of Eqs. (5.7) are not self-adjoint, it is necessary to consider complex-valued functions, and hence it is necessary to introduce a Hilbert space $\mathscr{H}$ on the set of complex-valued vector functions

$$\mathbf{a} = \begin{pmatrix} \mathbf{u} \\ \theta \end{pmatrix} = \begin{pmatrix} u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k} \\ \theta \end{pmatrix}$$

defined and square integrable in the domain V and satisfying the condition that the bilinear form determining the inner product $\langle\cdot,\cdot\rangle$ in $\mathscr{H}$ is meaningful, i.e.

$$\forall\mathbf{a},\mathbf{b}\in\mathscr{H}:\ \langle\mathbf{a},\mathbf{b}\rangle = \left\langle\begin{pmatrix}\mathbf{u}\\ \theta\end{pmatrix}, \begin{pmatrix}\mathbf{v}\\ \xi\end{pmatrix}\right\rangle = \int_V \left(\mathbf{u}\cdot\bar{\mathbf{v}} + \theta\bar{\xi}\right)\,dV\ \in\ \mathbb{C}.$$

Here the bar denotes the complex conjugation.

The coupled equations of motion and heat conduction (5.7) along with the boundary conditions (5.8) determine in $\mathscr{H}$ a *pencil of differential operators*:

$$\mathscr{L}_\nu = \begin{pmatrix} \nabla^2 - \nu^2\mathbf{I} + A\nabla\nabla\cdot & -B\nabla \\ -D\nu\nabla\cdot & \nabla^2 - C\nu \end{pmatrix}. \tag{5.9}$$

which can rewrite in polynomial form with operator coefficients

$$\mathscr{L}_\nu = \mathscr{L}_0 + \nu\mathscr{L}_1 + \nu^2\mathscr{L}_2, \tag{5.9′}$$

$$\mathscr{L}_0 = \begin{pmatrix} \nabla^2 + A\nabla\nabla\cdot & -B\nabla \\ \mathbf{0} & \nabla^2 \end{pmatrix}, \quad \mathscr{L}_1 = \begin{pmatrix} \mathbf{0} & 0 \\ -D\nabla\cdot & -C \end{pmatrix}, \quad \mathscr{L}_2 = \begin{pmatrix} -\mathbf{I} & 0 \\ \mathbf{0} & 0 \end{pmatrix}.$$

The domain of the operator $\mathscr{L}_\nu$ is denoted by $\mathscr{D}$ and may be defined as:

$$\mathscr{D} = \left\{\mathbf{a} \;\middle|\; \mathscr{B}_1\mathbf{a} = \mathbf{0} \wedge \mathscr{B}_2\mathbf{a} = \mathbf{0}\right\} \subset \mathscr{H},$$

$$\mathscr{B}_1\begin{pmatrix}\mathbf{u}\\ \theta\end{pmatrix} = \begin{pmatrix} \mathbf{u}\cdot\mathbf{n}\otimes\mathbf{n} + \mathbf{n}\cdot\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathrm{T}}\right)\cdot(\mathbf{I} - \mathbf{n}\otimes\mathbf{n})\big|_{\Gamma_1} \\ \mathbf{n}\cdot\nabla\theta\big|_{\Gamma_1} \end{pmatrix},$$

$$\mathscr{B}_2\begin{pmatrix}\mathbf{u}\\ \theta\end{pmatrix} = \begin{pmatrix} \mathbf{n}\cdot\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathrm{T}}\right) + \mathbf{n}\left((A-1)\nabla\cdot\mathbf{u} - B\theta\right)\big|_{\Gamma_2} \\ \mathbf{n}\cdot\nabla\theta\big|_{\Gamma_2} \end{pmatrix},$$

where $\mathscr{B}_1$ and $\mathscr{B}_2$ are called the operators of boundary conditions.

The operator pencil $\mathscr{L}_\nu$ is associated with the adjoint pencil $\mathscr{L}_\nu^*$ defined in the domain $\mathscr{D}^* \subset \mathscr{H}$ and satisfying the following relations:

$$\forall\mathbf{u} \in \mathscr{D}, \quad \forall\mathbf{v} \in \mathscr{D}^* \quad \langle\mathscr{L}_\nu\mathbf{u}, \mathbf{v}\rangle = \langle\mathbf{u}, \mathscr{L}_\nu^*\mathbf{v}\rangle\,.$$

The adjoint pencil can be defined explicitly [10] in following form

$$\mathscr{L}_\nu^* = \begin{pmatrix} \nabla^2 - \bar\nu^2\mathbf{I} + A\nabla\nabla\cdot & D\bar\nu\nabla \\ B\nabla\cdot & \nabla^2 - C\bar\nu \end{pmatrix}, \quad \mathscr{D}^* = \left\{\mathbf{a} \;\middle|\; \mathscr{B}_1^*\mathbf{a} = \mathbf{0} \wedge \mathscr{B}_2^*\mathbf{a} = \mathbf{0}\right\} \subset \mathscr{H},$$

$$\mathscr{B}_1^* = \mathscr{B}_1, \quad \mathscr{B}_2^*\begin{pmatrix}\mathbf{u}\\ \theta\end{pmatrix} = \begin{pmatrix} \mathbf{n}\cdot(\nabla\mathbf{u} + (\nabla\mathbf{u})^*) + \mathbf{n}\left((A-1)\nabla\cdot\mathbf{u} + \bar\nu D\theta\right)\big|_{\Gamma_2} \\ \mathbf{n}\cdot\nabla\theta\big|_{\Gamma_2} \end{pmatrix}.$$

which can be rewritten in form similar to (5.9′), i.e.

$$\mathscr{L}_\nu^* = \mathscr{L}_0^* + \bar\nu\mathscr{L}_1^* + \bar\nu^2\mathscr{L}_2, \quad \mathscr{L}_0^* = \begin{pmatrix} \nabla^2 + A\nabla\nabla\cdot & 0 \\ B\nabla\cdot & \nabla^2 \end{pmatrix}, \quad \mathscr{L}_1^* = \begin{pmatrix} \mathbf{0} & D\nabla \\ \mathbf{0} & -C \end{pmatrix}.$$

The pair of mutually adjoint operators $\mathscr{L}_\nu$ and $\mathscr{L}_\nu^*$ generates mutually adjoint eigenvalue problems (generalized Sturm–Liouville problems)

$$\mathscr{L}_\nu\mathbf{e} = \mathbf{0}, \quad \mathscr{L}_\nu^*\mathbf{e}^* = \mathbf{0}. \tag{5.10}$$

Since the domain V is bounded and the differential operator (5.9) is regular, the non-trivial solutions of problems (5.10) gives countable sequences of complex-valued

vector functions $\{\mathbf{e}_i\}$, $\{\mathbf{e}_i^*\}$, $i = 1, 2, \ldots$, which correspond to the sequence of generalized eigenvalues $\{\nu_i\}$. Due to the fact that the operator pencils are not self-adjoint, the eigenvalues are located on the complex plane in a rather complicated way, but since the coefficient D is positive and small, one can show that the eigenvalues form three subsequences one of which consists of numbers on the negative real semiaxis and has the limit point $-\infty$, and the other two sequences are located near the imaginary axis in the negative half-plane and have the limit points $\pm i\infty$. The same reasoning shows that the function systems $\{\mathbf{e}_i\}$ and $\{\mathbf{e}_i^*\}$ are complete in $\mathscr{H}$. Moreover, the functions $\mathbf{e}_i$ and $\mathbf{e}_j^*$ $i \neq j$ satisfy the biorthogonality conditions [10]:

$$\left\langle \mathscr{L}_1 \mathbf{e}_i, \mathbf{e}_j^* \right\rangle + (\nu_i + \nu_j) \left\langle \mathscr{L}_2 \mathbf{e}_i, \mathbf{e}_j^* \right\rangle = 0, \tag{5.11}$$

which can explicitly be written as

$$-\int_V \left[(D\nabla\cdot\mathbf{u}_i + C\theta_i)\,\bar{\theta}_j^* + (\nu_i + \nu_j)\mathbf{u}_i\cdot\bar{\mathbf{u}}_j^* \right] dV = 0, \quad \mathbf{e}_i = \begin{pmatrix} \mathbf{u}_i \\ \theta_i \end{pmatrix}, \ \mathbf{e}_j^* = \begin{pmatrix} \mathbf{u}_j^* \\ \theta_j^* \end{pmatrix}. \tag{5.11$'$}$$

Thus, the function systems $\{\mathbf{e}_i\}$ and $\{\mathbf{e}_i^*\}$ form mutually biorthogonal bases in $\mathscr{H}$. According to [10], the solution of the initial–boundary value problem (5.7), (5.8) can be represented by the expansion

$$\begin{pmatrix} \mathbf{u} \\ \theta \end{pmatrix} = \sum_{i=1}^{\infty} \left[\left(\left\langle \begin{pmatrix} \mathbf{u}_0 \\ \theta_0 \end{pmatrix}, \mathscr{L}_1^* \mathbf{e}_i^* + \bar{\nu}_i \mathscr{L}_2 \mathbf{e}_i^* \right\rangle + \left\langle \begin{pmatrix} \mathbf{v}_0 \\ 0 \end{pmatrix}, \mathscr{L}_2 \mathbf{e}_i^* \right\rangle \right) e^{\nu t} \right.$$
$$\left. - \int_0^t \left\langle \begin{pmatrix} \mathbf{F} \\ \omega \end{pmatrix}, \mathbf{e}_i^* \right\rangle e^{\nu_i (t-\tau)} \right] \mathbf{e}_i N_i^{-1}, \tag{5.12}$$

where $N_i = \left\langle \mathscr{L}_1 \mathbf{e}_i, \mathbf{e}_i^* \right\rangle + 2\nu_i \left\langle \mathscr{L}_2 \mathbf{e}_i, \mathbf{e}_i^* \right\rangle$ are the normalizing factors.

4 Account the Relaxing of the Heat Flux

One of the drawbacks of the classical law (5.3) is the fact that the thermal signal transmission speed turns out to be infinite in applications. Nevertheless, it effectively describes heat propagation in a wide range of applications. Therefore, the Fourier law has to be modified only if the relaxation time and the external action pulse duration are quantities of the same order of magnitude [17, 18]. However in many additive processes material is attached in thin laminars under intense external forcing and heating. It means that accounting of the relaxation time may be significant in the thermomechanical models for growing solids.

In the last 200 years, various linear generalizations of the Fourier heat conduction law have been suggested (a survey and historic comments are given in [19]). In the literature, most frequently used is the Cattaneo–Jeffreys law, which can be written as

$$\tau_1 \frac{\partial \hat{\mathbf{h}}}{\partial \hat{t}} + \hat{\mathbf{h}} = -\Lambda \hat{\nabla} \hat{\theta}, \tag{5.13}$$

where τ_1 is the relaxation time.

In the present paper, we use the heat conduction law with two relaxation times τ_1 and τ_2, namely,

$$\tau_1 \frac{\partial \hat{\mathbf{h}}}{\partial \hat{t}} + \hat{\mathbf{h}} = -\Lambda \left(\hat{\nabla} \hat{\theta} + \tau_2 \hat{\nabla} \frac{\partial \hat{\boldsymbol{\theta}}}{\partial \hat{t}} \right). \tag{5.14}$$

Now let us reformulate the statement of the problem (5.4). The heat conduction equation in system (5.4) has to be replaced by the following one

$$\Lambda \left(\hat{\nabla}^2 \hat{\theta} + \tau_2 \hat{\nabla}^2 \frac{\partial \hat{\theta}}{\partial \hat{t}} \right) - \rho c \left(\tau_1 \frac{\partial^2 \hat{\theta}}{\partial \hat{t}^2} + \frac{\partial \hat{\theta}}{\partial \hat{t}} \right) - (3\lambda + 2\mu)\alpha \hat{\theta}_0 \left(\tau_1 \hat{\nabla} \cdot \frac{\partial^2 \hat{\mathbf{u}}}{\partial \hat{t}^2} + \hat{\nabla} \cdot \frac{\partial \hat{\mathbf{u}}}{\partial \hat{t}} \right) + \tau_1 \frac{\partial \hat{\omega}}{\partial \hat{t}} + \hat{\omega} = 0. \tag{5.15}$$

The boundary conditions, corresponding to heat conduction and initial temperature rate distribution $\hat{\vartheta}_0$, may be written as

$$\left[C \left(\tau_1 \frac{\partial \hat{\theta}}{\partial \hat{t}} + \hat{\theta} \right) - D \hat{\Lambda} \mathbf{n} \cdot \left(\tau_2 \hat{\nabla} \frac{\partial \hat{\theta}}{\partial \hat{t}} + \hat{\nabla} \hat{\theta} \right) \right] \Bigg|_{\partial V} = 0, \qquad \left. \frac{\partial \hat{\theta}}{\partial \hat{t}} \right|_{t=0} = \hat{\vartheta}_0.$$

In dimensionless variables the mention above equations have the following form

$$\nabla^2 \theta + E \nabla^2 \dot{\theta} - F \ddot{\theta} - C \dot{\theta} - G \nabla \cdot \ddot{\mathbf{u}} - D \nabla \cdot \dot{\mathbf{u}} + \omega^{\mathrm{CJ}} = 0. \tag{5.16}$$

The coefficients E, F, G and inhomogeneous term ω^{CJ} can be written as

$$E = \frac{\tau_2 \sqrt{\mu}}{R_0 \sqrt{\rho}}, \quad F = \frac{\tau_1 c \mu}{\Lambda}, \quad G = \frac{\tau_1 (3\lambda + 2\mu)\alpha \mu}{\rho \Lambda}, \quad \omega^{\mathrm{CJ}} = \frac{R_0^2}{\Lambda T_0} \left(\tau_1 \frac{\partial \hat{\omega}}{\partial \hat{t}} + \hat{\omega} \right).$$

The boundary conditions have the dimensionless form

$$\left(G \dot{\theta} + C \theta - H \mathbf{n} \cdot \nabla \dot{\theta} - D \nabla \mathbf{n} \cdot \nabla \theta \right) \Big|_{\partial V} = 0,$$
$$G = C \frac{\tau_1}{\Lambda R_0} \sqrt{\frac{\mu}{\rho}}, \quad H = D \frac{\tau_2}{R_0^2} \sqrt{\frac{\mu}{\rho}}. \tag{5.17}$$

The initial data in dimensionless form are formulated as

$$\theta|_{t=0} = \frac{\hat{\theta}_0}{T_0}, \quad \dot{\theta}\Big|_{t=0} = \frac{\hat{\vartheta}_0}{T_0} R_0 \sqrt{\frac{\rho}{\mu}}.$$

The initial-boundary problems in framework of Cattaneo–Jeffreys type thermoelasticity generate quadratic pencils of the form

$$\mathscr{L}_\nu^{\mathrm{CJ}} = \mathscr{L}_0^{\mathrm{CJ}} + \nu \mathscr{L}_1^{\mathrm{CJ}} + \nu^2 \mathscr{L}_2^{\mathrm{CJ}}. \tag{5.18}$$

The operator coefficients are formulated as

$$\mathscr{L}_0^{\mathrm{CJ}} = \begin{pmatrix} \nabla^2 + A\nabla \otimes \nabla\cdot & -B\nabla \\ \mathbf{0} & \nabla^2 \end{pmatrix}, \quad \mathscr{L}_1^{\mathrm{CJ}} = \begin{pmatrix} \mathbf{0} & 0 \\ -D\nabla\cdot & E\nabla^2 - C \end{pmatrix}, \quad \mathscr{L}_2^{\mathrm{CJ}} = \begin{pmatrix} -\mathbf{I} & 0 \\ -G\nabla\cdot & -F \end{pmatrix}.$$

We recall that the differential operators $\mathscr{L}_0^{\mathrm{CJ}}$, $\mathscr{L}_1^{\mathrm{CJ}}$, and $\mathscr{L}_2^{\mathrm{CJ}}$ have a common domain of definition $\mathscr{D}^{\mathrm{CJ}}$ determined by the boundary condition operator $\mathscr{B}^{\mathrm{CJ}}$:

$$\mathscr{B}^{\mathrm{CJ}} \begin{pmatrix} \mathbf{u} \\ \theta \end{pmatrix} = \begin{pmatrix} [\nabla\mathbf{u} + (\nabla\mathbf{u})^*]\cdot\mathbf{n} + \mathbf{n}(A-1)\nabla\cdot\mathbf{u} - B\theta \\ E\nu\theta + C\theta - F\nu\mathbf{n}\cdot\nabla\theta - D\mathbf{n}\cdot\nabla\theta \end{pmatrix}\Big|_{\partial V}$$

For the initial-boundary problem under study, the adjoint pencil has the form

$$\mathscr{L}_\nu^{\mathrm{CJ}*} = \mathscr{L}_0^{\mathrm{CJ}*} + \bar{\nu}\mathscr{L}_1^{\mathrm{CJ}*} + \bar{\nu}^2\mathscr{L}_2^{\mathrm{CJ}*}, \quad \mathscr{L}_1^{\mathrm{CJ}*} = \begin{pmatrix} \mathbf{0} & D\nabla \\ \mathbf{0} & E\nabla^2 - C \end{pmatrix}, \quad \mathscr{L}_2^{\mathrm{CJ}*} = \begin{pmatrix} \mathbf{I} & G\nabla \\ \mathbf{0} & -F \end{pmatrix}.$$

Note that $\mathscr{L}_0^{\mathrm{CJ}*}$ and $\mathscr{L}_0^*$ are identical. The domain of the adjoint pencil is:

$$\mathscr{D}^{\mathrm{CJ}*} = \{\mathbf{y} \mid \mathscr{B}^{\mathrm{CJ}*}[\mathbf{y}] = \mathbf{0}\},$$

$$\mathscr{B}^{\mathrm{CJ}*} \begin{pmatrix} \mathbf{u} \\ \theta \end{pmatrix} = \begin{pmatrix} [\nabla\mathbf{u} + (\nabla\mathbf{u})^*]\cdot\mathbf{n} + \mathbf{n}(A-1)\nabla\cdot\mathbf{u} + \bar{\nu}D\theta \\ E\bar{\nu}\theta + C\theta - F\bar{\nu}\mathbf{n}\cdot\nabla\theta + B\mathbf{n}\cdot\nabla\theta \end{pmatrix}\Big|_{\partial V}.$$

5 Representations of the Eigenfunctions

Let us focus on system (5.4). The representation of functions of the family $\{\mathbf{e}_i\}$ (and respectively, of $\{\mathbf{e}_i^*\}$) which is consistent with the boundary conditions (5.8) has the form

$$\mathbf{e} = \begin{pmatrix} \mathbf{u} \\ \theta \end{pmatrix} = \begin{pmatrix} \mathbf{i}\,\sin(nx)\cos(my)a(z) + \mathbf{j}\,\cos(nx)\sin(my)b(z) + \mathbf{k}\,\cos(nx)\cos(my)c(z) \\ \cos(nx)\,\cos(my)\,d(z) \end{pmatrix},$$

where, for brevity, the following notation is used: $n = \frac{n'\pi}{a}$, $m = \frac{m'\pi}{b}$, $n', m' \in \mathbb{N}$, and $\mathbb{N}$ is the set of positive integers; the functions $a(z)$, $b(z)$, $c(z)$, and $d(z)$ are determined by solving the system of linear differential equations

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1+A & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .\mathbf{U}'' + \begin{bmatrix} 0 & 0 & -An & 0 \\ 0 & 0 & -Am & 0 \\ An & Am & 0 & -B \\ 0 & 0 & -D\nu & 0 \end{bmatrix} .\mathbf{U}'$$

$$- \begin{bmatrix} (1+A)n^2+m^2+\nu^2 & Amn & 0 & -Bn \\ Amn & (1+A)m^2+n^2+\nu^2 & 0 & -Bm \\ 0 & 0 & n^2+m^2+\nu^2 & 0 \\ D\nu n & D\nu m & 0 & n^2+m^2+C\nu \end{bmatrix} .\mathbf{U}=\mathbf{0}. \tag{5.19}$$

Here $\mathbf{U} = (a(z), b(z), c(z), d(z))^{\mathrm{T}}$ is a formal four-component vector function, and the prime $'$ denotes the derivatives with respect to the variable z. Note that system (5.19) is not symmetric, because the original problem is not self-adjoint.

Although the order of the system of Eqs. (5.19) is rather high, the fundamental system of its solutions has the simple form

$$\mathbf{U}_1 = \begin{bmatrix} -m \\ n \\ 0 \\ 0 \end{bmatrix} e^{\zeta z}, \quad \mathbf{U}_2 = \begin{bmatrix} -m \\ n \\ 0 \\ 0 \end{bmatrix} e^{-\zeta z}, \quad \mathbf{U}_3 = \begin{bmatrix} -\zeta \\ 0 \\ n \\ 0 \end{bmatrix} e^{\zeta z}, \quad \mathbf{U}_4 = \begin{bmatrix} \zeta \\ 0 \\ n \\ 0 \end{bmatrix} e^{-\zeta z},$$

$$\mathbf{U}_5 = \begin{bmatrix} m \\ n \\ -\xi \\ \frac{2D\nu^2}{P} \end{bmatrix} e^{\xi z}, \quad \mathbf{U}_6 = \begin{bmatrix} m \\ n \\ \xi \\ \frac{2D\nu^2}{P} \end{bmatrix} e^{-\xi z}, \quad \mathbf{U}_7 = \begin{bmatrix} m \\ n \\ -\eta \\ \frac{2D\nu^2}{Q} \end{bmatrix} e^{\eta z}, \quad \mathbf{U}_8 = \begin{bmatrix} m \\ n \\ \eta \\ \frac{2D\nu^2}{Q} \end{bmatrix} e^{-\eta z},$$

$$\zeta = \sqrt{\nu^2 + n^2 + m^2}, \quad \xi = \sqrt{n^2 + m^2 + \nu\frac{Q + 2\nu}{2(A+1)}}, \quad \eta = \sqrt{n^2 + m^2 + \nu\frac{P + 2\nu}{2(A+1)}},$$

$$P = R + \sqrt{R^2 + M}, \; Q = R - \sqrt{R^2 + M}, \; R = AC + BD + C - \nu, \; M = 4\nu BD. \tag{5.20}$$

By substituting the fundamental solutions into the boundary conditions corresponding to the upper and lower faces of the parallelepiped ($z = 0, h$), one obtains a system of eight linear equations for the constants of integration of system (5.19). By equating the determinant of the coefficients of this system with zero, one obtains an equation for the eigenvalues ν_i. This determinant can be written in the very concise form

$$\left[e^{2\zeta h}-1\right]\Big[P^2\eta L^2\sinh(\zeta h)\cosh(\xi h)\sinh(\eta h)$$
$$+M\xi L^2\sinh(\zeta h)\sinh(\xi h)\cosh(\eta h)$$
$$-2\zeta\xi\eta\left(L-\nu^2\right)\left(P^2+M\right)\cosh(\zeta h)\sinh(\xi h)\sinh(\eta h)\Big]=0, \quad (5.21)$$

where $L=\nu^2+2(n^2+m^2)$.

The eigenfunctions can be classified by expanding the left-hand side of Eq. (5.21). The first class consists of the eigenfunctions associated with the zeros of the first factor on the left-hand side in Eq. (5.21); i.e.,

$$e^{2\zeta h}-1=0 \quad \Rightarrow \quad 2\zeta h=2\pi i k, \quad k=1,2,\ldots.$$

The corresponding eigenvalues and eigenfunctions can be written as

$$\nu^{(I)}_{nmk}=i\sqrt{n^2+m^2+\left(\frac{k\pi}{h}\right)^2}, \quad n,m,k\in\mathbb{N},$$

and the corresponding eigenfunctions have the form

$$\mathbf{e}^{(I)}_{nmk}=\begin{pmatrix}\mathbf{u}\\ \theta\end{pmatrix}=\begin{pmatrix}-\mathbf{i}\,m\sin(nx)\cos(my)\cos\left(\frac{k\pi z}{h}\right)+\mathbf{j}\,n\sin(nx)\cos(my)\cos\left(\frac{k\pi z}{h}\right)\\ 0\end{pmatrix}. \quad (5.22)$$

The eigenfunctions of the second class determine the coupled thermoelastic vibrations. The corresponding eigenvalues can be determined as complex-values zeros of the second factor of the left-hand side of Eq. (5.21). It can be done only numerically. For the initial approximations one may take the real roots of the equation for uncoupled heat conduction and the pure imaginary roots of the equation for elastic vibrations, which are determined by the equations

$$\sinh(\eta h)=0, \quad L^2\sinh(\zeta h)\cosh(\xi h)-2\zeta\xi\left(L-\nu^2\right)\cosh(\zeta h)\sinh(\xi h)=0.$$

Note that for $n=m=0$ and $L=\nu^2$, the second factor in Eq. (5.21) may be factorized

$$\sinh(\zeta h)\left[P^2\eta\cosh(\xi h)\sinh(\eta h)+M\xi\sinh(\zeta h)\sinh(\xi h)\cosh(\eta h)\right]=0. \quad (5.23)$$

The complex-valued eigenfunctions can be written in the following form:

$$\mathbf{e}^{(II)}_{nmk}=\begin{pmatrix}\mathbf{i}\,n\Big[Y\zeta\cosh(\zeta z)+W\cosh(\xi z)-V\cosh(\eta z)\Big]\sin(nx)\cos(my)\\ +\mathbf{j}\,m\Big[Y\zeta\cosh(\zeta z)+W\cosh(\xi z)-V\cosh(\eta z)\Big]\cos(nx)\sin(my)\\ +\mathbf{k}\Big[\eta V\sinh(\eta z)-Y(n^2+m^2)\sinh(\zeta z)-\xi W\sinh(\xi z)\Big]\cos(nx)\cos(my)\\ 2D\nu^2\left[\frac{W}{P}\cosh(\xi z)-\frac{V}{Q}\cosh(\eta z)\right]\cos(nx)\cos(ny)\end{pmatrix},$$

where the following notation is used

$$
\begin{aligned}
Y &= 2L\left[\xi \sinh(h\xi)\cosh(h\eta) - \eta\cosh(h\xi)\sinh(h\eta)\right],\\
V &= 4[n^2+m^2]\zeta\xi\sinh(h\xi)\cosh(h\zeta) - L^2\cosh(h\xi)\sinh(h\zeta),\\
W &= 4[n^2+m^2]\zeta\eta\sinh(h\eta)\cosh(h\zeta) - L^2\cosh(h\eta)\sinh(h\zeta).
\end{aligned}
\tag{5.24}
$$

The eigenvalues of the operator (5.18), generated in the framework of the thermoelasticity of Cattaneo–Jeffreys type can be found as roots of Eq. (5.21), with the following equations for terms

$$
R = A\frac{C+F\nu}{1+E\nu} + B\frac{D+G\nu}{1+E\nu} + \frac{C+F\nu}{1+E\nu} - \nu, \quad M = 4\nu B\frac{D+G\nu}{1+E\nu}. \tag{5.20$'$}
$$

The eigenfunctions corresponded to the eigenvalues ν_i of Cattaneo–Jeffreys problem can be calculated by

$$
\mathbf{e}^{\mathrm{CJ}(II)}_{nmk} = \begin{pmatrix}
\mathbf{i}\, n\Big[Y\zeta\cosh(\zeta z) + W\cosh(\xi z) - V\cosh(\eta z)\Big]\sin(nx)\cos(my)\\
+\mathbf{j}\, m\Big[Y\zeta\cosh(\zeta z) + W\cosh(\xi z) - V\cosh(\eta z)\Big]\cos(nx)\sin(my)\\
+\mathbf{k}\Big[\eta V\sinh(\eta z) - Y(n^2+m^2)\sinh(\zeta z) - \xi W\sinh(\xi z)\Big]\cos(nx)\cos(my)\\
2\frac{D+F\nu}{1+E\nu}\nu^2\left[\frac{W}{P}\cosh(\xi z) - \frac{V}{Q}\cosh(\eta z)\right]\cos(nx)\cos(ny)
\end{pmatrix}
$$

with coefficients (5.24) and (5.20′).

6 Numerical Simulation

As was already noted, the effects of coupled mechanical and thermal fields play a significant role for bodies of micron dimensions. Therefore, it is of special interest to model the process of thermoelastic accretion for bodies of such scales.

Consider the thermoelastic process of growth of a microscopic copper crystal which is represented as a growing parallelepiped with the initial dimensions $\hat{a} = R_0$, $\hat{b} = 2R_0$, and $\hat{h} = 4R_0$. Here R_0 is the characteristic dimension equal to $1\,\mu$m (i.e., $R_0 = 10^{-6}$ m). The numerical simulation is based on the following physical and mechanical characteristics: $\lambda = 89.4708$ GPa, $\mu = 40.9531$ GPa, $\rho = 8{,}960\,\mathrm{kg/m^3}$, $\alpha = 16.4\cdot 10^{-6}$ 1/K, $\Lambda = 385$ W/(m·K), and $c = 385$ J/(kg·K). Reference temperature is $T_0 = 20^\circ$ C$= 293$ K.

For full coupled thermoelasticity thermal conductivity of with Cattaneo–Jeffreys type we use $\hat{\tau}_1 = \hat{\tau}_2 = 1$ ps according [17, 18].

Consider the "fast" accretion process. It is assumed that the growth is uniform, i.e., $\Delta h = h_{k+1} - h_k = \mathrm{const}$ and $\Delta\tau = \tau_{k+1} - \tau_k = \mathrm{const}$. The dimensionless rate of the material attachment is $v_a = \frac{\Delta h}{\Delta\tau} = 0.1$, which corresponds to the dimensional

rate $\hat{v}_a = 213\,\text{m/s}$. This rate value is associated with the characteristic time the growing body height doubling $\hat{\tau}_a = 18.71\,\text{ns}$. Note that τ_a is approximately five times greater than the period of the fundamental mode of the body vibrations at the beginning of the accretion process (3.658 ns) and is five times less than the basic relaxation time (91.27 ns), i.e., the value inverse to the real eigenvalue that is the least in absolute value. The given rate value $\hat{v}_a$ is associated with different accretion scenarios corresponding to different Δh. There is an analogy with different partitions in the theory of integration. A decrease in Δh leads to a decrease in $\Delta\tau = \frac{\Delta h}{v_a}$, and the process of discrete accretion approaches a continuous process. Therefore, the numerical simulation of such a limit process is of great interest.

For further computations, it is assumed that the excess dimensionless temperature of the attached body $\mathfrak{B}_{k+1} \setminus \mathfrak{B}_k$ is equal to $\theta\big|_{t=\tau_k} = 1$, and the excess temperature of the body at the beginning of the growth $\mathfrak{B}_0$ is equal to $\theta\big|_{t=0} = 0$.

It is also assumed that there are no mass forces $\hat{\mathbf{F}}$ and no internal heat sources $\hat{\omega}$, and at the initial time moment, the growing body was free from stresses and at rest. The computational algorithm implementation according to the solution representation constructed in the paper allows one to determine the stresses, temperature, and velocities of points of the discretely built-up body at any time moment.

Figure 5.2a shows the dependence of the temperature θ at the middle points of the parallelepiped cross-sections ($x = \frac{a}{2}$, $y = \frac{b}{2}$) on the coordinate z and time t, which corresponds to the accretion scenario with successively attached 15 layers ($N = 15$) of dimensionless thickness $\Delta h = 0.2$ ($\Delta\hat{h} = 0.2\,\mu\text{m}$). For the accepted accretion rate $v_a = 0.1$, one has the quantity $\Delta\tau = \frac{\Delta h}{v_a} = 2$ ($\Delta\hat{\tau} = 0.09355\,\text{ns}$), and the parallelepiped height increases from $\hat{h} = 4\,\mu\text{m}$ to $\hat{h}_{15} = 7\,\mu\text{m}$. For these parameters, the accretion process has the same typical time parameters as the heat transfer process. In particular, this is illustrated by the graph; namely, the temperature on the growth boundary smoothly varies from the initial temperature of the body to the temperature of the attached layers. Figure 5.2b illustrates the temperature distribution for the rate of growth that is ten times greater than v_a. The effective (step-averaged) temperature of the growing boundary is close to the temperature of the attached layers. Figure 5.2c shows the temperature distribution for the rate of growth that is ten times less that v_a. Figure 5.3 shows the temperature on the moving boundary (thin line) and its value averaged over the step (thick line).

An analysis of the temperature behavior on the growth boundary shows that, depending on the accretion rate, the boundary can be considered as an isothermal boundary (for high values of the accretion rate) or a boundary with variable "effective" temperature determined in the process of solving the problem.

The eigenvalues corresponded to three above mentioned thermoelastic models are shown in Tables 5.1 and 5.2. It easy to see the difference between the eigenvalues computed in the framework of the theory of thermal stresses (ν_0) and corresponded eigenvalues computed from the equations of full coupled thermoelasticity (ν and ν^{CJ}) is rather significant (it is about 1.2–1.8 %).

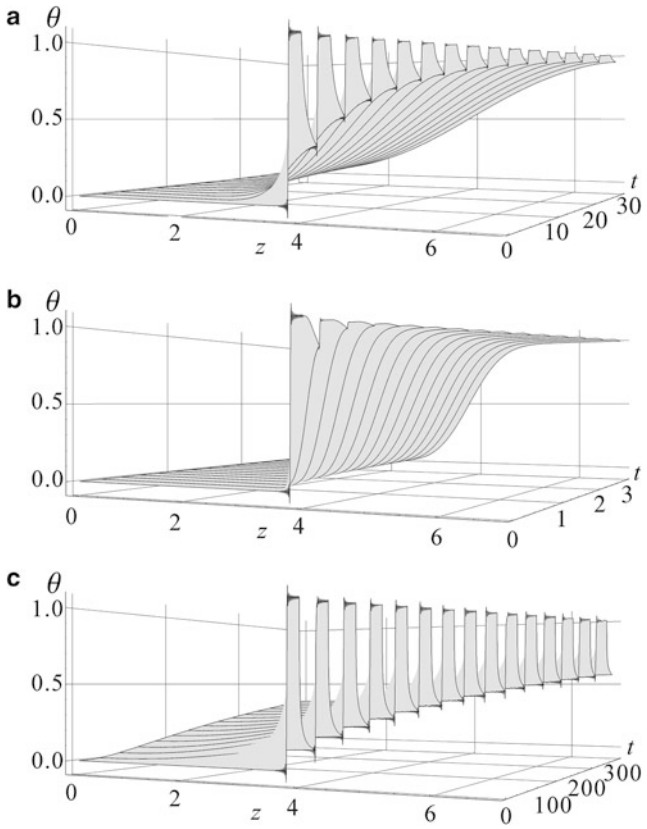


Fig. 5.2 The dependence of the temperature for the different rate of growth: (**a**) v_a, (**b**) $10v_a$, (**c**) $0.1v_a$

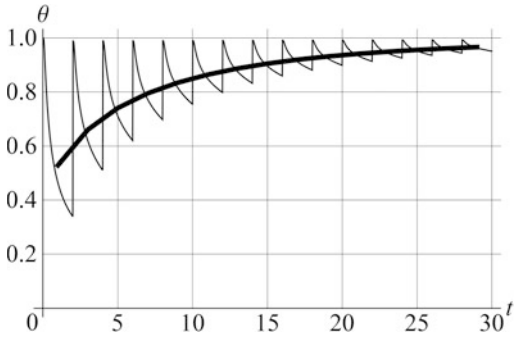


Fig. 5.3 The temperature on the moving boundary (*thin line*) and its value averaged over the step (*thick line*)

Table 5.1 Real-valued eigenvalues of the operator pencil $\mathscr{L}_\nu$ and $\mathscr{L}_\nu^{\mathrm{CJ}}$

	n	m	Theory of thermal stresses ν_0	Full coupled thermoelasticity with thermal conductivity of		$\frac{\nu-\nu_0}{\nu_0}100$	$\frac{\nu^{\mathrm{CJ}}-\nu}{\nu}100$
				Fourier type ν	Cattaneo–Jeffreys type ν^{CJ}		
1	0	0	−0.032202	−0.0317187	−0.0317188	1.52364	0.000284852
2	0	0	−0.128808	−0.126877	−0.126878	1.52188	0.00113900
3	0	1	−0.161010	−0.158295	−0.158298	1.71489	0.00142429
4	0	1	−0.257616	−0.253571	−0.253577	1.59519	0.00227899
5	0	0	−0.289818	−0.285481	−0.285489	1.51895	0.00256378
6	0	1	−0.418626	−0.412254	−0.412270	1.54549	0.00370456
7	0	0	−0.515231	−0.507543	−0.507566	1.51489	0.00456036
8	1	0	−0.547433	−0.538559	−0.538585	1.64778	0.00484682
9	0	2	−0.547433	−0.538559	−0.538585	1.64778	0.00484682
10	1	0	−0.644039	−0.633843	−0.633879	1.60870	0.00570314
11	0	2	−0.644039	−0.633843	−0.633879	1.60870	0.00570314
12	0	1	−0.644039	−0.634365	−0.634402	1.52498	0.00570234
13	1	1	−0.676241	−0.665369	−0.665408	1.63409	0.00598907
.....							
430	2	2	−11.8825	−11.7256	−11.7383	1.3379	0.108117
431	1	4	−11.8825	−11.7256	−11.7383	1.3379	0.108117

Table 5.2 Complex-valued eigenvalues of the operator pencil $\mathscr{L}_\nu$

	n	m	Theory of thermal stresses ν_0	Full coupled thermoelasticity with thermal conductivity of Fourier type ν
1	0	0	$-0.80333i$	$-5.49380\cdot10^{-5}-0.80942i$
2	0	1	$-1.47639i$	$-9.81100\cdot10^{-5}-1.47714i$
3	0	1	$-1.95012i$	$-5.97739\cdot10^{-6}-1.95018i$
4	0	0	$-2.40998i$	$-5.13872\cdot10^{-4}-2.42824i$
5	0	1	$-2.49713i$	$-1.13153\cdot10^{-5}-2.49730i$
6	0	1	$-2.98597i$	$-1.59008\cdot10^{-4}-2.98946i$
7	0	1	$-3.29092i$	$-8.70426\cdot10^{-4}-3.31273i$
8	1	0	$-3.30391i$	$-2.41040\cdot10^{-5}-3.30402i$
9	0	1	$-3.53557i$	$-3.31433\cdot10^{-4}-3.54296i$
10	1	1	$-3.65211i$	$-2.38101\cdot10^{-5}-3.65221i$
11	1	0	$-3.67909i$	$-1.79510\cdot10^{-5}-3.67919i$
12	0	1	$-3.96753i$	$-9.44726\cdot10^{-4}-3.98784i$
13	1	1	$-3.99111i$	$-2.33138\cdot10^{-5}-3.99122i$

The difference between eigenvalues computed by means of conventional relations of Fourier type thermoelasticity and by means of relations with account of heat inertia is not so significant ($2 \cdot 10^{-4}\% \ldots 0.1\%$). This difference increases with an eigenvalue number. Hence one have to take into account heat inertia for modelling of high frequency and short impulse heating.

The complex-valued eigenvalues that are situated in the neighborhood of the imaginary axis are listed in Table 5.2.

Acknowledgements This work was supported by the Russian Science Foundation under Grant 14-19-01280.

References

1. Arutyunyan, N.K., Manzhirov, A.V., Naumov, V.E.: Contact Problems in Mechanics of Growing Solids. Nauka, Moscow (1991) [in Russian]
2. Manzhirov, A.V.: The general non-inertial initial-boundaryvalue problem for a viscoelastic ageing solid with piecewise-continuous accretion. J. Appl. Math. Mech. **59**(5), 805–816 (1995). doi:10.1016/0021-8928(95)00095-X
3. Manzhirov, A.V., Parshin, D.A.: Accretion of a viscoelastic ball in a centrally symmetric force field. Mech. Solids **41**(1), 51–64 (2006)
4. Lychev, S.A., Lycheva, T.N., Manzhirov, A.V.: Unsteady vibration of a growing circular plate. Mech. Solids **46**(2), 325–333 (2011). doi:10.3103/S002565441102021X
5. Kuznetsov, S.I., Manzhirov, A.V., Fedotov, I.: Heat conduction problem for a growing ball. Mech. Solids **46**(6), 929–936 (2011). doi:10.3103/S0025654411060124
6. Manzhirov, A.V., Lychev, S.A.: The mathematical theory of growing solids: finite deformations. Dokl. Phys. **57**(4), 160–163 (2012). doi:10.1134/S1028335812040015
7. Levitin, A.L., Lychev, S.A., Manzhirov, A.V., Shatalov, M.Y.: Nonstationary vibrations of a discretely accreted thermoelastic parallelepiped. Mech. Solids **47**(6), 677–689 (2012). doi:10.3103/S0025654412060106
8. Lychev, S.A., Manzhirov, A.V.: The mathematical theory of growing bodies: finite deformations. J. Appl. Math. Mech. **77**(4), 421–432 (2013). doi:10.1016/j.jappmathmech.2013.11.011
9. Levitin, A.L., Lychev, S.A., Saifutdinov, I.N.: Transient dynamical problem for a accreted thermoelastic parallelepiped. In: Proceedings of the World Congress on Engineering 2014 (WCE 2014), London, 2–4 July 2014. Lecture Notes in Engineering and Computer Science, pp. 1196–1201 (2014)
10. Lychev, S.A., Senitskii, Y.E.: Nonsymmetric finite integral transformations and their applications to viscoelasticity problems. Vestnik Samar. Gos. Univ. Estestvennonauchn. Ser. **Special Issue**, 16–38 (2002) [in Russian]
11. Lychev, S.A.: Coupled dynamic thermoviscoelasticity problem. Mech. Solids **43**(5), 769–784 (2008). doi:10.3103/S0025654408050129
12. Nowacki, W.: Theory of Elasticity. PWN, Warsaw (1970)
13. Lychev, S.A., Manzhirov, A.V., Joubert, S.V.: Closed solutions of boundary-value problems of coupled thermoelasticity. Mech. Solids **45**(4), 610–623 (2010). doi:10.3103/S0025654410040102
14. Zhilin, P.A., Il'icheva, T.P.: Spectra and oscillation mode shapes of a rectangular parallelepiped obtained using three-dimensional theory of elasticity and theory of plates. Mech. Solids **15**(2), 94–103 (1980)

15. Lychev, S.A., Manzhirov, A.V.: Differential operators associated with the equations of motion and nondissipative heat conduction in the Green–Naghdi theory of thermoelasticity. J. Phys. Conf. Ser. **181**, 012096 (2009). doi:10.1088/1742-6596/181/1/012096
16. Polyanin, A.D., Lychev, S.A.: Various representations of the solutions of systems of equations of continuum mechanics. Dokl. Phys. **59**(3), 148–152 (2014). doi:10.1134/S1028335814030069
17. Marciak-Kozlowska, J., Kozlowski, M.: The thermal inertia of materials heated with a laser pulse faster than relaxation time. Int. J. Thermophys. **17**(5), 1099–1111 (1996). doi:10.1007/BF01441998
18. Wang, X., Xu, X.: Thermoelastic wave indused by pulsed laser heating. Appl. Phys. A. **73**(1), 107–114 (2001). doi:10.1007/s003390000593
19. Joseph, D.D., Preziosi, L.: Heat waves. Mod. Phys. **61**(1), 41–73 (1989)

Chapter 6
Developing of a 1-D Combustion Model and Study of Engine Performance and Exhaust Emission Using Ethanol-Gasoline Blends

Simeon Penchev Iliev

Abstract The problem with crude oil depletion has arisen in the last years. There has been intensive research to find out alternative to fossil fuels. The main group is biofuels, produced from crops and wastes. Alcohols are an important category of bio-fuels. Ethanol is a good candidate to be an alternative fuel since it is a liquid and has several physical and chemical properties similar to those of gasoline fuels. That is why this study is aimed to develop the 1-D combustion model of four-stroke spark ignited engine for predicting the effect of various fuel types on engine performances and fuel consumption on various engine operating conditions. AVL Boost was used as a computational fluid dynamics simulation tool to analyze the performance and emissions characteristics for different blends of ethanol and gasoline (by volume).

Keywords Alternative fuels • Engine simulation • Engine performance • Ethanol blends • Exhaust emissions • Internal combustion engine

1 Introduction

Fuel additives are very important because many of these additives can be added to fuel in order to improve its efficiency and its performance. Some of the most important additives to improve fuel performance are oxygen containing organic compounds (oxygenates). Several oxygenates have been used as fuel additives, such as methanol, ethanol, tertiary butyl alcohol and methyl tertiary butyl ether [1].

Presently, ethanol is a prospective fuel for use in vehicle as an alternative to petroleum based fuels. Ethanol is advocated as the prospective fuel because it can be manufactured from natural products or waste materials, unlike gasoline which is a non-renewable energy resource [2, 3]. One of the importantfeatures is that the

S.P. Iliev (✉)
Department of Engines and Vehicles, University of Ruse, 8 Studentska Str., Ruse 7017, Bulgatia
e-mail: spi@uni-ruse.bg

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_6

ethanol can be used directly without requiring any major changes in the structure of the engine. Among the various alcohols, ethanol is known as the most suitable fuel for spark ignited (SI) engines.

On the combustion characteristics, the auto-ignition temperature and flash point of ethanol are higher than those of gasoline, which makes it safer for transportation and storage. The latent heat of evaporation of ethanol is between three and five times higher than that of gasoline; this makes the temperature of the intake manifold lower, and increases the volumetric efficiency. The heating value of ethanol is lower than that of the gasoline. Therefore, we need 1.6 times more alcohol fuel to achieve the same energy output. The stoichiometric AFR (air–fuel ratio) of ethanol is about 2/3 that of the gasoline, so the required amount of air for complete combustion is lesser for alcohol [4].

There is plenty of literature to various blends of ethanol and gasoline. Palmer [5] studied the effect of using various blend rates of ethanol–gasoline fuels in engine tests. Results indicated that 10 % ethanol addition increases the engine power output by 5 %, and the octane number can be increased by 5 % for each 10 % ethanol added. He indicated that 10 % of ethanol addition to gasoline could reduce the concentration of CO emissions up to 30 %. Bata et al. [6] studied different blend rates of ethanol–gasoline fuels in engines, and found that the ethanol could reduce the CO and UHC emissions to some degree. The reduction of CO emissions are apparently caused by the wide flammability and oxygenated characteristic of ethanol.

Kim et al. [7] estimated that the potential for ethanol production is equivalent to about 32 % of the total gasoline consumption worldwide, when used in 85 % ethanol in gasoline for a midsize passenger vehicle.

The gasoline engine performance theory linked together with computer modeling of the engine thermodynamics in engine simulations is a great challenge, as the latter make the most complete use of the former and the models used are becoming widespread. Engine modeling is a very large subject, in part because of the range of engine configurations possible and the variety of alternative analytical techniques or sub-models, which can be applied in overall engine models Engine modeling is a fruitful research area and as a result many research laboratories have produced their own engine thermodynamics models with varying degrees of complexity, scope and ease to use [8–11].

Engine simulation is becoming an increasingly important engineering tool for time and cost efficiency in the development of internal combustion engines (ICEs). Most of results that are obtained by simulation are rather difficult to be obtained experimentally. The use of Computational Fluid Dynamics (CFD) simulations allow researchers to understand flow behavior and quantify important flow parameters such as mass flow rates or pressure drops, provided that the CFD tools have been properly validated against experimental results. For reasons such as the aforementioned, CFD simulations have become a valuable tool in helping both the analysis and design of the intake and exhaust systems of an ICEs. Many processes in the engine are 3-D but it requires greater knowledge and large computational time. Thus simplified 1-D simulation is often used. There are several components

that manifest a complex three-dimensional flow behavior, such as turbo machinery or manifolds which cannot be simulated properly by 1-D codes, and thus require viscous, 3-D codes.

Hence, it is a right choice to save computational time by simulating the complex components by means of a 3-D code and modeling the rest of the system with a 1-D code, i.e. the ducts. In this way, a coupling methodology between the 1-D and the 3-D code in the respective interfaces is required, and has become the objective of numerous authors [14–17].

In 1-D simulation, equations for conservation of mass, momentum, and energy are solved in time and in one dimension along the main flow direction in the engine pipes. Additional models, correlations, or measurements are needed in 1-D capturing 3-D phenomena such as flow over valves and combustion [12, 13].

The present paper aims to develop the 1-D combustion model of four-stroke port fuel injection (PFI) gasoline engine for predicting the effect of ethanol–gasoline (E0, E5, E10, E20, E30 and E50) fuel blends on the performance of SI engine. For this purpose, simulation of calibrated gasoline engine model was used as basic operating condition and the laminar burning velocity correlations of ethanol–gasoline fuel blends for calculating the changed combustion duration. The engine performances: torque and specific fuel consumption were compared and discussed.

2 Theoretical Study

2.1 Simulation Setup

The 1-D engine simulation model is developed by using the software AVL BOOST and has been employed to study the engine performance working on ethanol-gasoline blends.

The engine model used in this simulation was performed on a four stroke, four cylinder spark ignition engine with port fuel injection. The gasoline engine model was calibrated by AVL and its layout is shown in Fig. 6.1 with engine specification shown in Table 6.1.

The pre-processing step of AVL Boost enable the user to model a 1-Dimensional engine test bench setup using the predefined elements provided in the software toolbox. The various elements are joined by the desired connectors to establish the complete engine model using pipelines.

In Fig. 6.1, E1 represent the engine while C1, C2, C3 and C4 represent the number of cylinders of the engine. MP1 to MP18 represent the measuring points. PL1, PL2, PL3 and PL4 represent the plenum. SB1 and SB2 are for the system boundary. The flow pipes are numbered 1–34. CL1 represent the cleaner. R1 to R10 represent flow restrictions, CAT1 represent catalyst and I1 to I4 represent fuel injectors.

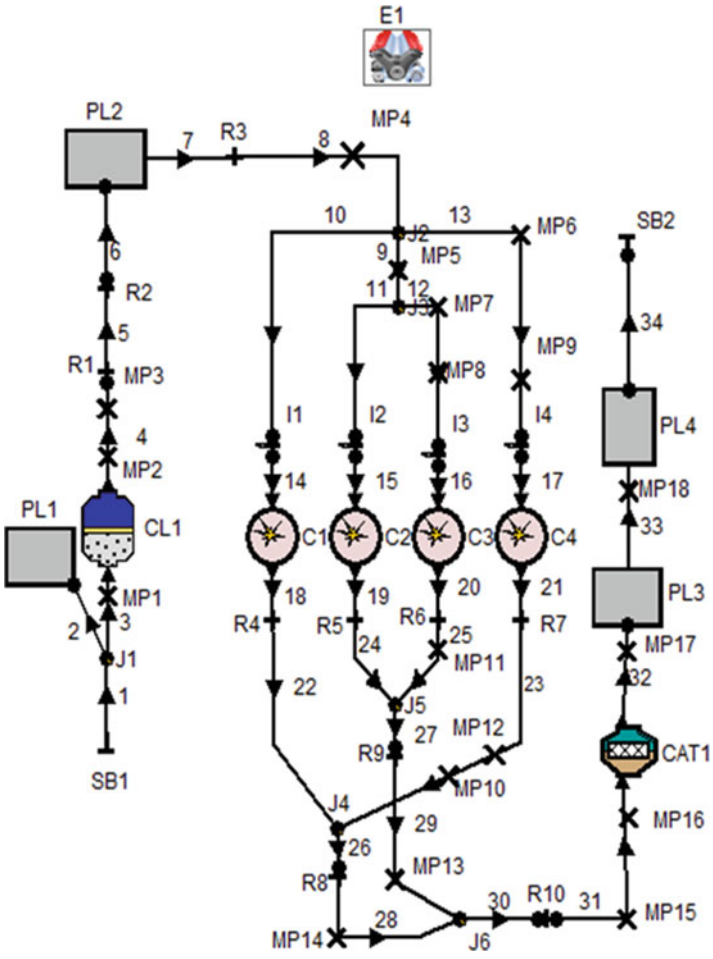


Fig. 6.1 Layout of gasoline engine model

Table 6.2 presents a comparison of properties of gasoline and ethanol. As shown in Table 6.2, compared with ethanol and gasoline, ethanol has a higher elemental oxygen content, heat of evaporation, octane number and a lower heating value, elemental molecular weight, elemental carbon and hydrogen content and stoichiometric air/fuel ratio (AFR).

2.2 Combustion Model

For the current study Vibe two zone model was selected for the combustion analysis. This model divides the combustion chamber into unburned and burned gas regions

Table 6.1 Engine specification

Engine parameters	Value
Bore (mm)	86
Stroke (mm)	86
Compression ratio	10.5
Connection rod length (mm)	143.5
Number of cylinder	4
Piston pin offset (mm)	0
Displacement (cc)	2,000
Intake valve open (deg)	20 BTDC
Intake valve close (deg)	70 ABDC
Exhaust valve open (deg)	50 BBDC
Exhaust valve close (deg)	30 ATDC
Piston surface area (mm^2)	5,809
Cylinder surface area (mm^2)	7,550
Number of stroke	4

BTC before top death center, *ABDC* after bottom death center, *BBDC* before bottom death center, *ATDC* after top death center

Table 6.2 Comparison of fuel properties

Properties	Gasoline	Ethanol
Chemical formula	C_8H_{15}	C_2H_5OH
Molecular weight	111.21	46.07
Oxygen content, wt%	–	34.73
Carbon content, wt%	86.3	52.2
Hydrogen content, wt%	24.8	13.1
Stoichiometric AFR	14.5	8.94
Lower heating value, Mj/kg	44.3	27
Heat of evaporation, Kj/kg	305	840
Research octane number	96.5	111
Motor octane number	87.2	92
Vapor pressure (psi at 37,7 °C)	4.5	2

[18]. However the assumption that burned and unburned charges have the same temperature is dropped. Instead the first law of thermodynamics is applied to the burned charge and unburned charge respectively.

$$\frac{dm_b u_b}{d\alpha} = -p_c \frac{dV_b}{d\alpha} + \frac{dQ_F}{d\alpha} - \sum \frac{dQ_{Wb}}{d\alpha} + h_u \frac{dm_b}{d\alpha} - h_{BB,b} \frac{dm_{BB,b}}{d\alpha}. \quad (6.1)$$

$$\frac{dm_u u_u}{d\alpha} = -p_c \frac{dV_u}{d\alpha} - \sum \frac{dQ_{Wu}}{d\alpha} - h_u \frac{dm_B}{d\alpha} - h_{BB,u} \frac{dm_{BB,u}}{d\alpha}. \quad (6.2)$$

where *dmu* is the change of the internal energy in the cylinder, $p_c \frac{dV}{d\alpha}$ is the piston work, $\frac{dQ_F}{d\alpha}$ is the fuel heat input, $\frac{dQ_W}{d\alpha}$ is wall heat loses, $h_u \frac{dm_b}{d\alpha}$ is the enthalpy flow from the unburned to the burned zone due to the conversion of a fresh charge to combustion products. Heat flux between the two zones is neglected. $h_{BB} \frac{dm_{BB}}{d\alpha}$ is the enthalpy due to blow by, *u* and *b* in the subscript are unburned and burned gas.

In addition the sum of the volume changes must be equal to the cylinder volume change and the sum of the zone volumes must be equal to the cylinder volume.

$$\frac{dV_b}{d\alpha} + \frac{dV_u}{d\alpha} = \frac{dV}{d\alpha}. \tag{6.3}$$

$$V_b + V_u = V. \tag{6.4}$$

The amount of mixture burned at each time setup is obtained from the Vibe function. For all other terms, like wall heat losses etc., models similar to the single zone models with an appropriate distribution of the two zones are used [19].

2.3 Emissions Model

The NOx formation model in AVL Boost is based on Pattas and Hafner [19] which incorporates the well known Zeldovich mechanism [20]. The rate of NOx production was estimated by using the following Eq. (6.5):

$$r_{NO} = C_{PPM} C_{KM} (2,0) \left(1 - \alpha^2\right) \left(\frac{r_1}{1 + \alpha A K_2} + \frac{r_4}{1 + A K_4} \right). \tag{6.5}$$

where $\alpha = \frac{C_{NO.act}}{C_{NO.equ}} \frac{1}{C_{KM}}$, $AK_2 = \frac{r_1}{r_2 + r_3}$, $AK_4 = \frac{r_4}{r_5 + r_6}$.

In the above equation, C_{PPM} denotes Post Processing Multiplier, C_{KM} denotes Kinetic Multiplier, C denotes molar concentration in equilibrium and r_i denotes reactions rates of Zeldovich mechanism.

The NOx formation model in AVL Boost is based on Onorati et al. [21].

$$r_{CO} = C_{Const} (r_1 + r_2) (1 - \alpha) \tag{6.6}$$

where, $\alpha = \frac{C_{CO.act}}{C_{CO.equ}}$.

In the above equation, C denotes molar concentration in equilibrium and r_i denotes reactions rates based on the model.

In a spark ignition engine the unburned hydrocarbons have different sources. A complete description of their formation process cannot yet be given and definitely the achievement of a reliable predictive model within a thermodynamic approach is prevented by the fundamental assumptions and the requirement of reduced computational times. Nevertheless a phenomenological model which accounts for

the main formation mechanisms and is able to capture the HC trends as function of the engine operating parameter may be proposed. The following major sources of unburned hydrocarbons can be identified in spark ignition engines [22]:

1. A fraction of the charge enters the crevice volumes and is not burned since the flame quenches at the entrance.
2. Fuel vapor is absorbed into the oil layer and deposits on the cylinder wall during intake and compression. The following desorption takes place when the cylinder pressure decreases during the expansion stroke and complete combustion cannot take place any more.
3. Quench layers on the combustion chamber wall which are left as the flame extinguishes prior to reaching the walls.
4. Occasional partial burning or complete misfire occurring when combustion quality is poor.
5. Direct flow of fuel vapor into the exhaust system during valve overlap in PFI engines.

The first two mechanisms and in particular the crevice formation are considered to be the most important and need to be accounted for in a thermodynamic model. Quench layer and partial burn effect cannot be physically described in a quasidimensional approach, but may be included by adopting tunable semiempirical correlations.

The process of formation of unburned hydrocarbons in the crevices is described by assuming that, the pressure in the cylinder and in the crevices is the same and that the temperature of the mass in the crevice volumes is equal to the piston temperature.

The mass in the crevices at any time is given by Eq. (6.7):

$$m_{crevice} = \frac{pV_{crevice}M}{RT_{piston}}. \tag{6.7}$$

In the above equation, $m_{crevice}$ is the mass of unburned charge in the crevice, p is cylinder pressure, $V_{crevice}$ is total crevice volume, M is unburned molecular weight, R gas constant and T_{piston} piston temperature.

A second significant source of hydrocarbon is the presence of lubricating oil in the fuel or on the walls of the combustion chamber. In fact, during compression, the fuel vapor pressure increases so, by Henry's law, absorption occurs even if the oil was saturated during the intake. During combustion the fuel vapor concentration in the burned gases goes to zero so the absorbed fuel vapor will desorb from the liquid oil into the burned gases. Fuel solubility is a positive function of the molecular weight, so the oil layer contributed to HC emissions depending on the different solubility of individual hydrocarbons in the lubricating oil.

The assumptions made in the development of the HC absorption/desorption are the following:

1. The oil film is at the same temperature as the cylinder wall;
2. Fuel is constituted by a single hydrocarbon species, completely vaporized in the fresh mixture;

3. Oil is represented by squalane ($C_{30}H_{62}$), whose characteristics are similar to the SAE5W20 lubricant;
4. Transverse flow across the oil film is negligible;
5. Diffusion of the fuel in the oil film is the limiting factor, since the diffusion constant in the liquid phase is 104 times smaller than the corresponding value in the gas phase.

Under these hypotheses the radial distribution of the fuel mass fraction in the oil film can be determined by solving the diffusion equation (6.8):

$$\frac{\partial w_F}{\partial t} - D\frac{\partial^2 w_F}{\partial r^2} = 0 \tag{6.8}$$

In the above equation, w_F denotes mass fraction of the fuel in the oil film, t denotes time, r denotes radial position in the oil film (distance from the wall), D denotes relative (fuel-oil) diffusion coefficient.

AVL Boost evaluates the position of the flame front at every time-step and accumulates only HCs that are desorbed into the burned gases, since any HC released into the unburned mixture would be burned by the propagating flame front.

For the current study Vibe two zone model was selected for the combustion analysis. This model divides the combustion chamber into unburned and burned gas regions.

3 Result and Discussion

The present study concentrated on the emissions and performance characteristics of the ethanol-gasoline blends. Different concentrations of the blends 0 % Ethanol (E0), 5 % Ethanol (E5), 10 % Ethanol (E10), 20 % Ethanol (E20), 30 % Ethanol (E30) and 50 % Ethanol (E50) by volume were analyzed using AVL BOOST at full load conditions for the speeds ranging from 1,000 to 6,500 rpm in the steps of 500 rpm. The results are divided into different subsections based on the parameter analyzed.

3.1 Engine Performance Characteristics

The results of the brake power, torque, and specific fuel consumption for ethanol gasoline blended fuels at different engine speeds are presented here.

Figure 6.2 shows the influence of ethanol gasoline blended fuels on engine brake power.

The brake power is one of the important factors that determine the performance of an engine. The variation of brake power with speed was obtained at full load conditions for E5, E10, E20, E30, E50 and pure gasoline E0 using the CFD results. When the ethanol content in the blended fuel was increased, the engine brake power

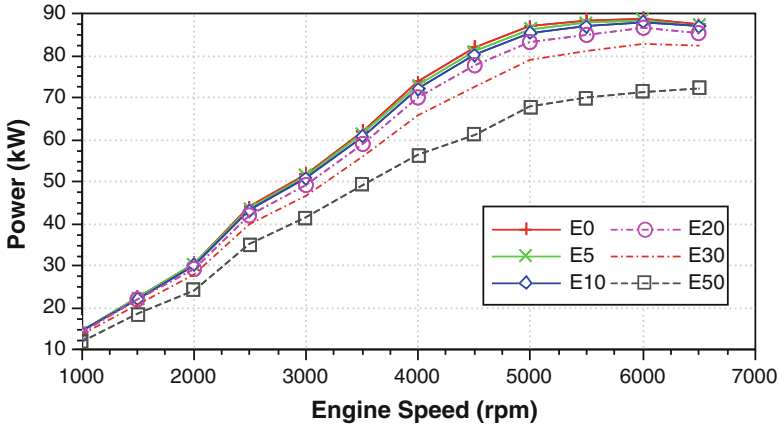


Fig. 6.2 Influence of ethanol gasoline blended fuels on engine brake power

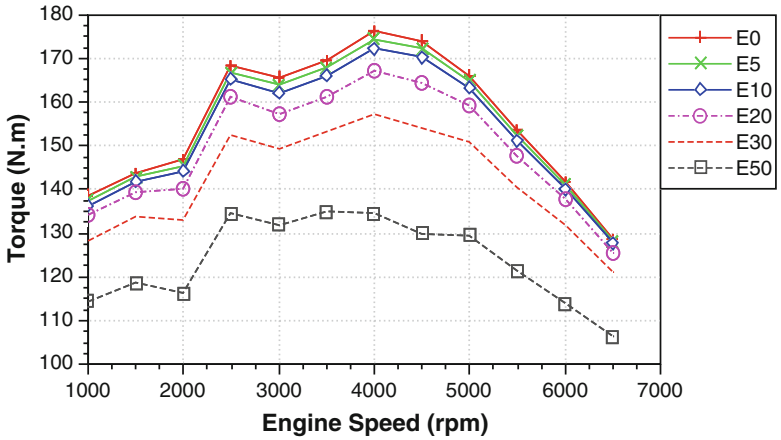


Fig. 6.3 Influence of ethanol gasoline blended fuels on engine toque

decreased for all engine speeds. The brake power of gasoline was higher than those of E5-E50 for all engine speeds. The heat of evaporation of ethanol is higher than that of gasoline (see Table 6.2), providing air-fuel charge cooling and increasing the density of the charge. Adding ethanol to blended fuel causes the equivalence ratio of blended fuel approaches to stoichiometric condition which can lead to a better combustion. On the other hand, the ethanol heating value is lower than that of gasoline and it can be neutralize the previous positive effects. As a result, a lower power output is obtained.

Figure 6.3 shows the influence of ethanol gasoline blended fuels on engine torque. The increase of ethanol content (E5-E50) decreased the torque of the engine. The brake torque of gasoline was higher than those of E5-E50.

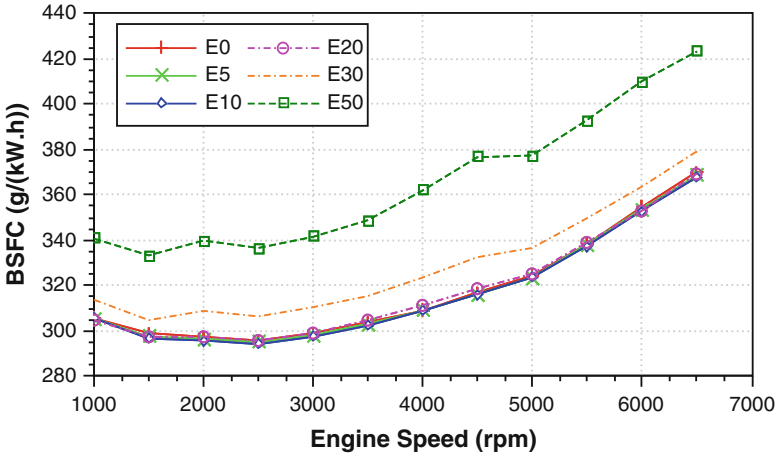


Fig. 6.4 Influence of ethanol gasoline blended fuels on engine brake specific fuel consumption

Because of the existence of oxygen in the ethanol chemical component (see Table 6.2), and the increase of ethanol, lean mixtures are produced that decrease equivalence air-fuel ratio to a lower value and due to the presence of oxygen which has entered the combustion chamber makes the burning more efficient.

Figure 6.4 indicates the variations of the BSFC for ethanol gasoline blended fuels under various engine speeds. As shown in this figure, the BSFC increased as the ethanol percentage increased. The reason It is well known that heating value and stoichiometric air-fuel ratio are the smallest for this fuel which means that for specific air-fuel equivalence ratio, more fuel is needed. The highest specific fuel consumption is obtained at E50 ethanol-gasoline blend.

Also, a slight difference exists between the BSFC when using gasoline and when using ethanol gasoline blended fuels (E5, E10 and E20). The lower energy content of ethanol gasoline blended fuels causes some increment in BSFC of the engine when it is used without any modification.

3.2 Engine Emissions Characteristics

The effect of the ethanol gasoline blends on CO emissions for different engine speeds is shown in Fig. 6.5. It can be seen that when ethanol percentage increases, the CO concentration decreases. This can be explained by the enrichment of oxygen owing to the ethanol, in which an increase in the proportion of oxygen will promote the further oxidation of CO during the engine exhaust process. Another significant reason for this reduction is that ethanol (C_2H_5OH) has less carbon than gasoline (C_8H_{18}).

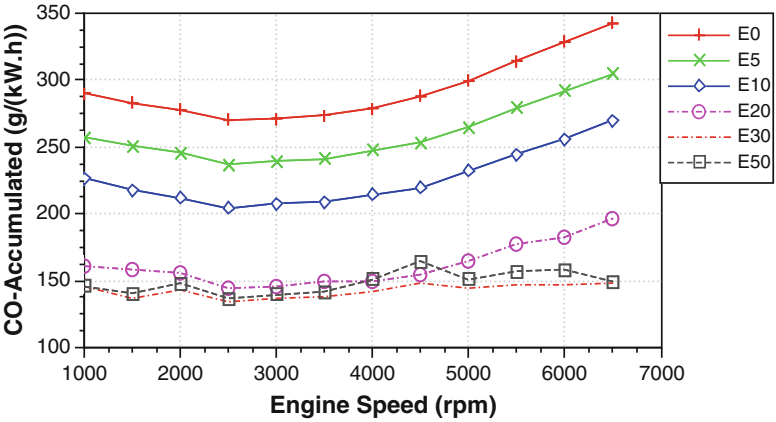


Fig. 6.5 Influence of ethanol gasoline blended fuels on CO emissions

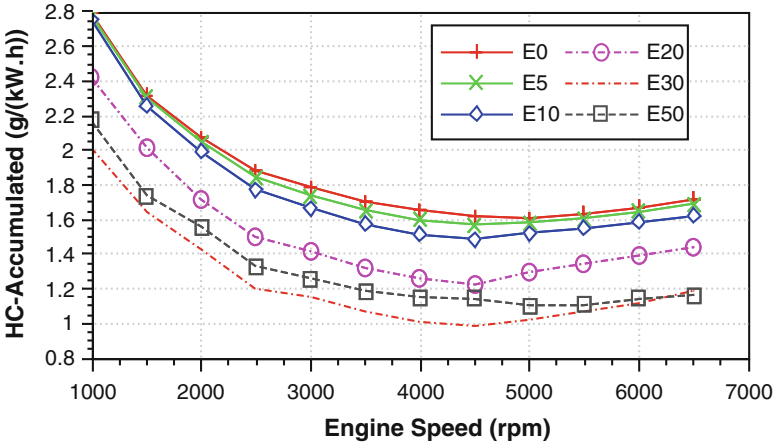


Fig. 6.6 Influence of ethanol gasoline blended fuels on HC emissions

The effect of the ethanol gasoline blends on HC emissions for different engine speeds is shown in Fig. 6.6. It can be seen that when ethanol percentage increases, the HC concentration decreases. The concentration of HC emissions decreases with the increase of the relative air-fuel ratio. The reason for the decrease of HC concentration is similar to that of CO concentration described above.

The effect of the ethanol gasoline blends on NOx emissions for different engine speeds is shown in Fig. 6.7. It can be seen that when ethanol percentage increases, the NOx concentration increase. When the combustion process is closer to stoichiometric, flame temperature increases, therefore, the NOx emissions are increased.

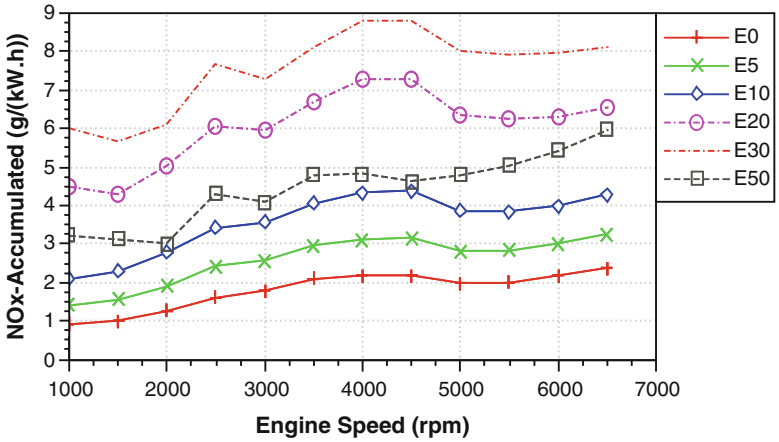


Fig. 6.7 Influence of ethanol gasoline blended fuels on NOx emissions

4 Conclusion

The present chapter demonstrates the influences of ethanol addition to gasoline on SI engine performance and emission characteristics.

Ethanol gasoline blended fuels show lower brake power and brake torque and higher BSFC than those of gasoline.

When ethanol percentage increases, the CO and HC concentration decreases.

Ethanol gasoline blends the significant increase NOx emissions with the increase of ethanol percentage. When the combustion process is closer to stoichiometric, flame temperature increases, therefore, the NOx emissions are increased.

To some extent, additional work will be needed to confirm the theoretical model using ethanol-gasoline blends because the model is calibrated with gasoline. This document is a revised and expanded version of the paper that was presented at WCE 2014 [23].

Acknowledgments The present work has been produced with the financial assistance of the European Social Fund under Operational Programme "Human Resources Development". The contents of this document are the sole responsibility of the "Angel Kanchev" University of Ruse and can under no circumstances be regarded as reflecting the position of the European Union or the Ministry of Education and Science of the Republic of Bulgaria. Project № BG051PO001-3.3.06-0008 "Supporting Academic Development of Scientific Personnel in Engineering and Information Science and Technologies" We are also eternally grateful to AVL-AST, Graz, Austria for granting use of AVL-BOOST under the university partnership program.

References

1. Al-Hasan, M.: Effect of ethanol-unleaded gasoline blends on engine performance and exhaust emission. Energy Convers. Manag. **44**, 1547–1561 (2003)
2. Chen, C.S., Rao, P.S.C., Delfino, J.J.: Oxygenated fuel induced to solvent effect on the dissolution of polynuclear aromatic hydrocarbons from contemned soil. Chemosphere **60**, 1572–82 (2005)
3. Rocha, M.S., Moriera, J.R.S.: A simple impedance method for determining ethanol and regular gasoline mixtures mass content. Fuel **84**, 447–52 (2005)
4. Pikunas, A., Pukalskas, S., Grabys, J.: Influence of composition og gasoline-ethanol blends on parameters of internal combustion engines. J. KONES Intern. Combust. Engines **10**, 3–4 (2003)
5. Palmer, F.: Vehicle performance of gasoline containing oxygenates. In: International conference on petroleum based and automotive applications, Institution of Mechanical Engineers Conference Publications, MEP, pp. 33–46. London (1986)
6. Bata, R.M., Elord, A.C., Rice, R.W.: Emissions from IC engines fueled with alcohol–gasoline blends: a literature review. Trans. ASME **111**, 424–431 (1989)
7. Kim, S., Dale, B.E.: Global potential bioethanol production from wasted crops and crop residues. Biomass Bioenergy **26**(4), 361–375 (2004)
8. Shamekhi, A., Khtibzade, N., Shamekhi, A.H.: Performance and emissions characteristics of a bi-fuel SI engine fueled by CNG and gasoline. ASME paper, ICES2006-1387 (2006)
9. Pourkhesalian, A.M., Shamekhi, A.H., Salimi, F.: Alternative fuel and gasoline in an SI engine: a comparative study of performance and emissions characteristics. Fuel **89**(2010), 1056–1063 (2010)
10. Chum, H., Overend, R.: Biomass and renewable fuels. Fuel Process. Technol. **71**, 187–195 (2010)
11. Egebäck, K., Henke, M., Rehnlund, B., Wallin, M., Westerholm, R.: Blending of ethanol in gasoline for spark ignition engines – problem inventory and evaporative measurements. Rapp. MTC 5407. AVL MTC Motortestcenter AB, Haninge (2005)
12. Iliev, S., Hadjiev, K.: Analysis of engine speed effect on the four stroke GDI engine performance. Proc. Manuf. Syst. **7**, 229–234 (2012)
13. Onorati, A., Montenegro, G., D'Errico, G.: Prediction of the attenuation characteristics of IC engine silencers by 1-D and multi-D simulation models. SAE technical paper series 2006, Tech. Rep. 2006-01-1541 (2006). doi:10.4271/2006-01-1541
14. Montenegro, G., Onorati, A.: Modeling of silencers for IC engine intake and exhaust systems by means of an integrated 1D-multiD approach. SAE Int. J. Engines **1**(1), 466 (2009). doi:10.4271/2008-01-0677
15. Montenegro, G., Onorati, A., Piscaglia, F., D'Errico, G.: Integrated 1D-Multi-D fluid dynamic models for the simulation of ICE intake and exhaust systems. SAE technical paper series 2007, Tech. Rep. 2007-01-0495 (2007). doi:10.4271/2007-01-0495
16. Lindström, F., Ångström, H.E., Kalghatgi, G., Möller, C.E.: An empirical SI combustion model using laminar burning velocity correlations. SAE2005-01-2106 (2005)
17. Bayraktar, H., Durgun, O.: Development of an empirical correlation for combustion durations in spark ignition engines. Energy Convers. Manag. **45**, 1419–1431 (2004)
18. Heywood, J.B.: Internal Combustion Engine Fundamentals. McGraw-Hill, New York (1988)
19. AVL List Gmbh: AVL Boost – Theory, Version 2013.2, Document N:01.0114.2013.2 (2013)
20. Bowman, C.T.: Kinetics of pollutant formation and destruction in combustion. Prog. Energy Combust. Sci. **1**(1), 33–45 (1975)
21. Onorati, A., Ferrari, G., D'Errico, G.: 1D unsteady flows with chemical reactions in the exhaust duct-system of S.I. engines: predictions and experiments. SAE paper no. 2001-01-0939 (2001)

22. D'Errico, G., Ferrari, G., Onorati, A., Cerri, T.: Modeling the pollutant emissions from a S.I. engine. SAE paper no. 2002-01-0006 (2002)
23. Iliev, S.: Developing of a 1-D combustion model and study of engine characteristics using ethanol-gasoline blends. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering WCE 2014, pp. 1268–1273. London, 2–4 July 2014

Chapter 7
An Approach to Modeling of Additive Manufacturing Technologies

Alexander V. Manzhirov and Sergei A. Lychev

Abstract Mathematical modeling of additive manufacturing technologies is aimed at improving the performance of device, machine, and mechanism parts. These technologies include stereolithography, electrolytic deposition, thermal and laser-based 3D printing, 3D-IC fabrication technologies, etc. They are booming nowadays, because they can provide rapid low-cost high-accuracy production of 3D items of arbitrarily complex shape (in theory, from any material). However, deformation and strength problems for products manufactured with these technologies yet remain to be solved. The fundamentally new mathematical models considered in the paper describe the evolution of the end product stress-strain state in additive manufacturing and are of general interest for modern technologies in engineering, medicine, electronics industry, aerospace industry, and other fields.

Keywords Additive manufacturing technologies • Deformation • Growing solids • Incompatibility • Residual stresses • Strength

1 Introduction

Mathematical modelling of a variety of natural phenomena and technological processes requires the consideration of the *material evolution* and *remodelling* of a solid. It may be associated either with the material points accession and annihilation or with the redistribution of internal constraints in the interior of the solid. If such a change takes place simultaneously with deformation of a whole body, it creates a *growing solid* whose properties differs significantly from the conventional ones. Models of the processes of winding and welding, vapor deposition, photopolymerization, ion implantation could act as examples [4, 7, 15]. In such processes the change of the material composition of a body can occur either by the addition

A.V. Manzhirov (✉) • S.A. Lychev
Institute for Problems in Mechanics, Vernadsky Ave 101 Bldg 1, Moscow 119526, Russia
Bauman Moscow State Technical University, 2nd Baumanskaya Str 5/1, Moscow 105005, Russia
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoye shosse 31, Moscow, 115409, Russia
e-mail: manzh@inbox.ru; lychevsa@mail.ru

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_7

of macroscopic volumes which are associated with locally thermostatic states described by statistical parameters such as temperature, distortion, tension, or by implementing of individual atoms or molecules (extra substance by the terminology of [10]) that leads from the macroscopic point of view to the evolution of the distributed defects in the boundary layer. Winding and welding are examples of the first case and the ion implantation is an example of the second case. Sometimes both cases have to be taken into account. Thus, in the process of vapor deposition the adherence of atomic clusters happens which consist of a big number of interrelated atoms as well as the adherence of individual atoms of ions that bombard the growth surface.

It should be emphasized that growth is often closely associated with the processes of defects formation. In particular the vapor phase deposition causes continuous formation of defects in a growing structure. This can be easily shown by the estimation of crystal growth rate. Actually, an atom which condenses from a vapor on a crystal surface of some regularized atomic structure is weakly coupled with this surface and evaporates again with high probability. But if the atomic plane on the growth surface is unfinished the atoms that hit the edge of the plane are strongly coupled. For this reason the unfinished atomic plane being completed quickly and crystal growth stops until a sufficiently large nucleus of a new atomic plane appears. One could estimate the probability of appearance of such a nucleus as well as obtain an estimation of the crystal growth rate. This result would be many orders of magnitude smaller as compared with that observed in the experiment. An explanation of this paradox can be obtained by assuming that a large number of defects are continuously formed on the surface of a crystal and they play the role of nucleuses for islands of atomic planes which grow independently [9]. In the case of such a growth the force interactions between these islands arise and result to the field of residual stresses.

In the pioneering work by E. Kröner [10] it has been shown that residual stresses in *simple materials* [17] may be presented in terms of the incompatibility of the field of local distortion which are defined in the reference description by methods of non- Euclidean geometry. Thus, the geometric language used in the theory of smooth manifolds can be used not only for the description of solids with distributed defects but also for growing solids.

Investigation of the stress-strain state for growing solids has been carried out in numerous works [1, 2, 6, 11, 12, 20]. One may highlight a number of trends in generalization of the classical continuum mechanics that has been utilized in this papers. One of them is developed in the framework of the theory of *inhomogeneity* (structural heterogeneity) arising due to the special connection of parts of a body but not due to the different physical properties of the material of these parts [16, 18]. This structural inhomogeneity arises also in bodies formed from one the same material which are homogeneous in the classical sense. To separate these two inhomogeneities we use the term *material uniformity* [17] for the latter one.

The growth of a solid is usually considered as the process of joining of an additional material to the solid which deforms during this process. It is assumed that the additional material can have the form of material surfaces, threads, and drops and can be deposited to the main body in some stressed state. Moreover the growing body together with additional material may be considered as a single one represented by a multiply solid and the process of growth may be treated as the generating constrains providing that the order of connectivity of a such multiply solid decreases. This leads to the change of topology of a body. In particular boundary points become internal. Here the process of gluing of paper sheets bundle may be mentioned as example. Prior to gluing we have multiply connected set which may be of sufficiently large order of connectivity but after gluing we have simply connected set. If each sheet would be subjected to deformation in the process of gluing from a standard (uniform) state then after gluing no one could find any smooth deformation of a final body that transforms all the sheets to the standard state simultaneously. The reaction of a local part of a body to external loadings in the case defined by elasticity tensor is different for different points of this body in any configuration which defines the immersion in Euclidean space. Thus the body is inhomogeneous even if it consists of one the same material (i.e. material uniform). This example illustrates the existence of a special type of inhomogeneity which is studied in [5, 14, 16].

Thus, an inhomogeneity appears as a result of the growth process and it is connected with the different physical and mechanical properties of material. One may say that it is appears in special scenarios of assembling. To describe the solid reaction to external actions one may either treat it as a non-uniform resulting in a complicated way of describing constitutive equations or somehow reconstruct its global natural configuration and use it as the reference. For simple materials it can be done if one allow embedding of the reference shapes in space with more free definition of the geometric properties (e.g. in the space of affine connectivity) and define the global natural form with additional geometric parameters such as torsion, curvature, and nonmetricity of above mentioned connectivity.

It should be noted that the description of inhomogeneity in solids can be constructed without ideas of non-Euclidean geometry. Of course, if configuration is the embedding of a body into the Euclidean space only then inhomogeneous solid have no global natural configuration, i.e. any configuration is not free from residual stress. At the same time we have to use the stressed configuration as a reference which complicates the formulation of constitutive relations. In particular, they have one more tensor argument – “implant” [6, 14, 20] which characterizes the initial local deformation. However, the geometric meaning of the implant becomes clear if we consider it as an initial (“assembly”) local deformation of the element in natural state which leads directly to the notion of local transformation of the natural frame used in the geometry of the space with absolute parallelism and, thus, introduces the concept of non-Euclidean geometry. Therefore we prefer to use geometric language from the beginning.

2 Body as a Smooth Manifold

In what follows we use the concept of a body as an abstract smooth manifold that is an open subset of some topological space endowed with special *material connection*. This conception allows one to describe the phenomenon of inhomogeneity in materially uniform bodies in rather elegant geometric way. The foundations of the theory of inhomogeneity have been laid down in the milestone work of W. Noll [16] and developed by C.-C. Wang, M. Epstein, G. Maugin [5, 14, 18]. Since inhomogeneity arises as a result of the process of accretion this geometrical approach seems to be effective for the considered problems.

We treat a body $\mathfrak{B}$ as a smooth manifold without boundary. It means that $\mathfrak{B}$ is a set endowed with topology that satisfy the axiom of separability and may be covered by finite number of open overlapped sets $U_k \subset \mathfrak{B}$ that are homeomorphic to open subsets in $\mathbb{R}^n$. This homeomorphism is established by co-ordinate charts $\chi_k : U_k \to \mathbb{R}^n$ which compositions in an intersection $U_k \cap U_p$, i.e. mapping $\chi_k \circ \chi_p^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ are continuous and have sufficient derivatives. Note that n may be equal to 1, 2, 3, depends on the class of the body: fiber, membrane or solid. We refer to n as to *dimension of body*. Material points as elements of the set $\mathfrak{B}$ may be identified by means of coordinates which are the values of charts χ_k. The collection of charts $\{\chi_k\}_{k=1}^n$ define atlas (order n) of the manifold. If a manifold may be covered by an atlas of first order, then this manifold is *trivial*. The necessity for a non-trivial atlases is clear for one- or two- dimensional solids (sphere is simple example). It seems that is not so in three dimensional case. Actually, three dimensional solid that present itself by imbedding into Euclidean space can be modelled by trivial manifold covered by single chart which values are not more than *Cartesian coordinates* of points that constitute the body. But this impression is erroneous! In fact the structure of atlas have to be conformed with *material connection* (see below) and this conformation may require non-trivial atlases, that was shown on examples by C.-C. Wang (one of them is famous "Mëbius crystall") [18]. Of course, all such bodies have non-trivial structure of inhomogeneity. It should be noted that these bodies may be created by appropriate growing process that "sews" three dimensional body using two dimensional surfaces. Thereby the notion of atlas plays significant role in the theory of growing bodies.

Connection in general is a rule that determines the transformation of vector during its transferring along the path (curve) on $\mathscr{B}$ that carries the vector from one fiber to another. Linear (affine) connection determines the linear transformation under infinitesimal transferring, i.e. mapping $\nabla : T_{\mathfrak{X}}(\mathfrak{B}) \to \mathscr{L}(T_{\mathfrak{X}}(\mathfrak{B}), T_{\mathfrak{X}}(\mathfrak{B}))$. In a local chart $\nabla_{\partial_\alpha}\partial_\beta = \Gamma^\gamma_{\alpha\beta}\partial_\gamma$, where $\Gamma^\gamma_{\alpha\beta}$ are Christoffel symbols of the connection. A linear connection is said to be compatible with a metric $\boldsymbol{g}$ of the manifold if the inner product of any two vectors remains the same when those vectors are parallel transported along any curve. It can be shown that the connection ∇ is compatible with $\boldsymbol{g}$ if and only if $\forall \boldsymbol{u} \nabla_{\boldsymbol{u}} \boldsymbol{g} = \boldsymbol{0}$. We consider an n-dimensional manifold $\mathfrak{B}$ with the metric $\boldsymbol{g}$ and a connection ∇. Then $(\mathfrak{B}, \nabla, \boldsymbol{g})$ is called a Riemann–Cartan manifold.

The torsion of a connection is a map $\mathfrak{T} : T_{\mathfrak{X}}(\mathfrak{B}) \times T_{\mathfrak{X}}(\mathfrak{B}) \to T_{\mathfrak{X}}(\mathfrak{B})$ defined by

$$\mathfrak{T}(\boldsymbol{u}, \boldsymbol{v}) = \nabla_{\boldsymbol{u}} \boldsymbol{v} - \nabla_{\boldsymbol{v}} \boldsymbol{u} - [\boldsymbol{u}, \boldsymbol{v}].$$

In components in a local chart $\mathfrak{T}^{\alpha}_{\beta\gamma} = \Gamma^{\alpha}_{\beta\gamma} - \Gamma^{\alpha}_{\gamma\beta}$. The connection is said to be symmetric if it is torsion-free, that is, $\nabla_{\boldsymbol{X}} \boldsymbol{Y} - \nabla_{\boldsymbol{Y}} \boldsymbol{X} = [\boldsymbol{X}, \boldsymbol{Y}]$.[1] The Riemannian curvature is a map $\mathfrak{R} : T_{\mathfrak{X}}(\mathfrak{B}) \times T_{\mathfrak{X}}(\mathfrak{B}) \to \mathscr{L}(T_{\mathfrak{X}}(\mathfrak{B}), T_{\mathfrak{X}}(\mathfrak{B}))$ defined by

$$\mathfrak{R}(\boldsymbol{u}, \boldsymbol{v}) = \nabla_{\boldsymbol{u}} \nabla_{\boldsymbol{v}} - \nabla_{\boldsymbol{v}} \nabla_{\boldsymbol{u}} - \nabla_{[\boldsymbol{u}, \boldsymbol{v}]}.$$

A metric-affine manifold is a manifold equipped with both a connection and a metric: $(\mathfrak{B}, \nabla, \boldsymbol{g})$. If the connection is metric compatible the manifold is called a Riemann–Cartan manifold. If the connection is torsion free but has curvature $\mathfrak{B}$ is called a Riemannian manifold. If the curvature of the connection vanishes but it has torsion $\mathfrak{B}$ is called a Weitzenbock manifold. If both torsion and curvature vanish, $\mathfrak{B}$ is a flat (Euclidean) manifold.

Let us explain in a few words the conception of *material connection*. It implements the idea of local uniform reference configuration that carries an infinitesimal neighborhood of material point to some uniform, typically natural, strain state. In most simple but frequently studied cases it is possible to carry the whole body to the uniform state by some global configuration. In this instance the connection appears to be Euclidean and the theory becomes trivial. In general there is no opportunity to transform by smooth mapping infinitesimal neighborhoods of all material points to the uniform state simultaneously. Due to this a materially uniform body, i.e. a body which material points are all of the same kind, is endowed with some "artificial" (or one may say structural) inhomogeneity. From the mechanical point of view such bodies have no one shape free from residual stresses. The only way to return the neighborhood of all material points to a natural state and hereupon relax the residual stresses is to cut the body into an infinite number of parts and allow them to deform independently (Fig. 7.1). This fictitious process in some sense is reciprocal to the process of growing. One can find a detailed statement of the theory in [11, 13].

3 Growing Body

Now let us give a precise definition of a growing body. We consider growing bodies that may be presented as continuous families of nested bodies. Recall that according to given above definitions manifolds which represent bodies have no boundaries. Certainly, the physical bodies have them. The boundary points are not include into open set $\mathfrak{B}$, but in union they compose the set $\partial\mathfrak{B} = \overline{\mathfrak{B}} \setminus \mathfrak{B}$ that represent the

[1] It can be shown that on any Riemannian manifold $(\mathfrak{B}, \boldsymbol{g})$ there is a unique linear connection ∇ that is compatible with $\boldsymbol{G}$ and is torsion-free. This connection is Levi-Civita connection.

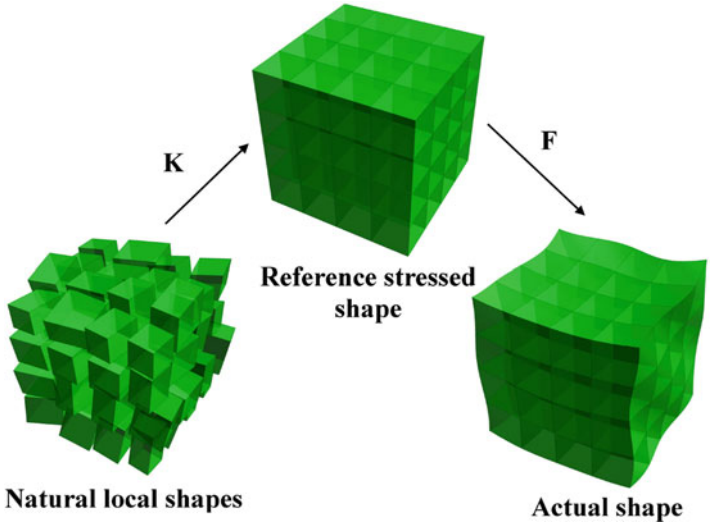


Fig. 7.1 Configurations and deformations

boundary of the body $\mathfrak{B}$. We can assume that $\partial\mathfrak{B}$ is smooth manifold with one less dimension then $\mathfrak{B}$. Finally we believe that inclusion of material can be represented as a continuous adjunction of *material surfaces* (in the sense of [8]) to the boundary $\partial\mathfrak{B}$. These considerations can be summarized as follows:

A *layerwise* growing body is a continuous monotone (with respect to inclusion) one–parametric family of manifolds

$$\mathfrak{C} = \{\mathfrak{B}_\alpha\}_{\alpha\in I}\,, \quad \forall\alpha < \beta \in I \quad \mathfrak{B}_\alpha \subset \mathfrak{B}_\beta, \tag{7.1}$$

where $I = (a, b) \subset \mathbb{R}$ is an open interval, that holds the following property

$$\forall\alpha \in I \quad \forall\mathfrak{X} \in \mathfrak{B}_\alpha \setminus \overline{\mathfrak{B}}_a \quad \exists\gamma \in I \quad \mathfrak{X} \in \partial\mathfrak{B}_\gamma. \tag{7.2}$$

We refer ro $\mathfrak{B}_a$ as to initial body and to $\mathfrak{B}_b$ as to final body. Since the family (7.1) represents some evolution process we shall refer to α as to *evolution parameter*. The property (7.2) states that any internal point belongs to adjoint part of a body in some stage of growing was the element of boundary manifold, so the topological dimension of its neighborhood was changed. This fact is specific for growing bodies.

We point out the particular class of growth, namely *complete surface growth* that satisfy the following condition

$$\forall\alpha < \beta \in I \quad \partial\mathfrak{B}_\beta \notin \mathfrak{B}_\alpha,$$

or, equivalently, $\forall\alpha < \beta \in I\ \partial\mathfrak{B}_\beta \cap \partial\mathfrak{B}_\alpha = \emptyset$. This property ensures that for every $\mathfrak{B}_\alpha \in \mathfrak{C}$ a continuous projection map from $\mathfrak{B}_\alpha \setminus \overline{\mathfrak{B}}_a$ onto interval $I' \subset I$ exists,

so one can interpret the manifold $\mathfrak{B}_\alpha \setminus \overline{\mathfrak{B}}_a$ as trivial bundle over interval I' whose fibers are manifolds $\partial\mathfrak{B}_\gamma$, $\gamma \in I'$.

The idea of continuity may be formalized by means of some measure (such as volume or mass) on manifolds $\mathfrak{B}_\alpha$. If a certain measure mes(...) has been introduced then the *continuity* of family (7.1) is equivalent to the following property:

$$\forall\varepsilon \quad \exists\delta \quad \forall\alpha < \beta \in I \quad \beta - \alpha < \delta \Rightarrow \mathrm{mes}(\mathfrak{B}_\beta \setminus \mathfrak{B}_\alpha) < \varepsilon.$$

It is also possible to interprets the layerwise character of growth in terms of measure as follows

$$\lim_{\beta\to\alpha} \frac{\mathrm{mes}(\mathfrak{B}_\beta \setminus \mathfrak{B}_\alpha)}{\beta - \alpha} = k,$$

$$0 < k < \infty, \quad \exists\Omega \subset \mathfrak{B}_\beta \setminus \mathfrak{B}_\alpha \quad \dim\Omega = \dim\mathfrak{B}_\alpha - 1.$$

The latter condition states that the infinitesimal increment, which is the set difference between two nearby instances of family $\mathfrak{C}$ that represent the growing body, are asymptotically equivalent to the body of one less dimension. For example, if $\mathfrak{B}_\alpha$ are three dimensional manifolds, then Ω is two dimensional and it mechanical response may be described by the relations suitable for membranes, shells, etc.

Certainly, the given definition does not cover all possible ways of surface growth that can be realized as continuous process of joining surfaces, fibers or drops. The appropriate classification may be obtained on the geometrical basis. In this framework we treat a growing body as a bundle of smooth manifold; the topological structure of bundle, particulary the dimensions of base and typical fiber, depends on the accretion process. See [11, 13] for details.

It is conventional in rational mechanics [17, p. 35] that bodies present themselves in physical space $\mathscr{E}$ as shapes $\mathscr{B}_\alpha \subset \mathscr{E}$. On the one hand shapes $\mathscr{B}_\alpha$ are connected subsets of the physical space, on the other hand every shape is an image of a configuration $\varkappa_\alpha : \mathfrak{B}_\alpha \to \mathscr{B}_\alpha$ and belongs to the class of admissible configurations that endow the body $\mathfrak{B}_\alpha$ with the structure of smooth manifold. We associate two shapes with each element of the family (7.1): the reference shape $\mathscr{B}_\alpha^R = \varkappa_\alpha^R\mathfrak{B}_\alpha$ and the actual shape $\mathscr{B}_\alpha = \varkappa_\alpha\mathfrak{B}_\alpha$. Thereby the family (7.1) induces the corresponding families of reference and actual configurations as well as families of reference and actual shapes. We use exactly such definitions of configurations and shapes hereinafter. Note that the reference shape in general is not stress free.

Note that conventional notation for the position of material point $\mathfrak{X}_\alpha$ in reference configuration is $T\mathscr{E} \ni \boldsymbol{X}_\alpha = \varkappa_\alpha^R(\mathfrak{X})$ and $\boldsymbol{x}_\alpha$ for the position in actual configuration, i.e. $T\mathscr{E} \ni \boldsymbol{x}_\alpha = \varkappa_\alpha(\mathfrak{X})$. The composition of configurations $\varphi_\alpha = \varkappa_\alpha \circ (\varkappa_\alpha^R)^{-1}$, i.e. $\varphi_\alpha : \boldsymbol{X}_\alpha \mapsto \boldsymbol{x}_\alpha$ is *deformation*. The derivative of mapping φ, that exists due to the differentiability of configurations and their inverse, is so-called "deformation gradient" (actually it is not gradient at all): $\boldsymbol{F} = \partial x^m/\partial X^n \boldsymbol{e}_m \otimes \boldsymbol{e}^n \equiv \partial x^m/\partial X^n \partial_m \otimes dx^n$. We assume that $J = \det \boldsymbol{F} > 0$ always.

In general, the configuration of a growing body is a smooth map of the material manifold endowed with a material connection onto the physical manifold which connectivity is essentially different.

We suppose that bodies $\mathfrak{B}_\alpha$ are materially uniform, simple and elastic [17], so their response may be defined by the response functional

$$\boldsymbol{T}_\alpha = \mathfrak{S}(\boldsymbol{H}_\alpha). \tag{7.3}$$

Here $\mathfrak{S}(\ldots)$ is a response functional which is nonlinear in general case. We assume that $\mathfrak{S}(\ldots)$ does not depend on evolution parameter α explicitly. The $\boldsymbol{T}_\alpha$ is some kind of stress tensor field (to be definite we use the term *the Cauchy stresses*). Assume that $\mathfrak{S}$ satisfy the principle of frame indifference and has been calibrated

$$\mathfrak{S}(\boldsymbol{0}) = \boldsymbol{0}, \quad \lim_{\det H \to 0} |\mathfrak{S}(\boldsymbol{H})| = \infty.$$

The $\boldsymbol{H}_\alpha$ is a smooth tensor field that represents the local distortion. It may be written in the form of multiplicative decomposition

$$\boldsymbol{H}_\alpha = \boldsymbol{F}_\alpha \circ \boldsymbol{K}_\alpha, \tag{7.4}$$

where $\boldsymbol{F}_\alpha$ is the conventional deformation gradient, i.e. linearization of mapping $\gamma_\alpha : \mathscr{B}_\alpha^R \to \mathscr{B}_\alpha$ that can be represented by the relative gradient $\nabla_{\varkappa_\alpha^R}$ as follows [17]:

$$\boldsymbol{F}_\alpha = \nabla_{\varkappa_\alpha^R} \gamma_\alpha. \tag{7.5}$$

It is important to note that the tensor field $\boldsymbol{F}_\alpha$ is compatible in the following sense: a vector field which gradient gives $\boldsymbol{F}_\alpha$ exists. Note that this property is not valid in general for the second factor in the right hand side of relation (7.4), namely the smooth tensor field $\boldsymbol{K}_\alpha$. This field was named in [14] as "implant field". Indeed, $\boldsymbol{K}_\alpha$ is a field of linear transformations upon undistorted noncompatible infinitesimal parts that joins them without gaps in the global reference configuration.

The implant field $\boldsymbol{K}_\alpha$ by virtue of its incompatibility induces inhomogeneity that can be presented by a non-Euclidean material connection that is a certain type of affine connection admissible on the manifold $\mathfrak{B}_\alpha$. In abstract terms this connection may be defined as a field of operators Γ_α that maps some tangent vector $\boldsymbol{h}$ onto a linear operator [16]:

$$\Gamma_\alpha \boldsymbol{h} = \boldsymbol{K}_\alpha^{-1} \nabla_{\boldsymbol{K}_\alpha} (\boldsymbol{K}_\alpha \boldsymbol{h}) \boldsymbol{K}_\alpha,$$

It is clear that material connection in fact is the geometric representation of an implant field $\boldsymbol{K}_\alpha$. Actually the non-Euclidean features of such connection are represented by the tensor field of torsion $\mathfrak{T}_\alpha$:

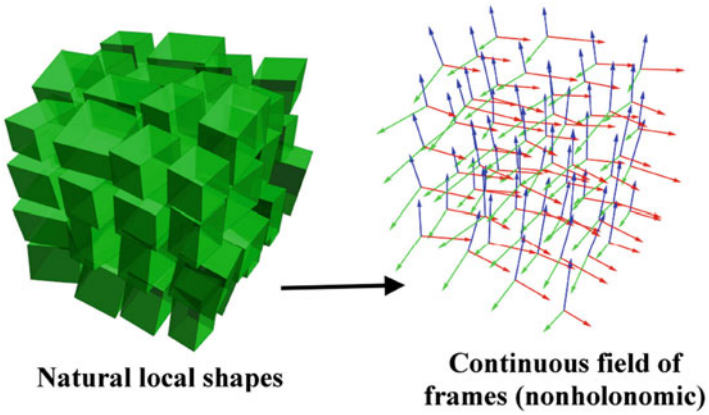


Fig. 7.2 Correspondence between local shapes and space with absolute parallelism

$$\mathfrak{T}_\alpha(\boldsymbol{h}, \boldsymbol{p}) = (\Gamma_\alpha \boldsymbol{p})\boldsymbol{h} - (\Gamma_\alpha \boldsymbol{h})\boldsymbol{p} - [\boldsymbol{h}, \boldsymbol{p}].$$

Here $[\cdot, \cdot]$ denotes the Lie brackets that represents commutator of tangent vector fields [3, 16]. One may express these relations in terms of natural frame ∂_ν induced by a certain coordinate map on manifolds $\mathscr{B}_\alpha$, say $\varkappa_\alpha^R$, in the form [14]

$$\boldsymbol{e}_\beta = (\boldsymbol{K}_\alpha)^\nu_{\cdot\beta}\partial_\nu, \quad (\Gamma_\alpha)^\beta_{\gamma\nu} = \left(\boldsymbol{K}_\alpha^{-1}\right)^\rho_{\cdot\gamma,\nu}(\boldsymbol{K}_\alpha)^\beta_{\cdot\rho},$$
$$(\mathfrak{T}_\alpha)^\beta_{\gamma\nu} = (\Gamma_\alpha)^\beta_{\gamma\nu} - (\Gamma_\alpha)^\beta_{\nu\gamma},$$

where $\boldsymbol{e}_\beta$ combine to form a nonholonomic system of frames which is corresponding to an "implantation" by $\boldsymbol{K}_\alpha$.

The connection turns the manifold $\mathfrak{B}_\alpha$ into a Cartan space (space with absolute parallelism, teleparallel space [3]) (Fig. 7.2). Indeed, by means of operators Γ_α we define on $\mathfrak{B}_\alpha$ the rule of parallel transport. The implant field $\boldsymbol{K}_\alpha$ defines some arbitrary affine transformations of natural frame ∂_ν in all points of $\mathfrak{B}$, so the field of frames $\boldsymbol{e}_\beta$ becomes nonholonomic. The rule of parallel transport may be stated as follows. A vector is transported parallely if its projections on frames $\boldsymbol{e}_\beta$ remain invariant. Due to the fact that the field $\boldsymbol{K}_\alpha$ defines inhomogeneity this situation has vivid physical interpretation, i.e. the observer adopted to the moving frame sees no inhomogeneities in the same way as a *geodesic observer* does not 'feel' any gravitation field in general relativity – see e.g. [14]. Actually, one can represent local shapes as continuous family of stressed reference shapes (Fig. 7.3).

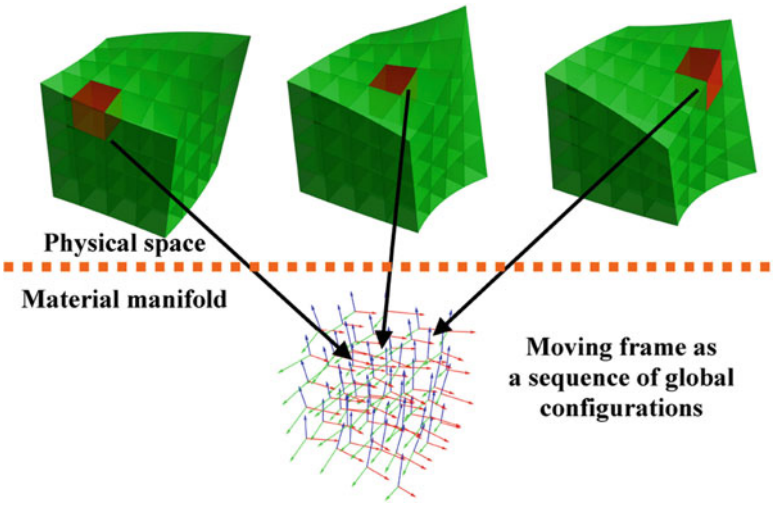


Fig. 7.3 Representation for local shapes as family of reference shapes

It has been pointed out above that the response functional $\mathfrak{S}$ is explicitly independent on α. Formally it means that for every $\alpha < \beta \in I$ in a fixed internal material point $\mathfrak{x} \in \mathfrak{B}_\alpha$ the response of a body $\mathfrak{B}_\alpha$ is equal to the response of its successor $\mathfrak{B}_\beta$. From the physical point of view it signifies that properties of already accreted material do not change during continuation of the growth process. Let the parameters of "implantation" corresponding to a material point be determined completely at the moment of its accretion and unchanged later. In other words the inhomogeneity appears due to growth process on the boundary and does not develop further in bulk. In terms of [5] it means that remodelling does not occur in bulk. Hereby we exclude processes like plasticity or shrinkage from consideration. We shall refer to growing bodies under such condition as to nonrearranged:

A layerwise growing body is *nonrearranged* if

$$\forall \alpha < \beta \in I \quad \forall \boldsymbol{x} \in \mathfrak{B}_\alpha \quad \mathfrak{T}_\beta(\boldsymbol{x}) = \mathfrak{T}_\alpha(\boldsymbol{x}). \tag{7.6}$$

Note that the property of nonrearrangement in bulk is typical for models of surface growth unlike so-called volume growth [5].

It is not difficult to take into account natural geometric definition of Cartan connection and its torsion $\mathfrak{T}_\alpha$, but this may require some additional calculations. So in many cases it is preferable to deal directly with the implant field $\boldsymbol{K}_\alpha$ that induces certain type of connection. In order to do this we can choose a *hereditary* family of reference shapes, i.e. family that satisfies the following relation:

$$\forall \alpha < \beta \in I \quad \mathscr{B}_\alpha = \varkappa^R_{\beta|\alpha} \mathfrak{B}_\alpha,$$

where $\varkappa^R_{\beta|\alpha}$ is the restriction of mapping $\varkappa^R_\beta$ onto the domain $\mathfrak{B}_\alpha \subset \mathfrak{B}_\beta$. This means that the family of reference shapes can be treated as a sequence of continuations of the set $\mathscr{B}^R_a$ up to the set $\mathscr{B}^R_b$. From the mechanical point of view it means that the shape $\mathscr{B}^R_a$ continuously increases through the addition of flux of material surfaces to its boundary while the state of already adhered material particles remains unchanged. Obviously these shapes are not stress–free and can be in equilibrium only under action of special external fields of bulk forces and surface forces on the boundary (Eshelby forces). In the issue the condition (7.6) may be represented in terms of implant $\boldsymbol{K}_\alpha$:

$$\forall \alpha < \beta \in I \quad \forall \boldsymbol{x} \in \mathfrak{B}_\alpha \quad \boldsymbol{H}_\alpha(\boldsymbol{x}) = \boldsymbol{H}_\beta(\boldsymbol{x}).$$

Furthermore, if the family $\mathfrak{C}$ (7.1) permits differentiation with respect to the parameter α, the condition (7.6) may be written also in terms of field equation:

$$\dot{\boldsymbol{K}} = \boldsymbol{0}, \tag{7.7}$$

where the symbol $\boldsymbol{K}$ denotes mapping $\boldsymbol{K} : I \ni \alpha \mapsto \boldsymbol{K}_\alpha$ and dot denotes differentiation over the parameter α.

4 Boundary-Value Problem in Hyperelastic Case

Introduce elastic potential $W^{\varkappa_R}_{\varkappa}$ that is an elastic energy per unit volume in the reference state $\varkappa_R$ and can be interpreted as a function of arguments $\boldsymbol{F}$, $\boldsymbol{K}$, $\mathfrak{X}$ [14]

$$W^{\varkappa_R}_{\varkappa}(\boldsymbol{K}, \boldsymbol{F}, \mathfrak{X}) = J^{-1}_{\boldsymbol{K}} W^{\varkappa_C}_{\varkappa}(\boldsymbol{H}, \mathfrak{X}) = J^{-1}_{\boldsymbol{K}} W^{\varkappa_C}_{\varkappa}(\boldsymbol{K}\cdot\boldsymbol{F}, \mathfrak{X}).$$

One can express the Piola stress tensor $\boldsymbol{T}^{\varkappa_R}_{\varkappa}$ relevant to $\varkappa_R$ by formula

$$\boldsymbol{T}^{\varkappa_R}_{\varkappa} = \frac{\partial W^{\varkappa_R}_{\varkappa}}{\partial \boldsymbol{F}^*} = J^{-1}_{\boldsymbol{K}} \boldsymbol{K}^* \cdot \frac{\partial W^{\varkappa_C}_{\varkappa}}{\partial \boldsymbol{H}^*}.$$

The stress tensor $\boldsymbol{T}^{\varkappa_C}_{\varkappa}$ relevant to $\varkappa_C$ can be defined fiber-by-fiber as follows

$$\boldsymbol{T}^{\varkappa_C}_{\varkappa} = \frac{\partial W^{\varkappa_C}_{\varkappa}}{\partial \boldsymbol{H}^*}.$$

The boundary value problem for an accreted solid is determined by the equations of equilibrium in $V(t)$ with boundary $\Omega(t)$ which parametrically depends on time, i.e.

$$\nabla_{\varkappa_R} \cdot \left[J^{-1}_{\boldsymbol{K}} \boldsymbol{K}^* \cdot \frac{\partial W^{\varkappa_C}_{\varkappa}(\boldsymbol{H}, \mathfrak{X})}{\partial \boldsymbol{H}^*}\bigg|_{\boldsymbol{H}=\boldsymbol{K}\cdot\boldsymbol{F}} \right] + \boldsymbol{b} = \boldsymbol{0}, \tag{7.8}$$

and boundary conditions formulated on $\Omega(t)$:

$$\boldsymbol{n}_{\varkappa_R}\cdot\left[J_{\boldsymbol{K}}^{-1}\boldsymbol{K}^*\cdot\frac{\partial W_{\varkappa}^{\varkappa_C}(\boldsymbol{H},\mathfrak{X})}{\partial \boldsymbol{H}^*}\Big|_{\boldsymbol{H}=\boldsymbol{K}\cdot\boldsymbol{F}}\right]\Bigg|_{\Omega(t)}=\boldsymbol{p}.$$

At the first glance a formal statement of the boundary value problem differs from the classical one only by the fact that the boundary of domain depends parametrically on time. However, there is more profound difference: the elastic potential depends on the tensor field of distortion the determination of which requires additional conditions. Particular form of these conditions depends on the geometrical structure of joined elements, that is in essence on the structure of the bundle of material manifold. If the growth of a body occurs due to continuous influx of prestressed material surfaces to this body then this condition can be written in the form

$$\boldsymbol{P}_{\varkappa_R}\cdot\left[J_{\boldsymbol{K}}^{-1}\boldsymbol{K}^*\cdot\frac{\partial W_{\varkappa}^{\varkappa_C}(\boldsymbol{H},\mathfrak{X})}{\partial \boldsymbol{H}^*}\Big|_{\boldsymbol{H}=\boldsymbol{K}\cdot\boldsymbol{F}}\right]\cdot\boldsymbol{P}_{\varkappa_R}\Bigg|_{\Omega(t)}=\mathscr{T}.$$

Here $\boldsymbol{P}=(\boldsymbol{E}-\boldsymbol{n}\otimes\boldsymbol{n})$ is a projector onto the tangent plane to $\Omega(t)$. This equation expresses the fact that the fibers align with the specified tension determined by the surface tensor $\mathscr{T}$, i.e., two-dimensional tensor of second rank defined in the tangent space of the adhering material surface.

The equation for distortion tensor $\boldsymbol{K}$ provided that the increase is the result of continuous adherence of prestressed surfaces can be obtained from the relations of the theory of material surfaces (theories of solids with a material boundary [8]). Effects of material surface adhering leads to an infinitesimal change of the stress–strain state of an accreted solid but as an elementary act of adhesion occurs during an infinitely small time interval the rate of stress state is finite. This rate can be found from the equations of contact interaction between the spatial body and adhering material surface. The equation of physical boundary equilibrium (which is a bounding surface of the body in its actual state from the geometrical point of view and a thin film undergoing a membrane state of stress from mechanical point of view) can be written as

$$\nabla_s\cdot\mathscr{T}+\boldsymbol{b}_s=\boldsymbol{n}_{\varkappa_R}\cdot\boldsymbol{T}_{\varkappa}^{\varkappa_R}\Big|_{\Omega(t)},$$

where ∇_s is the surface nabla operator, $\boldsymbol{b}_s$ is the surface density of external forces acting on $\Omega(t)$. To complete the boundary value problem formulation the condition on the curvilinear boundary $\partial\Omega(t)$ of the surface $\Omega(t)$ must be specified. This conditions may be of the form

$$\tilde{\boldsymbol{n}}\cdot\boldsymbol{T}\Big|_{\partial\Omega(t)}=\tilde{\boldsymbol{f}}.$$

Here $\tilde{\boldsymbol{n}}$ is an external unit normal to the curve $\partial\Omega(t)$ in the tangent plane and $\tilde{\boldsymbol{f}}$ is a linear density of forces distributed on $\partial\Omega(t)$.

5 Example: Deformation of the Growing Elastic Sphere

As an example we shall consider the problem of centrosymmetric deformation of a growing elastic sphere. Material manifold can be represented in an open ball, which allows a natural bundle on concentric spheres. Each sphere corresponds to a fiber, and the base of the bundle is an open interval.

Assume that the material is incompressible and, consequently, only isochoric deformation of the spatial configuration of the body are permitted. This assumption allows us to take advantage of the well-known Rivlin–Ericksen universal solution to describe the deformation of each fiber [19]. Let $\boldsymbol{e}_r$, $\boldsymbol{e}_\varphi$, $\boldsymbol{e}_\theta$ be the physical basis corresponded to spherical coordinates. Assume that each fiber has a natural configuration and it is a sphere. The joining of such fibers is not of a configuration of a body, because it does not present a connected set. After proper deformation the union of these spheres consist a solid sphere (a ball). Thus, the configuration can be achieved, does not free from stress. Clearly, such a deformation may be chosen in many ways, and the choice of the reference configuration contains arbitrariness. However, all such stressed configuration can be transformed into one another diffeomorphically. The corresponding distortion field has the form

$$\boldsymbol{K} = \xi_{-\alpha}^2 \boldsymbol{e}_r \otimes \boldsymbol{e}_r + \xi_{-\alpha}^{-1} \left(\boldsymbol{e}_\varphi \otimes \boldsymbol{e}_\varphi + \boldsymbol{e}_\theta \otimes \boldsymbol{e}_\theta\right); \quad \xi_{-\alpha} = (r^3 - \alpha)^{1/3}/r.$$

Here $r = r(\mathfrak{X}^3)$ is a spatial (radial) coordinate, $\alpha = \alpha(\mathfrak{X}^3)$ is the distortion parameter, $\mathfrak{X}^3$ is a material coordinate of a fiber. One can interpreted the field $\boldsymbol{K}$ as a deformation gradient defined by mapping $r = (r_0^3 + \alpha)^{1/3}$ and calculated for the infinitesimal spherical fiber, which has a natural configuration when $r = r_0$. Then α can be considered as a parameter to isochoric inflating of the fiber.

We now choose one of these configurations as the reference and denote it by the symbol $\varkappa_R$. The deformation of the reference configuration to the actual one $\varkappa$ is determined by isochoric diffeomorphism $\varkappa_R \to \varkappa$. Under the conditions of incompressibility and central symmetry this mapping belongs to the family $R = (r^3 + A)^{1/3}$, where A is the parameter of the family. The gradient of this transformation has the form

$$\boldsymbol{F} = \xi_A^{-2} \boldsymbol{e}_r \otimes \boldsymbol{e}_r + \xi_A \left(\boldsymbol{e}_\varphi \otimes \boldsymbol{e}_\varphi + \boldsymbol{e}_\theta \otimes \boldsymbol{e}_\theta\right); \quad \xi_A = (r^3 + A)^{1/3}/r.$$

Complete local deformation of each fiber from its natural state can be found as:

$$\boldsymbol{H} = \boldsymbol{F}\cdot\boldsymbol{K} = \eta^2 \boldsymbol{e}_r \otimes \boldsymbol{e}_r + \eta^{-1} \left(\boldsymbol{e}_\varphi \otimes \boldsymbol{e}_\varphi + \boldsymbol{e}_\theta \otimes \boldsymbol{e}_\theta\right), \quad \eta = (r^3 - \alpha)^{1/3}/(r^3 + A)^{1/3}.$$

The corresponding Cauchy–Green strain tensor may be written as

$$\boldsymbol{G} = \boldsymbol{H}^* \cdot \boldsymbol{H} = \eta^4 \boldsymbol{e}_r \otimes \boldsymbol{e}_r + \eta^{-2} \left(\boldsymbol{e}_\varphi \otimes \boldsymbol{e}_\varphi + \boldsymbol{e}_\theta \otimes \boldsymbol{e}_\theta \right).$$

Invariants of $\boldsymbol{G}$ can be found as follows:

$$I_1(\boldsymbol{G}) = \eta^4 + 2\eta^{-2}, \quad I_2(\boldsymbol{G}) = \eta^{-4} + 2\eta^2.$$

The equations of equilibrium in the absence of body forces are of the form $\nabla \cdot \boldsymbol{T} = \boldsymbol{0}$, where $\boldsymbol{T}$ is Cauchy stress tensor:

$$\boldsymbol{T} = T_{RR} \boldsymbol{e}_R \otimes \boldsymbol{e}_R + T_{\theta\theta} \left(\boldsymbol{e}_\theta \otimes \boldsymbol{e}_\theta + \boldsymbol{e}_\varphi \otimes \boldsymbol{e}_\varphi \right).$$

Here ∇ is the spatial Hamiltonian operator defined in the current configuration $\varkappa$. The boundary conditions are formulated in the form

$$\boldsymbol{n} \cdot \boldsymbol{T} \Big|_{\Gamma_1} = b\boldsymbol{n}, \quad \boldsymbol{n} \cdot \boldsymbol{T} \Big|_{\Gamma_2} = \boldsymbol{0}. \tag{7.9}$$

Here b is the intensity of the normal load on the inner surface of the ball Γ_1, and Γ_2 is the surface of growth. One more condition expresses the fibers that the layers are attached with a given preload f, i.e.

$$\boldsymbol{e}_\theta \cdot \boldsymbol{T} \cdot \boldsymbol{e}_\theta \Big|_{\Gamma_2} = T_{\theta\theta} \Big|_{\Gamma_2} = f.$$

In spherical coordinates subject to the central symmetry the equilibrium equations and boundary conditions are the follows

$$\frac{\partial T_{RR}}{\partial R} + \frac{2}{R} \left(T_{RR} - T_{\theta\theta} \right) = 0, \quad T_{RR} \Big|_{r=r_0} = b, \quad T_{RR} \Big|_{r=r_1} = 0.$$

Here R_0 is an inner and R_1 is an outer radius of the ball in the current configuration $\varkappa$. For incompressible isotropic hyperelastic material Cauchy stress tensor can be represented by the formula

$$\boldsymbol{T} = -p\boldsymbol{E} + 2\frac{\partial W}{\partial I_1}\boldsymbol{G} - 2\frac{\partial W}{\partial I_2}\boldsymbol{G}^{-1},$$

where $p = p(R)$ is the pressure determined from the equilibrium conditions, W is the specific density of the stored elastic energy. If $\varkappa_R$ is a reference configuration, then elastic potential should be written in the form

$$W = J_{\boldsymbol{K}}^{-1} W_{\varkappa_C} \left(\boldsymbol{K} \cdot \boldsymbol{F}, \mathfrak{X} \right),$$

where $W_{\varkappa C}$ is the potential which calibration is carried out locally for each fiber with respect to the natural state of it. Due to the incompressibility $J_{\boldsymbol{K}} = 1$ then the term $J_{\boldsymbol{K}}^{-1}$ has no effect on further derivations.

Substitution of the tensor $\boldsymbol{T}$ into the equilibrium equation leads to the equation in the pressure p

$$-\nabla p + 2\left(\boldsymbol{G}\cdot\nabla\frac{\partial W}{\partial I_1} - \boldsymbol{G}^{-1}\cdot\nabla\frac{\partial W}{\partial I_2} + \frac{\partial W}{\partial I_1}\nabla\cdot\boldsymbol{G} - \frac{\partial W}{\partial I_2}\nabla\cdot\boldsymbol{G}^{-1}\right) = \boldsymbol{0}.$$

Using the calculated expressions and taking into account the fact, that, by virtue of central symmetry $\nabla p = \boldsymbol{e}_R dp/dR$, we arrive at the equation with respect to $p(R)$, integral of which has the form

$$p(R') = 2\left[\eta^4\frac{\partial W}{\partial I_1} - \eta^{-4}\frac{\partial W}{\partial I_2}\right] + 4\int\limits_{R_0}^{R'}\left\{\frac{\partial W}{\partial I_1}\left(\eta^4 - \eta^{-2}\right) - \frac{\partial W}{\partial I_2}\left(\eta^{-4} - \eta^2\right)\right\}\frac{dR}{R} + p_0.$$

The upper limit of integration $R' \in (R_0, R_1)$ is the radial coordinate, which determines the position of a spherical fiber in the current configuration $\varkappa$, p_0 is the constant of integration, which corresponds to a constant hydrostatic pressure. By the change of variables $R \mapsto r$ the resulting expression can be transformed to the form:

$$p(r') = 2\left[\eta^4\frac{\partial W}{\partial I_1} - \eta^{-4}\frac{\partial W}{\partial I_2}\right] - 4q(r_0, r') + p_0,$$

$$q(r_0, r') = \int\limits_{r_0}^{r'} r^2(r^3 + A)^{-1}(\eta^{-4} - \eta^2)\left(\eta^2\frac{\partial W}{\partial I_1} + \frac{\partial W}{\partial I_2}\right)dr.$$

Radial stresses may be calculated according to the formula $T_{RR}(r') = q(r_0, r') - p_0$. It follows from the second boundary condition (7.9) that $p_0 = q(r_0, r_1)$. Consequently, $T_{RR}(r') = q(r_1, r')$. The circumferential stress $T_{\theta\theta}$ can be found from the equations of equilibrium and expressed in terms of T_{RR} as follows:

$$T_{\theta\theta} = T_{RR} + \frac{R}{2}\frac{\partial T_{RR}}{\partial R}.$$

In the boundary fiber $T_{\theta\theta}$ must be equal to a predetermined preload value f, i.e.

$$T_{\theta\theta}\Big|_{\Gamma_2} = T_{\theta\theta}\Big|_{r=r_1} = T_{RR}\Big|_{r=r_1} + \frac{(r^3+A)^{1/3}}{2}\frac{\partial T_{RR}}{\partial R}\Big|_{r=r_1} =$$
$$= 2\left[(r^3 + A)^{-1}(\eta^{-4} - \eta^2)\left(\eta^2\frac{\partial W}{\partial I_1} + \frac{\partial W}{\partial I_2}\right)\right]_{r=r_1} = f. \qquad (7.10)$$

From this equation it can be determined the relation $\alpha = \alpha(r_1, A, f)$. From the first boundary condition (7.9), which may be represented as $T_{RR}\Big|_{r=r_0} = b$, parameter A may be determined. We emphasize that $A = A(r_1)$, while $\alpha = \alpha(r)$.

Distortion tensor field induces a connection on a material manifold, which thus becomes a flat space with affine connection (i.e., the space of zero curvature) with nontrivial torsion. Nonzero components of the torsion tensor are given by

$$\mathfrak{T}^{\varphi}_{\cdot r\varphi} = \mathfrak{T}^{\theta}_{\cdot r\theta} = -\mathfrak{T}^{\varphi}_{\cdot \varphi r} = -\mathfrak{T}^{\theta}_{\cdot \theta r} = \frac{1}{3r^3 - 3\alpha}\frac{\partial \alpha}{\partial r}.$$

Acknowledgements This work was supported by the RSF under Grant 14-19-01280.

References

1. Aruytunyan, N.K., Drozdov, A.D., Naumov, V.E.: Mechanics of Growing Viscoelastoplastic Solids. Nauka, Moscow [in Russian] (1987)
2. Aruytunyan, N.Kh., Manzhirov, A.V., Naumov, V.E.: Contact Problems in Mechanics of Growing Solids. Nauka, Moscow [in Russian] (1991)
3. Choquet-Bruhat, Y., Dewitt-Morette, C., Dillard-Bleick, M.: Analysis, Manifolds and Physics. Part 1. Basics. North-Holland, Amsterdam (1982)
4. Choy, K.L.: Chemical vapour deposition of coatings. Prog. Mater. Sci. **48**, 57–170 (2003)
5. Epstein, M.: The Geometrical Language of Continuum Mechanics. Cambridge University Press, New York (2010)
6. Epstein, M., Maugin, G.A.: Thermomechanics of volumetric growth in uniform bodies. Int. J. Plast. **16**(7), 951–978 (2000)
7. Gibson, I., Rosen, D.W., Stucker, B.: Additive Manufacturing Technologies. Rapid Prototyping to Direct Digital Manufacturing. Springer, New York (2009)
8. Gurtin, M.E., Murdoch, A.I.: A continuum theory of elastic material surfaces. Arch. Ration. Mech. Anal. **27**, 291–323 (1975)
9. Harrison, W.A.: Solid State Theory. Dover, New York (1980)
10. Kröner, E.: Allgemeine kontinuumstheorie der versetzungen und eigenspannungen. Arch. Ration. Mech. Anal. **4**(1), 273–334 (1967)
11. Manzhirov, A.V., Lychev, S.A.: Residual stresses in growing bodies. In: Topical Problems in Solid and Fluid Mechanics, pp. 66–79. Elite Publishing House, Delhi (2011)
12. Manzhirov, A.V., Lychev, S.A.: Mathematical modeling of additive manufacturing technologies. In: Proceedings of the World Congress on Engineering 2014, London, vol. II, pp. 1404–1409 (2011)
13. Manzhirov, A.V., Lychev, S.A.: The mathematical theory of growing solids: Finite deformations. Dokl. Phys. **57**(4), 160–163 (2012)
14. Maugin, G.A.: Material Inhomogeneities in Elasticity. Chapman and Hall, London/New York (1993)
15. Nastasi, M., Mayer, J.W.: Ion Implantation and Synthesis of Materials. Springer, Berlin/Heidelberg/New York (2006)
16. Noll, W.: Materially uniform simple bodies with inhomogeneities. Arch. Ration. Mech. Anal. **27**(1), 1–32 (1967)
17. Truesdell, C.: A First Course in Rational Continuum Mechanics: General concepts. Academic, Boston (1991)

18. Wang, C.C.: On the geometric structure of simple bodies, or mathematical foundations for the theory of continuous distributions of dislocations. Arch. Ration. Mech. Anal. **27**(1), 33–94 (1967)
19. Wang, C. C., Truesdell, C.: Introduction to Rational Elasticity. Noordhoff International Publishing, Leyden (1973)
20. Yavari, A.: A geometric theory of growth mechanics. J. Nonlinear Sci. **20**(6), 781–830 (2010)

Chapter 8
Analysis of the Cutting Forces in Manufacturing Stainless Steel Femoral Heads Using Finite Element Method

Nikolaos I. Galanis and Dimitrios E. Manolakos

Abstract In this chapter it is examined the use of finite elements in predicting the cutting forces of machined parts of stainless steel AISI 316 L through turning. The process used the high speed machining technique, which is continuously improved and it has be found application in more and more manufacturing processes like aerospace industry, in die and mold companies and in the last years also in bioengineering in manufacturing hip joint implants. The cutting forces, which were measured through experimental process, were compared with predicted ones from the finite element modulation, and it was exported that they can be predicted with good precision when machining with the FEM model.

Keywords AISI 316L • Cutting forces • FEM analysis • Femoral heads • High speed machining • Surface roughness

1 Introduction

Turning is the material processing operation, where a cutting tool is used to remove an unwanted material to produce a desired product, and is generally performed on a lathe. In recent decades, considerable improvements were achieved in turning, enhancing machining of difficult-to-cut materials and producing better surface finish. Various techniques, such as high-speed machining (HSM), have been in use for considerable time [1]. This is not really a new technology. First investigations have been performed by Salomon in the twenties. However, these investigations were only ballistic analysis. The improvement of machine tools and controls made HSM possible in machining operations in the seventies [2]. However, this technique

N.I. Galanis (✉) • D.E. Manolakos
School of Mechanical Engineering, National Technical University of Athens, 9th Iroon Polytechniou Av. Zografou, Athens 15780, Greece
e-mail: ngalanis@central.ntua.gr; manolako@central.ntua.gr

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_8

and all metal machining processes are characterised mainly by quick changes in quantified elements. Individual changes do not occur in isolation and they influence each other. The analysis of these changes requires study of the complicated complete systems in their real situations. The study of cutting processes such as turning and facing, from dynamic aspects, is very important. The trend towards the measurement of cutting forces in machining leads to many theoretical and practical problems. Theoretical problems associate mainly with the choice of a suitable technique to measure, and the statistical methods to analyse the components of cutting force to be determined in realtime. Practical problems involve the errors and uncertainties relating to the measurement system used [3].

The cutting operation is controlled by the parameters v_c (cutting speed), d (cutting depth) and f (feed rate). The results obtained include specific quality of a machining surface but also cutting forces and tool wear. Only the knowledge of such results could help us to choose an appropriate set of work piece-cutting force-machine tool for a projected industrial production [4]. In this paper, it will be presented the effect of the aforementioned parameters on the cutting forces, during the HSM manufacturing of metallic femoral head, from stainless steel AISI 316 L. The forces were measured by a series of experimental measurements.

Furthermore there were analysed by the Finite Element Method using Advant-Edge™ and compared with the experimental results.

The results were analysed through the Analysis Of Variance (ANOVA) in order to eliminate the fault factor and they were also evaluated according to the produced surface roughness.

2 Cutting Forces

Material removal described so far is known as orthogonal, producing only two cutting forces; when turning these is axial and tangential. Tangential cutting force is by far the greater and axial cutting force is the force required to keep the cutting edge in contact with the. Oblique cutting introduces a third cutting force, radial. It is known that the cutting edge is not perpendicular to the axis of rotation as in orthogonal turning. Tangential cutting force resists the rotation of the work, as relatively high speeds are used the bulk of power consumption lies here. Axial cutting force resists the travel of the tool, however this is a relatively low speed compared with rotation of the work, so for all practical purposes power consumption may be ignored [5].

Furthermore investigations have shown that increasing cutting speed leads to reduced cutting forces and better surfaces. The effect of the cutting speed increase on the cutting forces during the turning process is the reduction of forces. Tests proved that for every investigated steel the cutting force decreases down at approximately 450 N when the cutting speed increases up to 500 m/min with feed rate 0.1 mm/rev and cutting depth at 1 mm. For the steel with the larger grain sizes, higher forces are to be applied in each case of material separation. At cutting speeds above

approximately 800 m/min no further reduction of cutting forces was detected. Therefore, it can be assumed that above this cutting speed HSM conditions are present concerning the cutting forces for these steels [6, 7].

3 Finite Element Analysis

The method of finite elements (FEA) was developed from one special concept: A work that acts on a Form is divided through analysis in a big amount of small parts (elements), which shows the developing of an action on the part [8]. The method is a numerical analysis technique for obtaining approximate solutions to a wide variety of engineering problems. Although originally developed to study the stresses in complex airframe structures, it has since been extended and applied to the broad field of continuum mechanics [9]. Through the years it has become a powerful tool for the numerical solution of a wide range of engineering problems. Applications range from deformation and stress analysis of automotive, aircraft, building. In this method of analysis, a complex region defining a continuum is discredited into simple geometric shapes called finite elements. The material properties and the governing relationships are considered over these elements and expressed in terms of unknown values at element corners [10].

Looking at the literature for metal cutting with FEM, it is observed that a large part of them describe the simulation results of the chip formation process during orthogonal machining [11–15]. Furthermore there are also papers, that describe the stresses during the machining [16–18], tool wear [19] and of course cutting forces [20–23]. In all these projects, there were utilised a number of softwares like Marc, Abaqus, Deform 2D/3D, Nike, AdvantEdge, etc.

Predicted results may vary with software and with the input data so the choice of the software is of extreme importance. Finite element software, specific for machining operations, was chosen to simulate the metal cutting process (in this case, a turning operation). Therefore, AdvantEdge™, supplied by Third Wave Systems, was used in this study. This commercial software package was built from the start with metal cutting operations in mind, allowing simulating turning, drilling, milling, micro machining, etc. in either two or three dimensions. It uses adaptive meshing to help improving the quality and the accuracy of the predicted results and it also supports several workpiece material libraries. Unfortunately, the solver cannot be controlled by the user but fast setups for several simulations can be done easily because of the easy to use software interface [24].

4 Material and Method

For this investigation, there were a number of cuttings took place, which variables was based on the design of experiments methodology [25, 26]. Eight experiments represent 2^3 factorial designs with added ten points in the middle edges and faces

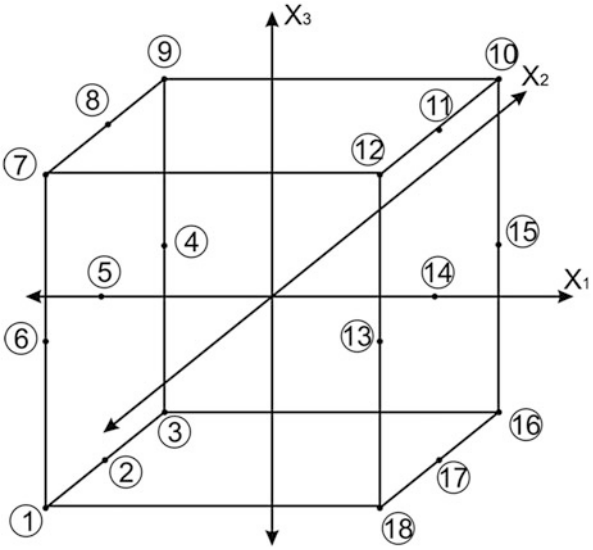


Fig. 8.1 Representation of a 2^3 factorial design with added parameters

Table 8.1 Level independent variables

Factor, coding (unit)	Low (−1)	Centre (0)	High (1)
Cutting speed, v_c (m/min)	264	352	440
Feed, f (mm/rev)	0.06	0.08	0.12
Depth of cut, d (mm)	0.1	0.15	0.2

of the representation cube, Fig. 8.1, was taken. Taking into account three different levels for each variable, as shown in Table 8.1, there were taken the experimental conditions for 18 experiments, shown in Table 8.3.

4.1 Workpiece and Cutting Tool Material

The materials, which was manufactured in a CNC lathe OKUMA Lb 10II, was from AISI 316 L steel, which hardness was 79HRB, as shown its properties in Table 8.2. Medically, the uses of stainless steels like 316 L, although their high Fe contents render them non-compatible with magnetic resonance imaging (MRI) and to be poor fluoroscopic materials. In spite of their limitations and a myriad of materials have been chosen (like titanium), stainless steels are still favoured, as evidenced by the fact that seven out of the eight implants approved by the US Food and Drug Administration are made of stainless steels [27].

Regarding the use of coolants, 316 L better results in tool wear and piece roughness are achieved when external coolant (emulsion 4–5 % of a mineral oil

Table 8.2 Material properties

Material properties	AISI 316 L stainless steel
Physical	
Density	8 g/cc
Mechanical	
Hardness, Rockwell B	79
Tensile strength, ultimate	560 MPa
Tensile strength, yield	290 MPa
Elongation of break	50 %
Modulus of elasticity	193 GPa
Poisson 's ratio	0.29

in water) is used [28]. Furthermore, the development of new materials for cutting tools, such as TiN-coated cemented tungsten carbides, has also led to better control of the material removal process [29–31]. Therefore, a coated tool from SECO specification: DNMG 110404 – M3 with TP 2000 coated grade was used for the manufacturing process. This has rhombic shape with cutting edge angle 55° and are intended for general turning on steels and alloyed steels, as they are coated with four layers of Ti (C, N) + Al_2O_3 + Ti (C, N) + TiN. The rake angle mounted in the toolholder is $\gamma - 5°$ and inclination angle is λ_s −9.5°. The tool cutting edge angle is κ 93°.

Small cylindrical parts with diameter 30 mm and length 28 mm with a conical hole were used to manufacture the spheres, femoral heads, Fig. 8.2.

4.2 Measurements

During the procedure there were measured the acting forces to the tool. In order to accomplish these measurements a Kistler dynamometer 9257A was used, as shown in Fig. 8.2b. This is a three-component piezoelectric dynamometer platform. The force data were recorded by a specifically designed, very compact multi-channel microprocessor controlled data acquisition system with a single A/D converter preceded by a multiplexer. The results are recorded in Table 8.3.

4.3 Finite Elements Analysis

After the experiments, a number of analyses were accomplished according to the experimental conditions. Through AdvantEdge™ software there were simulated all experiments in order to predict the cutting forces and to compare the experimental results with the results from the analysis. The software inputs parameters are shown in Table 8.4.

Table 8.3 Experimental conditions and results

No of cut	Depth mm	Speed (m/min)	Feed (mm/rev)	Coding			Cutting force (N)				Predicted cutting force (N)			
				X1	X2	X3	9°	30°	60°	90°	9°	30°	60°	90°
1	0.2	264	0.08	1	−1	0	500	159	85	75	485	146	95	75
2	0.2	264	0.12	1	−1	1	800	280	150	129	750	230	165	118
3	0.2	264	0.06	1	−1	−1	330	105	61	52	312	116	68	58
4	0.2	352	0.06	1	0	−1	235	82	46	40	220	88	52	44
5	0.2	352	0.08	1	0	0	380	120	69	58	350	128	75	64
6	0.2	352	0.12	1	0	1	660	210	115	99	620	180	125	90
7	0.2	440	0.12	1	1	1	510	190	93	81	485	160	98	78
8	0.2	440	0.8	1	1	0	310	100	53	47	341	110	60	50
9	0.2	440	0.06	1	1	−1	200	59	35	32	219	70	50	35
10	0.1	440	0.06	−1	1	−1	110	35	18	16	122	35	27	21
11	0.1	440	0.08	−1	1	0	160	48	30	24	178	50	32	28
12	0.1	440	0.12	−1	1	1	275	85	49	42	295	78	52	40
13	0.1	352	0.12	−1	0	1	330	109	61	51	352	90	64	46
14	0.1	352	0.12	−1	0	0	190	60	34	30	210	64	37	32
15	0.1	352	0.06	−1	0	−1	125	39	24	20	129	44	26	22
16	0.1	264	0.06	−1	−1	−1	160	50	32	26	156	55	35	29
17	0.1	264	0.08	−1	−1	0	245	78	41	39	230	75	39	35
18	0.1	264	0.12	−1	−1	1	430	135	79	68	400	118	86	61

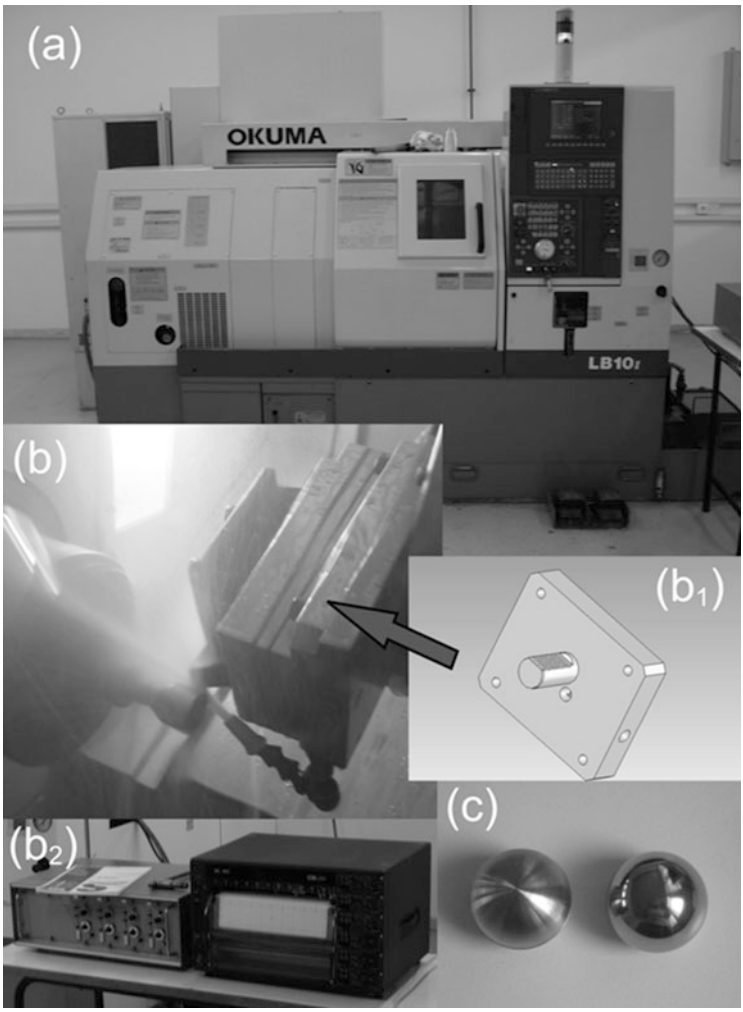


Fig. 8.2 Experimental procedure (**a**) the CNC Lathe which was used, (**b**) inside the cabin where there are the device to support the dynamometer (b_1), which sends its signal to amplifiers (b_2) and (**c**) the final femoral head

5 Results and Discussion

In Table 8.3 it is presented all the results that exported from this project, the experimental and the predicted results from the AdvantEdge™ software. However, during the process of manufacturing femoral heads, because of the spherical shape of the manufactured parts, the cutting depth is changing, as the tool follows its cyclic route. For this reason during the prediction process with the FEM, there must be taken, several cutting depths into account. For this analysis, there was examines the

Table 8.4 Software input parameters

Workpiece	
Workpiece length	5.0 mm
Workpiece height	3.0 mm
Workpiece material	AISI 316 L
Tool	
Rake angle	6.0°
Rake face length	2.0 mm
Relief angle	10.0°
Relief face length	2.0 mm
Cutting edge radius	0.04 mm
Material	Carbide-general
Coating (3 layers)	TiN: 0.01 mm
	Al_2O_3: 0.02 mm
	TiC: 0.01 mm
Process	
Depth of cut	0.1 and 0.2 mm
Length of cut	3.0 mm
Feed	0.06, 0.08 and 0.12 mm/rev
Cutting speed	264, 352 and 440 m/min
Friction coefficient	Default
Coolant	Used
Simulation	
Maximum number of nodes	24,000
Maximum element size	0.1 mm
Minimum element size	0.02 mm

forces when the tool was at 9°, 30°, 60° and 90° according to the axis of revolution of the parts. If there is a comparison between the results in order to find the error, it will be two pairs, the error between numerical and experimental values and the error between predicted and experimental values. The errors can be found using the following formulas:

$$Error = \frac{Experimentalvalue - Pr\ edictedvalue}{Experimentalvalue} \times 100\ \ \% \tag{8.1}$$

From the comparison of the values it is exported that the error is not greater than 11.5 % for the majority of the experiments. There are two values which have error 16.5 % and 31.5 %. For this two pair it can be concluded that there were a mistake during the procedure or a measurement fault, something which affect the result. For all the others the different is normal, because with the Finite Element analysis which is predicted the values there will be an error less than 15 %. A clear view of these results it can be exported from the graphs below, Figs. 8.3 and 8.4, where there the results for the two cutting depth, 01 and 0.2 mm.

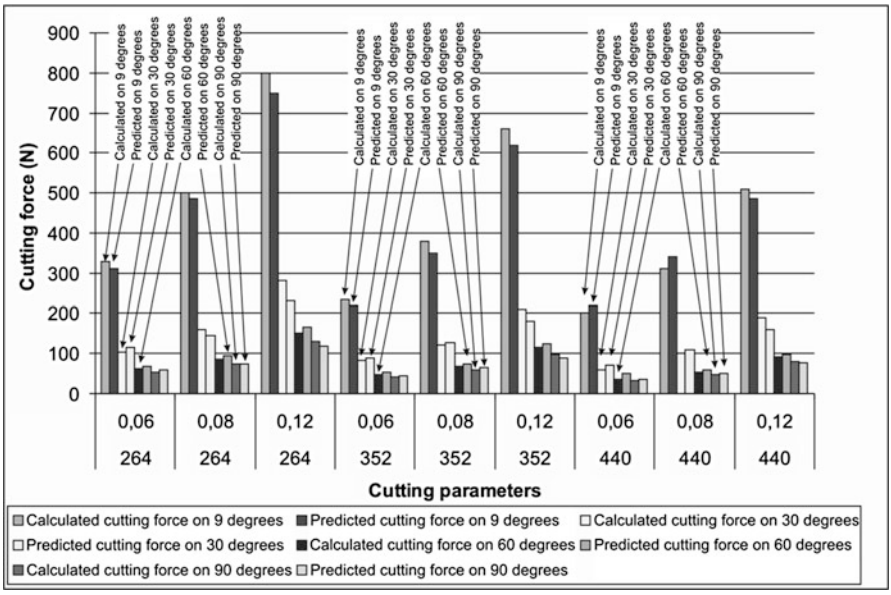


Fig. 8.3 Comparison of cutting forces between experimental and predicted values for cutting depth of 0.2 mm

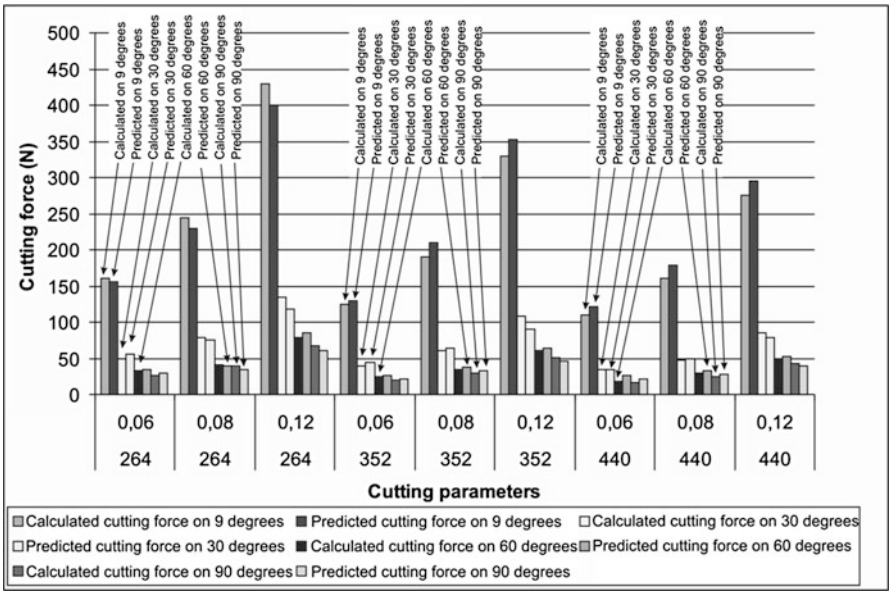


Fig. 8.4 Comparison of cutting forces between experimental and predicted values for cutting depth of 0.1 mm

Table 8.5 ANOVA of the experimental results

Source	DF	SS	MS	F	P
Cutting depth	1	4900.5	4900.5	53.10	0.000
Cutting speed	2	1834.8	917.4	9.94	0.003
Feed rate	2	7057.4	3258.7	38.23	0.000
Error	12	1107.6	92.3		
Total	17	14900.3			
$S = 9.60710$ $R^2 = 92.57$ % $R^2(adj) = 89.47$ %					

Table 8.6 ANOVA of the predicted results

Source	DF	SS	MS	F	P
Cutting depth	1	4933.6	4933.6	67.61	0.000
Cutting speed	2	1309.8	654.9	8.15	0.006
Feed rate	2	4333.4	2166.7	26.69	0.000
Error	12	875.7	73.0		
Total	17	11452.4			
$S = 8.5423$ $R^2 = 92.35$ % $R^2(adj) = 89.17$ %					

In order to compare better the results of the two methods, numerical and experimental an analysis of variance (ANOVA) was held between the forces and the cutting parameters separately for each method. The variation between the groups of cutting parameters represents systematic variation due to the effect on the forces. The between-groups variation is often called the effect variance and the within-groups variance is often called the error variance. In statistical terms, the analysis will tell whether the groups differ significantly or not. If a result is statistically significant, it tells that the group means are too different to have been that way by chance alone. Two levels of significance, $p < 0.05$ and $p < 0.01$, are typically employed in statistics. These mean that the probability of getting that result alone is less than 5 % and 1 % perceptively. These give pretty good confidence that the result obtained is a true reflection of an actual difference [32–34]. Through this analysis, it was examined an important aspect of statistical modelling, which distinguishes it from mere function approximation, the interpretability of results [35]. In Table 8.5 it is shown the analysis of the experimental Forces and in Table 8.6 the analysis of predicted Forces.

It is obvious that their confidence interval for three is very good, over 90 % and too close to numerical, with 91.45 %. Also the calculated F – values, which shows the ratio of the mean square of parameters to mean square of pure error are close. For numerical results and predicted ones, the F-value for cutting speed is the same for both of them. Hence all three analyses are found to be adequate.

In Fig. 8.5 there is the comparison of the distribution of temperature and chip between four analyses, with the first 440 m/min cutting speed, 0.06 mm/rev feed rate and 0.2 cutting depth at position of 60°, the second with 264 m/min cutting speed, 0.06 mm/rev feed rate and 0.2 cutting depth at position of 9°, the third 352 m/min cutting speed, 0.08 mm/rev feed rate and 0.2 cutting depth at position of 30° and the forth with 440 m/min cutting speed, 0.12 mm/rev feed rate and 0.1 cutting depth at position of 9°.

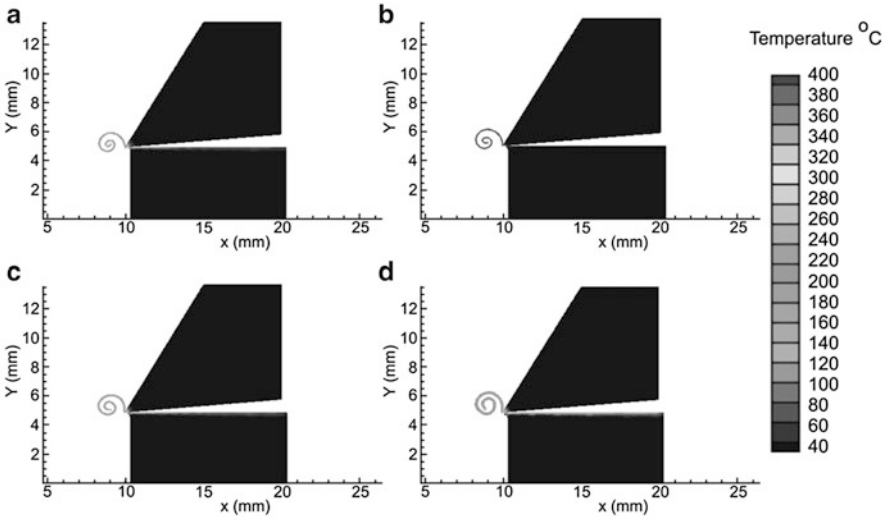


Fig. 8.5 Distribution of temperature and chip between (**a**) 440 m/min cutting speed, 0.06 mm/rev feed rate and 0.2 cutting depth at position of 60°, (**b**) with 264 m/min cutting speed, 0.06 mm/rev feed rate and 0.2 cutting depth at position of 9°, (**c**) 352 m/min cutting speed, 0.08 mm/rev feed rate and 0.2 cutting depth at position of 30° and (**d**) with 440 m/min cutting speed, 0.12 mm/rev feed rate and 0.1 cutting depth at position of 9°

From this comparison it can be seen the form of the chip that is produced, when the cutting speed and feed rate changes.

As it can be seen also, the values of the forces are decreasing as the cutting speed increases and the cutting depth and feed rate is decreasing. The highest decrease was for cutting speed of 352 m/min, where for the same depth and feed rate, there were at about 25 % decrease. For cutting speed 440 m/min the amount of change was smaller. The effect of cutting speed can be attributed to the fact that as speed decreases, the shear angle decreases and the friction coefficient increases. Both effects increase the cutting force [36]. From the other hand the increase of cutting depth and feed rate, increases the amount of the removed material, so is increasing the resistance to the tool, which means the increase of the cutting speed. To this increase, it can be outputted that the cutting depth affects the forces directly proportional as it quite double, when the depth is twice increased. However the proportion of the feed rate is not direct, according to the results, which lead to the choice of the k_2 factor in the numerical model.

So in these experiments, the forces for the manufacturing with cutting speed 440 m/min, cutting depth 0.1 mm and feed rate 0.06 mm/rev are the smallest from all. So in these conditions it can be made easier and with less tool wear the manufacturing of the spheres of femoral heads. As the feed rate becomes bigger up to 0.12 mm/rev the forces increases and with the cutting depth the relevance between them is proportional. The highest cutting force were logically at 264 m/min cutting speed with cutting depth 0.2 mm and feed rate 0.12 mm/rev, and its value 129 N.

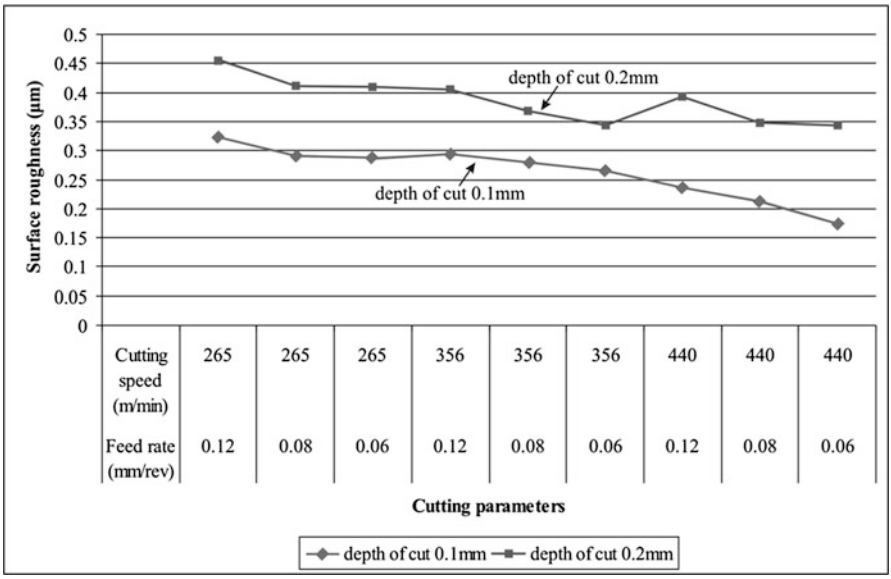


Fig. 8.6 Surface roughness for measuring spheres

5.1 Surface Roughness

To fully characterize the process quality, all balls were measured after their fabrication in an Atomic Force Microscope (AFM). This technique was used for two reasons. Firstly, we have to measure very low and accurate values of roughness, and secondly, it is a quite good method for measuring spherical surfaces, as it can measure a very small surface, i.e., 50 × 50 μm, considering it as a flat portion of the sphere. It was found that the surface quality becomes better as the cutting speed increases.

However, their values are so small and only the picks take values close to 1.3 μm, not significantly affecting in the results for the mean roughness Rα. From the graph of Fig. 8.6, surface roughness seems to become lower, as the cutting speed increases. However, with reducing both the feed rate and the cutting depth, the roughness improves and takes the values closer to the international standards limits for femoral heads [37, 38].

6 Conclusion

To summarize, this chapter refers to the forces that act on the tool during the manufacturing of femoral heads with high speed turning.

1. The predicted values from the finite element software in this investigation have a deflection not more than 12 %.
2. The higher the cutting speed and lower cutting depth and feed rate the lower are the cutting forces that acts on the system tool and manufactured part.
3. The highest decrease was for cutting speed of 352 m/min, where for the same depth and feed rate, there were at about 25 %
4. Expansion of tool life, because small cutting forces cause less tool wear, and of course to better surface quality which plays a very important role to the manufacturing of femoral heads according to the strict regulations of ISO 7206.
5. An increase in tool life was throughout apparent due to small cutting forces and better surface quality.

References

1. Ahmed N, Mitrofanov AV, Babitsky VI, Silberschmidt VV (2007) Analysis of forces in ultrasonically assisted turning. J Sound Vib 308:845–854
2. Toenshoff, H.K., Friemuth, T., Andrae, P., Lapp, C.: High speed cutting-fundamentals and machine tool development, 10th International Conference on Precision Engineering, Yokohama (2001)
3. Audy J (2006) An appraisal of techniques and equipment for cutting force measurement. J Zhejiang Univ Sci A 11:1781–1789
4. Remadna M, Rigal JF (2006) Evolution during time of tool wear and cutting forces in the case of hard turning with CBN inserts. J Mater Process Technol 178:67–75
5. Tlusty G (2000) Manufacturing Process and Equipment. Prentice Hall, Upper Saddle River
6. Trent E, Wright P (2000) Metal Cutting. Butterworth-Heinemann, Oxford
7. Brinksmeier, E., Mayr, P., Lübben, T., Pouteau, P., Diersen, P.: Influence of material properties on surface integrity and chip formation in high speed turning. In: 3rd International Conference on Metal Cutting and High Speed Machining, Metz (2001)
8. Buck KE, Scharpf DW, Stein E, Wunderlich W (1973) Finite Elemente in der Statik. Verlag von Wilhelm Ernst & Sohn, Berlin/München/Düsseldorf
9. Huebner KH (1975) The Finite Element Method for Engineers. Wiley, New York/London/Sydney/Toronto
10. Chandrupatla TR, Belegundu AD (1991) Introduction to Finite Elements in Engineering. Prentice – Hall International Editions, Englewood Cliffs
11. Shih AJ (1996) Finite element analysis of the rake angle effects in orthogonal metal cutting. Int J Mech Sci 38(1):1–17
12. Carroll JT, Strenkowsk JS (1988) Finite element models of orthogonal cutting with application to single point diamond turning. Int J Mech Sci 30(12):899–920
13. Xie JQ, Bayoumi AE, Zbib HM (1998) FEA modeling and simulation of shear localized chip formation in metal cutting. Int J Mach Tools Manuf 38:1067–1087
14. Galanis, N.I., Manolakos, D.E.: Finite Element Analysis of the Cutting Forces in Turning of Femoral Heads from AISI 316L Stainless Steel, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering, WCE 2014, pp. 1232–1237, London, (2014), 2–4 July 2014
15. Fang G, Zeng P (2005) Three-dimensional thermo–elastic–plastic coupled FEM simulations for metal oblique cutting processes. J Mater Process Technol 168:42–48

16. Uhlmann E, Graf M, Zettier R (2007) Finite element modeling and cutting simulation of Inconel 718. Ann CIRP 56(1):61–64
17. Zhang L (1999) On the separation criteria in the simulation of orthogonal metal cutting using the finite element method. J Mater Process Technol 89–90:273–278
18. Baeker M, Roesler J, Siemers C (2002) A finite element model of high speed metal cutting with adiabatic shearing. Comput Struct 80:495–513
19. Kose E, Kurt A, Seker U (2008) The effects of the feed rate on the cutting tool stresses in machining of Inconel 718. J Mater Process Technol 196:165–173
20. Yen Y-C, Söhner J, Lilly B, Altan T (2004) Estimation of tool wear in orthogonal cutting using the finite element analysis. J Mater Process Technol 146:82–91
21. Hashimura M, Ueda K, Dornfeld D, Manabe K (1995) Analysis of three-dimensional burr formation in oblique cutting. Ann ClRP 44(1):27–30
22. Mouazen AM, Nemenyi M (1999) Finite element analysis of subsoiler cutting in non-homogeneous sandy loam soil. Soil Tillage Res 51:1–15
23. Bil H, Kilic SE, Tekkaya AE (2004) A comparison of orthogonal cutting data from experiments with three different finite element models. Int J Mach Tools Manuf 44:933–944
24. Baeker M (2006) Finite element simulation of high-speed cutting forces. J Mater Process Technol 176:117–126
25. Davim JP, Maranhγo C (2009) A study of plastic strain and plastic strain rate in machining of steel AISI 1045 using FEM analysis. Mater Design 30:160–165
26. Dabnun MA, Hashmi MSJ, El-Baradie MA (2005) Surface roughness prediction model by design of experiments for turning machinable glass–ceramic (Macor). J Mater Process Technol 164–165(2005):1289–1293
27. Ross PJ (1996) Taguchi Techniques for Quality Engineering. Mc Graw-Hill, New York
28. Lo KH, Shek CH, Lai JKL (2009) Recent developments in stainless steels. Mater Sci Eng R 65:39–104
29. Gandarias A, López de Lacalle LN, Aizpitarte X, Lamikiz A (2008) Study of the performance of the turning and drilling of austenitic stainless steels using two coolant techniques. Int J Mach Mach Mater 3(1/2):1–17
30. M' Saoubi R, Outeiro JC, Changeux B, Lebrun JL, Morao Dias A (1999) Residual stress analysis in orthogonal machining of standard and resulfurized AISI 316 L steels. J Mater Process Technol 96:225–233
31. Outeiro JC, Umbrello D, M' Saoubi R (2006) Experimental and numerical modelling of the residual stresses induced in orthogonal cutting of AISI 316 L steel. Int J Mach Tools Manuf 46:1786–1794
32. Umbrello D, MSaoubi R, Outeiro JC (2007) The influence of Johnson–Cook material constants on finite element simulation of machining of AISI 316 L steel. Int J Mach Tools Manuf 47:462–470
33. Turner JR, Thayer JF (2001) Introduction to Analysis of Variance. Sage, Thousand Oaks
34. Hinkelmann K, Kempthorne O (1994) Design and Analysis of Experiments. Wiley, Canada
35. Schimek MG (2000) Smoothing and Regression – Approaches, Computation and Application. Willey, Canada
36. Kalpakjian S (1997) Manufacturing Processes for Engineering Materials. Addison Wesley Longman Inc., Canada
37. Galanis NI, Manolakos DE (2009) Surface roughness of manufactured femoral heads with high speed turning. Int J Mach Mach Mater 5(4):371–382
38. Galanis NI, Manolakos DE (2010) Surface roughness prediction in turning of femoral head. Int J Adv Manuf Technol 51(1):79–86

Chapter 9
Flowslide Investigations Test Rig Design

Giandomenico Di Massa, Luca Pagano, Stefano Pagano, Michele Russo, Riccardo Russo, and Maria Claudia Zingariello

Abstract The paper describes a mechanical slope developed to investigate rainfall-induced shallow landslides in loose coarse-grained soils, which can evolve into high-speed flowslides with propagation capabilities even in areas with very low gradients, putting people's lives and property at risk. The mechanical slope is the main component of an experimental plant built at the University of Naples. It consists of two inclinable parts: the upper one where the soil sample is deposited and the flowslide is generated, and the lower part that allows flowslide behaviour to be observed. The monitoring system implemented consist of load cells which measure the sample weight changes, tensiometers and Time Domain Reflectometry (TDR), which measure, respectively, soil suction and water content at different depths, Particle Image Velocimetry (PIV) and a Laser scanner techniques which provide the sample surface movements.

Keywords Instrumentation • Mechanical simulation • Physical model • Rainfall-induced landslides • Test rig • Two link mechanism

1 Introduction

Natural slopes in loose soils are often subject to wetting-induced instabilities involving shallow layers [1, 2]. These phenomena frequently pose hazards due to their rapid kinematics, especially when the soil grain-size distribution lies between that of gravelly sands at coarsest and sandy-silt at finest [3]. In this grain-size range capillarity effects are significant, with resulting suction levels being suitable for contributing to slope stability [4–7]. Instability upon wetting is usually induced by

G. Di Massa • S. Pagano (✉) • M. Russo • R. Russo
Dipartimento di Ingegneria Industriale, Università degli Studi di Napoli Federico II, via Claudio n. 21, Naples 80125, Italy
e-mail: stefano.pagano@unina.it

L. Pagano • M.C. Zingariello
Dipartimento di Ingegneria Civile, Edile e Ambientale, Università degli Studi di Napoli Federico II, via Claudio n. 21, Naples 80125, Italy

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_9

rainfall [8, 9], even if there may be other contributory factors, such as water lost from hidden underground conduits or water infiltration generated by human activities.

Precipitation events causing instabilities may vary according to the soil type involved. Highly pervious slopes made of gravelly sands become unstable under the effects of very intense, usually short-lasting, precipitation, without any effect exerted by antecedent rainfall. On the other hand, slopes in sandy-silty soils lose stability under the effects of prolonged wet periods followed by a major event of significant intensity [3, 8, 10].

In order to investigate what relationship establishes between precipitation history and landslide triggering under different conditions involving soil types, soil states (essentially porosity), slope inclination and slope thicknesses, a physical model was built. It consists of a mechanical slope with adjustable inclination and with the possibility of controlling or measuring the main factors that contribute to triggering a landslide. The paper presents the design of the prototype, describes its main mechanical features and shows its flexibility in changing slope factors.

2 Soil Sample Cycle

The plant consists essentially of a variable inclination mechanical slope and of other components that allow the soil sample to be reused for the execution of other tests [11]. The sample cyclically executes the following path (Fig. 9.1). It is deposited on the upper part (Part-A) of the mechanical slope (Figs. 9.2 and 9.3) by means of a hopper which moves slowly (0.2–8 cm/s) at a constant distance from Part-A. The material flows through an adjustable slit placed at the bottom of the hopper. The height of the hopper can be regulated to take account of the properties of the material to be deposited and the desired degree of compaction.

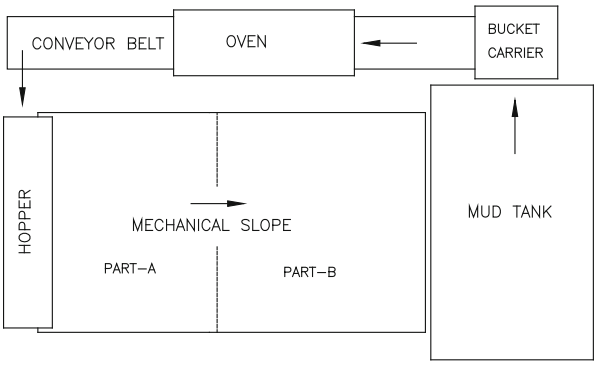


Fig. 9.1 Plant scheme and sample cycle

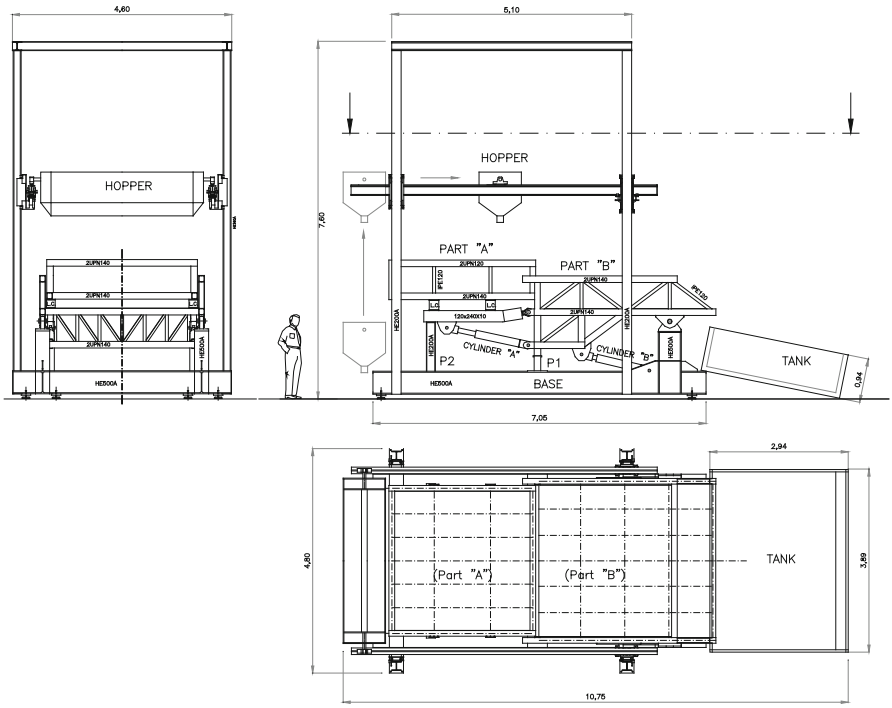


Fig. 9.2 Mechanical slope and deposition system

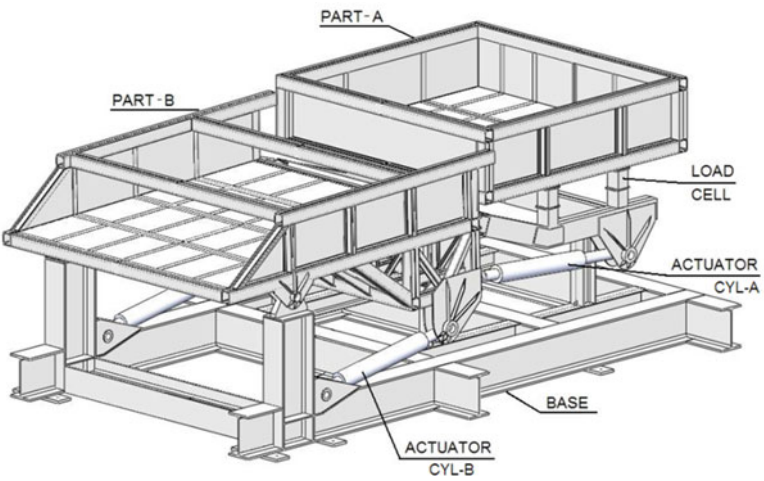


Fig. 9.3 Mechanical slope

The mechanical slope can be tilted after or before the deposition step; in the first case the hopper translates onto tilted guides, parallel to part-A, in order to keep the distance between the hopper and part-A constant (Fig. 9.4).

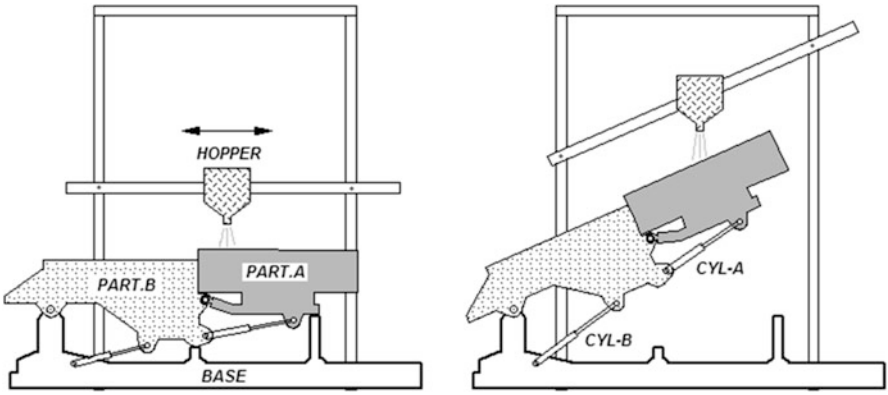


Fig. 9.4 Deposition phase

After the deposition phase the sample is subjected to artificial rainfall until it collapses. Then the collapsed soil flows onto Part-B and is collected in the mud tank. From this tank, the sample is raised, by means of a bucket carrier, onto a conveyor belt and is transported into a passing oven where the sample is dried. Then the sample can be re-deposited on part-A for a new test.

3 The Mechanical Slope

The mechanical slope consists of a base on which is articulated a two-link-mechanism; the two moving parts have the following plan dimensions (Fig. 9.3): Part-A, 3 × 3 m; Part-B, 4 × 3 m.

Part-A is instrumented to measure the weight changes of the sample. Both parts have side walls 0.70 m high to confine the sample under test. Their frames are made of steel beams (mainly wide flange beams and channel section shapes) joined by welding. The base is made of two principal horizontal H-section beams (HE500A) 7.10 m wide, connected crosswise by four other H-section beams.

Onto the base the following are welded:

- Two pillars (HE500A) on which part-B is pin-connected;
- The supports for the two lower hydraulic cylinders (cyl-B);
- Two pairs of H-section pillars, P1 and P2 (Fig. 9.2), placed under part-B and part-A respectively, to sustain the mechanical slope at rest in the horizontal configuration;
- The support for the columns of the frame sustaining the moving hopper.

Part-A and part-B are built with rectangular hollow section beams, obtained by welding two U-shaped beams (UPN140 or UPN120); the bottom and side walls of both parts are coated with 6 mm-thick steel plates.

The base and part-B are hinged with two steel pivots (ϕ95) and rolling bearings. Part-A is connected to part-B by means of two spherical joints and can be tilted with respect to part-B by using two hydraulic cylinders (Cyl-A).

Four load cells (Fig. 9.3) are interposed between the tank of part-A its supporting frame to measure the changes in sample weight due to the water absorbed during a test.

By powering the lower cylinders (Cyl-B) the entire mechanical slope can be tilted, with respect to the horizontal, up to 45°; the other pair of cylinders (Cyl-A) allows part-A to be tilted with respect to part-B in the range: +15°; −45°. The maximum inclination speed of the two elements is approximately 0.1°/s.

To define the cylinder thrust and the power required to lift the slope, the symmetry of the system was considered and one-half of the mechanical slope was analysed.

The following overall masses are considered:

– Part-A, including the load cells and the upper cylinders: 2,400 kg;
– Part-B, including the lower cylinders: 2,800 kg;
– soil sample on part-A: 12,600 kg.

The maximum thrust exerted by the cylinders and the forces acting in the pin joints connecting the slope elements were evaluated for different configurations of the mechanical slope.

Figure 9.5 reports the thrust required to tilt the slope up to 45°. The thrust has its maximum value, equal to about 300kN, when the mechanical slope is horizontal. The same figure reports the force acting in the pinned connection with the base.

Figure 9.6 reports the thrust required for the upper cylinders and the force in the pin connection between the two parts; the diagram considers the absolute part-A rotation ranging in the interval 0–50°, with part-B tilted by 45°. Even in this case the thrust reaches its maximum value when part A is in the horizontal position; in this configuration the thrust arm, with respect to the pin joint, is minimal.

The mechanical slope is tilted by means of two pairs of single-rod hydraulic cylinders [12, 13] (Bosch Rexroth - mod. CDM1); the lower ones (cyl-B) have a bore of 160 mm and a piston rod diameter of 110 mm; the maximum thrust is equal to 344 kN with an oil pressure of 170 bar. The upper cylinders (cyl-A) have a bore of 125 mm and piston rod diameter of 90 mm and can exert a thrust of 211 kN at 170 bar. The two pairs of cylinders are powered by the same oil pump with 22 cm^3 of displacement and maximum flow rate of 32 l/min; the pump is driven by an 11 kW electric motor. Synchronization between each pair of cylinders is guaranteed by the mechanical connections and by the flow divider valves that provide the same oil flow from the single source into the two actuators.

4 The Hopper

The hopper (Figs. 9.2 and 9.10) is made of sheet steel and is supported by two ϕ80 pins, engaged on two sliders guided on two beams. The sliders allow the hopper to

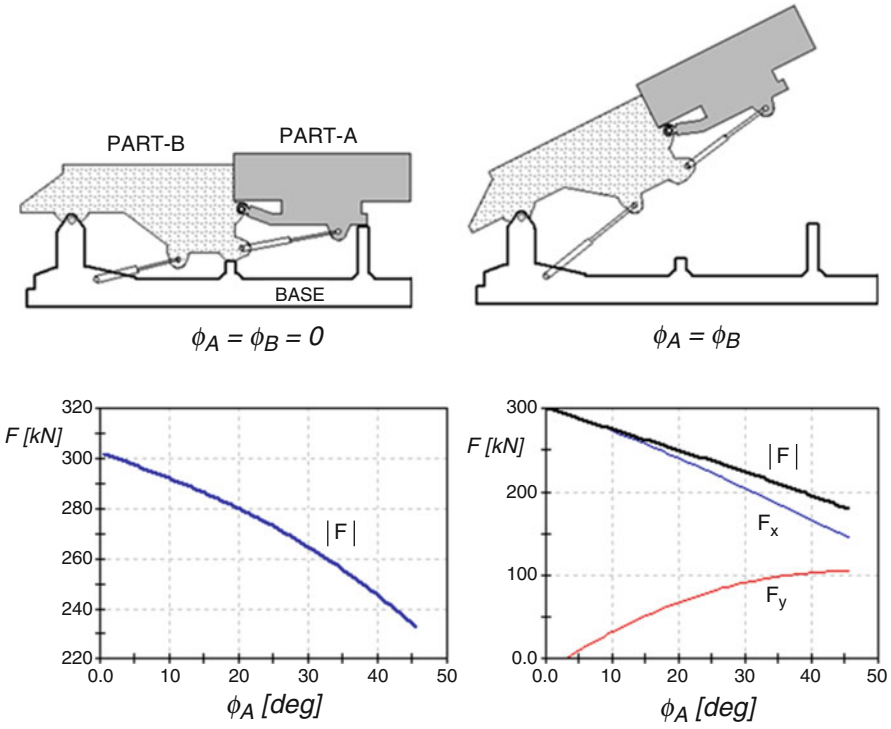


Fig. 9.5 Cylinder-A thrust and force acting in the pinned connection with the base (horizontal, Fx, and vertical, Fy, components)

translate, forward and backward, along the supporting beams during the deposition step; the translating movement is driven by two electric motors and two lead-screw systems.

The two sustaining pin axes are displaced above the hopper centre of mass such that the hopper always assumes the same vertical orientation even when the supported beams are inclined.

The two beams can climb the four columns to adjust the hopper height to achieve the suitable distance from part-A; they can also rotate to adapt the inclination to that of part-A. The vertical translational motion is driven by four electrical motors placed on the vertical sliders; each motor has a gear-wheel spliced on the motor shaft, engaged with a toothed rack welded on the column (Figs. 9.8 and 9.9).

The hopper has a volume of about 1.5 m^3 and can be loaded with a mass soil sample of about 3,000 kg. The whole sample can thus be deposited in several successive stages.

The amplitude of the opening can be regulated and, in order to facilitate soil exit, four mechanical vibrators are fixed on the hopper's external surface.

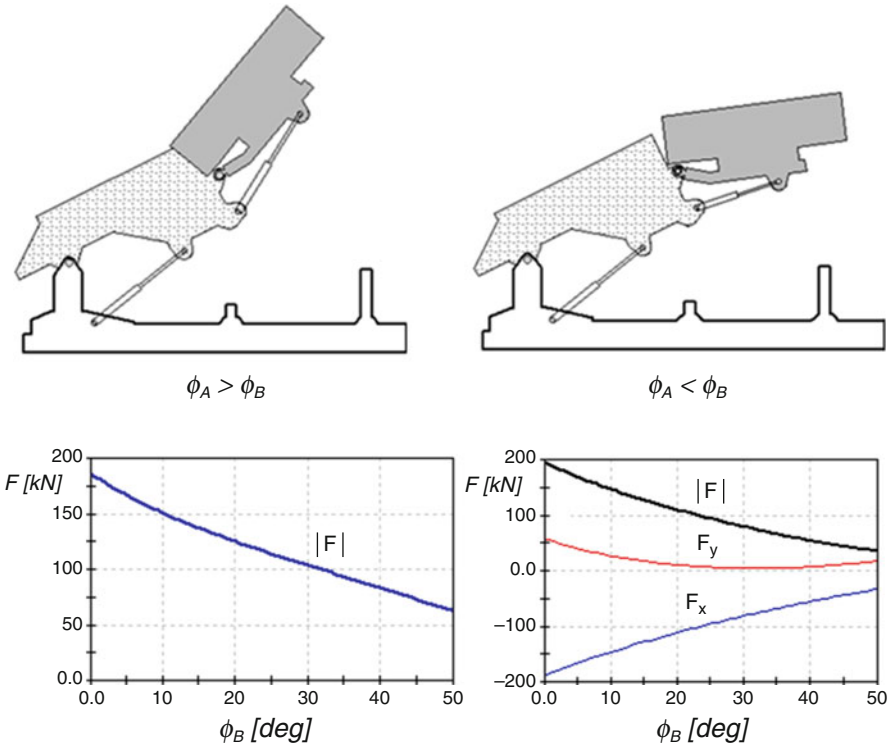


Fig. 9.6 Cylinder-B thrust and force acting in the pin connection with part-B (horizontal, Fx, and vertical, Fy, components)

5 Mechanical Slope Stress Analysis

The mechanical slope was modelled with a FEM code to perform linear static analysis and buckling linear analysis.

Figure 9.11 shows the mechanical slope in the lowest configuration, corresponding to the beginning of the lifting stage, subjected to the structural weight and sample weight of 126 kN distributed on the part-A surface. The wind action was not considered because the plant was set up inside a laboratory. This configuration, as shown in Figs. 9.5 and 9.6, involves the maximum thrusts of the hydraulic cylinders.

The main elements of the structure were modelled by linear beam element, with six degrees of freedom; the four hydraulic cylinders were instead modelled as rods, because of their terminal spherical hinges connections. Finally part-A and part-B inner linings, made of steel sheets, were modelled by means of bi-dimensional elements having membrane and bending behaviour. The base is vertically constrained in correspondence of the levelling feet (Fig. 9.7). The model

Fig. 9.7 Mechanical slope under construction: the connection between the base and part-B and the lower cylinders can be clearly distinguished. The base is supported by adjustable feet

Fig. 9.8 Mechanical slope back view. The load cells are placed between part-A tank and its supporting frame. The cylinder placed between part-A and part-B can also be seen

has about 3,000 degree of freedoms. Interpretation and consistency of the results were checked by means of simplified bi-dimensional schemes.

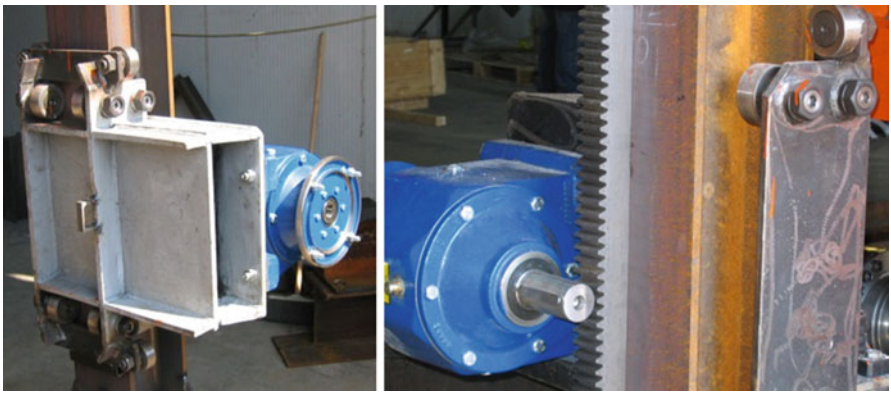


Fig. 9.9 The vertical slider allows the hopper to climb the column on which a toothed rack is welded (in the photo the gear-wheel is not yet spliced on the motor shaft)

Fig. 9.10 The hopper

Figure 9.11 shows the FEM model and the coloured map of the vertical displacement; Fig. 9.12 reports the fibre stress representation (combined axial and bending stresses) of the main elements.

6 The Instrumentation

The mechanical slope is equipped with a control/acquisition systems to govern the hydraulic actuators and to detect the instrumentation signals that allows to follow the evolution of the sample characteristics induced by the rain [14].

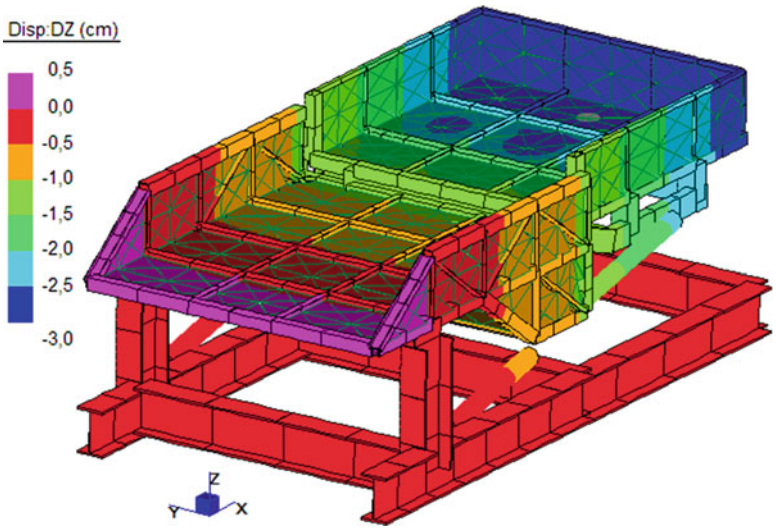


Fig. 9.11 Mechanical slope FEM model

As stated above, four load cells (MTI Multi-Component Transducers MC8-6-20000) are installed between the frame and the tank of part-A (Fig. 9.3). Each cell measures continuously in time the six reaction force components; as a result, changes in weight of part A may be obtained; such changes may be used to derive the sample unit weight during soil deposition, the changes in the water mass stored by the sample during the test (as a result of rain, seepage processes and run off), and the losses of soil mass associated to the occurrence of landslides or limited earth flows.

Once the deposition phase had been completed, the sample was wetted for about a week to reduce suction until the suction level targeted at the beginning of the test had been reached. The test was then performed tilting part-B by 30° with respect to the base and part-A by 40°. Part-A and Part-B inclination time histories are controlled by means of linear position transducers integrated in the hydraulic cylinders; tilting angles can be also checked through two removable clinometers or accelerometers.

This mechanical slope configuration makes it easier to identify post-failure behaviour by maximizing differences in time needed to cover the trench between a rapid flowslide and a slow-drying one. Rainfall intensity was assigned equal to 30 mm/h.

Increments in weight of the sample during the test indicate that the sample stores water. Importantly, the small drops in weight correspond to water lost in run-off from the sample surface and the emptying of tubes when rainfall has been stopped in order to facilitate working with the laser scanner, as explained below.

Water storage capability under constant rainfall intensity however declines with time, as indicated by the decrease in the derivative of the curve. This reduction

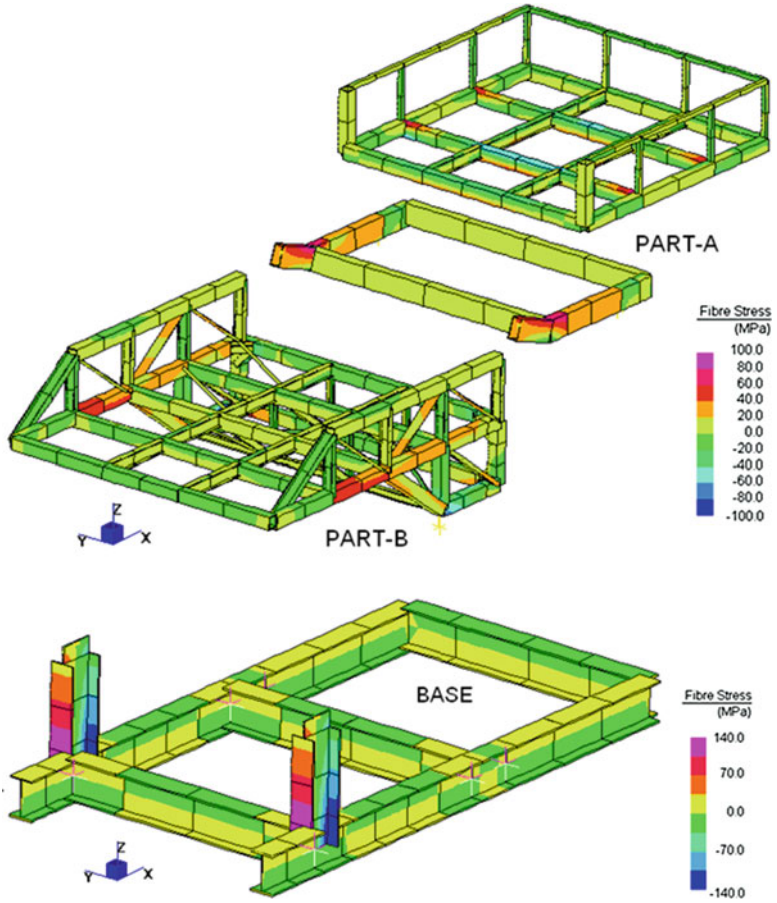


Fig. 9.12 Fibre stress representation of the main groups of elements

lowers the hydraulic gradients (driving the water drops within the sample), acting at the top surface between the exterior and interior of the sample. In the initial stages, an additional contribution to the same effect is provided by the time needed for the seeping water to reach the draining boundary downstream. Soil permeability is known to increase during the wetting process. In the initial stages, while the water has not yet drained through the permeable boundary, soil permeability increments should enhance water adsorption.

However, Fig. 9.13 indicates that permeability effects are not as substantial as those produced by the gradient reductions. Consistent with what is expressed by Fig. 9.13, initially the rainfall appears to the naked eye to be fully absorbed by the sample surface and, then increasingly rejected by it, with run-off being enhanced.

The mechanical slope is also equipped with a digital laser scanner to detect, over time, the position of the sample surface, moved by a robotized system above the

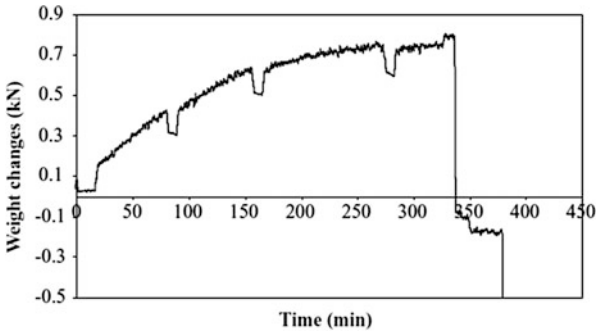


Fig. 9.13 Sample weight changes

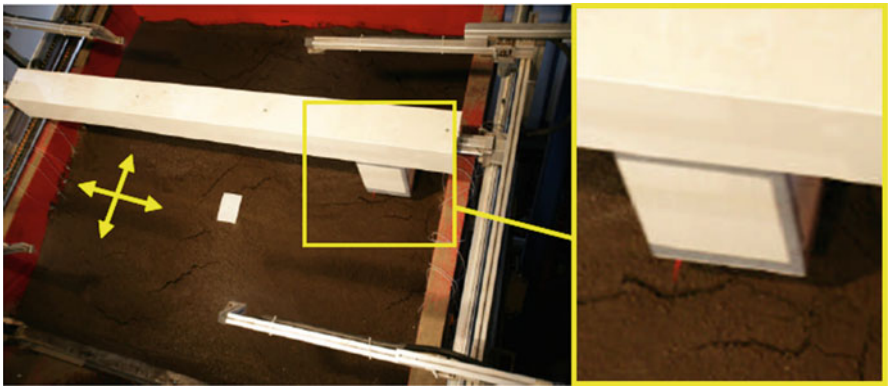


Fig. 9.14 Laser scanner apparatus on part-A

sample surface, in a plane parallel to part-A. The robotized system consists of a bar driven longitudinally by an electrical motor along two slide guides fixed on part-A border (Fig. 9.14). Another electrical motor moves the scanner transversally, along the bar.

To be strictly, the instrument generates a band of laser rays forming longitudinally on the impacted surface strip of points which are detected by the CCD sensor. Each strip is made of a large number of monitored points, being the point resolution of 0.25 mm, and is about 100 mm large.

The laser scanner sensor has a working depth ranging from $z = 120$ to $z = 220$ mm.

To determine the sample surface displacement and velocity, a video acquisition system is employed and the data of interest are derived through the Particle Image Velocimetry (P.I.V.) analysis that compares different frames at two different instant of time.

The system consists of a video camera (Basler A404K), having a pixel resolution of 0.8 Megapixel and a sample period up to 25 frames per second. The videocamera points perpendicularly to the sample surface three meters far from it (Fig. 9.15).

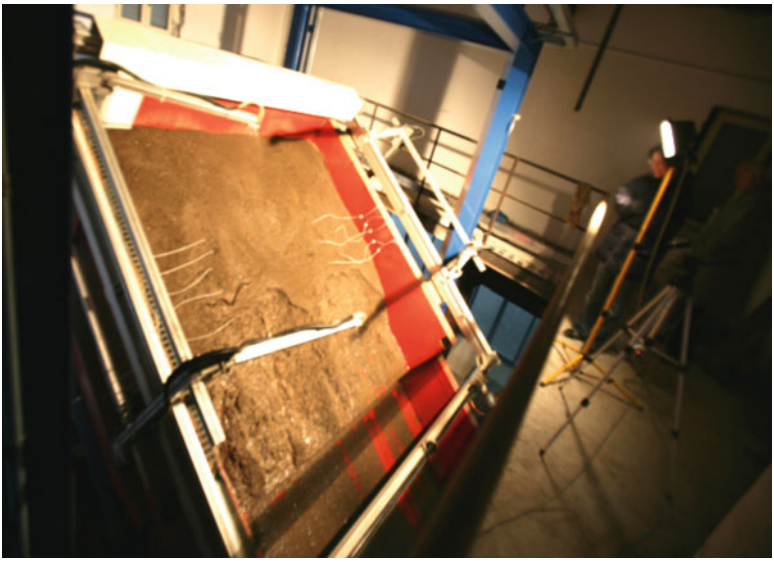


Fig. 9.15 P.I.V. apparatus

The technique enable one to individuate exactly the triggering time and to follow the kinematics of the landslide. However, the accuracy in terms of displacements, of about 2 cm, may be considered satisfactorily only for large mass movements characterizing the post-failure stages.

Finally, a number of tensiometers [15] and of TDR rod-probes [16] have been adopted to measure within the sample soil suction and water content, respectively.

For pyroclastic soils suction may be measured quite reliably through the small tip tensiometers developed by the *Soil Moisture*, since hardly suction values exceed 80 kPa, i.e. the full scale of such devices. These devices however suffer of time lag problems for high suction values. The greater the suction value the longer is the time needed for the tensiometer to equilibrate a change in suction. The time lag arises significant when suction exceeds 30 kPa, while becomes satisfactorily short below this value.

As well known, the TDR technique adopts the measurement of the soil permittivity (dielectric constant) to derive the soil water content. Soil permittivity is obtained by measuring the travel time of an electrical pulse generated by a reflectometer within a cable connected to a metallic probe inserted within the soil. The TDR device used in this experimentation is the TDR100 reflectometer developed by *Campbell* connected with 8 m long cables to a number of TDR 3-rod probes, each one 30 cm long.

Both tensiometer and TDR measurements have been acquired through a data logger (*Campebell Scientific* CR1000).

Small tip tensiometers and TDR probes have been installed within the sample during the deposition process at three or four different depths.

7 Conclusions

In this paper, a mechanical slope that allows to simulate rainfall-induced flowslides and its instrumentation useful for the measurement of the main variables involving the phenomenon was presented.

It was explained how the device allows the main factors affecting such phenomena to be taken into account. The result of a test regarding the measure of the amount of water absorbed is reported together with possible experimental procedures that may be adopted to perform the tests.

References

1. Toll, D.G., Lorenco, S.D.N., Mendes, J., Gallipoli, D., Evans, F.D., Augarde, C.E., Cui, Y.J., Tang, A.M., Rojas Vidovic, J.C., Pagano, L., Mancuso, C., Zigariello, C., Tarantino, A.: Soil suction monitoring for landslides and slopes. Quart. J. Eng. Geol. Hydrogeol. **44**(1), 23–33 (2011). doi:10.1144/1470-9236/09-010
2. Pagano, L., Zingariello, M.C., Vinale, F.: A large physical model to simulate flow-slides in pyroclastic soils. 1st European Conference on Unsaturated Soils, 1–4 luglio 2008, pp. 205–213, Durham (2008)
3. Pagano, L., Picarelli, L., Rianna, G., Urciuoli, G.: A simple numerical procedure for timely prediction of precipitation-induced landslides in unsaturated pyroclastic soils. Landslide **7**(3), 273–289 (2010). doi:10.1007/s10346-010-0216-x
4. Rianna, G., Pagano, L., Urciuoli, G.: Rainfall patterns triggering shallow flowslides in pyroclastic soils. Eng. Geol. **174**, 22–35 (2014)
5. Rianna, G., Pagano, L., Urciuoli, G.: Investigation of soil-atmosphere interaction in pyroclastic soils. J. Hydrol. **510**, 480–492 (2014)
6. Pagano, L., Reder, A., Rianna, G.: Processi di infiltrazione ed evaporazione nei terreni piroclastici illustrati attraverso la selezione di alcuni eventi rappresentativi. Rivista Italiana di Geotecnica **48**, 56–70 (2014)
7. Calabresi, G., Colleselli, F., Danese, D., Giani, G., Mancuso, C., Montrasio, L., Nocilla, A., Pagano, L., Reali, E., Sciotti, A.: Research study of the hydraulic behaviour of the Po river embankments. Can. Geotech. J. **50**, 947–960 (2013)
8. Pagano, L., Reder, A., Rianna, G.: Experiments to investigate the hydrological behaviour of volcanic covers, The Third Italian Workshop on Landslides (Naples, October 2013). Procedia Earth Planet Sci, Elsevier, doi: 10.1016/j.proeps.2014.06.013, 9, 14–22 (2014)
9. Pagano, L., Reder, A., Rianna, G.: Prediction of suction evolution of silty pyroclastic covers in flume tests and field monitoring, Paper describing predicted results for Round Robin Test of IWL2013 – The Third Italian Workshop on Landslides (Naples, October 2013). Procedia Earth Planet Sci. Elsevier, 9, 214–221 (2013)
10. Rianna, G., Pagano, L., Urciuoli, G.: A physical model to investigate the influence of atmospheric variables on soil suction in pyroclastic soils. In: Proceedings of the Second European Conference on Unsaturated Soils, E-UNSAT 2012: Unsaturated Soils: Research and Applications, Part 2, 221–227 (2012). doi: 10.1007/978-3-642-313431_28
11. Di Massa, G., Pagano, L., Pagano, S., Russo, M., Russo, R., Zingariello, C.: A mechanical slope for flowslide investigations, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 1376–1380, London. 2–4 July 2014

12. Pagano, S., Russo, R., Strano, S., Terzo, M.: Modelling and control of a hydraulically actuated shaking table employed for vibration absorber testing, ASME 2012. In: 11th Biennial Conference on Engineering systems Design and Analysis (ESDA) – Advanced Computational Mechanics – Nantes. July 2–4 2012
13. Pagano, S., Russo, R., Strano, S., Terzo, M.: Non-linear modelling and optimal control of a hydraulically actuated seismic isolator test rig. Mech. Syst. Signal Process. **35**(1–2), 255–278 (2013). doi:10.1016/j.ymssp.2012.09.002
14. Pagano, S., Russo, M., Strano, S., Terzo, M.: Seismic isolator test rig control using high-fidelity non-linear dynamic system modelling. Meccanica **49**(1), 169–179 (2014). doi:10.1007/s11012-013-9783-y
15. Rojas, J.C., Pagano, L., Zingariello, M.C., Mancuso, C., Giordano, G., Passeggio, G.: A new high capacity tensiometer: first results. CRC Press – Taylor & Francis Group, London I European Conference on Unsaturated Soils, pp. 205–211. Durham, 2–4 July 2008
16. Pagano, L., Reder, A., Rianna, G.: Calibration of TDRs and heat dissipation probes in pyroclastic soils, The Third Italian Workshop on Landslides (Naples, October 2013). Procedia Earth Planetary Sci.. Elsevier. **9**, 171–179 (2014). doi: 10.1016/j.proeps.2014.06.016

Chapter 10
Influence of Wire EDM on Fracture Toughness of TI6AL4V

Daniel M. Madyira and Esther T. Akinlabi

Abstract Grade 5 titanium (Ti6Al4V) is considered as the workhorse titanium alloy. It is widely termed an aerospace alloy and is a relatively new engineering material. One of the major challenges in the use of this aerospace material is its machinability. Its high strength which is maintained at elevated temperatures, low thermal conductivity, low elastic modulus and high reactivity with oxygen is a perfect recipe for machining challenges. This leads to high tool wear and long production times. Such challenges can be overcome by electrical discharge machining (EDM). Given that titanium is usually applied to mission critical components (gears, shafts, wing sections etc.), it is important to understand the possible effect of wire EDM (WEDM) on their structural performance. The purpose of this chapter is to present critical issues related to the effect of wire EDM on the fracture toughness of this aerospace material. EDM and WEDM processes are discussed. Their effects on the structural integrity of Ti6Al4V are then demonstrated through fracture toughness measurements. Four specimens were produced using wire EDM. This includes the pre-crack which is usually introduced by fatigue cycling. Obtained results indicate a slight decrease in fracture toughness compared to that reported in literature. It was also concluded that wire EDM technique can be used as an alternative to fatigue pre-cracking in fracture toughness testing of Ti6Al4V.

Keywords Compact tension specimen • Electrical discharge machining • Grade 5 titanium alloy • Fracture toughness • Ti6Al4V • Wire EDM

1 Introduction

Electrical discharge machining (EDM) is a non-traditional material removal technology that was developed in the 1940s. Since then, it has been predominantly restricted to special applications such as the manufacture of high accuracy com-

D.M. Madyira (✉) • E.T. Akinlabi
Department of Mechanical Engineering Science, University of Johannesburg, Auckland Park 2006, P.O. Box 524, Johannesburg, South Africa
e-mail: dmadyira@uj.ac.za; etakinlabi@uj.ac.za

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_10

ponents such as dies and molds [1]. It has also been used as a finishing process for certain aerospace and automotive components. Unlike traditional material removal processes such as turning, milling, grinding and honing, EDM is a non-contact process. It has the potential to reduce problems associated with conventional metal removal processes such as high residual stresses, chatter and vibration which may affect the quality and structural integrity of the finished product [2]. These benefits make EDM eminently suitable for producing aerospace and automotive components from advanced high strength alloys such as 7075-T6 aluminum and Ti6Al4V. The only condition for using EDM is that the material must be conductive. This chapter presents the theory of EDM and wire EDM as applied to the production of components from high strength alloys. This will be followed by a discussion on fracture toughness and how it is affected by a manufacturing process applied to a component and lastly, the influence of wire electrical discharge machining (WEDM) on the fracture behavior of Ti6Al4V.

2 Electrical Discharge Machining

In the EDM process, material removal is achieved by repeated electrical discharges between an electrode and a work piece in the presence of a dielectric fluid. Most applications use copper electrodes and deionized water as the dielectric fluid but hydrocarbon dielectrics are also viable alternatives. EDM is therefore a thermoelectric process hence the need for both the work piece and the electrode to be good electrical and thermal conductors. The process involves an electrode being moved towards the work piece until the gap is small enough for the applied voltage to ionize the dielectric and cause an electrical discharge to occur across the gap. Material removal is achieved by a complex mix of generated high temperatures and the erosive effect of the electrical discharge. Li et al. [3] suggest that the maximum temperature attained in the process zone during EDM can be as high as 40,000 K. Such temperatures inevitably lead to localized melting of the material which is immediately quenched and washed away by the dielectric fluid. At the same time, vaporization of the dielectric fluid can also occur hence the need to use environment neutral dielectrics such as water. Hydrocarbons tend to be damaging to the environment and must therefore be avoided.

The performance of the EDM process depends on such factors as the discharge power density, pulse duration, electrical and thermal conductivity of the materials being processed, and motion of the tool, tool-work piece gap and rate of tool wear. Over the almost half a century of EDM use, a number of improvements have been introduced aimed at better productivity and quality. The most prominent enhancements have been the introduction of tool and wire vibration [4], work piece vibration [5], ultrasonic assistance and dry EDM [6].

3 Wire Electrical Discharge Machining

Wire EDM is a variation of EDM among other variations that include EDM milling, Ultrasonic EDM (UEDM) and conventional EDM. In WEDM, the electrode is a wire that is supplied from a reel. Mohd Abbas et al. [1] suggest that WEDM makes up about 24 % of EDM activities. Therefore, it is becoming a key enabling manufacturing technique. Figure 10.1 presents a schematic representation of the WEDM process.

Some of the advantages of wire EDM include low work holding forces, low cutting forces, high accuracy process, capability to produce complex shapes, ability to produce small holes of the order of 50 μm diameter, minimal tool wear, ability to machine materials with high hardness (i.e. difficult to machine) and can handle small corners and narrow slots and no burrs are generated during machining. The negative side of WEDM is that these machines are very expensive compared to other similar sized tools, high skilled operators are required to run the machines, the process can only be applied to electrical conductors, the material removal rate (MRR) is low compared to turning and milling and the process requires de-ionized water. Furthermore, the high localized temperatures produce a heat affected zone (HAZ) which can negatively affect the surface and hence the structural integrity of the component [8].

Li et al. [3] have shown that varying discharge energy density during WEDM has a direct effect on the resultant surface integrity of Inconel 718. Deionized water dielectric and an uncoated brass wire were used in the process. The surface microstructure changed from coral reef topography at high discharge energy density to one with random micro voids at low energy density. A largely isotropic surface topology was reported in which average surface roughness decreased from 3.75 to 1.25 μm for corresponding energy density values. The white layer was found to be discontinuous in the thickness range of 13.3–3.3 μm. Micro hardness was found to be very low at the surface due to the white layer and increased to parent material

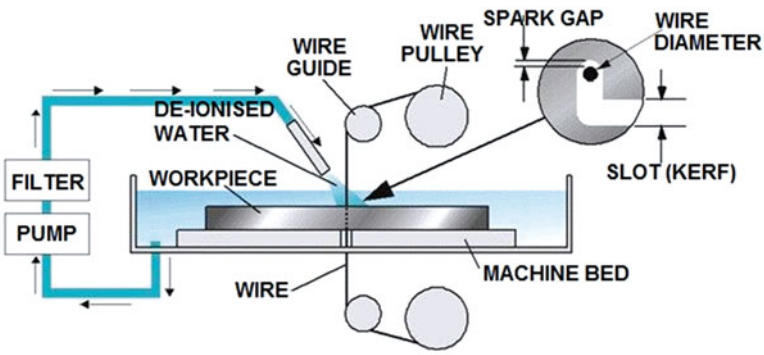


Fig. 10.1 Illustration of the WEDM process [7]

hardness within 20 μm from the surface. No surface micro cracks were reported for the range of discharge energies investigated.

4 Fracture Mechanics Theory

Fracture can be defined as the separation of a component into two or more pieces as a result of externally applied loading. Fracture is a problem that society has lived with since the dawn of man-made structures. The cause of failure of structures has been identified to arise from two main aspects. The first being negligence during design, construction or operation of a structure. The second is the application of new design or new materials which produce unexpected and undesirable result. This chapter is more concerned about the second aspect. However there could also be cases where during construction of a component for an established design, new manufacturing techniques are applied for expedience. This is a first aspect issue that can lead to unintended behavior resulting for example from new induced residual stresses which had not been anticipated during initial design analysis. The second aspect is much more difficult to prevent especially in cases where new materials are introduced and unanticipated behavior is obtained. New designs and new materials should therefore only be introduced into service after extensive testing. This will not completely eliminate failure but will reduce failure frequency.

History is dotted with a lot of catastrophic failures resulting from this second failure aspect. The Liberty Ships were the first to have an all welded hull [9]. Because they were introduced during the Second World War, there was limited time to investigate material behavior using this new manufacturing technique with its inherent material modifications. A significant number of these ships suffered catastrophic fracture in service resulting from brittle material behavior in cold sea waters. However, today almost all ships have welded hulls but do not fracture because the material behavior under those conditions is now well understood, ironically, as a result of data collected from the Liberty Ship failures.

The fracture behavior of materials has been investigated for many decades. The first qualitative work on fracture is that of Leonardo da Vinci who reported the inverse relationship between iron wire strength and wire length [10]. However, Griffith is considered the father of fracture mechanics due to his contribution and quantified his findings using stress and energy analysis of the cracked components. He provided the link between the fracture stress and flaw size [11]. According to the Griffith theory, a crack becomes unstable when the strain energy change that results from a small increment of crack is sufficient to overcome the surface energy of the material. Instability in this case refers to the fast facture of the component. This theory agreed well with brittle materials such as glass and can be expressed mathematically as [9]:

$$G = \frac{\pi \sigma^2 a}{E}$$

Where G is the energy release rate (rate of energy change per crack area extension), σ is the remotely applied stress in an infinite plate, 'a' is half the crack size of a crack of size 2a in an infinite plate and E is the Young's modulus of elasticity of the plate material. G characterizes the behavior of the cracked component and hence, can be considered as the crack driving force. For a given stress field, as the crack grows, the crack driving force increases proportionately. Similarly, for a given crack size there is a corresponding increase of G with increase in stress. For a given crack size 'a', the energy release rate at the point of crack instability is known as the critical energy release rate (G_c) which can be identified as a material property. The G_c value can be measured experimentally for a cracked component by monitoring the stress at the point of crack instability. This theory agreed well with brittle materials but failed for ductile materials. In addition, the difficulty of measuring and integrating this approach into engineering design required an approach with an analytical and mathematical approach that can be used in day-to-day engineering analysis. Therefore, a mathematically analyzable parameter called the stress intensity factor, K, was introduced by Irwin [12]. This was the birth of linear elastic fracture mechanics (LEFM). The stress distribution around a crack in a linear elastic material was determined to be [9]:

$$\sigma_{ij} = \frac{K_I}{\sqrt{2\pi r}} \mathrm{f}(\theta)$$

Where σ_{ij}, are the stresses around the crack, r is the distance from the crack tip and θ is the angular position relative to the direction of crack growth. K_I is the stress intensity factor. The stress field ahead of the crack tip is therefore governed and characterized by the magnitude of the stress intensity factor. It is also worth noting the stress singularity as r becomes very small. This is obviously not realistic since metallic materials would redistribute this stress through yielding. Issues related to extensive plastic deformation ahead of the crack tip are addressed through elastic-plastic fracture mechanics (EPFM). For global elastic behavior and small scale localized plasticity in the crack tip area, the crack driving force in terms of the crack tip stress intensity factor is given by [9]:

$$\mathrm{K_{(I,II,III)}} = \mathrm{Y}\ \sigma\ \sqrt{\pi\ \mathrm{a}}$$

Where K is the stress intensity factor which can be considered the crack driving force and Y is dimensionless constant depending on component geometry. The subscripts I, II and III refer to the mode I (peel or opening), mode II (in plane shear or sliding) and mode III (out of plane shear or tearing) deformation modes. In a similar fashion to the energy release rate, at the point of crack instability, $K = K_c$ where K_c is the critical stress intensity factor that can be considered a material property. For plane stress conditions, the energy release rate is related to the stress intensity factor by the equation [9]:

$$\mathrm{G} = K^2\ E'$$

Where $E' = 1/E$ for plane stress conditions and $(1 + \upsilon)$/E for plane strain conditions. For linear elastic behavior, the choice of which parameter to use to characterize fracture behavior is therefore, a matter of choice and convenience. The mathematical analysis used in normal engineering design favors the use of the stress intensity factor and hence is the most common.

5 Fracture Toughness Measurement

Fracture toughness is a material property that can be characterized by a number of different parameters ranging from Charpy Impact Energy (CIE), Energy Release Rate (ERR), Stress Intensity Factor (SIF) and the J integral among others [13]. These parameters represent the resistance of the material to crack growth through the quantitative change in energy that occurs at the point of fracture as presented in the previous section. In this chapter, the stress intensity factor concept was used to determine the fracture toughness of Ti6Al4V after WEDM. The experimental determination of fracture toughness in terms of stress intensity factor is therefore discussed.

Fracture toughness measurement is conducted in accordance with ASTM Standard E1820-11 [14]. In this method, the mode I (opening mode) critical stress intensity factor (K_{IC}) is measured using the compact tension specimen. The size of the specimen depends on the material being tested. Figure 10.2 shows the variation of stress intensity with specimen thickness.

As the specimen transitions from plane stress to plane strain conditions through increase in thickness, there is a corresponding reduction in stress intensity factor. Once the full plane strain conditions are attained, the stress intensity factor becomes independent of geometry and hence is dependent on the component material. Therefore the critical SIF as a material property is obtained under plane strain conditions. The dimensions of the specimen required to attain plane strain conditions are specified in the ASTM E1820-11 standard as [14]:

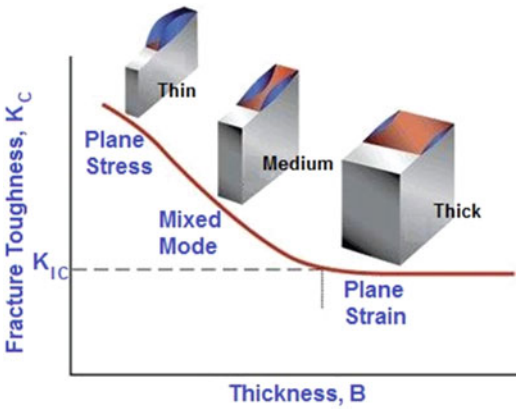


Fig. 10.2 Variation of stress intensity factor with thickness [15]

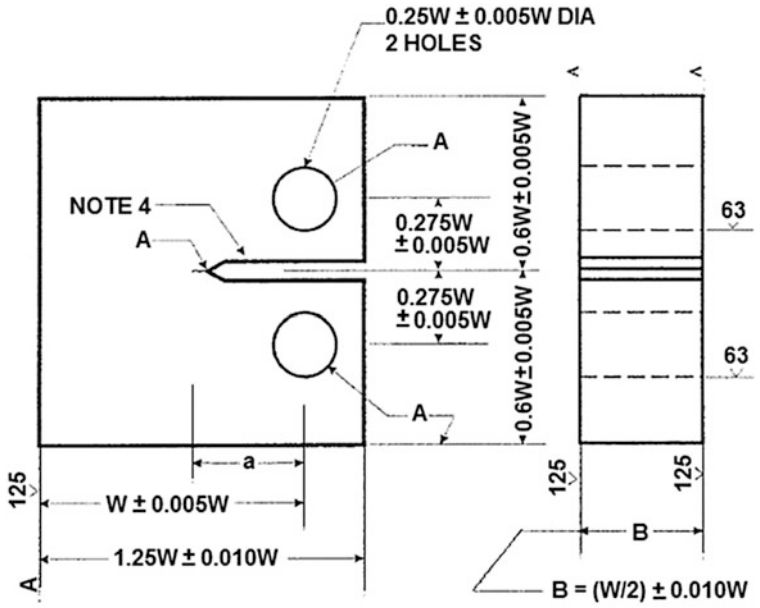


Fig. 10.3 Compact tension specimen geometrical specifications [14]

$$B,\ a,\ W\ \geq\ 2.5\left(\frac{K_{IC}}{\sigma_Y}\right)^2$$

Where B is the specimen thickness, 'a' is the crack size, W is the width, K_{IC} is the critical stress intensity factor (fracture toughness) and σ_Y is the yield strength of the material. For new materials, an iterative process is required in which K_{IC} is estimated in the determination of specimen geometry. The experimentally determined value is then used to confirm the validity of specimen geometry. If the new geometry does not meet this condition, retesting is required until the specimen meets the standard specifications. The full specimen geometrical specifications are shown in Fig. 10.3.

The critical stress intensity factor (material fracture toughness) is then computed from [14]:

$$K_{IC} = \frac{P_{max}}{(B\ B_N\ W\ )^{0.5}}\ f\left(\frac{a}{W}\right)$$

Where P_{max} is the maximum force recorded in the tensile test, B and B_N are equal in the absence of side groves and:

$$f\left(\frac{a}{W}\right) = \frac{\left(\left(2+\frac{a}{W}\right)\left[0.88+4.64\left(\frac{a}{W}\right)-13.32\left(\frac{a}{W}\right)^2-14.72\left(\frac{a}{W}\right)^3-5.6\left(\frac{a}{W}\right)^4\right]\right)}{\left(1-\frac{a}{W}\right)^{\frac{3}{2}}}$$

6 WEDM Effect Fracture Toughness

Titanium as an element is the ninth most abundant on Earth and makes 1 % of the Earth's crust [16]. Titanium alloys have unique mechanical properties making them most suited to challenging applications. The density of titanium alloys is 4,500 kg/m^3 compared to 7,800 kg/m^3 for steel and 2,800 kg/m^3 for aluminum while tensile strength is about 1,000 MPa, steel 600 MPa and aluminum 450 MPa. Therefore, titanium exhibits superior strength-to-weight. This is additional to its excellent corrosion resistance. Furthermore, these properties are maintained at relatively high temperatures making titanium alloys good candidates for use in high temperature applications such as burners and compressor and turbine blades and discs [17]. Ti6Al4V is the workhorse alloy of the titanium family of alloys and has nominal composition of 6 % aluminum and 4 % vanadium the remainder being largely titanium. This is a two phase material with both alpha (hexagonal close packed (hcp)) and beta (body centered cubic (bcc)) phases. The alloying element aluminum is the alpha phase stabilizer while vanadium stabilizes the beta phase.

Despite these benefits, engineering use of titanium alloys is still limited to specialized applications such as aerospace and biomedical fields. This is a result of high primary and secondary processing costs which make these alloys unaffordable for common engineering applications. Primary processing is complicated by the high reactivity of titanium with other elements including oxygen. Therefore, nothing much can be done, using the current technology, to reduce primary processing costs. However, secondary processing costs such as machining can be reduced by using such techniques as high speed machining (HSM) which increase the material removal rate (MRR). HSM has been highly successful in the production of aerospace aluminum components such as wing sections. In addition, HSM produces better surface finish leading to better quality products when compared to conventional machining. Extensive research work is currently underway to understand the effect of HSM on the structural integrity of Ti6Al4V [18]. Recently, wire EDM has also become a viable option for economically processing difficult-to-machine materials such as Ti6Al4V. Titanium is classified as "difficult" to machine by traditional methods due to poor thermal conductivity, high strength that is maintained at elevated temperatures, high chemical reactivity and low modulus of elasticity.

The poor thermal conductivity of titanium alloys affects the surface integrity after WEDM. Recently, Nourbakhsh et al. [19] have reported on the effect of varying WEDM parameters on surface integrity of titanium components. WEDM was applied to Ti6Al4V using deionized water dielectric and three different wires. Standard uncoated high-speed brass wire was used as baseline study and compared to zinc coated brass wire. Using Taguchi design of experiments, they concluded that the cutting speed is a function of pulse width and peak current. Surface roughness was reported to increase with pulse width and peak current while decreasing with wire tension. Increase in wire tension leads to a reduction in wire vibration and hence lower surface roughness. In general, uncoated wire produced the worst

surface finish. Furthermore, the uncoated brass wire produced surfaces with more cracks, craters and melted drops.

Antar et al. [20] reported similar performance of coated and uncoated wire on WEDM of Ti6Al2Sn4Zr6Mo titanium alloy. Using a five axis machine with deionized water dielectric and two coated copper core wires, they reported interesting results when compared to the standard brass wire. Roughness decreased from about 3.8 μm for rough cut to about 0.6 μm for finishing cuts. There was no major difference in roughness between coated and uncoated wires. However, the uncoated wire produced almost twice the size of the white layer thickness. This is significant as most micro cracks are found in this layer. In terms of residual stresses, all wires produced tensile residual stresses of about 150 MPa during rough cutting. This decreased to about 50 MPa for finishing cuts. It was interesting that the uncoated wire produced compressive residual stresses for the finishing cut. This is significant as compressive residual stresses are associated with favorable fatigue performance attributed to crack closure. Finally, almost 70 % improvement in productivity was reported for coated wires compared to uncoated wires during rough cutting. This was due to increased sparking gap due to rapid melting and vaporization of zinc coating. The zinc particles also led to the improvement in dielectric ionization. Parameters of greatest import on structural integrity are surface roughness, heat affected zone and residual stresses. Surface roughness determines the level of stress concentrations on the surface of loaded components. The heat affected zone implies that the surface and subsurface layers see modified mechanical properties while residual stresses affect the load carrying capacity and fatigue performance of the components in service.

Previous research has therefore proven the significant effect of the WEDM process on surface integrity of the produced components. This is amplified during WEDM of titanium alloys due to poor thermal conductivity. However, little work has been published on the effect of WEDM or EDM on structural integrity of produced components. Important structural integrity indicators include fracture toughness which will be investigated for Ti6Al4V. The purpose of this investigation was to determine the effect of WEDM on the fracture toughness of Ti6Al4V.

6.1 Description of Material Used

The grade 5 titanium alloy used for this investigation was supplied in round bar form (65 mm diameter) by a Pretoria based company GEM Manufacturing (Pvt) Ltd. The chemical composition of the material as per supplier material certificate is given in Table 10.1.

The corresponding mechanical properties for this material are given in Table 10.2.

Table 10.1 Chemical composition of grade Ti6Al4V used in this investigation

	Al	V	Fe	C	N	O	H	Ti
Content (%)	6.4	4.2	0.03	0.01	0.01	0.18	0.003	Balance

Table 10.2 Mechanical properties of Ti6Al4V alloy used in this investigation

Yield strength (MPa)	Tensile strength (MPa)	Elongation (%)	Area reduct. (%)
885–910	980–1,010	14–18	43–45

Fig. 10.4 WEDM system. (**a**) WEDM machine. (**b**) Wire reel. (**c**) Actspark control console

6.2 *Specimen Preparation*

Fracture toughness testing was conducted according to ASTM E1820-11 [14]. Computed size of the specimen according to Fig. 10.3 was W = 40 mm and B = 20 mm. Four specimens were machined from a 65 mm diameter round bar using a Xenon WEDM machine (see Fig. 10.4). The machine is controlled by Actspark software. Deionised water dielectric was used with a 250 μm diameter uncoated brass wire for profile cutting followed by pre crack cutting using100 μm diameter wire.

The preparation was done in three phases. The first stage involved WEDM of the profile of the specimen using a 250 μm wire. This was followed by introducing a pre crack using a 100 μm diameter wire. Finally, the specimens were then separated from the main bar. Figure 10.5 shows the WEDM of fatigue pre-crack.

Specimens with extensometer mounting bracket holes are shown in Fig. 10.6.

6.3 *Equipment*

Tensile loading of the specimens was conducted using Instron 1195 testing rig driven by Bluehill 2 software. This is a screw type machine with a 100 kN load cell. The specimen mounting arrangement with extensometer mounted is shown in Fig. 10.7.

Fig. 10.5 WEDM of fatigue pre-crack

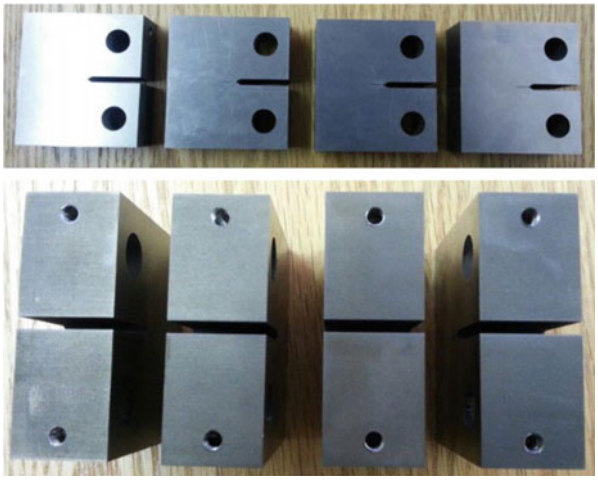


Fig. 10.6 Machined Ti6Al4V compact tension specimens

6.4 Experimental Protocol

Tests were done under displacement control at a rate of 1.6 mm/min in line with ASTM E1820-11 recommendations. Displacement was measured using the extensometer shown in Fig. 10.7. In cases where initial fast fracture did not lead to complete separation of the specimen, the specimen was un-mounted, cooled in liquid nitrogen and quickly mounted in the machined for fast brittle fracture to complete specimen separation.

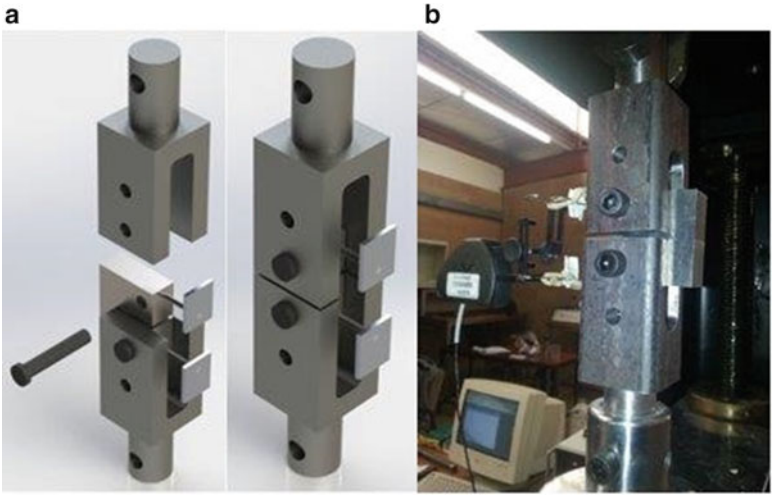


Fig. 10.7 Specimen mounting arrangement. (**a**) 3D model. (**b**) Actual machine

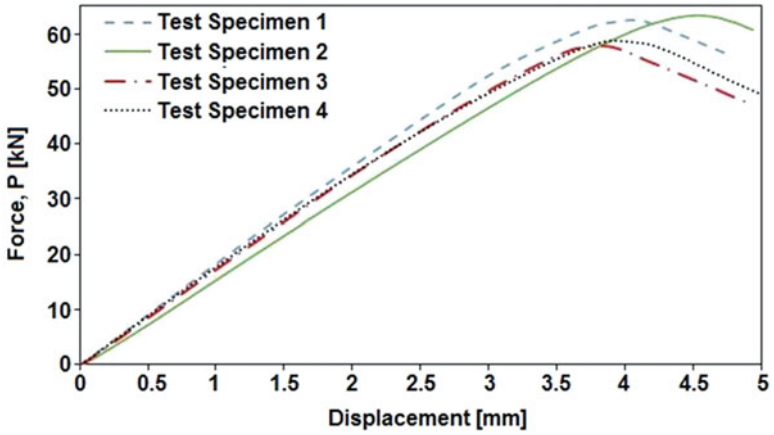


Fig. 10.8 Force-displacement response for the four specimens

6.5 Fracture Toughness Measurements

The load displacement response for all the four specimens is given in Fig. 10.8. Maximum loads obtained are given in Table 10.3.

Fracture toughness values obtained are given in Fig. 10.9 and compared to the average toughness value for the material as reported in the literature. The average value obtained experimentally was 111.82 MPa.m$^{0.5}$ with a standard deviation of 6.38 MPa.m$^{0.5}$. This compares to 117 MPa.m$^{0.5}$ reported in literature. Confirmation of plane strain fracture was done by observing fracture surfaces using optical microscope. Figure 10.10 shows typical fracture topography of all the specimens tested. A large portion of the fracture surface underwent plane strain fracture as evidenced by the flat fracture surface.

Table 10.3 Maximum fracture forces

Specimen	Maximum force (kN)	Test time (s)
1	62.5	154.44
2	63.4	165.078
3	57.9	120.252
4	58.8	149.592

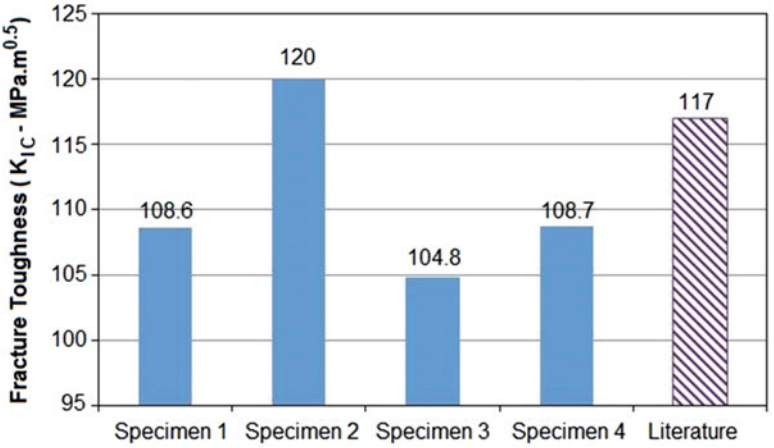


Fig. 10.9 Comparison of the fracture toughness values

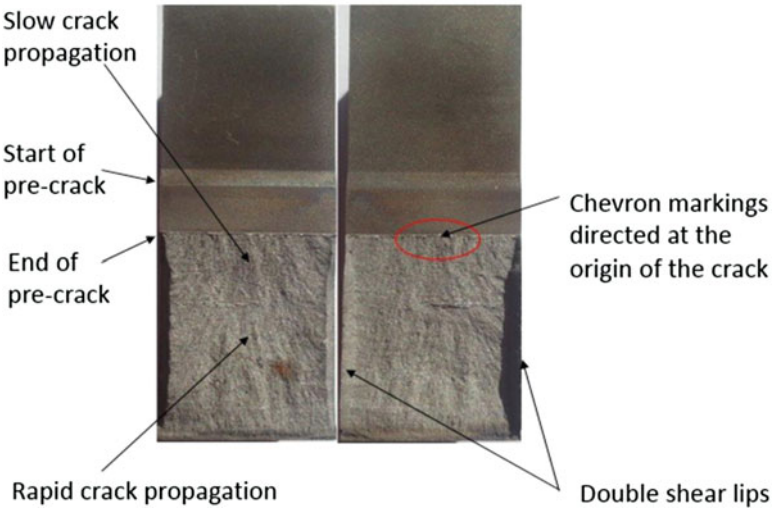


Fig. 10.10 Typical fracture surface after test

Specimens had shear lips of 1 mm size associated with plane stress failure. This was 10 % of the plane strain region. This agrees with the standard. The crack tip radius was measured to be 200 μm which is twice the size of the wire used to generate the pre-crack.

Compact tension specimens produced for this investigation were made using WEDM in a fraction of the time that it would take using fatigue. This reduced the cost significantly. The average fracture toughness measured using SIF was 111.82 MPa.m$^{0.5}$ which was lower than the value of 117 MPa.m$^{0.5}$ reported in literature. The specimens used in the tests satisfied the requirements of ASTM E1820-11 with the exception of fatigue pre-crack which was produced by WEDM instead of the recommended fatigue cycling.

This difference is insignificant for practical design as conservative safety factors are applied in most applications. However, this variance could also be due to the WEDM process. The order of magnitude of the crack tip radius expected using fatigue pre-cracking is 20 μm which is about the average grain sizes for these materials. This compares to 200 μm achieved using WEDM with a 100 μm diameter wire. The result is reduced stress concentration and hence, stress intensity in the crack tip region. This is expected to increase the fracture toughness. However, the modified properties of the material as a result of the HAZ would result in lower fracture load. Although no detailed analysis was conducted on the effect of WEDM on the microstructure or hardness of the material, the results indicate a definite reduction of mechanical properties. The combined effect of these key factors was the slightly reduced average critical stress intensity factor compared to the expected values as reported in literature. Finally, the fracture surfaces confirmed the attainment of plane strain conditions as stipulated by the standard (see Fig. 10.10).

7 Conclusion

Fracture toughness tests were conducted on Ti6Al4V compact tension specimens made using WEDM. The pre-crack was introduced by WEDM using a 100 μm diameter uncoated brass wire. The obtained average toughness value using four specimens was 111.82 MPa.m$^{0.5}$ which was 4.4 % lower than that reported in the literature. The fracture surfaces showed conformance with plane strain requirements. It can therefore be concluded that the WEDM technique has a negligible effect in facture toughness of Ti6Al4V and can be safely used to pre-crack compact tension specimens.

References

1. Mohd Abbas, N., Solomon, D.G., Bahari, M.F.: A review on current research trends in in electrical discharge machining (EDM). Int. J. Mach. Tools Manuf. **47**, 1214–1228 (2007)
2. Zhixin, J., Jianhua, Z., Xing, A.: Ultrasonic vibration pulse electro discharge machining of holes in engineering ceramics. J. Mater. Process. Technol. **53**, 881–816 (1995)

3. Li, L., Guo, Y.B., Wei, X.T., Li, W.: Surface integrity characteristics in wire-EDM of inconel 718 at different discharge energy. Proc. CIRP **6**, 220–225 (2013)
4. Guo, Z.N., Lee, T.C., Yue, T.M., Lau, W.S.: Study on the machining mechanism of WEDM with ultrasonic vibration of wire. J. Mater. Process. Technol. **69**, 212–221 (1997)
5. Egashira, K., Masuzawa, T.: Microultrasonic machining by the application of work piece vibration. CIRP Ann. Manuf. Technol. **48**, 131–134 (1999)
6. Yu, Z.B., Jun, T., Masanori, K.: Dry eletrical discharge machining of cemented carbide. J. Mater. Process. Technol. **149**, 353–357 (2004)
7. Mechanical Design Forum Electro-discharge wire cutting (EDWC): Mechanical design forum. [Online]. Available http://www.mechanicaldesignforum.com/content.php?16-Electro-discharge-wire-cutting-(EDWC) (2009). Accessed 1 Nov 2014
8. Madyira, D.M., Akinlabi, E.T.: Effects of wire electrical discharge machining on fracture toughness of grade 5 titanium alloy. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, WCE 2014, London, pp 1393–1398. 2–4 July 2014
9. Liu, A.F.: Structural Assessment Methods. ASM International, West Hills (1998)
10. Anderson, T.L.: Fracture Mechanics: Fundamentals and Applications. CRC Press, London (1995)
11. Griffith, A.A.: The theory of rupture. In: *Proceedings of the First International Congress of Applied Mechanics*, 1924
12. Irwin, G.R.: Relation fo stresses near a crack to the crack extension force. In: Ninth International Congress of Applied Mechanics (1957)
13. Kanninen, M.F., Popelar, C.H.: Advanced Fracture Mechanics. Oxford University Press, New York (1985)
14. ASTM International Standard test method for measurement of fracture toughness, ASTM E1820-11, ASTM International (2011)
15. NDT Resource Center.: Fracture Toughness, [Online]. Available http://www.ndt-ed.org/EducationResources/CommunityCollege/Materials/Mechanical/FractureToughness.htm. Accessed 7 Nov 2014
16. RMI Company. Titanium Alloy Guide, [Online]. Available: http://www.rtiintl.com/Titanium/RTI-Titanium-Alloy-Guide.pdf (2000). Accessed 7 Nov 2014
17. Moiseyev, N.: Titanium: Russian Aircraft and Aerospace Applications. CRC Press, London (2005)
18. Madyira, D.M., Laubscher, R.F., Janse van Rensburg, N., Henning, P.F.J.: High speed machining induced residual stresses in Grade 5 titanium alloy. J. Mater. Design. Appl. **227**, 208–215 (2013)
19. Nourbakhsh, F., Rajurkar, K.P., Malshe, A.P., Cao, J.: Wire eletro-discharge machining of titanium alloy. First CIRP Conf. Biomanuf. Proc. CIRP **5**, 13–18 (2013)
20. Antar, M.T., Soo, S.L., Aspinwall, D.K., Jones, D., Perez, R.: Productivity and workpiece surface integrity when WEDM aerospace alloys using coated wires. Proc. Eng. **19**, 3–8 (2011)

Chapter 11
Accuracy of Available Methods to Evaluate Vierendeel Failure Load

Pattamad Panedpojaman and Worathep Sae-Long

Abstract To suggest cost-effective designs of steel beams with circular opening under Vierendeel failure, computational results of the available methods such as SCI P100, SCI P355, Chung et al.'s formula and Panedpojaman and Rongram (PPM)'s formula for evaluating the load carrying capacity based on Vierendeel failure is investigated. The failure load are investigated, in terms of the normalized moment-shear interactions for the steel beams and the maximum shear loads for the cellular beams, and compared with finite element analysis. A total of 114 non-linear finite element models of steel beams with singular openings and cellular beams are used in this study that covers various beam section sizes and opening parameters. Comparing with the FE results, the available methods conservatively provide the failure loads for steel beams with a single opening and cellular beams. It is also found that sizes of the steel sections less affect the FE interaction curve's shape. The interactions are slightly degraded for the large opening ratio. Comparing with the FE analysis, SCI P100, SCI P355, Chung et al.'s formula and PPM's formula are conservative to evaluate the interaction up to 40%, 25%, 20% and 10% respectively, under the high shear ratios. However, the methods provide the less conservative interaction and the Vierendeel failure load under the high moment ratios.

Keywords Cellular beam • Circular opening • Design • Moment-shear interaction • Nominal strength • SCI • Vierendel bending resistance

1 Introduction

The performance of cellular beams varies depending on the geometric details of the beams such as web or flange thickness, opening diameters, openings spacing and beam depths, etc. Due to opening effect, various failure modes are expected to happen. Vierendeel mechanism is the most common failure for perforated steel beams as shown in Fig. 11.1a. Vierendeel mechanism is caused the failure due

P. Panedpojaman (✉) • W. Sae-Long
Department of Civil Engineering, Faculty of Engineering, Prince of Songkla University, 15 Karnjanavanit Road, Hat Yai, Songkhla 90110, Thailand
e-mail: ppattamad@eng.psu.ac.th; s.warathep@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_11

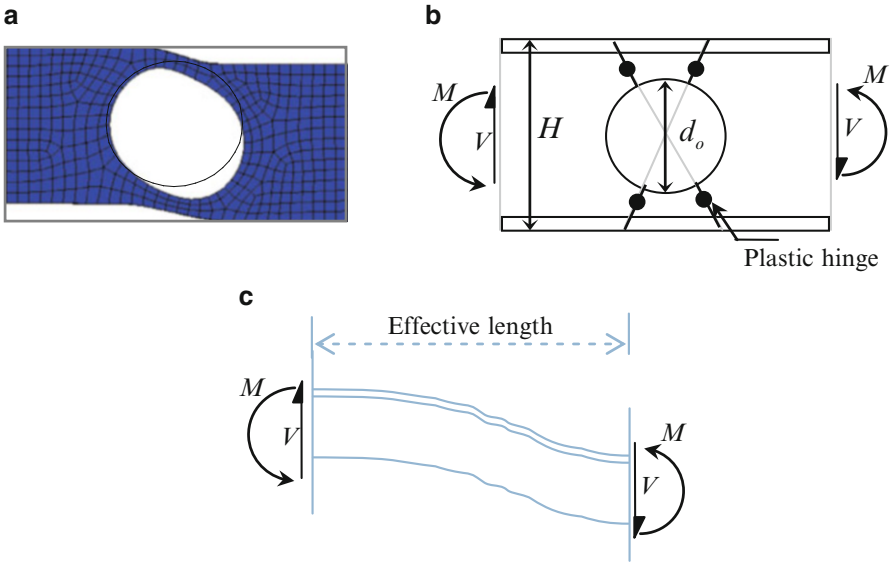


Fig. 11.1 Vierendeel mechanism of a steel beam with a circular opening: (**a**) Vierendeel failure, (**b**) plastic hinges and (**c**) Vierendeel bending and effective length

to the formation of four plastic hinges in the top and bottom tees as shown in Fig. 11.1b. The shear force, which transfers across the opening, causes some secondary moments (Vierendeel bending) in the top and bottom tee sections as shown in Fig. 11.1c. Interaction of the secondary moments with the global bending moments and their corresponding local axial force dominates the formation of plastic hinges in the tees [1].

To evaluate load carrying capacity of the beams based on Vierendeel mechanism, SCI P100 [2] proposed a linear interaction relationship between the local axial force and moment in the tees. The load carrying capacity depends on the location of the plastic hinge which considers at the critical section and angle to center line of the openings [3]. To simplify SCI P100's computation, SCI P355 [4] determines Vierendeel bending based on an assumed effective length of the opening [5, 6] as shown in Fig. 11.1c. To prevent the Vierendeel failure, Vierendeel bending resistance of the tees must be higher than Vierendeel bending.

A wide range of numerical studies [3, 7–10] were also used to investigate Vierendeel mechanism in terms of stress distributions and moment-shear interactions of the sections. Various empirical interaction formulae were proposed. The shear and moment capacity of the perforated sections are the key parameters in the numerical studies. Chung et al. [9] investigated the vertical shear area for computing the shear capacity of the sections and suggested that a percentage of the flange also contributed in carrying vertical shear force. However, the vertical shear area used

for predicting the load carrying capacity under Vierendeel failure in the literature [1–3, 7–9] differed from BS EN 1993-1-1 [11], normally used for recent designs. The shear area of BS EN 1993-1-1 is normally larger than the shear area used in the literatures. Different computations of the shear capacity can also affect evaluating Vierendeel failure load.

Level of conservatism of the available methods to evaluate the Vierendeel failure load is investigated to suggest cost-effective designs. SCI P100's computation [2] SCI P355's computation [4] and the empirical formula of Chung et al. [9] and Panedpojaman and Rongram [10] are scoped in this study. The computations are compared with finite element analysis of various steel beam models with circular openings. To minimize effects of the different computations on the shear capacity, all methods are derived based on the shear area according to BS EN 1993-1-1.

2 Beam Design Based on Vierendeel Failure

Consider a circular opening with diameter, d_o, formed in a steel beam with overall depth, H, as shown in Fig. 11.1b. The global bending moment and the global shear force at the centre of the web opening are M_{sd} and V_{sd}, respectively. For rolled steel beams with the compact section, the moment capacity of the perforated section $M_{o,rd}$ is given as follows:

$$M_{o,rd} = f_y W_{o,pl} \tag{11.1}$$

where

$$W_{o,pl} = W_{pl} - \frac{d_o^2 t_w}{4} \tag{11.2}$$

where W_{pl} and $W_{o,pl}$ are the plastic modulus of the imperforated section and the perforated section, t_w is the web thickness, and f_y is the yield stress of the steel. The shear capacity of the perforated section, $V_{o,rd}$, is given as follows:

$$V_{o,rd} = 0.577 f_y A_{vo} \tag{11.3}$$

where

$$A_{vo} = A_v - d_o t_w \tag{11.4}$$

where A_v and A_{vo} are the shear area of the imperforated section and the perforated section. This study applies the shear area according to BS EN 1993-1-1 as shown in Fig. 11.2.

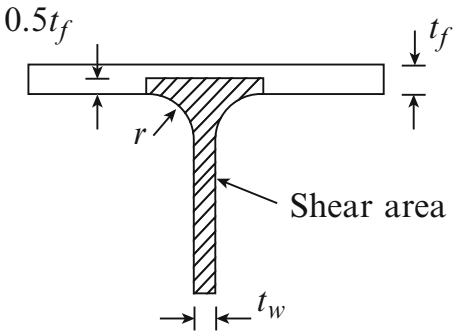


Fig. 11.2 Shear area of an imperforated tee according to BS EN 1993-1-1 [11]

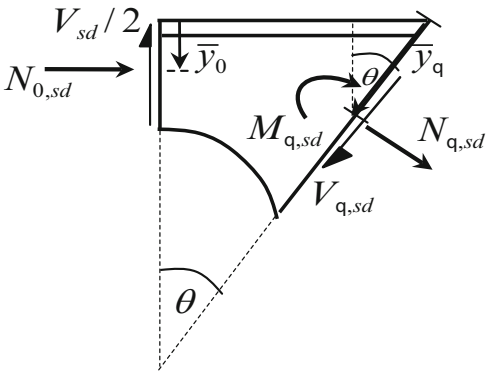


Fig. 11.3 Global actions on a perforated section at an angle θ from the center line

2.1 SCI P100

As a result of global actions on a perforated section at an angle θ from the center line as shown in Fig. 11.3, the sections are subjected to three co-existing actions [2]: axial force $N_{\theta,sd}$, shear force $V_{\theta,sd}$ and local bending moment $M_{\theta,sd}$. Limitation of the co-existing local axial and moment in the tee sections is evaluated according to a linear interaction formula [2, 12] as follows:

$$\frac{N_{\theta,sd}}{N_{\theta,rd}} + \frac{M_{\theta,sd}}{M_{\theta,rd}} \leq 1 \tag{11.5}$$

where

$$N_{\theta,sd} = N_{0,sd}\cos\theta - (V_{sd}/2)\sin\theta \tag{11.6}$$

$$M_{\theta,sd} = N_{0,sd}\left(\overline{y}_\theta\cos\theta - \overline{y}_0\right) + (V_{sd}/2)\left(H/2 - \overline{y}_\theta\cos\theta\right)\tan\theta \tag{11.7}$$

$$N_{0,sd} = \frac{M_{sd}}{d'} \tag{11.8}$$

$N_{\theta,rd}$ and $M_{\theta,rd}$ are the axial force capacity and the moment capacity of the tee section at an angle θ from the center line, V_{sd} and M_{sd} are the shear force and the moment subjected to the section at the center line, $N_{0,sd}$ is the axial force at the center line, $\overline{y}_0$ and $\overline{y}_\theta$ is the distance from the top edge to centroid of the tee section in the line of 0^o and θ, and d' is the distance between centroid of the top tee section and the bottom tee section in the center line.

In the presence of high shear force ($V_{sd}/V_{o,rd} > 0.5$), both the axial force and the moment capacities should be reduced in accordance with Eurocode 3: Part 1.1 [11] in terms of the reduced web thickness, t_w':

$$t_w' = t_w\sqrt{1-(V_{sd}/V_{o,rd})^2} \tag{11.9}$$

Evaluation of the load carrying capacity against the Viereendel action depends on the positions of the plastic hinges. The angle θ is found to vary from 0° for openings under pure moment to approximately 28° for openings under pure shear [9]. SCI P100 [2] conservatively recommended a typical θ value of 25° to evaluate the load carrying capacity.

2.2 SCI P355

As a simpler method, the circular opening is simplified as equivalent rectangular opening as shown in Fig. 11.4. However, the sections are still subjected to three co-existing actions. Vierendeel bending moment, $M_{v,sd}$, is directly considered due to transferring of the shear forces across an opening as

$$M_{v,sd} = V_{sd}l_{eff} \tag{11.10}$$

where l_{eff} is the effective length between the critical section.

$$l_{eff} = 0.45d_o \text{ for circular openings} \tag{11.11}$$

Residual Vierendeel bending resistance, $M_{v,rd}$, is considered as the critical location under effect of the axial force and the shear force as follows:

$$M_{v,rd} = M_{vn}\left[1-m^2\right] \tag{11.12}$$

$$M_{vn} = 4M_T \tag{11.13}$$

where m is the moment utilization ratio $M_{sd}/M_{o,rd}$, M_{vn} is Vierendeel bending resistance due to the 4 tee sections at opening corner as shown in Fig. 11.4, and

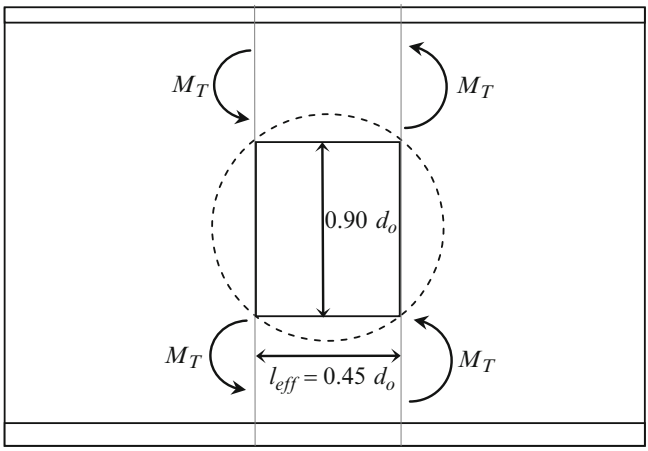


Fig. 11.4 Equivalent rectangular opening of circular opening for computing Vierendeel bending resistance

M_T is the plastic bending moment of the tee sections. In the presence of high shear force ($V_{sd}/V_{o,rd} > 0.5$), Vierendeel bending resistance is computed based on the reduced web thickness, t_w':

$$t_w' = t_w\left[1 - (2V_{sd}/V_{o,rd} - 1)^2\right] \tag{11.14}$$

Limitation of Vierendeel bending moment in the opening sections is evaluated according to $M_{v,sd} < M_{v,rd}$.

2.3 Empirical Formula

In order to provide a simple design of steel beams with circular web openings, researchers [9, 10] modified an empirical moment-shear interaction curve for solid rectangular plates to include an effect of Vierendeel mechanism. Chung et al. [9] proposed their empirical formulation as follows:

$$\left(\frac{v}{\overline{v}}\right)^2 + m^2 \leq 1 \text{ for } \overline{v} < 2/3 \tag{11.15}$$

$$\left(\frac{v - (\overline{v} - 2/3)}{2/3}\right)^2 + m^2 \leq 1 \text{ for } \overline{v} \geq 2/3 \tag{11.16}$$

where m is the moment utilization ratio $M_{sd}/M_{o,rd}$, v is the shear utilization ratio $V_{sd}/V_{o,rd}$ and $\overline{v}$ is the coupled shear capacity ratios. For a circular opening, $\overline{v}$ is equal

to 0.95 for d_o/H =0.50, 0.87 for d_o/H =0.67 and 0.8 for d_o/H =0.75. The shear utilization ratio, v, should not exceed the coupled shear capacity ratio $\bar{v}$.

Through an empirical study of the FE results, Panedpojaman and Rongram [10] found Chung's formulation provided conservative results to predict Vierendeel resistance. The formula for evaluating the moment-shear interaction curves was proposed as

$$(v/0.92)^2 + m^2 \leq 1 \tag{11.17}$$

for the openings with $d_o/H \leq 0.8$. Panedpojaman's formulation (PPM) is very simple and less conservative comparing with Chung's formulation. However, both formulations are mainly based on steel beams with single circular web openings, not cellular beams.

3 Finite Element Model

To evaluate accuracy of the available method for predicting the Vierendeel failure load, the computed maximum load of given cellular beams are compared with that of the FE results. Finite element (FE) models (ANSYS software [13]) of simply supported steel beam with a single opening and cellular beams are as shown in Fig. 11.5. Due to its symmetry, the FE model is involved only half the beam length with symmetric boundary conditions around the mid-span.

Stiffeners are provided at support and under each point of load introductions. Thickness of the stiffeners is same as the flange thickness. The FE models are simulated by the four-node shell element (SHELL181) for steel beams. The shell element is normally used to analyze thin to moderately thick steel structures under linear, large rotation, and/or large strain nonlinear applications. Based on a sensitivity study of FE element size on the load results, the element size of $H/15$ is chosen in the meshes. The structural analysis is controlled by time step analysis. Validation of the FE beam model can be found in [14, 15]. The Von Mises yield criterion with kinematic hardening is adopted to define the yield point of the steel. The steel grade S355 with yield stress of 355 MPa, the initial linear elastic modulus of 200 GPa and Poisson's ratio of 0.3 are employed throughout this FE investigation.

To investigate Vierendeel mechanism, the rolled steel beam sections, UB203 × 133 × 25, UB457 × 152 × 52, and UB914 × 305 × 201, with opening ratio d_o/H: 0.5 and 0.8 but varying the beam length are simulated. The models are with only a single opening to avoid an effect of adjacent openings as shown in Fig. 11.5a. A total of 60 non-linear finite element models of the beams with a single opening are used in this study. Cellular beams with various sizes and beam lengths as tabulated in Table 11.1 is used to study an accuracy of the available methods for the cellular beam cases. In Table 11.1, d is the original section depth before fabricating to cellular beams. A total of 54 non-linear finite element models of cellular beams as shown in Figs. 11.5b and 11.6 are used in this study. As normal uses in long span

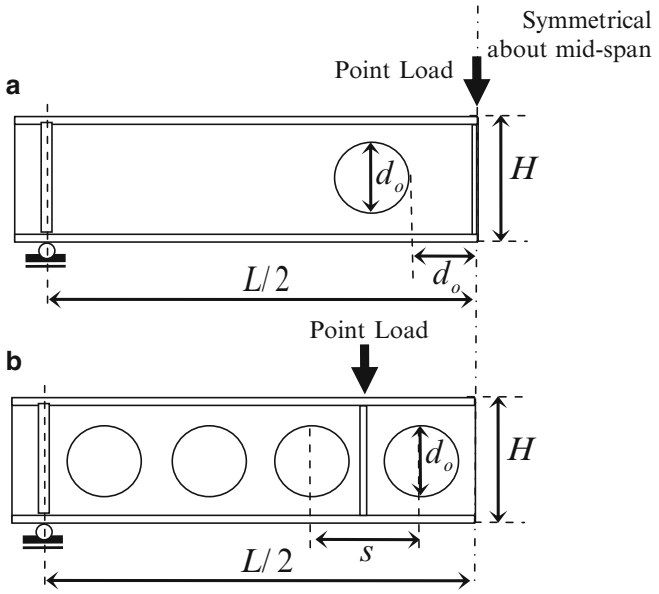


Fig. 11.5 Model for evaluating Vierendeel failure: (**a**) steel beam with a single opening and (**b**) cellular beams

Table 11.1 Investigated cellular beams (half span length)

Case	Rolled beam data					$\frac{d_o}{d}$	$\frac{s}{d_o}$	Total opening no.
	d (mm)	b (mm)	t_w (mm)	t_f (mm)	r (mm)			
1	900	300	18	34	28	0.8, 1.0, 1.2	1.3, 1.5, 1.7	10,20
2	800	300	14	26	28			
3	700	300	13	24	28			

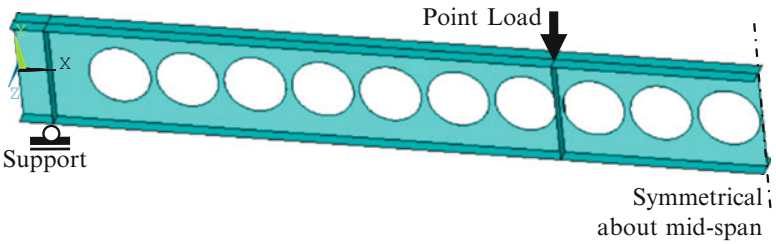


Fig. 11.6 Finite element model of cellular beams

cases, the cellular beams used in the investigation are under the high moment ratio about between 0.65 and 0.85 at the critical opening. As such high moment ratios, the shear failure and the web post buckling can be avoided. Note that the critical openings located near the point loads.

4 Analytical Result and Comparison

Stress distributions and Vierendeel failure behavior of all models are similar [10]. The models of UB457 × 152 × 52 are represented to describe their Von Mises stress distribution as shown in Fig. 11.7. Under a combination of moment and shear force, first yield appears in edge opening at an angle from the center line of about 0° to 28° of the low moment side. The angle of 0° is for the pure moment case whereas about 28° is for the pure shear case. Under combination of the shear and moment force, the beam continues to carry additional loading until the yield large propagates to cause extensive yielding in the tee sections. Under the yield propagations, Vierendeel failure occurs as shown in Fig. 11.1a. Variation of the angle at the first yield agrees well with the finding of [3].

This study also investigates Vierendeel's effects on overall behavior of the beams with a single opening in terms of the moment-shear interaction curve by varying the position of the opening and the beam length. The moment-shear interaction

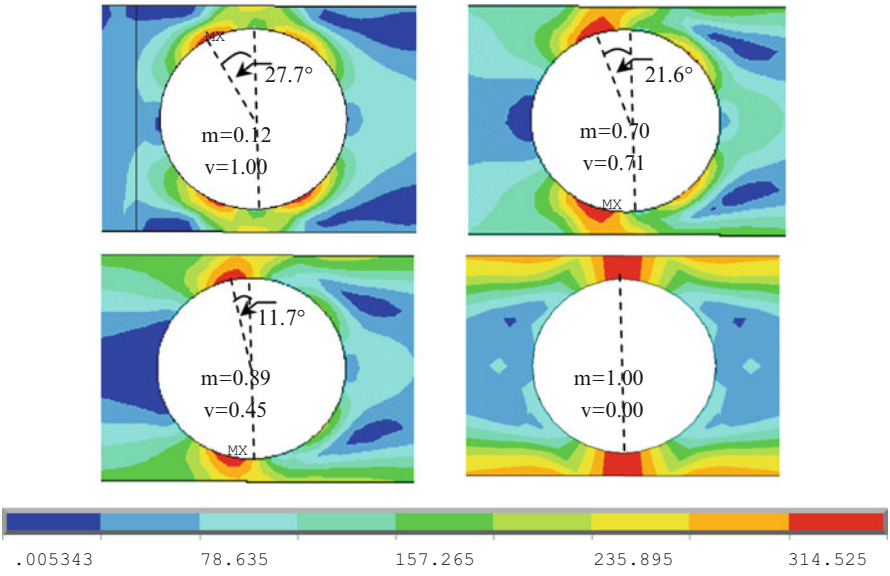


Fig. 11.7 Von Mises Stress distributions of cellular beams UB457 × 152 × 52 with do/H = 0.8

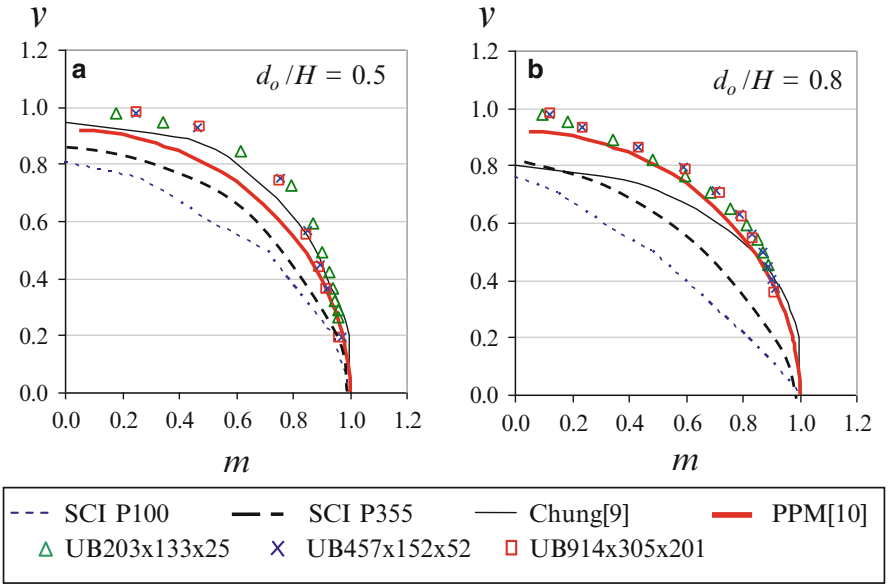


Fig. 11.8 Comparison of FE moment-shear interaction with: (**a**) the available methods for $d_o/H = 0.5$, (**b**) the available methods for $d_o/H = 0.8$

is represented in terms of a non-dimensional interaction curve relating the shear utilization ratio v and the moment utilization ratio m. The interactions of the investigated beams based various methods are plotted in Fig. 11.8a for d_o/H of 0.5 and in Fig. 11.8b for d_o/H of 0.8. The computational shear and moment capacity used in the computational curves are derived from the theory as in (1) and (3). However, the FE curves derive from the shear and moment capacity of the FE analysis based on the FE models [10].

The interaction curves are clearly non-linear and very similar to each other in shape for all beam sizes and all opening sizes. However, the interaction curves of the larger opening are slightly more critical comparing with the smaller opening. SCI's interactions are similar for various steel sections with a given opening ratio. Therefore, SCI's interactions are plotted as an average interaction. Comparing with the FE curves, SCI P100 method, SCI P355 method, Chung et al.'s formula and PPM formula provide conservative results up to 40%, 25%, 20% and 10%, respectively. However, the conservative results are lessened under the high moment ratio.

For cellular beams subjected to the point loads, the critical opening of Vierendeel failure is near the point loads at the support side. At that location, the openings are under high combination of the shear and moment force. The failure behavior of cellular beams is same as the beams with a single opening, as described above. Comparisons of the maximum shear loads due to Vierendeel failures between the FE models and the available methods are shown in Fig. 11.9. Comparing with the

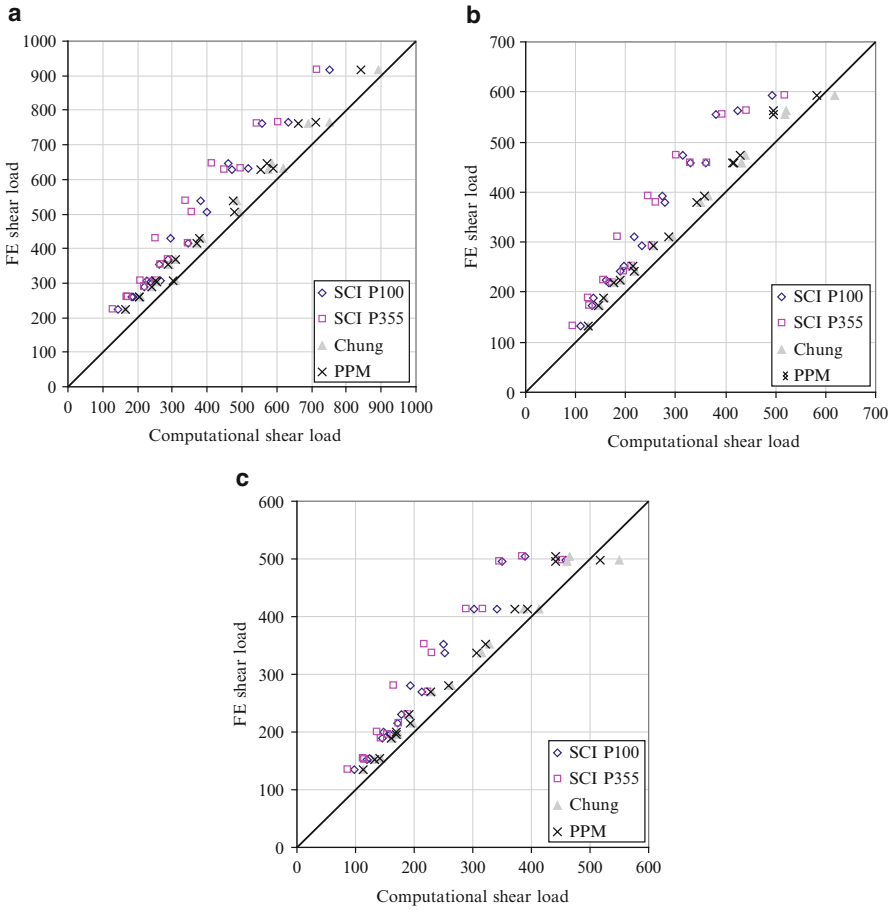


Fig. 11.9 Comparison of the maximum shear loads due to Vierendeel failures between the FE models and the available methods: (**a**) Case 1, (**b**) Case 2 and (**c**) Case 3

FE loads, SCI P100 method, SCI P355 method, Chung et al.'s formula and PPM formula provide conservative results with an average of 27%, 24%, 10% and 12%, respectively.

5 Conclusion and Discussion

Accuracy of the available methods, SCI P100, SCI P355, Chung et al.'s formula and Panedpojaman and Rongram's formula to evaluate the Vierendeel failure load are investigated. Computations of all methods are derived based on the shear area according to BS EN 1993-1-1. The computations are compared with the finite

element analysis of various steel beam models, the beams with a single opening and the cellular beams. This study investigates Vierendeel's effect on overall behavior of the perforated beams in terms of the normalized moment-shear interaction curve and the maximum shear loads.

Through the analytical results, the stress distributions and Vierendeel failure behavior of all models are similar. First yield appears in edge opening at an angle from the center line of about 0° for the pure moment cases to 28° for the approximately pure shear cases. The beam continues to carry additional loading until the yield large propagates to cause extensive yielding in the tee sections.

The FE moment-shear interactions of various steel sections with the same opening ratios are similarly in shape. However, for the large opening ratio, the interactions are slightly degraded. Comparing with the FE interactions, the current design methods are significantly conservative to evaluate the interaction, especially for SCI P100's method. Even though, SCI P100's method is a detailed computation, accuracy of the method to evaluate the interaction is not improved. However, to evaluate Vierendeel failure load, SCI P355's method provides the less conservative results with less computational steps.

Since the interactions of the perforated sections are similar, an empirical formula is reasonably suitable to evaluate the interaction. However, Chung et al.'s formula is very conservative under the high shear ratio but suitable under the low shear ratio. Panedpojaman and Rongram's formula provides the least conservative results in average for all shear ratios.

The cellular beam cases investigated in this study are under the moderate moment ratio, 0.65–0.85, to avoid other failure modes. Vierendeel failure behavior of cellular beams is same as the beams with a single opening. Comparison of the maximum shear loads due to Vierendeel failures with the FE loads, SCI P100 and SCI P355 method is also very conservative. Chung et al.'s formula and Panedpojaman and Rongram's formula provide less conservative results. Even though the empirical methods are developed based on steel beams with single opening, the methods can also be used to evaluate Vierendeel failure loads for cellular beams.

The available methods are investigated through a limit number of the beam models. Different beam sections and other related parameters should be further investigated to assure the accuracy and limitation of evaluating Vierendeel failure loads.

References

1. Sherbourne, A.N., Oostrom, J.V.: Plastic analysis of castellated beams – interaction of moment, shear and axial force. Comput. Struct. **2**, 79–109 (1972)
2. Ward, J.K.: Design of Composite and Non-composite Cellular Beams, The Steel Construction Institute Publication 100 (1990)
3. Chung, K.F., Liu, T.C.H., Ko, A.C.H.: Investigation on Vierendeel mechanism in steel beams with circular web openings. J. Construct. Steel Res. **57**, 467–490 (2001)

4. Ward, J.K.: Design of Composite and Non-composite Cellular Beams, The Steel Construction Institute Publication 355 (1990)
5. Redwood, R.G.: Design of Beams with Web Holes. Canadian Steel Industries Construction Council, Willowdale, Ontario, Canada (1973)
6. Lawson, R.M., Lim, J., Hicks, S.J., Simms, W.I.: Design of composite asymmetric cellular beams and beams with large web openings. J. Construct. Steel Res. **62**(6), 614–629 (2006)
7. Liu, T.C.H., Chung, K.F.: Practical design of universal steel beams with single web openings of different shapes. In: Proceedings of the Second European Conference on Steel Structures-Eurosteel 99, Prague (1999)
8. Liu, T.C.H., Chung, K.F.: Steel beams with large web openings of various shapes and sizes: finite element investigation. J. Construct. Steel Res. **59**, 1159–1176 (2003)
9. Chung, K.F., Liu, T.C.H., Ko, A.C.H.: Steel beams with large web openings of various shapes and sizes: an empirical design method using a generalised moment-shear interaction curve. J. Construct. Steel Res. **59**, 1177–1200 (2003)
10. Panedpojaman, P., Rongram, T.: Design equations for Vierendeel bending of steel beams with circular web openings. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, WCE 2014, pp. 1493–1498. London 2–4 July 2014
11. Eurocode 3.: Design of Steel Structures Part 1–1: General rules and rules for buildings. British Standards Institution, BS EN 1993-1-1, London (2005)
12. Eurocode 3.: Design of steel structures: Part 1–1: General rules and rules for buildings, 1992, and Amendment A2 of Eurocode 3: Annex N 'Openings in webs', British Standards Institution, ENV 1993-1-3, London (1998)
13. ANSYS release 11.0 documentation, Ansys Inc (2007)
14. Panedpojaman, P., Thepchatri, T.: Finite element investigation on deflection of cellular beams with various configurations. Int. J. Steel Struct. **13**(3), 487–494 (2013)
15. Panedpojaman, P., Thepchatri, T., Limkatanyu, S.: Novel design equations for shear strength of local web-post buckling in cellular beams, *Thin-walled Structures*, **76**, 92–104 (2014)

Chapter 12
Influence of Tibial Translation on Estimates of Patellar Tendon Force During Knee Flexion

Ahmed Imran

Abstract The mechanics of the knee was analyzed in the sagittal plane to study effects on patellar tendon force as influenced by placement of external flexing loads on the tibia and by translation of the tibial bone 7 mm anterior to the femoral bone during the flexion range 0–120°. Anatomical parameters and measurements for orientations and moment arms of the patellar tendon during the flexion range were estimated from experiments on cadaver knees available in the literature.

The analysis suggests that the force in the patellar tendon varies directly with the distance of external flexing load placed distal to the tibial surface. This effect is uniform throughout the joint flexion range. Further, 7 mm anterior translation of the tibia relative to the femur resulted in significantly reduced anterior component of net shear force for all flexion positions and all placements of loads distal to the joint line. This has relevance to ACL-deficient knees where large tibial translations may be necessary to compensate for the deficiency (ACL is anterior cruciate ligament).

Two critical factors that may require special attention to protect the ACL during rehabilitation exercises are the flexion angle and the position of flexing load below the tibial surface.

Keywords Knee biomechanics • Knee joint forces • Patellar tendon forces • Quadriceps contraction • Rehabilitation of anterior cruciate ligament • Tibial translation at the knee

1 Introduction

The patellar tendon transmits forces of the quadriceps muscles through the knee extensor mechanism to the tibia [1]. The tendon force provides extending moment at the knee as well as force components that can translate the lower bone (or tibia) relative to the upper bone (or femur) [1]. This relative translation of the bones is mainly in the sagittal plane and is resisted by the cruciate ligaments which tend to stretch in order to restrict the bony movements. Overstretching of the ligament

A. Imran (✉)
Ajman University of Science & Technology, Ajman, PO Box 346, UAE
e-mail: ai_imran@yahoo.com; ajac.ai_imran@ajman.ac.ae

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_12

fibers can lead to injuries. Other effects of the relative translation of bones is that the patellar tendon as well as the ligament fibers get reoriented and the contact position between the bones changes. These later effects can possibly result in changed moment arms of the internal load-bearing structures of the joint.

Several investigators have estimated moment arms and orientations of the patellar tendon at the knee using theoretical or experimental approach for the intact or replaced knee [2–10]. During the knee flexion, moment arm available to the patellar tendon force varies due to changing location and orientation of the tendon as well as due to changing location of the point from which the moment arm is measured [2–10]. In general, the moment arm of the tendon normally remains small and varies non-linearly from more than 4 cm in extension to less than 4 cm in high flexion [2]. The tendon orients anteriorly in extension and posteriorly in high flexion, with an overall variation of nearly 20° during 0–120° flexion [2].

In early–to–mid flexion range, the patellar tendon force has a component parallel to the tibial surface directed anteriorly, which can pull the tibia anterior to the femur, thus, stretching the anterior cruciate ligament (ACL) [1]. In mid–to–high flexion range, the patellar tendon force has a component parallel to the tibial surface directed posteriorly, which can pull the tibia posterior to the femur, thus, stretching the posterior cruciate ligament (PCL) [1]. These anterior or posterior force components of the tendon can be large in certain strenuous activities, like in playing football, and could lead to serious consequences for the subject as the knee ligaments, particularly the ACL, are injured frequently during situations involving large tibial translations [1, 11, 12]. Further, studies suggest that the positions and orientations of the external loads acting on the lower leg can significantly influence forces in the internal structures of the joint [3, 4].

The purpose of this study is to analyze the patellar tendon force as influenced by the relative translations of the tibial and femoral bones in the sagittal plane as well as by the placement of external flexing loads on the tibia during flexion of the knee.

2 Methods

Mechanics of the knee was analyzed in the sagittal plane during 0–120° flexion of the joint. With reference to Fig. 12.1, mechanical equilibrium of the joint was considered due to four types of forces, namely, a force in the patellar tendon (P), a ligament force (L), a tibio-femoral joint contact force (C) applied by the femur normal to the tibial surface and a flexing load (R) applied externally on the tibia 'M_R' cm below the joint line (or tibial articular surface) and acting parallel to the line.

The anatomical posterior direction was defined along the positive x-axis. Also, the tibial surface was taken to be flat and parallel to the x-axis.

Equilibrium of moment was given by (12.1); equilibrium of forces parallel to the tibial surface was given by (12.2) and equilibrium of forces perpendicular to the tibial surface was given by (12.3).

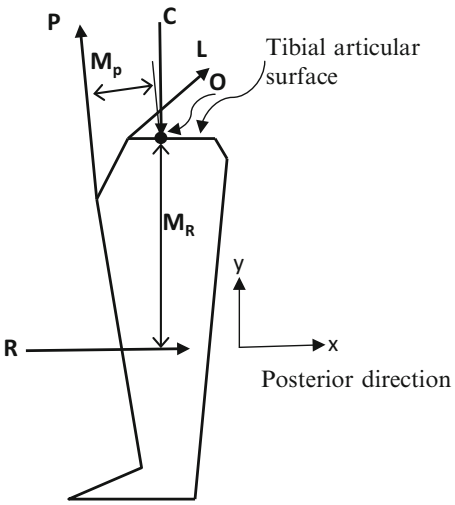


Fig. 12.1 Flexing moment due to the external load (*R*) on the tibia is balanced by an extending moment provided by the patellar tendon force (*P*). The external force is balanced by the internal forces P, L and C

In (12.1), the rotational contribution of the ligament force, due to either the ACL or the PCL, was ignored because the moment arms available to the ligaments through most of the flexion range are relatively much smaller, 1 cm or less [2]. Also, the frictional effect between the bones in a natural intact joint, being negligible [13], was ignored.

$$P \times M_P + R \times M_R = 0 \tag{12.1}$$

$$P \times \cos(\theta_P) + R \times \cos(\theta_R) + L \times \cos(\theta_L) = 0 \tag{12.2}$$

$$P \times \sin(\theta_P) + L \times \sin(\theta_L) + C \times \sin(\theta_C) = 0 \tag{12.3}$$

Where,

R, P, L and C are the forces as defined earlier in this section.

θ is the angle with positive x-axis for a force given by its respective subscript. By definition, $\theta_R = 0°$.

M is the moment arm from the tibio-femoral contact point (point O in Fig. 12.1) for a force given by its respective subscript. This definition is similar to that adopted for cadaver knee study in the literature [2].

Data used in this study was based on the experimental measurements of Herzog and Read [2]. Figures 12.2 and 12.3, respectively, give estimated values for moment arms and orientations of the patellar tendon extracted from the experimental study in Ref. [2] that was conducted on six cadaver knees and recorded at several positions during the knee flexion. In the experiment, as also in the present study, the orientation of the patellar tendon was taken with the posterior direction and the moment arm was taken about the tibio-femoral contact point. The estimates were

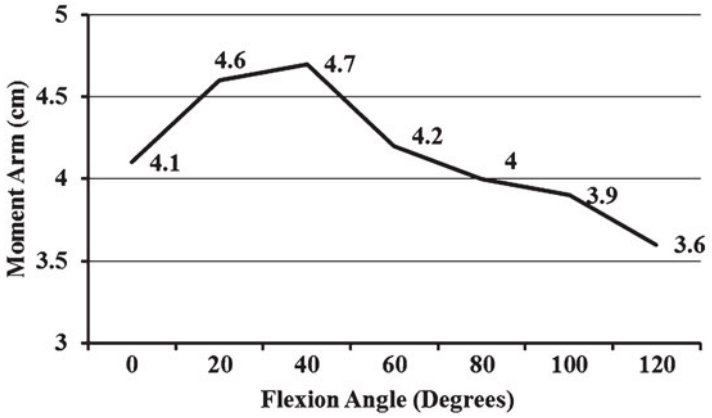


Fig. 12.2 Moment arms of the patellar tendon during flexion of the knee as estimated from Ref. [2]

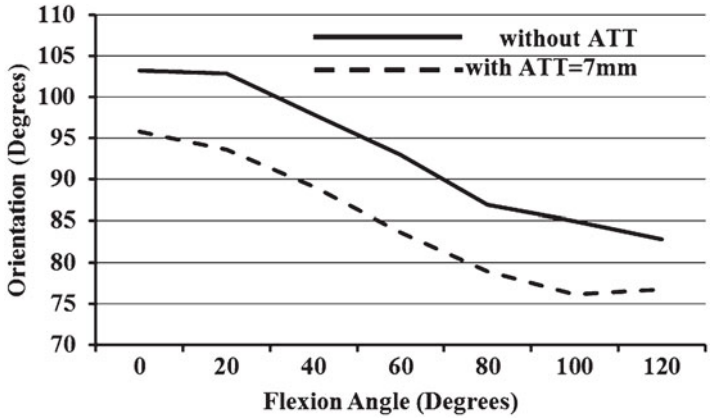


Fig. 12.3 Orientations of the patellar tendon during flexion of the knee without ATT and with 7 mm ATT as estimated from Ref. [2]

also made for orientations of the patellar tendon measured after anterior translation of the tibia relative to the femur (ATT).

Using the estimated moment arms from Fig. 12.2, values of patellar tendon force per unit external load (P/R) were calculated based on (12.1).

In (12.2), the ligament force L would arise due to stretching of the ACL if the tibia translated anteriorly (negative x-direction in Fig. 12.1) or due to the PCL if the tibia translated posteriorly. The net tangential force parallel to the tibial surface (T) was responsible for the tibial translation. T was calculated per unit R from Eq. (12.4).

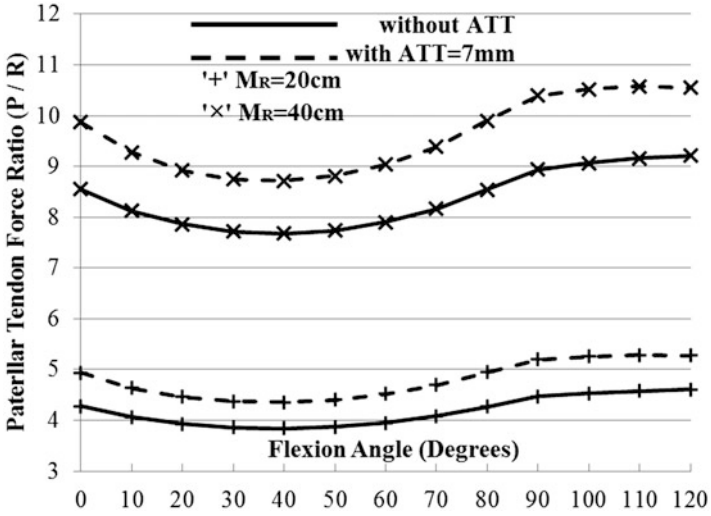


Fig. 12.4 The patellar tendon force per unit of the external load (P/R) plotted during 0–120° flexion with and without tibial translation given for values of $M_R = 20$ and 40 cm

$$T = [P \times \cos(\theta_P) + R \times \cos((\theta_R)\,] \tag{12.4}$$

The values for θ_P were obtained from Fig. 12.3, while $\theta_R = 0°$ was taken by definition. The calculations described above were performed during 0–120° flexion and repeated with different positions of R below the tibial surface taking $M_R = 20$ and 40 cm.

As mentioned earlier, relative translations of the bones have the potential to relocate and reorient the forces in the tendon and ligaments, the calculations were also performed with 7 mm ATT in order to analyze the possible effects on the tangential forces parallel to the tibial surface that might require contributions from the ACL.

3 Results

Figure 12.4 shows the P/R ratio calculated during flexion with $M_R = 20$ and 40 cm. The calculations were performed without ATT and with ATT = 7 mm.

T/R ratio was calculated during flexion with $M_R = 20$ and 40 cm. Figure 12.5 gives T/R ratio without tibial translation. Figure 12.6 gives T/R ratio with 7 mm anterior tibial translation.

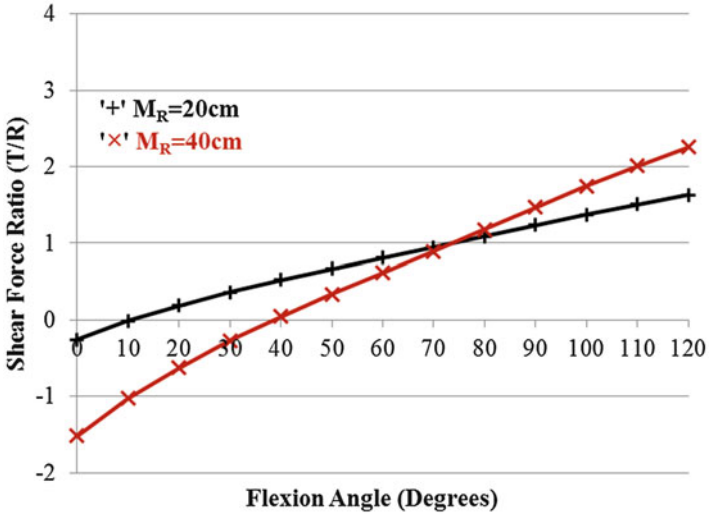


Fig. 12.5 The net tangential force arising due to the patellar tendon force and the external load was calculated as T/R during 0–120° flexion without tibial translation given for values of $M_R = 20$ and 40 cm

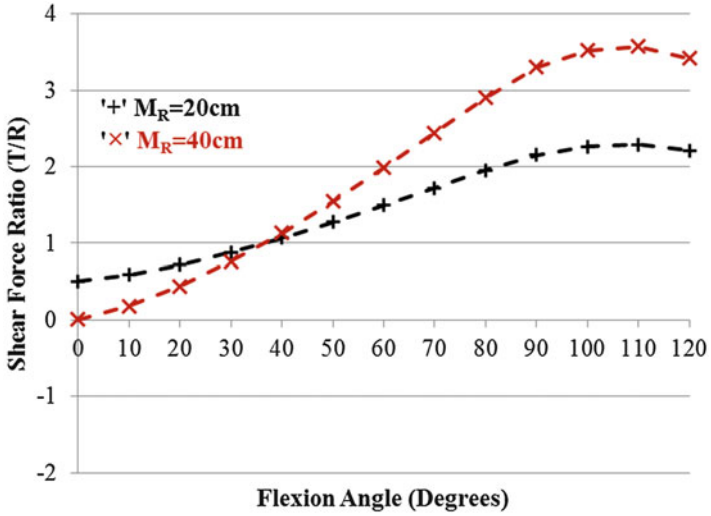


Fig. 12.6 The net tangential force arising due to the patellar tendon force and the external load was calculated as T/R during 0–120° flexion with 7 mm anterior tibial translation given for values of $M_R = 20$ and 40 cm

4 Analysis

4.1 The Role of the Patellar Tendon in Resisting a Flexing Moment due to an External Load Placed Distally on the Tibia

With reference to Fig. 12.2, the moment arm available to the patellar tendon was around 4 cm near extension, increased with flexion angle until around 40° to an average value of 4.7 cm and then decreased in higher flexion to remain below 4 cm after 80° flexion. In comparison, the moment arm of the external load could be around 40 cm or more during normal activities. As a consequence, the force P would be much larger than R and it would vary with the position of R below the tibial surface as suggested by (12.1) and demonstrated by the calculations in Fig. 12.4.

With increase in M_R, P/R ratio in Fig. 12.4 increased uniformly throughout the flexion range. The effect of ATT was to reduce the angle of P at all flexion positions of the joint. The reduction in angle was about 8° near full extension and about 6° at 120° flexion, varying nearly linearly. The P/R ratio increased due to ATT. The increase was more when the load was placed 40 cm compared to that when the load was placed 20 cm distally.

During quadriceps strengthening exercises, distal placement of R may be required for the muscle exercises to be effective. However, such distal placements could have detrimental effects on the ACL loading as suggested by the analysis given below.

4.2 Influence of Anterior Tibial Translation on Patellar Tendon Force

With reference to Fig. 12.5, the T/R ratio was negative in early flexion for both positions of R below the tibial surface. The T/R ratio became positive with increasing flexion. With reference to Fig. 12.6, 7 mm ATT resulted in positive values of T/R for all flexion positions, suggesting that no anteriorly directed shear force acted on the tibia. Translations of this magnitude or even more may occur in knees with ACL-deficiency providing balancing positions for the internal forces when sufficient ligament resistance is not available. This occurs at low flexion angles when the patellar tendon is oriented anteriorly in the unloaded state of the joint.

For external loads placed far distal to the tibial surface, the flexion range with negative values for T/R was increased in the absence of tibial translation. As shown by Fig. 12.5, for $M_R = 20$ cm, T/R was negative for the flexion range 0–10°, while for $M_R = 40$ cm, the ratio was negative for flexion less than 40°. This observation suggests that the quadriceps exercises performed with the external load placed far distal to the tibial surface could stretch the ACL up to nearly the mid flexion range.

Further, since most of the normal activities, like walking, jogging, stair climbing, involve low–to–mid flexion range at the most [14–16], anterior translation of the tibia and, thus, loading of the ACL would be expected during such activities.

5 Conclusions

The mechanics of the knee was analyzed in the sagittal plane to study the patellar tendon force during flexion. The influence of distal positions of the external flexing loads on the tibia and the influence of anterior translation of the tibia relative to the femur were analyzed in several flexion positions of the knee. The analysis suggests that the force in the patellar tendon varies directly with the distance of the external flexing load placed distal to the tibial surface. This effect is uniform throughout the joint flexion range. Further, anterior translation of the tibia relative to the femur resulted in significantly reduced anterior component of the net shear force, which has relevance to ACL deficient knees where large tibial translations may be necessary to partially compensate for the ACL-deficiency.

Two critical factors that may require special attention to protect the ACL during rehabilitation exercises are the flexion angle and the position of the flexing load below the tibial surface.

Acknowledgment This work was supported by research fund provided by the College of Engineering at Ajman University of Science & Technology, Ajman, UAE.

References

1. Masouros, S.D., Bull, A., Amis, A.: Biomechanics of the knee joint. Orthop. Trauma **24**(2), 84–91 (2010)
2. Herzog, W., Read, L.J.: Lines of action and moment arms of the major force-carrying structures crossing the human knee joint. J. Anat. **182**, 213–230 (1993)
3. Imran, A.: Analyzing the patellar tendon force during quadriceps muscle exercise. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 1313–1316. London, 2–4 July 2014
4. Imran, A.: Analyzing the role of patellar tendon force during flexion of the knee. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2013, WCE 2013, pp. 1929–1932. London, 3–5 July 2013
5. Imran, A., Huss, R., Holstein, H., O'Connor, J.: The variation in the orientations and moment arms of the knee extensor and flexor muscle tendons with increasing muscle force: a mathematical analysis. J. Engg. Med. **214**(H), 277–286 (2000)
6. Krevolin, J.L., Pandy, M.G., Pearce, J.C.: Moment arm of the patellar tendon in the human knee. J. Biomech. **37**(5), 785–788 (2004)
7. O'Brien, D.T., Reeves, N.D., Baltzopoulos, V., Jones, D.A., Maganaris, C.N.: Moment arms of the knee extensor mechanism in children and adults. J. Anat. **215**, 198–205 (2009)
8. Stagni, R., Fantozzi, S., Catani, F., Leardini, A.: Can patellar tendon angle reveal sagittal kinematics in total knee arthroplasty? Knee Surg. Sports Traumatol. Arthrosc. **18**(7), 949–954 (2010)

9. Sellis, E.K., Baltzopoulosb, V.: In vivo determination of the patella tendon and hamstrings moment arms in adult males using videofluoroscopy during submaximal knee extension and flexion. Clin. Biomech. **14**, 118–124 (1999)
10. Tsaopoulos, D.E., Baltzopoulos, V., Richards, P.J., Maganaris, C.N.: In vivo changes in the human patellar tendon moment arm length with different modes and intensities of muscle contraction. J. Biomech. **40**(15), 3325–3332 (2007)
11. Mommersteeg, T., Blankevoort, L., Huiskes, R., Kooloos, J., Kauer, J.: Characterization of the mechanical behavior of human knee ligaments: a numerical-experimental approach. J. Biomech. **29**(2), 151–160 (1996)
12. Klass, D., Toms, A.P., Greenwood, R., Hopgood, P.: MR imaging of acute anterior cruciate ligament injuries. Knee **14**, 339–347 (2007)
13. Unsworth, A., Dowson, D., Wright, V.: Some new evidence on human joint lubrication. Ann. Rheum. Dis. **34**, 277–285 (1975)
14. Deluzio, K.J., Wyss, U.P., Zee, P., Costigan, P.A., Sorbie, C.: Principal component models of knee kinematics and kinetics: normal vs. pathological gait patterns. Hum. Mov. Sci. **16**, 201–217 (1997)
15. Collins, J.J., O'Connor, J.: Muscle-ligament interactions at the knee during walking. Proc. Inst. Mech. Eng. H J. Eng. Med. **205**, 11–18 (1991)
16. Nakhaee, K., Farahmand, F., Salarieh, H.: Studying the effect of kinematical pattern on the mechanical performance of paraplegic gait with reciprocating orthosis. Proc. Inst. Mech. Eng. H J. Eng. Med. **226**(8), 600–611 (2012)

Chapter 13
Numerical Simulation of the Aerodynamic Loads on Trees During Storms

Edward Chern Jinn Gan and Salim Mohamed Salim

Abstract Fluid-structure interactions for a single tree and a pair of trees with varying spacing subjected to *gentle breeze* and *storm* wind conditions were evaluated using Computational Fluid Dynamics (CFD). The generated velocity and pressure fields are then analysed using Finite Element Analysis (FEA) to determine the likelihood of tree damage due to the aerodynamic loads induced by the two wind conditions. It is observed that the pressure difference between the windward and leeward sides of the trees are much larger during the *storm* condition resulting in greater mechanical stresses and deformation magnitudes experienced by the tree trunks. Increasing the spacing between neighbouring trees resulted in larger aerodynamic loads on the sheltered trees downstream.

Keywords Atmospheric boundary layer flows • CFD • FEA • Fluid-structure interaction • Tree spacing • Turbulence • Wind loads • Windbreak protection • Windthrow • Uprooting

1 Introduction

The impacts of tree windthrow are extensive as the damage it causes results in injuries and fatalities, significant economic losses due to reduction of timber value, destruction of public amenities, and may upset the balance of ecosystems in forests [1–3]. Windthrow is likely to occur during storms [4], and is the motivation of the present study to determine the key factors that could influence windthrow (see Fig. 13.1 for example of windthrow). Previous investigations have mainly focused

E.C.J. Gan (✉)
Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia
e-mail: edwardgancj@gmail.com

S.M. Salim
School of Engineering and Physical Sciences, Heriot-Watt University Malaysia, Putrajaya 62200, Malaysia
e-mail: m.salim@hw.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_13

Fig. 13.1 Uprooting of several roadside trees at Kelana Jaya, Malaysia after a severe tropical storm

on wind-tunnel (WT) experiments taking into account mechanical stability and other factors such as geographical, topographical, seasonal, and meteorological influences [5, 6].

Experiments have also been carried out on the aerodynamic loading of trees with varying tree configurations such as Novak et al. [7] and Gardiner et al. [8]. These studies evaluated the differences in turbulence statistics and wind loading due to the variations in tree density (i.e. spacing between adjacent trees) by comparing WT models with field measurement results. They observed that wider spaced trees were more likely to overturn as a result of increased aerodynamic loads.

A number of numerical simulations have been employed in the past to analyse the interaction between the atmospheric boundary layer flow and trees, with various modelling techniques explored. For example, Gross [9] and Tiwary et al. [10] investigated the airflow around a single tree and through a vegetative barrier that was represented by a row of bushes, respectively. These studies expressed the obstacles as porous media with the addition of drag coefficients and speed resistance factors. Others like Salim et al. [11, 12], Buccolieri et al. [13], and Gromke et al. [14] modelled trees as porous media and accounted for their internal structures through the use of WT acquired pressure loss coefficients, λ (m^{-1}). They investigated the influence of trees on the ambient airflow and consequently the pollutant dispersion in urban street canyons.

Recently, the authors [15] investigated the aerodynamic loads and resulting mechanical effects of a single tree subjected to different wind conditions by combining CFD and FEA analysis of the fluid-structure interaction. In the present work, the authors aim to expand the findings made previously in addition to establishing

Table 13.1 Values used in present study

Parameter	Value
Inlet velocity (ms^{-1})	4.7 (*Gentle breeze*), 24 (*Storm*)[a]
Tree species	*Acacia mangium*[b]
Tree spacing, *x/H*	0 (single tree), 1.5, 3

[a]*Storm* value is obtained from the extreme wind speed analysis for 100 year return period in Ipoh, Malaysia [16]
[b]*Acacia mangium* is a tree species commonly found at the roadsides of Malaysia [17]

the effects of tree spacing on windthrow of the sheltered tree (i.e. the downstream tree). Velocity and pressure fields around a single tree and between two trees with varying spacing, $x/H = 1.5$ and $x/H = 3$ are numerically predicted using standard k-ε turbulence closure scheme based on the steady-state Reynolds-averaged Navier-Stokes (RANS) equations. Typical Malaysian meteorological conditions [16] and roadside tree species [17] are taken into account where a baseline (*gentle breeze*) wind speed of 4.7 ms^{-1} and extreme wind speed of 24 ms^{-1} (for a severe tropical *storm*) are investigated. These are summarised in Table 13.1.

The imposed aerodynamic loads experienced by the trees, which manifests as the wind flow pressure field are then extracted from ANSYS FLUENT and imported into ANSYS Mechanical to perform static structural analysis based on Finite Element Analysis (FEA). This enables the prediction of the likelihood of windthrow to occur by considering both the prevailing wind conditions and structural properties of the trees.

2 Methodology

2.1 Computational Domain and Boundary Conditions

Numerical flow simulations are performed in ANSYS FLUENT, employing the simulation techniques and best practices of Salim et al. [11, 18, 19]. Validation and selection of a suitable numerical model are initially achieved using the simulations by Gross [9] as benchmark.

In order to mimic a typical urban atmospheric boundary layer flow, the inlet wind is assumed to follow the power law profile

$$u(z) = U_H \left(\frac{Z}{H} \right)^{\alpha} \tag{13.1}$$

where u is the vertical velocity profile, z is the vertical distance and the profile exponent, $\alpha = 0.3$ while u_H is the free-stream velocity (= 4.7 and 24 ms^{-1}) at reference height, $H = 0.12$ m for *gentle breeze* and *storm* inlet conditions,

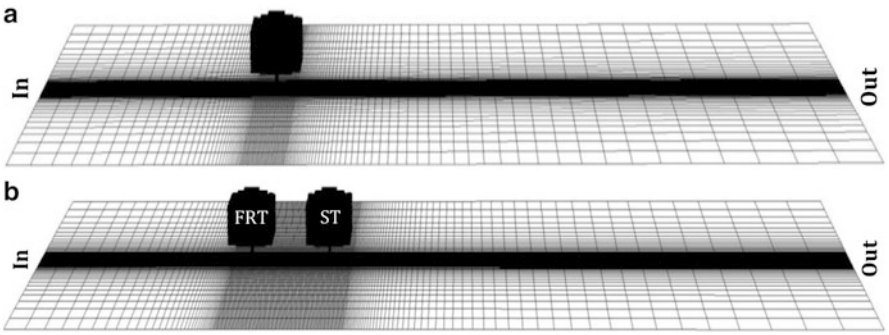


Fig. 13.2 Computational grids used in flow simulation of present study for (**a**) a single tree and (**b**) two trees with spacing, $x/H = 1.5$ and 3 (*FRT* front row tree, *ST* sheltered tree)

respectively. This follows the established settings used in [11, 13, 14] validated against WT experiments.

The computational grids illustrated in Fig. 13.2 are generated using approximately 1.1 million hexahedral elements for the single tree configuration and 2.4 million for the pair of trees. Resolution is enhanced progressively at the vicinity of the tree (at the centre) to maximise numerical accuracy at regions of high solution gradients and reduce computational cost. This complies with recommendations based on the wall y^+ approach [20].

An inlet boundary condition is defined at the entrance while a pressure outlet is set at the exit of the domain to evacuate air. The top and sides of the computational domain are defined as symmetry to represent slip condition while the floor is considered non-slip to correspond to typical conditions in open space. A summary of the 3D computational domain and implemented boundary conditions are illustrated in Fig. 13.3 for (a) a single tree and (b) two trees with spacing, $x/H = 1.5$ and 3.

2.2 Tree Modelling

The tree is geometrically idealised to resemble the *Acacia Mangium*, a common roadside tree in Malaysia [17].

The tree trunk is modelled as a solid obstacle while the crown is set to be permeable with a porosity of $P_{vol} \approx 96$ %. This is numerically defined as a momentum sink comprising of the viscous and inertial loss terms (refer to Eq. (13.2)).

$$S_i = -\left(\sum_{j=1}^{3} D_{ij}\,\mu v_j + \sum_{j=1}^{3} C_{ij}\,\frac{1}{2}\rho\,|v|\,v_j\right) \tag{13.2}$$

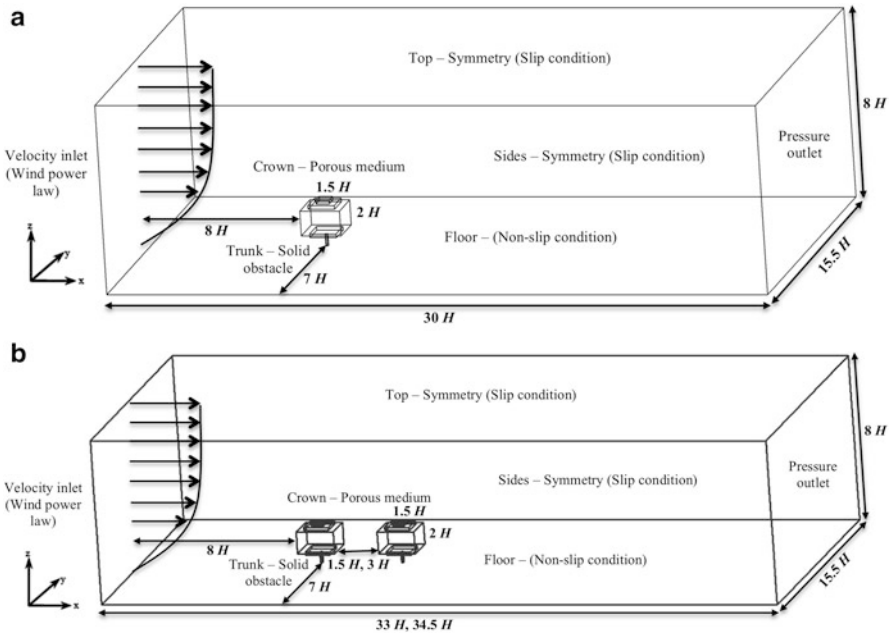


Fig. 13.3 Computational domains used in flow simulation of present study for around (**a**) a single tree and (**b**) two trees with spacing, $x/H = 1.5$ and 3

Table 13.2 Tree specifications

Parameter	Value
Tree species	*Acacia mangium*
Largest width of crown (m)	27
Largest height of crown (m)	27
Trunk height (m)	9
Trunk diameter (m)	2.8
Tree crown porosity, P_{vol} (%)	96
Pressure loss coefficient, λ (m^{-1})	200
Flexural modulus, E (MPa)	5,828[a]
Modulus of rupture, R (MPa)	62.28[a]

[a]Mechanical properties of the tree are obtained from field measurements [17]

where S_i is the source term for the ith (x, y or z) momentum equation while $|\upsilon|$ is the magnitude of velocity, C and D are prescribed matrices.

The porosity is defined by a pressure loss coefficient ($\lambda = 200\ m^{-1}$) and the tree specifications are summarised in Table 13.2.

Table 13.3 Solution methods

Parameter	Scheme
Scheme	Simple
Gradient	Least squares cell based
Pressure	Standard
Momentum	Second order upwind
Turbulent kinetic energy	Second order upwind
Turbulent dissipation rate	Second order upwind
Reynolds stresses	Second order upwind

Second order upwind scheme is selected to minimise numerical diffusion

2.3 Numerical Setup

The steady-state RANS solutions are obtained using standard *k*-ε turbulence model and the convergence criterion for all flow properties are set to 1×10^{-6} for both wind conditions. These are summarised in Table 13.3.

2.4 Mechanical Analysis

The resulting wind loads derived from ANSYS FLUENT are mapped, interpolated and exported to ANSYS Mechanical as face pressures. The material of the tree trunk is set to match the mechanical specifications from Table 13.2 with properties from field measurements [17] and the crown is assumed to be rigid. In order to investigate the aerodynamic effects resulting from interaction between the wind flow and structural behaviour, the tree is assumed to be massless for simplicity.

3 Results and Discussions

The numerical solutions obtained from CFD simulation and FEA analysis are presented and discussed in this section are based on non-dimensionalised distances on a 1:18 scale.

3.1 Flow Analysis Using CFD

Figure 13.4 presents the velocity contours for (a) a single tree and (b) around two trees with spacing, $x/H = 1.5$ and 3 comparing between *gentle breeze* and *storm* conditions. It is observed that the airflow tends to go around the tree rather than

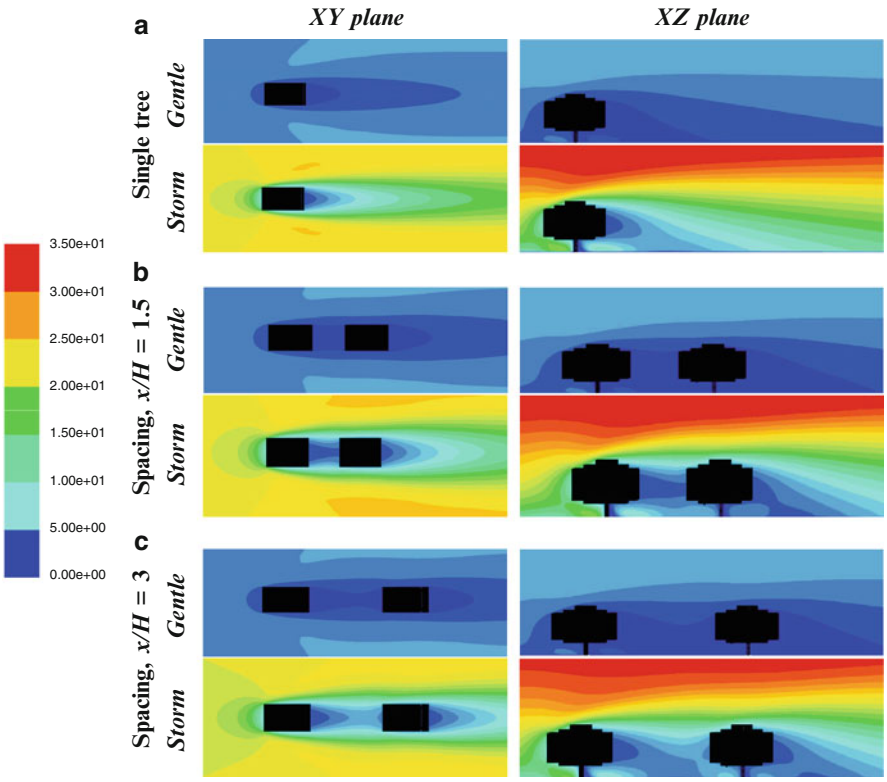


Fig. 13.4 Contours of the velocity magnitude for 4.7 ms^{-1} (*gentle breeze*) and 24 ms^{-1} (*storm*) around (**a**) a single tree, (**b**) two trees with spacing, $x/H = 1.5$ and (**c**) $x/H = 3$

through it. The magnitudes of the flow field and ensuing recirculation region are much stronger for the *storm* condition in comparison to the *gentle breeze*.

A large reduction in airflow velocity imposed on the sheltered tree compared to the front row tree was observed to occur for *gentle breeze* ($\approx$ 80 %) and *storm* conditions ($\approx$ 85 %) for tree spacing, $x/H = 1.5$. While, a much lower reduction was recorded for *gentle breeze* ($\approx$ 64 %) and *storm* conditions ($\approx$ 71 %) when the tree proximity was increased to $x/H = 3$. This implies that the increase of tree spacing would reduce the windbreak efficiency of the front row trees and cause higher resulting aerodynamic loads on the sheltered trees.

Figure 13.5 quantitatively illustrates the vertical velocity profiles that develop as a result of the two wind conditions for a single tree and two trees with spacing, $x/H = 1.5$ and 3. It is observed that the velocity difference between the windward and leeward sides of the *gentle breeze* are negligible compared to the *storm* condition for all the three situations. The larger velocity drop generated by the *storm* condition presents a larger pressure difference implying that larger wind loads are experienced.

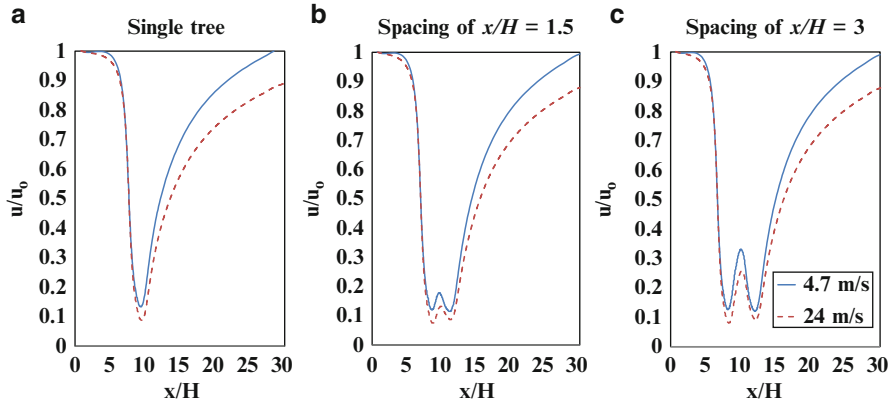


Fig. 13.5 Profiles of *x/H* velocity 1 *H* above ground along symmetry axis comparing between *gentle breeze* and *storm* inlet conditions around (**a**) a single tree, (**b**) two trees with spacing, $x/H = 1.5$ and (**c**) $x/H = 3$

An increase in tree spacing results in greater airflow velocity experienced by the shelter tree downstream of the prevailing wind because the airflow is given enough time to reattach and accelerate towards its initial velocity behind the upstream tree (at $x/H \approx 10$).

The profiles and contours of pressure fields for the two conditions are presented in Figs. 13.6 and 13.7, respectively. These support the observations derived from the velocity fields, further illustrating that the pressure difference experienced during the *storm* conditions are much larger than those during *gentle breeze*, resulting in greater wind forces on the trees. The recirculating region generated by the *storm* condition introduces a negative pressure field immediately behind the tree increasing the net force exerted. The negative pressure drop is attributed to the separation, reversal of airflow and recirculation.

When increasing tree spacing, the sheltered tree was found to experience greater pressure difference due to the diminishing windbreak protection offered by the front row (i.e. upstream) tree. For example, significant reductions in pressure difference was experienced by the sheltered tree in comparison to the front row tree for both *gentle breeze* ($\approx$ 69 %) and *storm* conditions ($\approx$ 67 %) for tree spacing, $x/H = 1.5$. This is in contrast to the minor reduction in pressure difference experienced during the *gentle breeze* ($\approx$ 56 %) and *storm* conditions ($\approx$ 64 %) when the spacing between the trees was increased to $x/H = 3$.

Table 13.4 summarises the aerodynamic data and demonstrates that during *storm* conditions, the wind loads on trees are much larger due to the greater pressure differences imposed. Furthermore, the increase in tree spacing resulted in a reduced windbreak protection for the sheltered trees; hence the likelihood of structural damage is significantly increased. This is discussed in the following section.

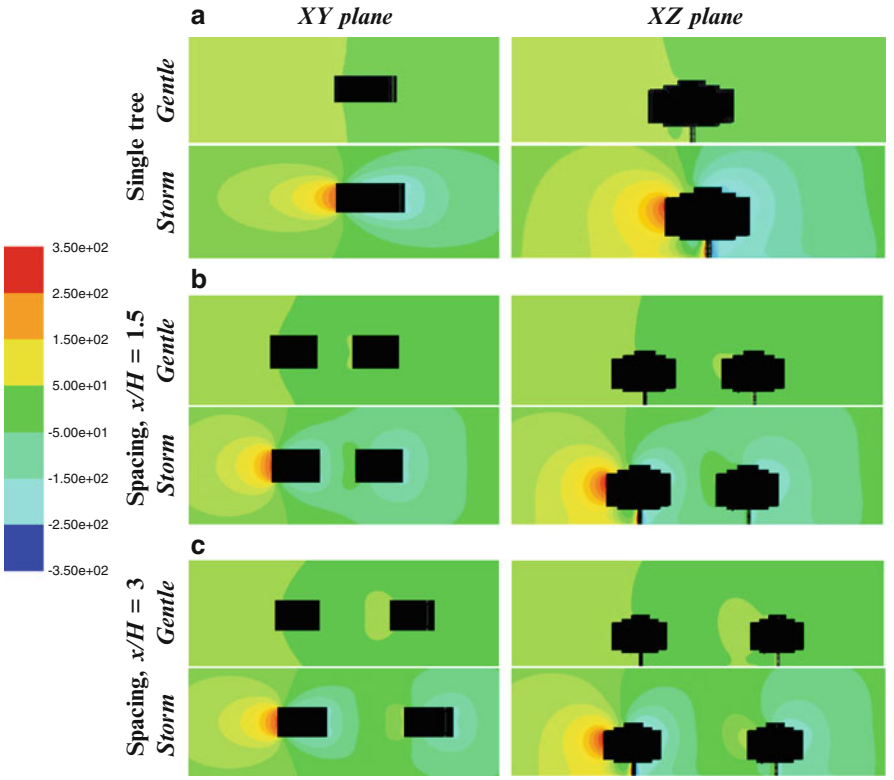


Fig. 13.6 Contours of the pressure magnitude for 4.7 ms^{-1} (*gentle breeze*) and 24 ms^{-1} (*storm*) around (**a**) a single tree, (**b**) two trees with spacing, $x/H = 1.5$ and (**c**) $x/H = 3$

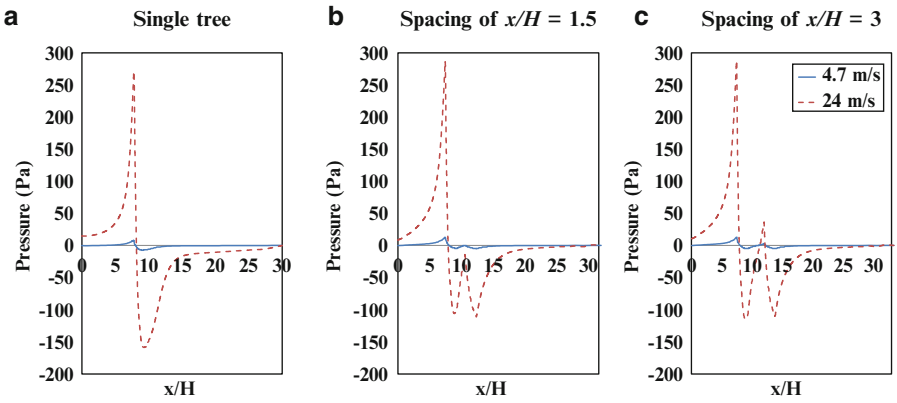


Fig. 13.7 Profiles of static pressure (Pa) $1H$ above ground along symmetry axis comparing between *gentle breeze* and *storm* inlet conditions around (**a**) a single tree, (**b**) two trees with spacing, $x/H = 1.5$ and (**c**) $x/H = 3$

Table 13.4 Summary of CFD results

Parameter	Value							
Tree spacing, *x/H*	1.5				3			
Inlet velocity (ms^{-1})	4.7 (*Gentle*)		24 (*Storm*)		4.7 (*Gentle*)		24 (*Storm*)	
Tree location	*FRT*	*ST*	*FRT*	*ST*	*FRT*	*ST*	*FRT*	*ST*
Imposed velocity (ms^{-1})	≈ 4	≈ 0.8	≈ 20.5	≈ 3.1	≈ 4	≈ 1.5	≈ 20.5	≈ 6
Pressure difference (Pa)	≈ 18	≈ 5.4	≈ 400	≈ 117	≈18	≈ 7.9	≈ 400	≈ 147

FRT front row tree, *ST* sheltered tree

3.2 Mechanical Analysis Using FEA

The pressure field that develops around the trees is utilised in the structural analysis to determine the mechanical effects and likelihood of windthrow.

Tree damage commonly occurs as a result of breakage or failure in the root anchorage, when the exerted forces exceed the structural resistance of the trunk or support system in the case of anchorage. In the present study, only the former is considered.

Winds typically induce a turning moment on the tree trunk hence flexural modulus rather than Young's modulus is considered. For the purpose of this investigation, failure by virtue of breakage of trunk is assumed to occur when the applied stresses due to wind loads exceed the modulus of rupture of the tree trunk. Additional forces due to gravity when the trunk deflects substantially are neglected in order to identify and quantify the aerodynamic effects only. The deformation values in this section are based on non-dimensionalised lengths employed for the computational domain.

Figure 13.8 demonstrates the deformation of the tree and resulting stresses when subjected to the two wind conditions. It can be seen that the *storm* condition produces a larger deflection on the tree as compared to the *gentle breeze*.

Generally, it is observed that the stresses are at a minimum at the core of the trunk and increases outward. This suggests that at the time of trunk damage, failure first occurs on the outside and rapidly propagates inwards. On the leeward side of the trunk (i.e. back of tree), the stresses can be seen to progressively increase as it moves farther from the ground (fixed support); where the magnitude of deflection is proportional to the resulting stresses.

It is evident that a smaller spacing (i.e. $x/H = 1.5$) offers a much better windbreak protection for the sheltered tree during a *storm*. The sheltered tree experiences much lower deformations and resulting stresses as compared to the front row tree.

In contrast for a larger spacing (i.e. $x/H = 3$), the front row tree presents no protection to the sheltered tree. While previous observations indicate that the pressure difference experienced by the sheltered tree is consistently lower than the front row tree, higher deformations were found to occasionally occur for the former at this spacing. As demonstrated in Fig. 13.8, the front row tree is observed to be

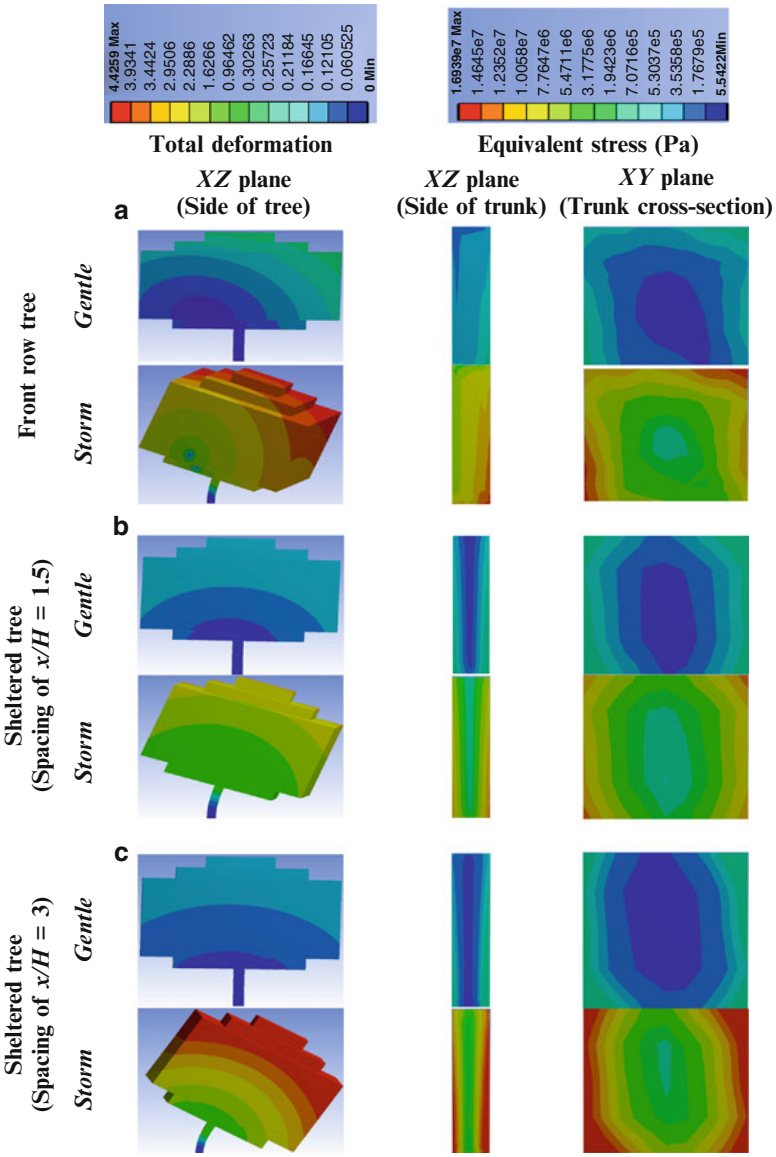


Fig. 13.8 Contours of the total deformation and equivalent stress (Pa) for 4.7 ms^{-1} (*gentle breeze*) and 24 ms^{-1} (*storm*) around (**a**) a single tree, (**b**) two trees with spacing, $x/H = 1.5$ and (**c**) $x/H = 3$

more twisted rather than bent, while a bigger deflection occurs in the sheltered tree, resulting in larger stresses. Therefore, while the aerodynamic loads imposed on the front row tree are higher, more of that energy is used to twist the trunk rather than deflect it.

Table 13.5 Summary of FEA results

Parameter	Value							
Flexural modulus, E (MPa)	5,828							
Rupture modulus, R (MPa)	62							
Tree spacing, *x/H*	1.5				3			
Inlet velocity (ms^{-1})	4.7 (*Gentle*)		24 (*Storm*)		4.7 (*Gentle*)		24 (*Storm*)	
Tree location	*FRT*	*ST*	*FRT*	*ST*	*FRT*	*ST*	*FRT*	*ST*
Maximum deformation	0.23	0.25	4.43	3.08	0.23	0.18	4.43	5.12
Von Mises stress (MPa)	0.88	1.15	16.94	14.2	0.88	0.89	16.94	25
Maximum elastic strain	1.5E-4	2.0E-4	2.9E-3	2.5E-3	1.5E-4	1.5E-4	2.9E-3	4.3E-3

FRT front row tree, *ST* sheltered tree

Table 13.5 summarises the results obtained from the structural analysis carried out in ANSYS Mechanical, where the flexural modulus and modulus of rupture are based on published data [17] (see Table 13.2). It can be concluded that windthrow by trunk breakage is more likely to occur during a *storm* as compared to a *gentle breeze*; with the resulting stresses are much closer to the modulus of rupture for the given material. In addition, smaller tree spacing (i.e. $x/H = 1.5$) offers better protection for sheltered trees by minimising the likelihood of windthrow as opposed to a larger distance.

4 Conclusions

The wind loads and mechanical effects of two different wind conditions on trees are investigated for a typical baseline scenario based on *gentle breeze* and a severe case of tropical *storm* conditions. The effects of different tree spacing are also examined to establish the windbreak protection offered on sheltered trees by the front row trees. Flow solutions (i.e. velocity and pressure fields) obtained from ANSYS FLUENT is used to perform structural analysis in ANSYS Mechanical to determine the deflection and stresses likely to be experienced by trees due to aerodynamic loading.

The pressure difference between the windward and leeward sides during the *storm* conditions are much larger than during *gentle breeze* because of a more prominent recirculating region which induces a negative flow and consequently pressure field. This implies that the wind forces exerted on the tree are larger and as a result the trunks experiences greater deformation and stresses, increasing the likelihood of structural failure. Better shelter protection is offered to the downstream trees at smaller tree spacing as a result of reduced airflow velocities and pressure differences.

Windthrow by trunk damage is shown to likely occur during a *storm* condition as the recorded maximum equivalent stresses due to bending are much closer to the rupture modulus of the material.

References

1. Ulanova, N.G.: The effects of windthrow on forests at different spatial scales: a review. For. Ecol. Manag. **135**, 155–167 (2000)
2. Duelli, P., Obrist, M.K., Wermelinger, B.: Windthrow-induced changes in faunistic biodiversity in alphine spruce forests'. For. Snow Lands. Res. **77**(1/2), 117–131 (2002)
3. Putz, F.E., Coley, P.D., Lu, K., Montalvo, A., Aiello, A.: Uprooting and snapping of trees: structural determinants and ecological consequences'. Can. J. For. Res. **13**(5), 1011–1020 (1983)
4. Sinton, D.S., Jones, J.A.: Extreme winds and windthrow in the Western Columbia River Gorge. Northwest Sci. **76**(2), 173–182 (2002)
5. Peltola, H.M.: Mechanical stability of trees under static loads'. Am. J. Bot. **93**(10), 1501–1511 (2006)
6. Brudi, E., Wassenaer, P.v.: Trees and statics: non-destructive failure analysis. How trees stand up and fall down, pp. 53–70 (2002)
7. Novak, M.D., Warland, J.S., Orchansky, A.L., Ketler, R., Greem, S.: Wind tunnel and field measurements of turbulent flow in forests. Part I: uniformly thinned stands'. Bound.-Lay. Meteorol. **95**, 457–495 (2000)
8. Gardiner, B.A., Stacey, G.R., Belcher, R.E., Wood, C.J.: Field and wind tunnel assessments of the implications of respacing and thinning for tree stability'. Forestry **70**(3), 233–252 (1997)
9. Gross, G.: A numerical study of the air flow within and around a single tree. Bound. Lay. Meteorol. **40**, 311–327 (1987)
10. Tiwary, A., Morvan, H.P., Colls, J.J.: Modelling the size-dependent collection efficiency of hedgerows for ambient aerosols'. Aerosol Sci. **37**, 990–1015 (2005)
11. Salim, S.M., Cheah, S.C., Chan, A.: Numerical simulation of dispersion in urban street canyons with avenue-like tree plantings: comparison between RANS and LES'. Build. Environ. **46**, 1735–1746 (2011)
12. Salim, S.M., Buccolieri, R., Chan, A., Di Sabatino, S., Cheah, S.C.: Large eddy simulation of the aerodynamic effects of trees on pollutant concentrations in street canyons'. Procedia Environ. Sci. **4**, 17–24 (2011)
13. Buccolieri, R., Salim, S.M., Leo, L.S., Sabatino, S.D., Chan, A., Ielpo, P., Gennaro, G., Gromke, C.: Analysis of local scale tree–atmosphere interaction on pollutant concentration in idealized street canyons and application to a real urban junction. Atmos. Environ. **45**, 1702–1713 (2011)
14. Gromke, C., Buccolieri, R., Sabatino, S.D., Ruck, B.: Dispersion study in a street canyon with tree planting by means of wind tunnel and numerical investigations – evaluation of CFD data with experimental data. Atmos. Environ. **42**, 8640–8650 (2008)
15. Gan, C.J., Salim, S.M.: Numerical analysis of fluid-structure interaction between wind flow and trees. Lecture notes in engineering and computer science. Proceedings of the world congress on engineering 2014, WCE 2014, pp. 1218–2014. London (2014)
16. Sapuan, M.S., Razali, A.M., Ibrahim, K.: Forecasting and mapping of extreme wind speed for 5 to 100-years return period in Peninsula Malaysia. Aust. J. Basic Appl. Sci. **5**(7), 1204–1212 (2011)
17. Sahri, M.H., Ashaari, Z., Kader, R.A., Mohmod, A.L.: Physical and mechanical properties of Acacia mangium and Acacia auriculiformis from different provenances'. Pertanika J. Trop. Agric. Sci. **21**(2), 73–81 (1998)
18. Salim, S.M., Ong, K.C.: Performance of RANS, URANS and LES in the prediction of airflow and pollutant dispersion. IAENG Transactions on Engineering Technologies, pp. 263–274. Springer, Netherlands (2013)
19. Salim, S.M., Buccolieri, R., Chan, A., Di Sabatino, S.: Numerical simulation of atmospheric pollutant dispersion in an urban street canyon: comparison between RANS and LES'. J. Wind Eng. Ind. Aerodyn. **99**(2–3), 103–113 (2011)
20. Salim, S.M., Ariff, M., Cheah, S.C.: Wall y^+ approach for dealing with turbulent flows over a wall mounted cube. Prog. Comput. Fluid Dyn. **10**(5/6), 341–351 (2010)

Chapter 14
Effecting Quench Agitation by Immersion Speed Variation of C30 Carbon Steel and Mechanical Properties Examination

Segun Mathew Adedayo, Adebayo Surajudeen Adekunle, and Tunji Ebenezer Oladimeji

Abstract A study of quench agitation by immersion speed variation was carried out on a C30 carbon steel material and examination of resulting mechanical properties. Quenching was carried out in a special extended height bath under conditions of constant bath temperature and a variable immersion speed. Material thermal history data was taken during the quench process and mechanical properties comprising of hardness and tensile strength of material were examined thereafter. Immersion speed variation was effected by a variable weight-force application acting on the quenched C30 specimen falling freely through an extended height quench bath. At immersion speeds of 0.106, 0.697, 0.853, 1.065 and 1.139 m/s; the yield strength of the material are 310.40, 496.12, 500.56, 565.40 and 579.92 MN/m^2 respectively while at a typical location of radius 15 mm on specimen mid-height the corresponding hardness values at the respective immersion speeds are 275, 293.40, 454.60, 408 and 594 VHN. There is an enhancement of mechanical strength with immersion speed increase.

Keywords Agitation • Extended height bath • Hardness • Immersion speed • Quenching • Tensile strength • Thermal history

1 Introduction

Quenching is a metal heat treatment process involving a controlled cooling of a metal from a temperature above the austenitic range to a low temperature in order to form certain desired microstructure and physical properties [1, 2]. Heat treatment is a multi-parameter process. Selection of the appropriate parameters helps in predicting possible behavior of metals. Some of these parameters are type of quenching medium, quenchant temperature and agitation level [3]. Quenching is an

S.M. Adedayo (✉) • A.S. Adekunle • T.E. Oladimeji
Department of Mechanical Engineering, University of Ilorin, Ilorin P. M. B. 1515, Nigeria
e-mail: amsegun@unilorin.edu.ng; adekunlebayor@gmail.com; tunjiebenezer51@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_14

essential process in developing the desired properties of many steel and aluminium alloys. Agitation, or forced circulation of the quenchant tends to shorten the material cooling time. Without agitation, natural convection within the quenchant and quenchant vaporization limit the heat transfer rate through the fluid film boundary at the surface of parts [4–6]. Obtaining a forced convection fluid regime greatly reduces the resistance to heat flow at the fluid film boundary layer [7]. This can be accomplished by mechanically moving the parts through the bath, pumping to re-circulate the quenchant, or mechanically inducing agitation/circulation of the fluid. Where control over the cooling rate is important, mechanical agitation provides the best approach at the lowest energy cost. The cooling rate is however influenced by such factors as bath temperature, quenchant type, immersion temperature, agitation and immersion speed [8]. The effect of agitation on materials had been studied by various researchers. Agitation during quenching generally enhances heat transfer at all cooling stages. Hardness of AISI 4135 steel across its entire thickness in a conventional oil quench process increased with agitation [9–11]. Quench severity is dependent on agitation rate, tank size, fluid viscosity, type and placement of agitators [12, 13]. Agitation affects the hardness and depth of hardening during the quench because of the rupture of the relatively unstable film boiling cooling process that always occur in vaporizable quenchants such as oil, water and aqueous polymers. An early breakdown of the vapour blanket results in transition into the nucleate boiling [2, 14]. This process results into a time reduction of the slow cooling stage thus resulting in rapid heat transfer. Agitation enhances recirculation of quenchant unto the metal surface with consequent higher temperature difference between quenchant and the surface. Application of direct quenching and tempering (DQ-T) combined with controlled rolling has been widely used in the production of low and medium carbon steel plates and rods. The steel produced by (DQ-T) has an advantage of attaining better combination of strength, toughness and weldability in comparison with the steels produced by conventional reheating and quenching (RQ) process. Hot deformation of austenite refines ferrite and ausformed martensite [14]. The influence of microstructure on mechanical properties of low alloy steels has been a subject of considerable research interest in physical metallurgy [15, 16]. Most researchers do not factor the immersion speed of materials into the quenchants; this works attempts to examine the effect of immersion speed of C30 carbon steel into the quenchant on its mechanical strength properties.

2 Experimental Methods

2.1 Quech Materials and Bath Specification

A C30 carbon steel material with chemical composition as indicated in Table 14.1 was preliminarily annealed at 900 °C and soaked for 1 h before machining to tensile and hardness tests specifications [17]. Figure 14.1a, b shows the as – quenched

Table 14.1 Chemical composition of the C30 steel

Symbol	% Weight
C	0.311
Si	0.028
S	0.026
P	0.007
Mn	0.742
Ni	0.202
Cr	0.172
Mo	0.017
V	0.001
Cu	0.035
W	0.004
As	0.010
Sn	0.001
Co	0.0057
Al	0.004
Pb	0.001
Zn	0.004
Fe	97.76

test pieces. The quench bath consists of an extended height square base tank of 305×305 mm base and 2,134 mm height. A separate immersion weight-force carrier is also constructed.

The weight-force carrier is a pyramidal shape with $300 \times 300\ mm^2$ base as shown at point of immersion end above the tank in Fig. 14.2. The pyramidal shape allows for a streamlined movement through the extended height quench tank bath. Different quench immersion speeds was effected by weights placed inside the pyramid carrier falling freely through the quench bath height. Weights of 3.35, 11.05, 13.05, 16.85 and 18.35 kg were used corresponding to speeds of 0.106, 0.697, 0.853, 1.065 and 1.139 m/s immersion speeds respectively. A calibration of weight-force against speed was done by noting the time it takes for specified weights to travel vertically through known liquid heights. The C30 carbon steel was heated in an electric Muffle furnace at the rate of 25 °C/min to a temperature of $860^{\pm 2}\,^{\circ}$C and soaked for duration of 1 h. Test material was removed from the furnace and attached to the weight carrier for immediate quench immersion with the corresponding weight. Each quench process was repeated three times and average values reported. Work-piece temperature during immersion was separately monitored using type K thermocouple and a SD card data logger digital thermometer Model MTM-380SD. The thermocouple was inserted in a hole of diameter 3 mm drilled on the top surface of the probe and quickly inserted within 2 s into the 2,000 ml of bioquenchants having a temperature of 25 °C at static condition. Cooling rates during quench process were obtained from the cooling curves. This was obtained by taking the slope at each temperature on the temperature versus timecurve of each immersion

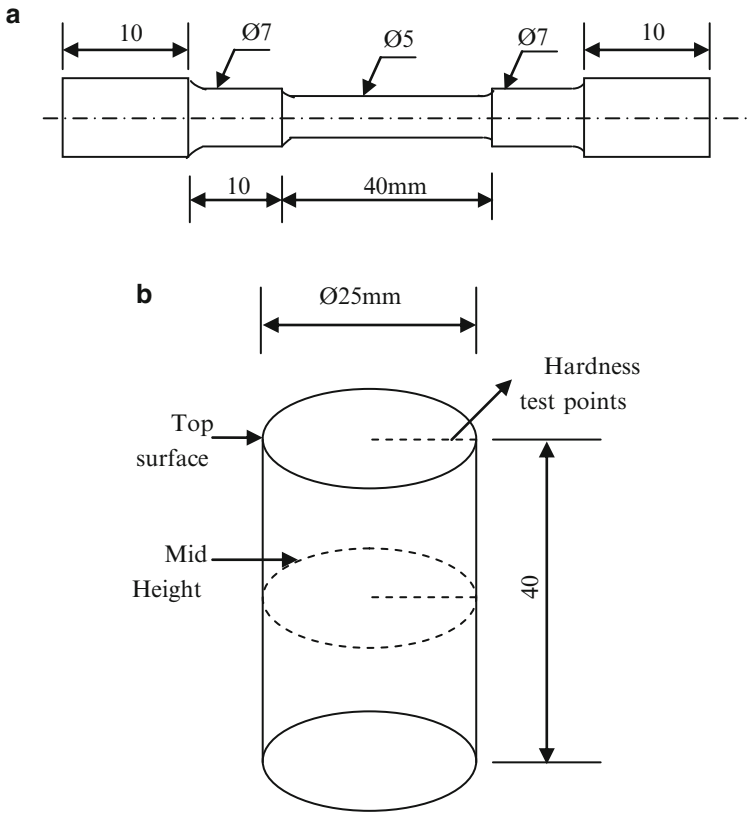


Fig. 14.1 (**a**) Tensile test specimen. (**b**) Hardness test specimen

speed tested. The work-piece was positioned at exterior side of pyramidal shape in a mesh thus allowing for free quenchant flow around the work-piece during immersion [17].

2.2 Mechanical and Micro-structural Property Test

Room temperature uni-axial tensile tests were performed on round tensile test specimens from quenched C30 carbon steel material machined to nominal length 40 mm and gauge diameter 5.5 mm with grip ends having diameter 10 mm. Hardness test specimens was machined to length 40 mm and diameter 25 mm. Computerized Universal Tensile Testing Machine (Model: 3369) was used to conduct the tensile tests. The samples were tested at a nominalstrain rate of 10–

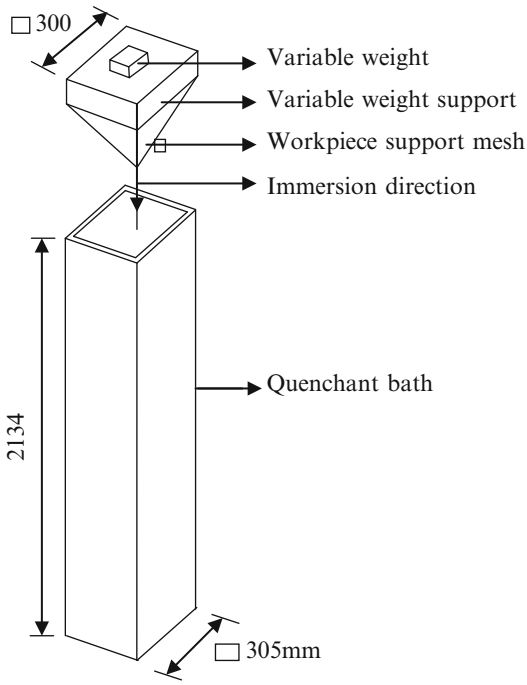


Fig. 14.2 Extended height quench bath

3/s until failure. Repeated tests were performed for each test condition to ensure reliability of results and average values reported. The tensile properties evaluated from the engineering stress – strain curves developed from the tension test are the yield strength (σ_y) and ultimate tensile strength (σ_u). Hardness measurement was taken along the radial direction at 5 mm intervals at exterior top or bottom surface and a sectioned mid – height surface (see Fig. 14.1b). This was done using Daheng microscope and Vickers hardness tester (LECO AT700 Microhardness Tester). Prior to testing, the steel specimens were smoothened using phenolic powder, grinded and polished to obtain a smooth surface finish. A direct load of 490.3MN (50.03 kg) was thereafter applied on the specimen for a dwell time of 10 s. Multiple hardness tests were performed on each sample and the average values reported. Optical micrographs of quenched specimens were carried out on sectioned surface, grinded with silicon carbide papers of 240, 320, 400 and 600 grit sizes. They were subsequently polished using a cloth impregnated with alumina until a mirror surface was obtained. Grinded surface were cleaned with water and ethanol. Etching with 2 % Nital was done and microstructures observed using a high powered Daheng optical microscope.

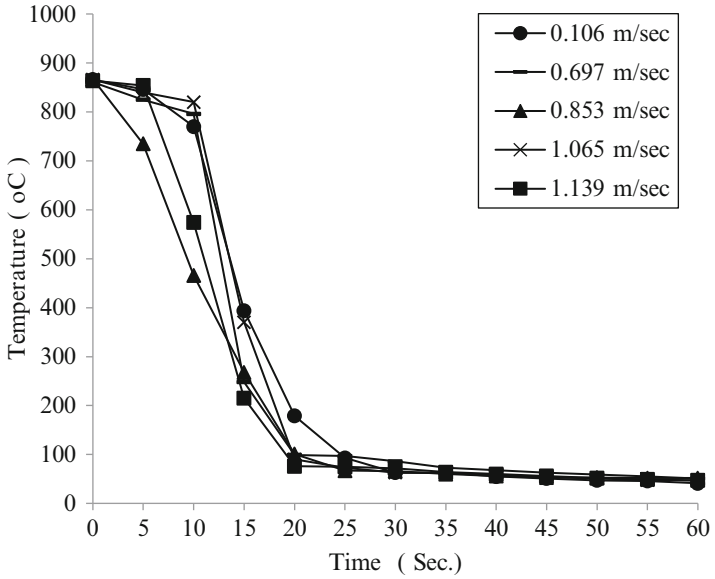


Fig. 14.3 Cooling curves at different immersion speed

3 Results and Discussion

3.1 Cooling Curves and Microstructures

Cooling curves under different immersion speeds are depicted in Fig. 14.3. They are obtained at an initial quenchant temperature of 25 °C under dynamic quenching conditions. Corresponding curves of the cooling rates versus time is also shown in Fig. 14.4. Two of the three stages of quenching mechanism; film boiling and convective cooling are evident in the cooling curves. With increasing immersion speed the film boiling stage time range diminished. Maximum cooling rates of 53.88, 63.80, 73.88, 90.76 and 97.70 °C/s. with corresponding immersion speeds of 0.106, 0.697, 0.853, 1.065 and 1.139 m/s respectively were observed. Cooling rates increases with immersion speed as a result of the rapid movement of workpiece through the quenchant. The rapid destruction of the film boiling stage allowed for an enhanced heat convection to the liquid medium.

The microstructures of the as-received, minimum immersion speed and maximum immersion speeds of the quenched C30 carbon steel are as shown in Fig. 14.5a–c. The as-received specimen comprised of ferrite and pearlite predominantly. The C30 carbon steel with highest speed of 1.139 m/s showed traces of martensitic structures while the lowest immersion speed specimen of 0.106 m/s showed mainly bainitic structures with a mixture of pearlite (see Fig. 14.5b). At maximum immersion speed of 1.139 m/s corresponding maximum cooling rate

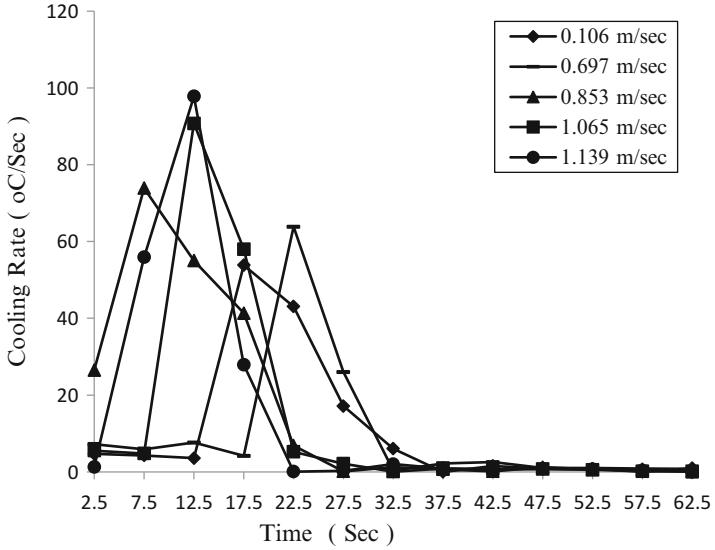


Fig. 14.4 Variation of cooling rate with immersion speed

of 97.7 °C/s occurred at temperature of 534.3 °C. This falls within the pearlitic transformation temperature and it is considered high enough to suppress diffusional transformation in this range such that transformation is delayed till the martensitic range of about 330–120 °C was reached.

3.2 Mechanical Properties

Figure 14.6 shows the yield strength and ultimate tensile strength of the C30 material at different immersion speeds. Material strength increased with immersion speed by as much as 19.30 % when immersion speed increased by 63.41 % (from 0.697 to 1.139 m/s.). This property enhancement is as a result of increased cooling rate resulting from a disruption of both film and nucleate boiling stages at higher speeds.

Figures 14.7 and 14.8 show the variation of material hardness with immersion speed across the radial direction taken at extreme heights and mid-height points of test specimen. In both cases hardness increased with immersion speed and it is attributable to the higher cooling rate resulting from the collapse of the film boiling and nucleate boiling stages. Hardness tend to increase with increasing radius at both mid and top surfaces. These can be explained in terms of higher severity of quenching towards the exterior cylindrical boundaries. At a typical radius of 20 mm on mid – height, an increase in immersion speed by 52.7 % increased microhardness by 71 %.

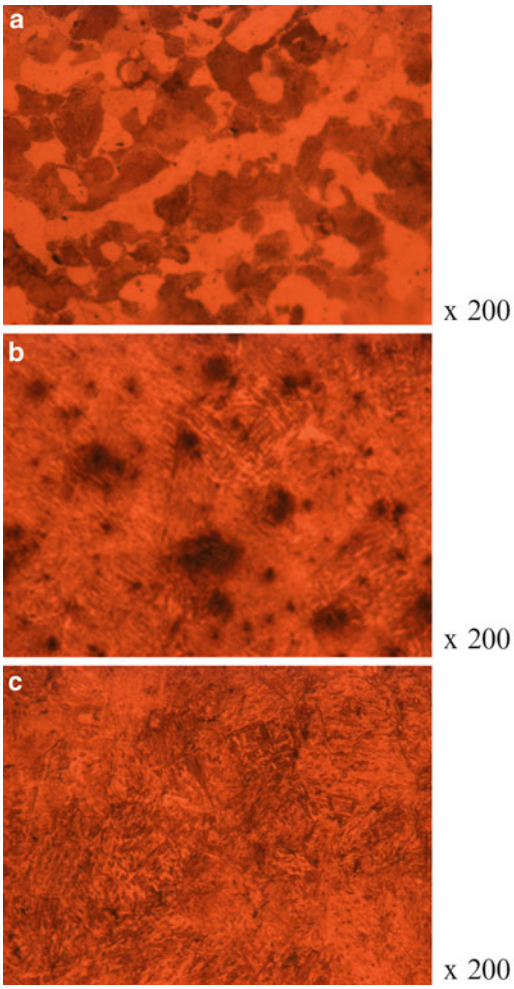


Fig. 14.5 (**a**) Microstructures of the as-received C30 steel material. (**b**) Microstructure of minimum immersion speed (0.106 m/s). (**c**) Microstructure at maximum immersion speed (1.139 m/s)

4 Conclusion

Variable immersion speed effects in quenching heat treatment process are investigated. To confirm these effects on a C30 carbon steel material, sample mechanical tests and microstructures are conducted. The main findings are summarized as follows.

Cooling rates increased with immersion speed due to rapid transformation of the film boiling to convective cooling.

At the highest immersion speed of 1.139 m/s, the C30 carbon steel material showed traces of presence of martensite microstructures.

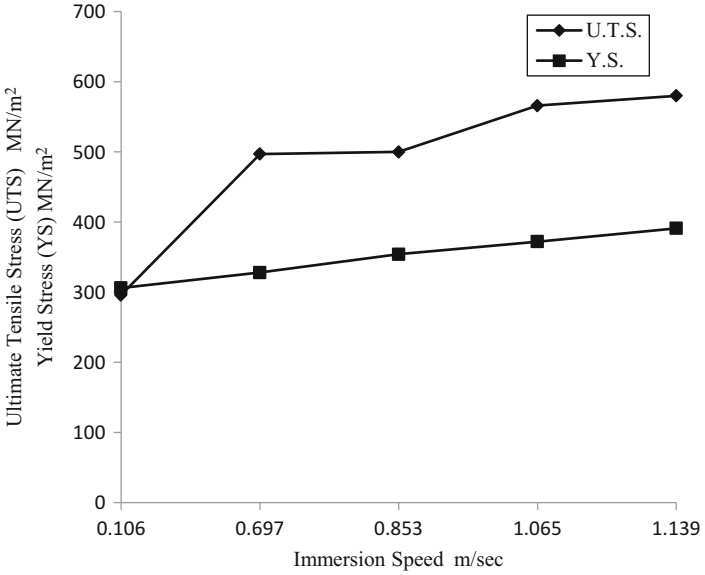


Fig. 14.6 Variation of material strength properties with immersion speed

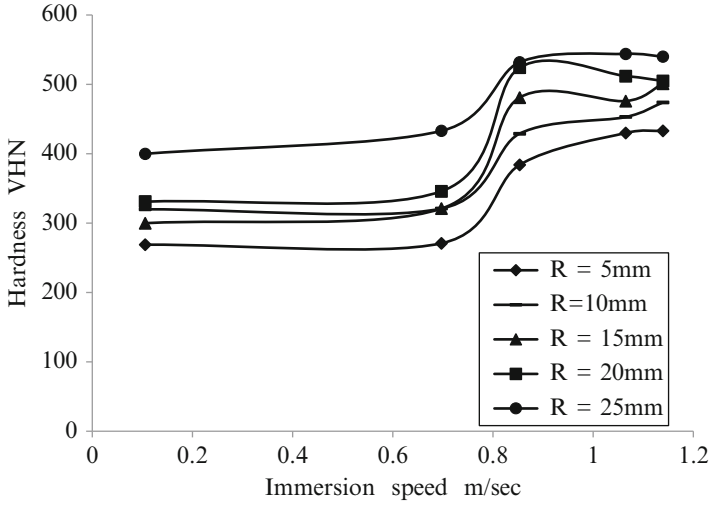


Fig. 14.7 Variation of hardness with immersion speed at extreme height

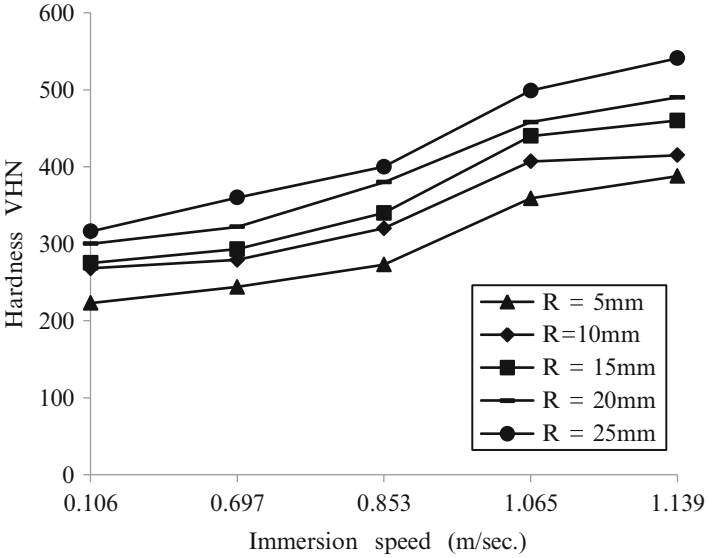


Fig. 14.8 Variation of hardness with immersion speed at mid-height

Material strength increased by as much as 19.30 % with 63.41 % increase in immersion speed.

Material hardness increased steadily with immersion speed and also increased along the geometrical radius of work material.

References

1. ASM Handbook: Heat Treating of Aluminium Alloys. Heat treating, vol. 4, 10th edn, pp. 823–873. ASM International, Materials Park (1995)
2. Cavaliere, P., Cerri, E., Leo, P.: Effect of heat treatment on mechanical properties and fracture behavior of a thixocast A356 aluminum alloy. J. Mater. Sci. **39**, 1653–1658 (2004)
3. Bohumil, T., Steven, D., Spanielka, J.S.: Effect of agitation work on heat transfer during cooling in oil isorapid 277HM. J. Mech. Eng. **58**(2), 102–106 (2012)
4. Caner, S., Hakan-Gur, C.: A simulation of the quenching process for predicting temperature, microstructure and residual stresses. J. Mech. Eng. **56**(2), 93–103 (2010)
5. Rometsch, P.A., Starink, M.J., Gregson, P.J.: Improvements in quench factor modeling. Mater. Sci. Eng. **339**(1–2), 255–264 (2003)
6. Totten, G.E., Bates, C.E., Clinton, N.A.: Quenching system design, handbook of quenchant and quenching technology, vol. 119, pp. 339–411. ASM International, Materials Park (1993)
7. Hanguang, F., Xiaole, C., Zhongze, D., Yingping, L., Zhenjun, F.: Effect of quenching temperature on structure and properties of centrifugal casting of high speed steel roll. J. Res. Dev. **6**(1), 15–19 (2009)
8. Fontecchio, M., Maniruzzaman, M., Sission Jr. R.D.: The effect of bath temperature and agitation rate on the quench severity of 6061 aluminium in distilled water. In: Proceedings of the 21st heat treating society conference, Indianapolis (2002)

9. Liscic, B., Totten, G.E., Howes, M.A.H. (eds.): Steel Heat Treatment Handbook, vol. 97. Marcel Dekker, New York (1997). Chap. 3
10. Ma Shuhui, Aparna, S.V., Makiko, T., Dareeell, K.R., Maniruzzaman, M.D., Sisson Jr. R.D.: Quenching–understanding, controlling and optimizing the process. In: Proceedings of the 4th international conference on quenching and the control of distortion. Beijing, 20–23 May 2003
11. Chen, N., Han, I., Zhang, W., Hao, X.: Enhancing mechanical properties and avoiding cracks by simulation of quenching connecting rods. Mater. Lett. **61**, 3021–3024 (2007)
12. Kim, C.K., Lee, D.G., Lee, S.: Correlation of microstructure and fracture properties of five centrifugal cast high speed steel rolls. Mater. Sci. Technol. **23**(9), 1065–1074 (2007)
13. Taraba, B., Spanielka, J.: Combined heat transfer coefficient calculation for cooled probe to 850 °C in quenching oil. In: The international conference of the Carpathian Euro – region specialists in Industrial Systems, North University of Baia Mare, Romania, pp. 281–286 (2010)
14. Lenumadi, F.Z., Chala, A., Ferhati, S., Chabane, F., Benramache, S.: Structural and mechanical behavior during quenching of 40CrMoV5 steel. J. Sci. Eng. **3**(1), 1–6 (2013)
15. Hasan, H.S., Peet, M.J., Jalil, J.M., Bhadeshia, K.H.D.: Heat transfer coefficients during quenching of steels. http://www.template-div-transfer_Hasan_2011.pdf (2014). Accessed 25 Feb 2014
16. Momoh, I.M., Akinribide, O.J., Ayanleke, J., Olowonubi, J., Olorunfemi, G.O., Oshodin, T.: Investigating the mechanical properties of post weld heat treated 0.33 % C low alloy steel. Int. J. Sci. Technol. **2**(6), 433–437 (2013)
17. Adedayo, S.M., Adekunle, A.S., Oladimeji, T.E.: Effect of immersion speed on mechanical properties of a quenched C30 carbon steel. Lecture notes, in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 1233–1237. London, 2–4 July 2014

Chapter 15
Ventilation Flow Through a Room Fitted with a Windcatcher Using a LES CFD Technique

Amirreza Niktash and B. Phuoc Huynh

Abstract In this study, air flow through a three-dimensional room fitted with a two-sided windcatcher is observed numerically, using a commercial computational fluid dynamics (CFD) software package. A LES (Large Eddy Simulation) method is used and the results are compared with those obtained previously by using a RANS (Reynolds Averaged Navier-Stokes) technique. Since LES is generally considered as more accurate but requiring more computational efforts than RANS, the results obtained in this work would verify the RANS results which have covered many more cases in addition to the one considered in this work. The LES and RANS results for the windcatcher configuration considered in this work are in good agreement. Specifically, it is confirmed that the two-canal centred position windcatcher with the bottom length of 10 cm provides full circulation for most part of the room and a large region of stable velocity in the acceptable range of indoor air speed for human comfort.

Keywords CFD • Circulation • Flow path traces • Human comfort • K-ε • LES • RANS • Simulation • Ventilation • Windcatcher

1 Introduction

A windcatcher is a structure fitted on the roof of a building for providing natural ventilation using wind power; it exhausts the inside stale air to the outside and supplies the outside fresh air into the building interior space working by pressure

A. Niktash (✉) • B.P. Huynh
Faculty of Engineering and IT, University of Technology (UTS), Sydney, Australia
e-mail: Amirreza.Niktash@uts.edu.au; Phuoc.Huynh@uts.edu.au

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_15

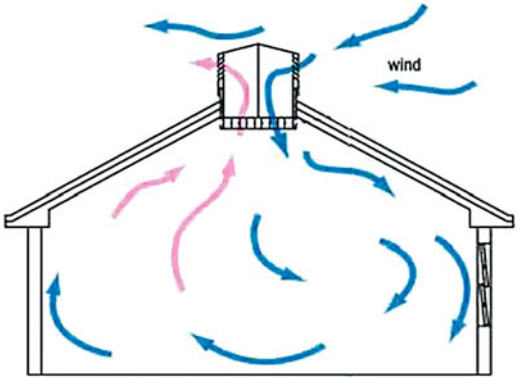


Fig. 15.1 Operation of a typical windcatcher

difference between outside and inside of the building and using two fundamental natural ventilation principles; namely the stacks effects and the wind driven ventilation (Fig. 15.1).

Natural ventilation has been known as one of the effective methods to provide healthy and comfortable interior environment by preventing moisture development in the air and reducing pollutants concentration effectively. Windcatcher is one of the green features for providing natural ventilation using the free driving force of wind power which has been employed over centuries in the hot arid regions of the Persian Gulf countries called as Baud-Geer and north African countries called as Malqaf.

Malqaf is a bidirectional windcatcher which is mounted on the top of a covered court yard (Fig. 15.2). It is normally combined with another architectural element known as Salasabil which is a wavy marble plate that is linked to a source of water [1].

Baud-Geer, on the other hand, is not only bidirectional and has diverse forms and various shapes but also they are used during summers and are closed during winters (Fig. 15.3). The cross sections of all windcatchers which have circular or squared shapes are divided internally into various segments to get one-sided, two-sided, three-sided, four-sided, hexahedral, and octahedral windcatchers [2, 3].

The low cost of windcatcher system (operational and maintenance cost) in comparison with mechanical ventilation system, being noiseless, durable, requiring no fossil energy, supplying clean air and using sustainable energy of wind have led to use of the windcatcher today as a passive and environmental friendly system.

The experimental studies of windcatcher systems for all different cases are obviously expensive or even impossible. Using computational fluid dynamics (CFD) for flow analysis in the buildings equipped with a windcatcher has become a creditable and reliable tool with reasonable accuracy [4, 5].

In this chapter, the optimized model of windcatcher which had been obtained from using RANS K-ε CFD technique is simulated by applying LES CFD method and its results are compared with the achieved results from RANS K-ε method [6]. A commercial CFD software known as CFD-ACE+ by ESI group is employed throughout the simulation.

Fig. 15.2 A typical Malqaf

Fig. 15.3 A typical Badgir

Table 15.1 Design configurations of windcatcher models

Configuration	Categories	Variations
1	Windcatcher's location	Central
2		Right-sided
3		Left-sided
4		Front
5		Back
6	Windcatcher's bottom shape	Two-rod
7		Flat
8		Two-canal
9	Windcatcher's bottom length	10 cm
10		20 cm
11		40 cm
12	Windcatcher's inlet velocity	0.5 m/s
13		1 m/s
14		3 m/s
15		4.5 m/s
16		6 m/s

2 Modelling and Computation

According to a previous numerical study which has been done by the same authors using the standard RANS K-ε method [6], various windcatcher models have been considered. Table 15.1 shows the design configurations of the windcatcher model which have been simulated by CFD-ACE+ software using the standard RANS K-ε method.

That previous study has resulted in the two-canal centred position windcatcher with the bottom length of 10 cm as the optimized two-sided windcatcher among the different models with various windcatcher's location, windcatcher's bottom shape and windcatcher's bottom length with the wind inlet velocity of 3 m/s.

Since LES method is very much time consuming due to its time-dependent flow calculations [7], only the optimized two-sided windcatcher that had been chosen via RANS K-ε CFD method is simulated by using LES CFD method.

Figure 15.4 shows the optimized two-sided windcatcher model in a three dimensional room with the length of 5 m, the width of 4 m, and the height of 3 m; it is assumed that wind blows from right to left.

In this model, the outdoor part of windcatcher's height has been assumed to be 2 m and height of indoor part is 90 cm with the cross sectional area of 80 cm $\times$ 80 cm; length of the windcatcher at its bottom (vertical distance from the roof to the windcatcher's canal) is considered as 10 cm (Fig. 15.5).

For the present simulation, unstructured triangle meshes have been used throughout the model due to obtain better accuracy.

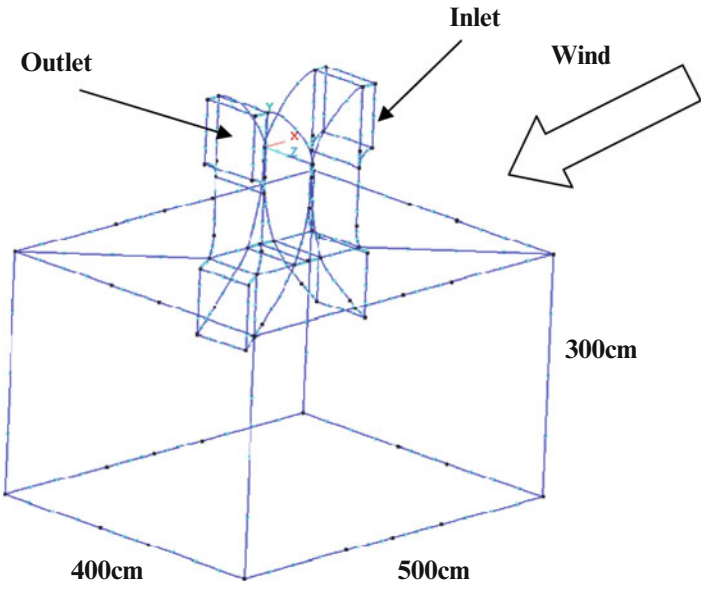


Fig. 15.4 A 3D modelled room fitted with a windcatcher

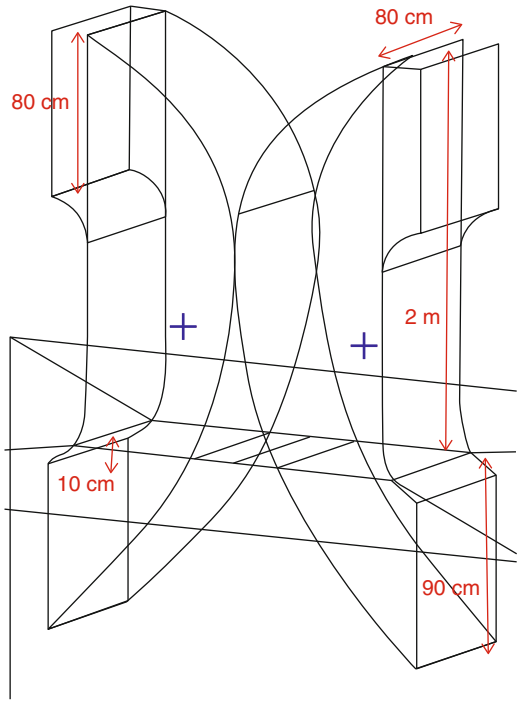


Fig. 15.5 The optimized two-sided windcatcher

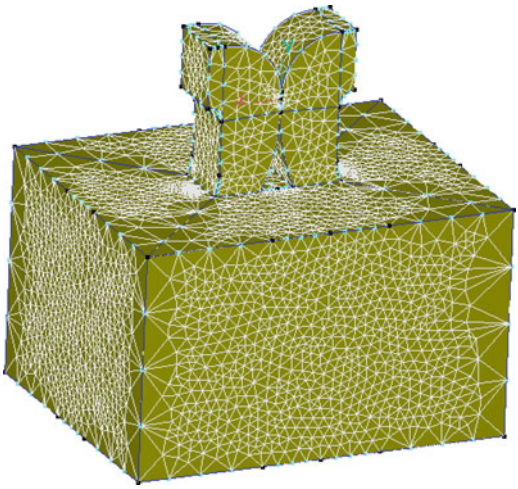


Fig. 15.6 Unstructured triangle meshes in the model

To save on computational efforts and reduced accumulated errors, mesh distribution is less dense in regions of expected near stagnant flow like the room corners.

On the other hand, dense mesh has been used in the living area which is far from the corners of the room (Fig. 15.6).

A grid-independence study has been done for different grid numbers in the model to make sure that the grid pattern used is adequate. Consequently, the total number of grids in the model is around 186,000 and the maximum and the minimum grid areas are about 2×10^{-2} m^2 and 1×10^{-4} m^2, respectively.

LES or "Large Eddy Simulation" is a simulation that directly solves the large scale motion but approximates the small scale motion using turbulence models.

LES requires a flow field where only the large scale components are present; this can be achieved via a filtering process. Within the finite volume method, it is rather sensible and natural to define the filter width as an average of the grid volume.

The flow eddies larger than the filter width are large eddies while eddies smaller than the filter width are small eddies which require modeling. When filtering is performed on the incompressible Navier-Stocks equations, a set of equations very similar to the RANS equations are obtained.

Similar to the RANS methods [1], there are additional terms where a modeling approximation must be introduced in LES method. In the context of LES, these terms are the sub-grid scale turbulent stresses which require sub-grid scale (SGS) models to close the set of governing equations.

For the unsolved subgrid scale turbulent stresses, these are modeled accordingly as:

$$\tau_{ij} - \frac{\delta_{ij}}{3}\tau_{kk} = -2\vartheta_T^{SGS}\overline{S}_{ij} \tag{15.1}$$

$$\overline{S}_{ij} = \frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i} \tag{15.2}$$

Where ϑ_T^{SGS} is the subgrid scale kinematic viscosity and $\overline{S}_{ij}$ is the strain rate of the large scale or resolved field. The form of the subgrid scale eddy viscosity μ_T^{SGS} (noting that $\vartheta_T^{SGS} = \mu_T^{SGS}/\rho$) can be derived by dimensional arguments and is given by:

$$\mu_T^{SGS} = C_s^2 \rho \Delta^2 \left|\overline{S}_{ij}\right| \tag{15.3}$$

$$\left|\overline{S}_{ij}\right| = \sqrt{2\,\overline{S}_{ij}\overline{S}_{ij}} \tag{15.4}$$

Where $\triangle$ is denoted by the grid filter width and the model constant varies between 0.0065 and 0.3 depending on the particular fluid flow problem [8]. In this work, Smagorinsky SGS model is applied and the averaged value is assumed for $\triangle$ which is 0.15325.

3 Results and Discussion

Uniform and stable air flow velocity in the occupied zone (the lower part of a room to the height 2 m) and full circulation across the room are two important human comfort factors for having proper indoor ventilation.

Figure 15.7 shows velocity magnitude along the room at level 1.2 m above the floor for the model (two-canal centred position windcatcher with the bottom length of 10 cm) by using LES method.

It is seen that the velocity magnitude at the level 1.2 m above the floor is approximately stable in the range of 0.48–0.52 m/s across the room at the distance of 1.1–4.8 m from the right wall which is about 3.7 m of 5 m as the total length of the room (74 % of the room's length); consequently, the model provides the uniform flow and results in the most region of the room having velocity in the acceptable range for human comfort; this helps have smooth and less turbulence air flow in the room and confirms the result achieved by using RANS K-ε method.

Some flow path traces of the model are shown in Fig. 15.8; according to the figure, it is seen that the model provides the room with the uniform flow distribution and almost full and complete circulation; although there are some small stagnation areas indicated as "A" and "B" in Fig. 15.8. The round shape of the canals leads to this uniform flow circulation and distribution in the room which verifies the RANS K-ε and LES methods' results as well.

Figure 15.9 compares the velocity magnitudes at level 1.2 m above the floor which has been achieved via RANS K-ε method for the two-canal centred position windcatcher with the bottom length of 10 cm [1] and the one obtained by LES method.

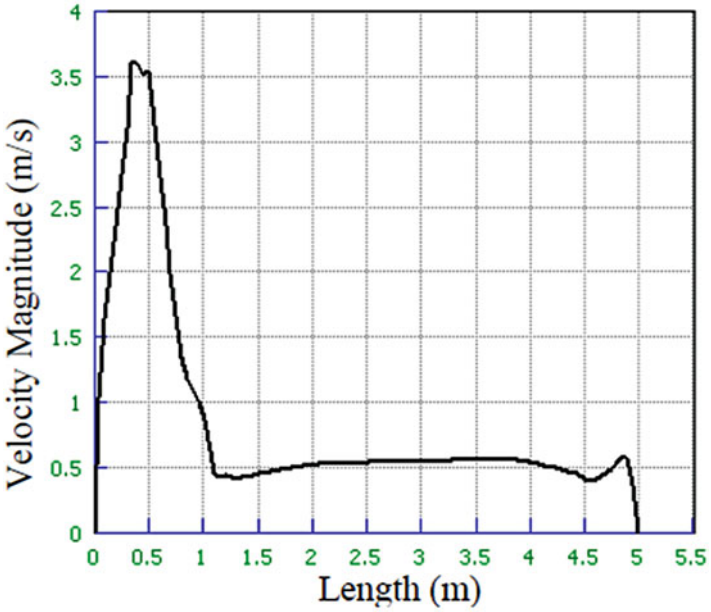


Fig. 15.7 Velocity magnitude for the model at level 1.2 m by using LES method

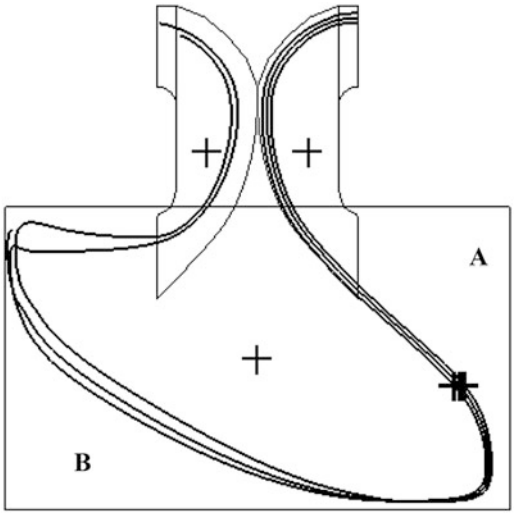


Fig. 15.8 Some traces of flow path corresponding to the model by using LES method

As seen from the graphs, there are only small differences in calculated velocity magnitude via these two computational methods while the trends of velocity changes are very similar to each other. The difference between the two methods' results is about 4.43 % [9].

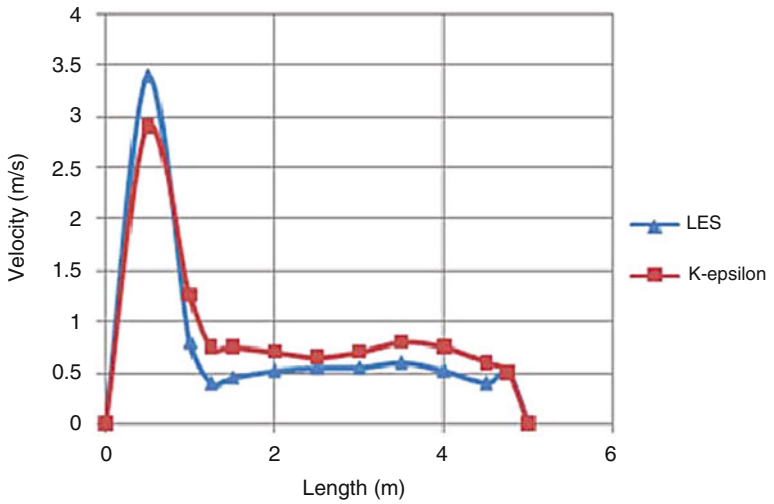


Fig. 15.9 Comparing the calculated velocities in LES and K-epsilon methods for 10 cm bottom length two-canal centred position two-sided windcatcher

4 Conclusion

As a result, LES method is a reliable and accurate technique for simulation and analyzing the ventilation flow thorough a room fitted by a windcatcher.

The simulation by LES confirms that two-canal centred position windcatcher with the bottom length of 10 cm provides full circulation for most part of the room and large region of stable velocity magnitude in the acceptable range of indoor air velocity for human comfort; it is in good agreement with RANS results obtained in previous study [6] as well.

RANS K-ε method is a useful method and good alternative for more accurate simulation methods such as LES where high accuracy is not required.

References

1. El-Shorbagy, A.M.: Design with nature: windcatcher as a paradigm of natural ventilation device in buildings. Int. J. Civil Environ. Eng. **10**, 26–31 (2009)
2. Elmualim, A.A., Awbi, H.B.: Wind tunnel and CFD investigation of the performance of windcatcher ventilation systems. Int. J. Vent. **1**(1), 53–64 (2002)
3. Bahadori, M.N.: Viability of wind towers in achieving summer comfort in the hot arid regions of the Middle-East. Renew. Energy J. **5**(5), 879–892 (1994)
4. Karakatsanis, C., Bahadori, M.N., Vickery, B.J.: Evaluation of pressure coefficients and estimation of air flow rates in buildings employing wind towers. Sol. Energy J. **37**, 363–374 (1986)

5. Elmualim, A.A.: Dynamic modelling of a windcatcher/tower turret for natural ventilation. Build. Serv. Eng. Res. Technol. **27**(3), 165–182 (2006)
6. Niktash, A.R., Huynh, B.P.: Numerical analysis of ventilation flow through a three dimensional room fitted with a two-sided windcatcher. In: Eighteenth Australasian fluid mechanics conference (AFMC), Australia (2012)
7. Davidson, L., Nielsen, P.V.: Large eddy simulations of flow in a three-dimensional ventilated room. In: Fifth international conference on air distributions in rooms, pp. 161–168. Yokohama (1996)
8. Tu, J., Yeoh, G.H., Liu, C.: Computational fluid dynamics – a practical approach, pp. 185–390. Butterworth-Heinemann, Amsterdam (2008)
9. Niktash, A., Huynh, B.P.: Simulation and analysis of ventilation flow through a room caused by a two-sided windcatcher using a LES method. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 1294–1297. London, 2–4 July 2014

Chapter 16
Visualising Dynamic Stall Around a Vertical Axis Wind Turbine Blade Through Particle Image Velocimetry

Okeoghene Eboibi, Jonathan Edwards, Robert Howell, and Louis Angelo Danao

Abstract The vertical axis wind turbine aerodynamics are highly complex and unsteady. Inherent in the operation of VAWTs is the presence of the dynamic stall phenomenon that has a major influence in the overall performance of the rotor. The acquisition of a reliable experimental flow field data set presents a means to increase the level of understanding of VAWT performance and flow physics through visualisations. The method developed in this study includes the setup of the PIV system in the wind tunnel, surface treatment of the VAWT blades, verification of test settings, and image processing and data analysis. The measurement of the flow fields around a VAWT blade at tip speed ratios of $\lambda = 2.5$ and 4 were carried out and the results show significant differences in the stalling characteristics between different λ with increased occurrence of deep and prolonged separation of flow from the blade surface at lower λ. In both cases, however, dynamic stall is observed. The data acquired is an invaluable reference for VAWT flow physics as well as validation of numerical models such as CFD.

Keywords Aerodynamics • Experiments • PIV • Stall • VAWT • Visualizations

O. Eboibi (✉)
Department of Agricultural Engineering, Delta State Polytechnics, Ozoro, Nigeria
e-mail: o.eboibi@gmail.com

J. Edwards
Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK
e-mail: j.m.edwards@sheffield.ac.uk

R. Howell
Department of Mechanical Engineering, University of Sheffield, Sheffield, UK
e-mail: r.howell@sheffield.ac.uk

L.A. Danao
Department of Mechanical Engineering, University of the Philippines, Quezon City, Philippines
e-mail: louisdanao@coe.upd.edu.ph

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_16

Nomenclature

A	Rotor swept area
c	Blade chord
CFD	Computational fluid dynamics
d_c	Obstacle characteristic dimension
d_d	Tracer particle diameter
FOV	Field of view
L	Blade span
LEV	Leading edge vortex
PIV	Particle image velocimetry
R	Rotor radius
Re	Reynolds number
S_k	Stokes number
TEV	Trailing edge vortex
U	Flow velocity
VAWT	Vertical axis wind turbine
α	Angle of attack
θ	Azimuth position
λ	Tip speed ratio
μ_f	Dynamic viscosity of the fluid
ρ_d	Tracer particle density
σ	Solidity
τ	Tracer particle response time
Φ_z	Z-vorticity

1 Introduction

The vertical axis wind turbine (VAWT) has attracted attention in the past decade as a candidate for energy generation in the urban environment. Some of the arguments to the effectiveness of the rotor is its tolerant nature to unsteady wind inflows including changing speeds and changing directions, an inherent characteristic of urban wind. However the VAWT design and operation are complex; currently, the VAWT is not often viewed as a viable alternative means of generating electricity. A fundamental issue in VAWT operation is the cyclic variation in the blade loading even in steady wind conditions. The presence of dynamic stall, a phenomenon that induces extended attachment of flow beyond static stall angles resulting in blade force coefficients that exceed static stall values, complicates the performance prediction of VAWTs using numerical models. There is disagreement between researchers on the overall effect of dynamic stall to the performance of the VAWT.

Early attempts to study the flow physics around VAWT blades were performed in water by Fujisawa and Takeuchi [1] followed by Fujisawa and Shibuya [2]. In both studies, the flow field was visualised by a tracer method with plastic microspheres of 30–50 μm diameter. The VAWT had a single straight blade of a NACA0018 section with chord c = 0.01 m, span L = 0.135 m, and rotor radius R = 0.03 m and made of acrylic resin. The blade was fixed on an end plate with no central shaft to facilitate visualisation all around the rotation. The experiment was carried out in water tunnels with maximum flow velocity U = 0.05 m/s giving a Reynolds number Re = 2,000. The light source was a set of stroboscopes triggered by a photosensor connected to the rotating end plate. The time interval between the two flashes was set to 2 or 3 ms depending on the tip speed ratio, λ, being tested.

Simao Ferreira et al. [3–6] conducted PIV experiments on a larger VAWT in a wind tunnel. The work was performed at $Re = 5 \times 10^5$ and 7×10^5 and λ = 2, 3, and 4. The flow was seeded using a fog machine with approximately 1 μm droplets. The particles were illuminated using a light sheet generated by a Nd:YAG laser (200 mJ/pulse) that was approximately 2 mm thick at the field of view (FOV). A narrowband green filter was used for daylight interference on a CCD camera with 1,374 × 1,040 pixels. The time interval between pulses was set to roughly 8–pixel displacement of particles within a 32 × 32 pixel interrogation window assuming local velocities are 4 times the free stream values. At each selected blade position around the rotation, 30–100 images were taken and analysed with an iterative multi–grid window deformation technique.

This paper builds up on the previous studies and presents an advanced and very detailed methodology for PIV visualisation of the near-blade flow field of a VAWT. The methodology was originally developed by Edwards, through his Ph.D. study [7] and used to validate a CFD model [8]. Danao et al. subsequently used the visualisations in their studies of VAWT performance in unsteady wind [9–11]. Most recently, Eboibi utilised [12] and advanced [13] the methods for his investigations into the effect of chord and solidity on VAWT performance. A collected account of the authors' methodologies is presented here, in high detail. This paper is envisaged as being of significant interest, and of substantial practical use, to the VAWT experimenter.

2 Methodology

The rotor used was a straight–bladed VAWT with three NACA0022 blades of c = 0.04 m each supported by two NACA0026 spokes of c = 0.03 m at 25 % and 75 % blade length positions. The rotor R is 0.35 m and L is 0.6 m giving the VAWT a solidity of σ = 0.34 following the conventional definition (σ = Nc/R) and a wind tunnel blockage ratio of 0.29 (2RL/A, where A: test section area).

2.1 Particle Image Velocimetry

The basic principle of particle image velocimetry (PIV) is to capture a pair of images of the flow under study with a specific time interval between them. Each image is generated by seeding the fluid domain with particles that are assumed to follow the flow dynamics and illuminating a plane of particles using some sort of light source such as a laser sheet. The measured small difference in particle positions between images is used to compute for the velocity flow field.

A Dantec Dynamics 2D PIV system was used for all visualisation tests. The system has a Litron Nano–S–65 Nd:YAG laser which emitted light with a maximum energy of 65 mJ per pulse at a wavelength of 520 nm. A 4 megapixel CCD camera was used to capture the images. A TSI 9306A Six Jet Atomiser generated tracer particles of olive oil that had an approximate size of 2 μm in diameter.

In PIV, particle motion is the measured quantity and is used to represent the fluid velocity field. Therefore, it is extremely important that the particles' tendency to attain velocity equilibrium with the fluid is achieved and can be properly quantified. The ability of tracer particles to follow the flow is measured using Stokes number S_k such that a value $<<0.1$ gives a tracing error of less than 1 % [7, 14]. S_k is defined via Eq. (16.1).

$$S_k = \frac{\tau U}{d_c} \tag{16.1}$$

$$\tau = \frac{\rho_d d_d^2}{18\mu_f} \tag{16.2}$$

where

τ: response time of the particle
U: velocity of the fluid under study
d_c: characteristic dimension of the obstacle
ρ_d: density of the tracer particle
d_d: diameter of the tracer particle
μ_f: dynamic viscosity of the fluid

The response time τ of the tracer particle should be faster than the smallest time scale of the flow and can be deduced using Eq. (16.2). In this study, olive oil (density ≈ 920 kg/m^3) was used as tracer particle. The τ was computed to be about 1.1×10^{-5}. The characteristic dimension used was the chord length of the blade while the velocity of the fluid was set to $U = 40$ m/s derived from the maximum computed local velocity of a comparable CFD simulation at $\lambda = 4$. These values led to a $S_k \approx 0.01$ which meant that the particles should follow fluid streamlines closely and avoid deviating from the flow during rapid changes in flow speed and direction.

Seeding was carried out by running the wind tunnel fan while introducing the particles upstream essentially seeding the entire laboratory room. This was found to

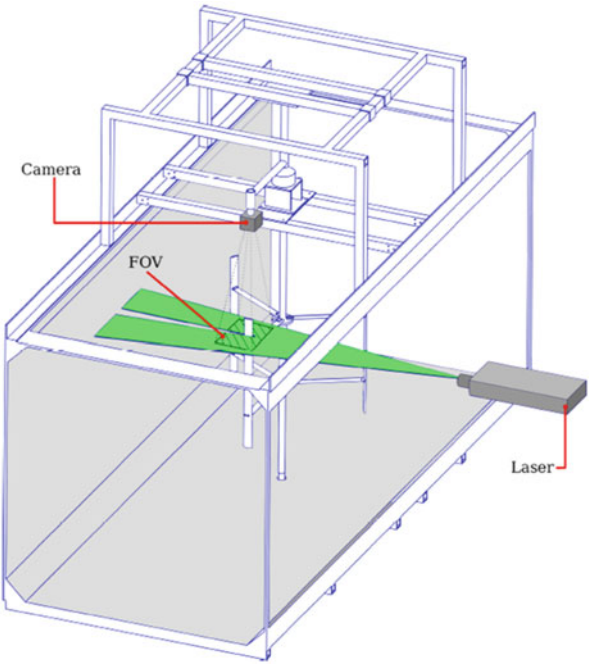


Fig. 16.1 The experiment setup showing the position of the laser sheet relative to the camera and the blade, and the location of the FOV

be the most effective way to achieve adequate and uniform seeding distribution and density. After every 30 min of testing, the seeding was topped–up for 1 min. The laser was mounted on an elevated platform outside the wind tunnel. The laser sheet plane was positioned midway between the support arm and the blade end, avoiding these disturbances to the blade flowfield. This was selected to be within the region that best represented a quasi–2D flow that can be compared to CFD results, yet avoided the support arm obscuring the most critical part of the flowfield.

2.2 Experimental Setup

The setup of the equipment for the PIV campaign is shown in Fig. 16.1. A camera rig was installed on top of the wind tunnel to capture the end view of the blade perpendicular to the laser sheet plane while a 3,000 pulse per rev encoder was mounted at the bottom to measure rotational velocity of the rotor. The camera was mounted on a rotating arm with the axis in line to the VAWT axis thereby permitting the positioning of the FOV to the desired azimuth, the position of a blade around the rotor. The synchronisation of the laser and the camera with the desired azimuth was achieved using an extra channel in the encoder that gave a once–per–rev pulse. The camera and the laser are perpendicular relative to each other and to the general direction of flow.

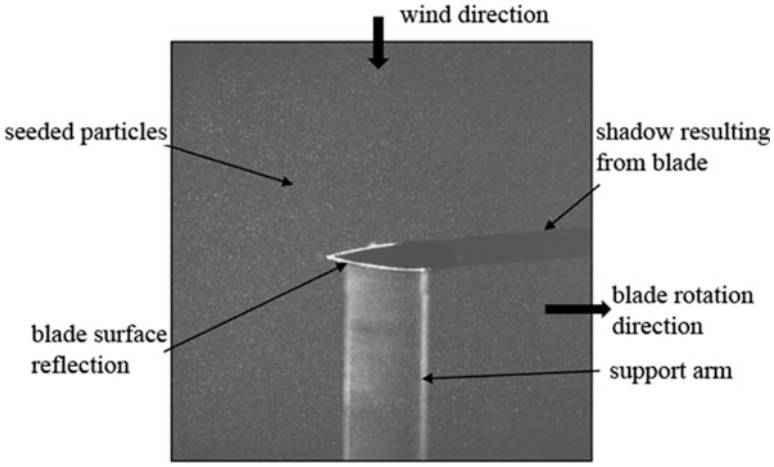


Fig. 16.2 Raw PIV data at a 90° azimuth and U = 6 m/s

2.3 Image Processing

A sample of one of an acquired raw image pair is presented in Fig. 16.2. It shows the 2,048 × 2,048 pixel FOV with an approximate area of 135 mm × 135 mm. Also shown are the flow direction, seeded particles, blade surface reflection, support arm and shadow region. Prior to processing the data, the acquired images were masked to remove areas in the raw image that were unwanted. It was necessary to remove these regions, such as near blade surface reflection and shadow region, since the vectors within these regions could be poorly correlated when processed without being masked out.

The captured image was split into interrogation windows and processed using cross correlation. To determine the velocity vector for each interrogation window, basic cross correlation function, which relates pixel intensity with particle distance, was used. The displacement with the highest correlation is the most likely particle movement within the interrogation window. The process was applied to the entire image and the output is shown in Fig. 16.3.

To remove poorly correlated vectors, a moving average filter with a 3 × 3 window was applied on the cross-correlated images. The moving average filter compares a velocity vector against surrounding vectors and replaces vectors with magnitudes exceeding 20 % of the neighbouring vectors. As the edge of an image is slightly further away from the camera aperture than the centre, the flow velocity scaling factor is not constant. However, in the case of the measurements detailed in this study the camera and FOV were 0.9 m apart, resulting in a negligible difference of only 0.3 % between the scaling factor for the centre and edge of the image.

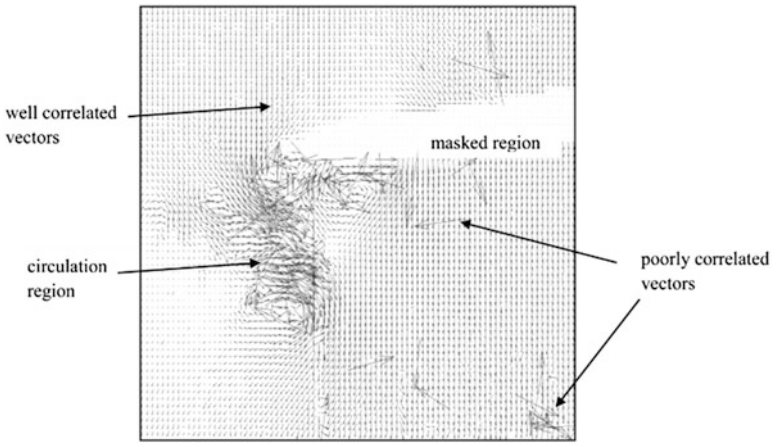


Fig. 16.3 Result of cross-correlation procedure on acquired data

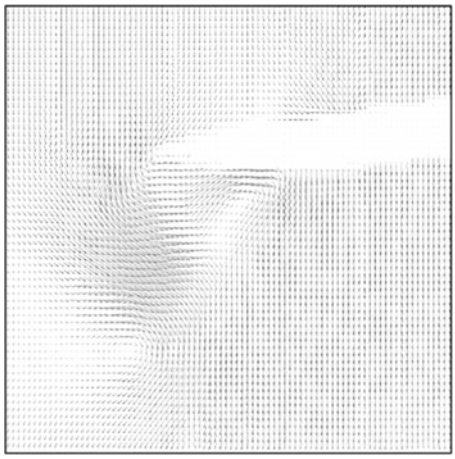


Fig. 16.4 Ensemble averaged velocity flow field for 90° azimuth

Multiple image pairs were ensemble averaged to get a representative picture of the velocity flow field for each azimuth position (Fig. 16.4). The number of image pairs captured for each azimuth position was determined after a systematic study of the effects of sample number on the ensemble averaged velocity plot. It was determined that beyond 50 image pairs there was little observed difference in the velocity plots and 50 was a sufficient number for the author's study of fundamental flow physics. It must be stressed however, that for future investigations, the number of image pairs should be reassessed for significantly different rotors or operating conditions, or for studies with more in-depth scope.

2.4 Blade Surface Treatment

Surface reflection from the blades due to the laser was unavoidable and had to be minimized. The intensity of emitted light from the laser and the image filtering devices used to limit these effects are a major consideration during PIV experiments. Despite the optimal camera position and laser intensity, surface reflection was still present in the images. Hence, it was necessary to treat the blade surface to achieve a tolerable level of reflection to prevent damage to the camera CCD and reduced the negative effects on the accuracy of the results.

The VAWT blades are aluminium and naturally have highly reflective surfaces even at very low laser intensity. In order to reduce the surface reflection, the blade surface was treated following the procedure detailed in [7], similar to that used previously by Pierce and Lu [15]. A thin layer of matt black paint was sprayed on the surface of the blade. Once dry, the blade was then coated in two thin layers of a clear varnish containing a small amount of Rhodamine 6G. Rhodamine 6G absorbs wavelengths of light close to 520 nm, the wavelength of the laser used in this investigation, and so is effective at limiting the laser light reflections. The blade surface was occasionally recoated after long periods of testing, as the Rhodamine coating gradually lost effectiveness.

2.5 Seeding Concentration

The correct seeding density is important in capturing complicated flow details especially in the recirculation or separation zones. To achieve this, seeding must be uniformly distributed and of the required concentration. The atomizer was run from 5 to 10 min at intervals of 1 min. The data were analysed based on the correlated vectors.

Figure 16.5 shows the effects of varying the seeding time on the rejected vectors. At the lowest seeding time of 5 min, 4.1 % of the vectors were poorly correlated, above the 3 % tolerable level. As seeding time is increased, a reduction in poorly correlated vectors is observed. As seen in the figure, 3.0 % of vectors are rejected at 8 min seeding while 2.6 % of vectors are rejected at 10 min seeding time. In consideration of time and small differences between the poorly correlated vectors between 8 and 10 min, 8 min seeding time was adopted for all the PIV testing presented in this study.

2.6 Laser Power

The intensity of the light sheet is directly related to the laser power. At the maximum laser power, the laser can emit highly concentrated light beams. For safety reasons,

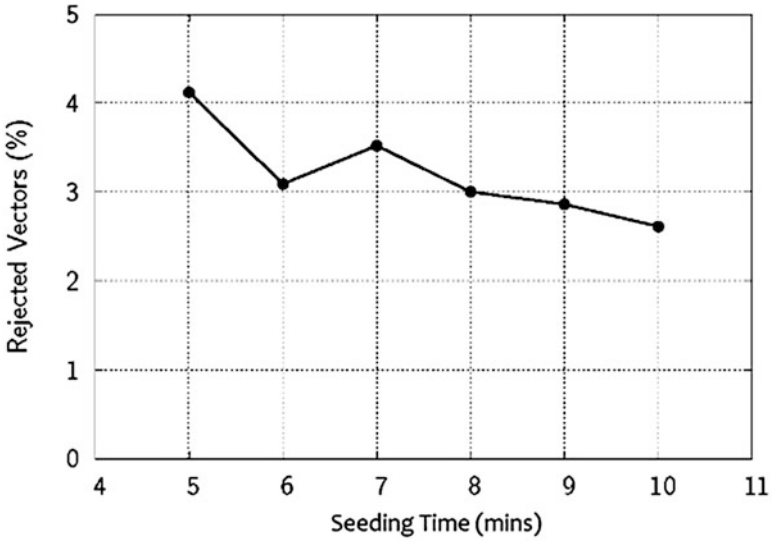


Fig. 16.5 The effects of seeding time on percentage of rejected vectors

and also to prevent the camera sensors from being damaged due to too high light sheet intensity, different levels of the laser power were tested to find a suitable power level. Images were acquired from 70 % to 100 % of the laser power at 10 % intervals – starting from low to high, for safety.

The data were analysed based on the quality of the correlation. Figure 16.6 shows the percentage of poorly correlated vectors for 70 % power level at above 20 %. At 80 % laser power level, poorly correlated vectors drop to 6 %. Further reduction in rejected vectors is observed as laser power is increased to 90 % and 100 %. There is very small difference in percentage of rejected vectors between 90 % and 100 % laser power. The 90 % laser power level is adopted for this study since the rejected vectors are less than 3 %.

2.7 Time Between Pulses

The time interval between image acquisitions is a critical parameter because there must be sufficient movement of the tracer particle within the interrogation window to accurately represent the flow velocity. Too short intervals would result in inaccurate particle displacement computations while too long intervals may result in particles having exited the window altogether. The time interval was varied from 5 to 150 μs. The poorly correlated vectors at 150 μs constitute 57 % of all vectors in the FOV (Fig. 16.7). The high number of rejected vectors is due to the particles that have already exited the window in the second image or new particles have entered into the window in the second image.

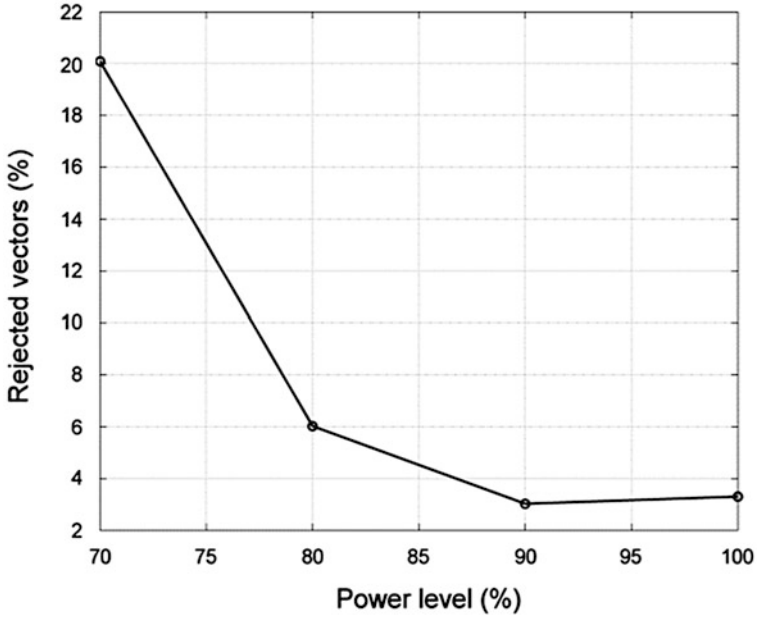


Fig. 16.6 The effects of laser power level on percentage of rejected vectors

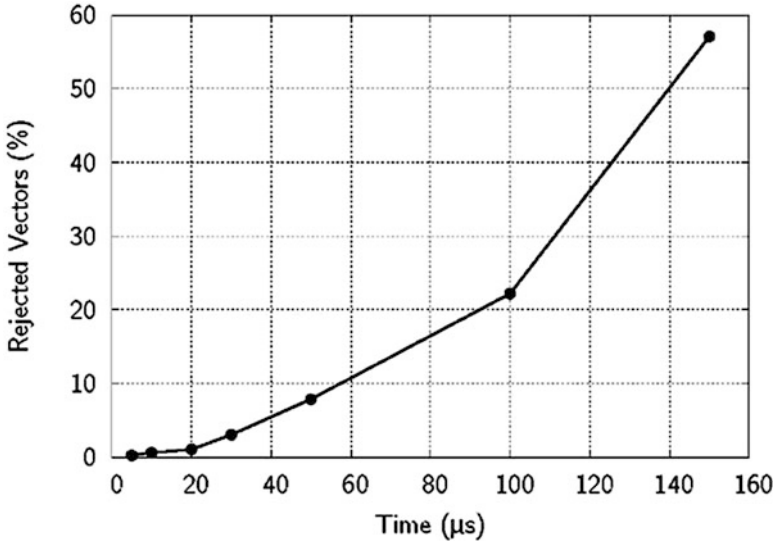


Fig. 16.7 The effect of time between laser pulses on percentage of rejected vectors

At 50 μs, the number poorly correlated vectors correspond to 8 % of rejected vectors. Low levels of poorly correlated vectors are observed at 5, 10 and at 20 μs time intervals, with rejected vectors at 3 % or lower. For this study 20 μs was chosen for the PIV testing.

3 Results and Discussion

One of the major characteristics of VAWT aerodynamics is the dynamic stall phenomenon that is usually expected in a broad range of operating conditions. When an aerofoil is under oscillating motion in a moving fluid, stalling can be considerably delayed beyond the static stall angle. A consequence of this is that static aerofoil analysis and data are no longer suitable because the forces on the blade exceed static stall values and large hysteresis are exhibited with respect to the instantaneous angle of attack. This is more prominent in oscillations with amplitudes in the order of the static stall angle [16]. Dynamic stall is characterised by the shedding of a vortex over the suction surface of an aerofoil under pitching motion in a stream of fluid. If the frequency, amplitude and maximum incidence are sufficiently high, an organised and clearly defined shedding of vortices is observed. Necessary to the analysis of VAWT performance is the proper understanding of the onset of blade stall including the dynamic stall and the subsequent reattachment of flow.

During operation, a VAWT blade experiences cyclic variations in angle of attack α. The blade may undergo stalled and unstalled conditions as well as interact with its own wake and that of other blades' within one rotation. As the VAWT rotates with angular velocity ω in a flow of wind speed U, the angle of attack α varies periodically between positive and negative values. The magnitude of the angle of attack is given by

$$\alpha = \tan^{-1}\left(\frac{\sin\theta}{\lambda + \cos\theta}\right) \tag{16.3}$$

where θ is the azimuth angle and λ is the tip speed ratio. The variation of α resembles a skewed sine wave as shown in Fig. 16.8. This perceived variation is relative to a reference frame attached to the rotating VAWT with its origin at the VAWT axis. As the tip speed ratio λ increases, the skewness of the α-variation reduces and the profile comes closer to a sine wave (zero skew).

A major objective of this study is to make available the stalling characteristics of VAWT blades at two different tip speed ratios through visualisations. The velocity flow field acquired in the PIV campaign does not provide very meaningful information due to difficulties in interpretation and identification of separation and stalled flow. To make more sense of the results, the velocity data was further processed to produce vorticity plots. Vorticity has been chosen since it shows a good indication of separation and reattachment, wake convection and interaction,

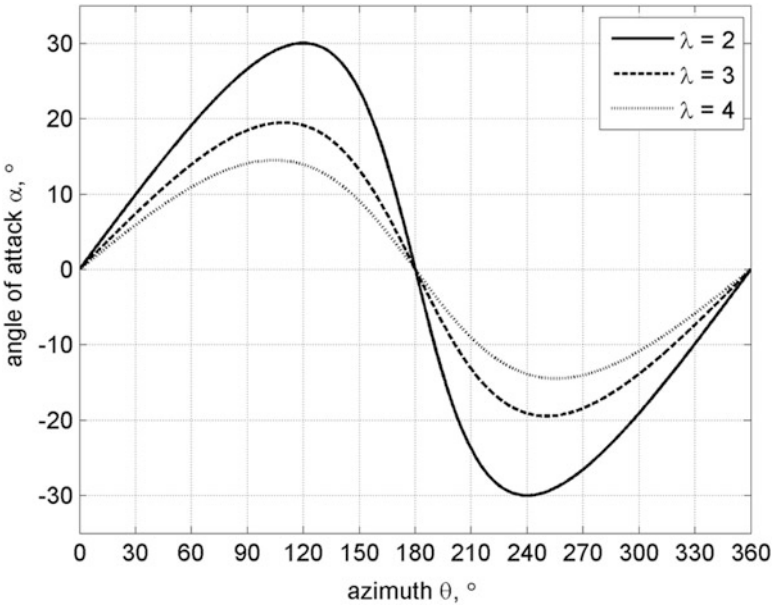


Fig. 16.8 Variation of angle of attack α at different tip speed ratios λ

and presence of shed vortices without the use of streamlines, which can be messy at times. Vorticity is defined as the curl of the velocity field. In two dimensions such as in 2D PIV, it is expressed through Eq. (16.4).

$$\Phi_z = \frac{\partial V}{\partial x} - \frac{\partial U}{\partial y} \tag{16.4}$$

Figure 16.9 shows the vorticity plot of 90° azimuth position. It can be seen that the suction side of the aerofoil is fully separated with a pair of vortices being shed. Red contours indicate anti-clockwise vorticity while blue contours show clockwise vorticity. There are spurious vorticity contours next to the shadow region that does not have velocity data and along the support arm edge that has strong laser light reflection. These are ignored in the analysis through the afore mentioned masking process.

3.1 Flow Field at λ = 2.5

Figure 16.10 shows the vorticity contours for different azimuth positions at U = 6 m/s and λ = 2.5. It was observed that the flow is fully attached from the start

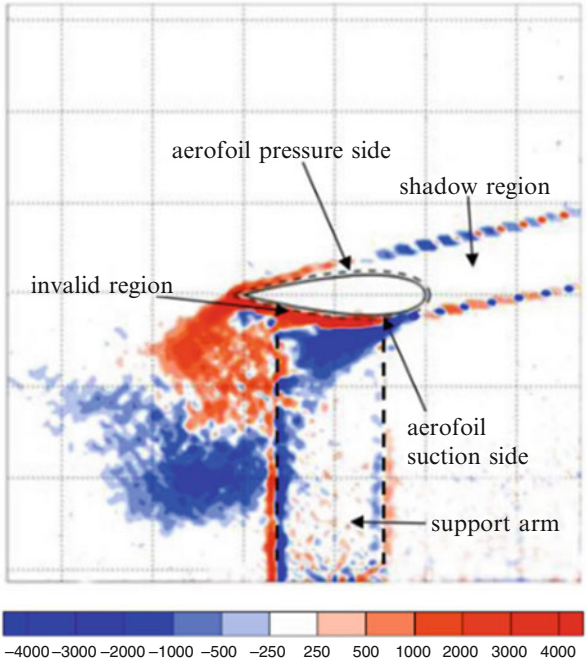


Fig. 16.9 Vorticity plot at U = 6 m/s, λ = 2.5 and θ = 90°

of rotation up to θ = 60° (Fig. 16.10a). At these conditions, classical static aerofoil studies dictate the angle of attack perceived by the blade to have exceeded static stall already ($\alpha_{stall} \approx 11°$) [11]. This is a good indication of the onset of dynamic stall. At θ = 70° (Fig. 16.10b), the growth of a leading edge vortex (LEV) is initiated and is seen to continue to develop until it fully forms and starts to detach from the blade (Fig. 16.10c). Following this, another vortex forms at the trailing edge (TEV) that facilitates the detachment of the LEV (Fig. 16.10d) and itself is detached from the blade surface. The size of the TEV and the LEV continues to increase more than the chord length of the blade until shedding of the pairs of vortices is seen at θ = 130° (Fig. 16.10e). Thereafter, the blade is seen to be under a regime of shedding of a pair of vortices. The onset of flow reattachment to the blade surface is observed at θ = 140° (Fig. 16.10f) where the depth of the stall is seen to reduce and the wake is narrowing. At Fig. 16.10g, blade stall is still deep with the separation point still within quarter chord from the leading edge. Once again, classical static aerofoil theory suggests that the angle of attack at this condition is below stall and the flow should be fully attached by this time. On the contrary, full reattachment only happens after θ = 190° (Fig. 16.10h) which is way past the halfway point in a full rotor revolution, at which α would be expected to be close to zero.

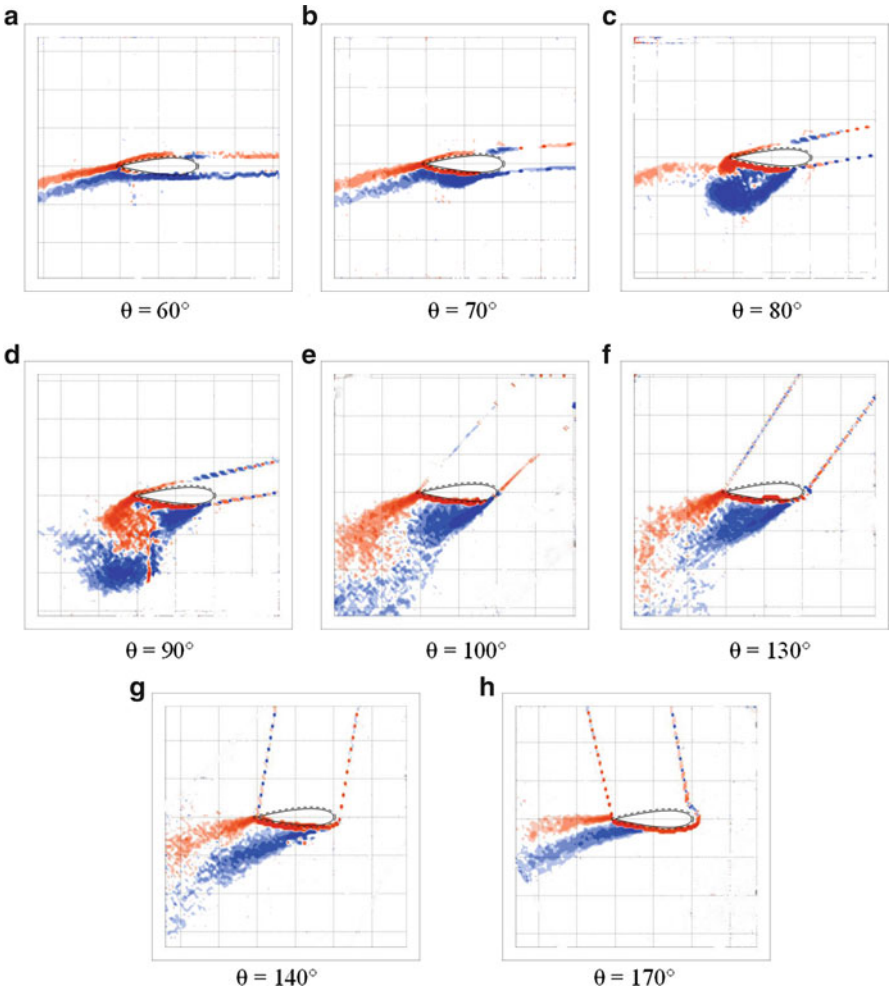


Fig. 16.10 Visualisations of vorticity at U = at 6 m/s and λ = 2.5

3.2 Flow Field at λ = 4.0

The flow is very different at λ = 4 when compared to λ = 2.5 since the onset of the blade stall development at λ = 4 is significantly delayed in the upwind section of the rotation. At the start of the rotation leading to θ = 100° (Fig. 16.11b) and up until θ = 120° (Fig. 16.11c), the flow is observed to be fully attached to the blade surface. The perceived α at θ = 120° azimuth should be greater than static stall conditions. This suggests that the blade is undergoing dynamic stall even at high

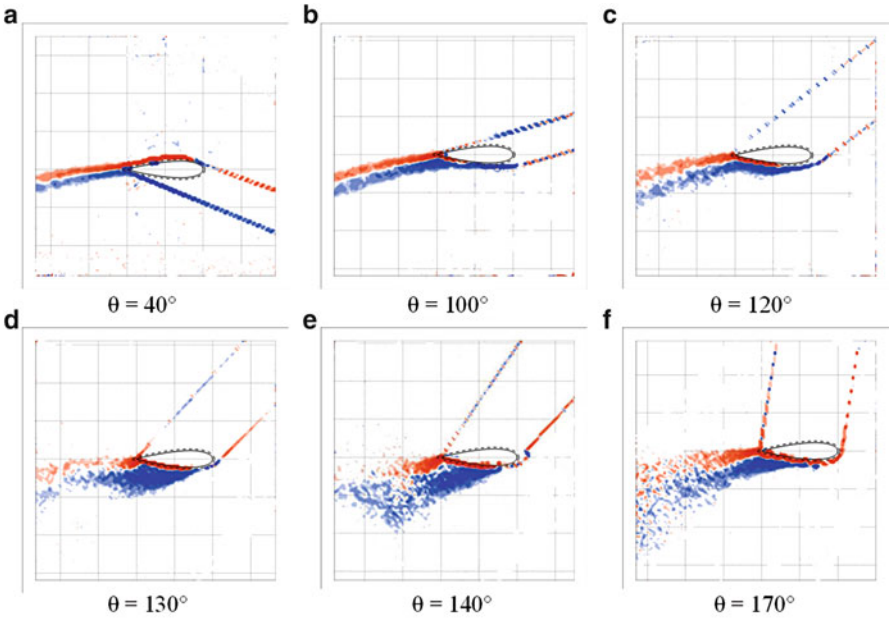


Fig. 16.11 Visualisations of vorticity at U = at 6 m/s and λ = 4.0

λ. Separation is initiated not from the leading edge but from the trailing edge, as observed at θ = 130° (Fig. 16.11d), separation progressively moves forwards into fully separated flow at θ = 140° (Fig. 16.11e). A very small trailing edge roll up is observed but does not develop into a chord-sized vortex as seen at lower λ. Delayed reattachment is also seen at this λ with partial separation still persistent at θ = 170° (Fig. 16.11f) where the expected α should be less than 5°.

4 Conclusion

A method has been developed and presented in the acquisition of visualization data for VAWT blade aerodynamics through Particle Image Velocimetry. A systematic characterization of parameters was carried out to establish the appropriate settings in the conduct of experiments including seeding concentration, laser intensity, and time interval. Results show the presence of dynamic stall at two different operating conditions with significant delay in the onset of stall and reattachment. This study provides valuable information relevant to the study of stall influenced VAWT performance as well as validation data for computational modelling work.

References

1. Fujisawa, N., Takeuchi, M.: Flow visualization and PIV measurement of flow field around a darrieus rotor in dynamic stall. J. Vis. **1**(4), 379–386 (1999)
2. Fujisawa, N., Shibuya, S.: Observations of dynamic stall on Darrieus wind turbine blades. J. Wind Eng. Ind. Aerodyn. **89**(2), 201–214 (2001)
3. Simão Ferreira, C.J., Dixon, K., Hofemann, C., van Kuik, G., van Brussel, G.J.W.: The VAWT in Skew: stereo-PIV and vortex modeling. Paper presented at the 47th AIAA aerospace aciences meeting, 5–8 Jan 2009. Orlando (2009)
4. Simão Ferreira, C.J., van Brussel, G.J.W., Scarano, F., van Kuik, G.: PIV visualization of dynamic stall VAWT and blade load determination. Paper presented at the 46th AIAA aerospace sciences meeting and exhibit, 7–10 Jan 2008. Reno (2008)
5. Simão Ferreira, C.J., van Kuik, G., Van Brussel, G.J.W., Scarano, F.: Visualization by PIV of dynamic stall on a vertical axis wind turbine. Exp. Fluids **46**(1), 97–108 (2009)
6. Simão Ferreira, C.J., van Zuijlen, A., Bijl, H., van Bussel, G., van Kuik, G.: Simulating dynamic stall in a two-dimensional vertical-axis wind turbine: verification and validation with particle image velocimetry data. Wind Energy **13**(1), 1–17 (2010)
7. Edwards, J.: The influence of aerodynamic stall on the performance of VAWT blades. Ph.D. thesis, University of Sheffield, Sheffield (2012)
8. Edwards, J.M., Danao, L.A., Howell, R.J.: Novel experimental power curve determination and computational methods for the performance analysis of vertical axis wind turbines. J. Sol. Energy Eng. **134**(3), 11 (2012)
9. Danao, L.A., Edwards, J., Eboibi, O., Howell, R.: A numerical investigation into the effects of fluctuating wind on the performance of a small scale vertical axis wind turbine. Eng. Lett. **21**(3), 149–157 (2013)
10. Danao, L.A., Edwards, J., Eboibi, O., Howell, R.: A numerical investigation into the influence of unsteady wind on the performance and aerodynamics of a vertical axis wind turbine. Appl. Energy **116**, 111–124 (2014)
11. Danao, L.A.M.: The influence of unsteady wind on the performance and aerodynamics of vertical axis wind turbines. Ph.D. thesis, University of Sheffield, Sheffield (2012)
12. Eboibi, O., Edwards, J., Howell, R., Danao, L.A.: Development of velocity flow field measurement method around a vertical axis wind turbine blade using particle image velocimetry. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, pp 1184–1189. London, 2–4 July 2014
13. Eboibi, O. The influence of blade chord on the aerodynamics and performance of vertical axis wind turbines. Ph.D. thesis, University of Sheffield, Sheffield (2013)
14. Raffel, M., Willert, C.E., Wereley, S.T., Kompenhans, J.: Particle Image Velocimetry: A Practical Guide, 2nd edn. Springer Berlin, Heidelberg (2007)
15. Pierce, A., Lu, F.: New seeding and surface treatment methods for Particle image velocimetry. Paper presented at the 49th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition (2011)
16. Mccroskey, W.J.: The Phenomenon of Dynamic Stall. Ames Research Center, Moffett Field (1981)

Chapter 17
Discrete and Continuous Growth of Deformable Cylinder

Sergei A. Lychev, Alexander V. Manzhirov, and Pavel S. Bychkov

Abstract The finite deformations of the growing cylinder of incompressible elastic material are under consideration. We assume that the deformations are axisymmetric and do not change along the axis of cylinder. The discrete and continuous types of growing are studied. The analytical solutions of the corresponding boundary-value problems are derived. The results of numerical simulation of both discrete and continuous growth are given.

Keywords Additive technologies • Continuous growth • Discrete growth • Finite deformations • Growing solids • Hyperelasticity

1 Introduction

Additive manufacturing technologies include stereolithography, electrolytic deposition, laser and thermal 3D printing, production of 3D integrated circuits and a number of other technologies [1–3]. Actually there is a real boom in the development of additive manufacturing technologies since they allow to reproduce a 3D object of arbitrarily complicated shape (theoretically from any material) with high accuracy and low expenses in short time. However, problems of deformation and strength of products manufactured using such technologies remain still unsolved.

It is essential that the residual stresses can occur in growing bodies through a variety of mechanisms. For example, in layer-by-layer welding technology a heat from parts being welded may cause localized expansion. When the finished melt cools the incompatible distortion appears that cause the residual stresses. Another example occurs during an additive manufacturing when thin film materials with different thermal and crystalline properties are deposed sequentially under different

S.A. Lychev (✉) • P.S. Bychkov
Institute for Problems in Mechanics, pr-t Vernadskogo 101, str. 1, Moscow, 119526 Russia
e-mail: lychevsa@mail.ru; bychkov@ipmnet.ru

A.V. Manzhirov
Institute for Problems in Mechanics, pr-t Vernadskogo 101, str. 1, Moscow 119526, Russia
Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5/1, Moscow 105005, Russia
e-mail: manzh@inbox.ru

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_17

process conditions. In general it is impossible to avoid residual stresses. It leads to undesirable consequences, such as a shape distortion, local discontinuity, loss of stability. In particular, the estimation (and minimization) of possible distortions in stereolithography, the analysis of stability of epitaxial thin-walled structures applicable in micro-electromechanical systems (MEMS) are significant.

In the design of mentioned above additive technologies it is often desirable to minimize distortion and residual stresses, or to produce structures with a predefined distributions of initial stresses. This may be achieved through mathematical modeling of the stress-strain state evolution of a growing body. The present paper deals with the development of this theory.

In the paper the concept of the *growth of a solid* is used. This concept refers to a new branch of continuum mechanics [4–8], therefore it seems appropriate here to clarify the definition of a growing solid. In a broad sense growing process defines the alteration of the body composition occurring in the course of deformation. The growing process may be accompanied by a change of topological properties of the body. We say that the altering of the body composition is the accession of new material points and (or) formation of new constrains between particles already included into the composition. It also should be noted that the change of topological properties can occur without the influx of material and can be caused by the transition of the boundary points into the interior.

In modern continuum mechanics there are a number of different approaches to the studying of the growth phenomenon. For today a large number of papers devoted to mechanics of growing solids have been published. References may be found in the review [9]. Some works of direct relevance to the issues discussed in the paper are mentioned below. In the paper [10] the volumetric grows, in particular the growth of biological tissues, is studied. Article [11] is devoted to the development of geometric methods adopted for the mechanics of incompatible strains arising as the result of the growing process. In the works [4–7, 12] growth is investigated as the continuous process of deposition of strained material surfaces to a deformable 3D body.

It is known that under certain additional assumptions on the continuity of functions defining the stress-strain state of adhered material surfaces the continuous growing process can be considered as the limit of a sequence of discrete processes. However, one can find only few examples concerning the finite deformations of growing solids and comparison of discrete and continuous growth. The aim of the paper is to give such example.

2 Common Definitions

In what follows the geometrical concept of a body that is represented in terms of smooth manifolds is used [13–17]. Let the *body* $\mathfrak{B}$ be the connected abstract subset of a topological space such that the image of $\mathfrak{B}$ may be imbedded into physical space as a region with regular boundary. Furthermore, assume that the body $\mathfrak{B}$ exhibit properties of differentiable manifold [18]. We say that $\mathfrak{p} \in \mathfrak{B}$ are material points.

Assume that they are *simple*, i.e. the local response of the body depends only on first deformation gradient [19].

In the classical continuum mechanics bodies are treated as permanent sets of material points. In mechanics of growing solids one consider the *evolution* of the set $\mathfrak{B}$. Note that the evolution of growing body in the abstract topological (material) space can be very complex and may be described in general in terms of discontinuous mappings [20]. However, under some restrictions on the smoothness of functions describing the growing process the evolution of the body can be presented by a continuous family of bodies ordered with respect to inclusion. In general this family can be associated with a smooth bundle. The dimension of a base of this bundle defines the class of a growing body [5]. In present paper we will considered the simplest class that corresponds to the three-dimensional bundle over one-dimensional base. It has the following interpretation in the terms of continuum mechanics. For a sufficiently large class of additive technologies the growing process may be modelled as a continuous influx of prestressed material surfaces [21] to a growing three-dimensional body. Due to this assumption the growing body can be represented by a one-parameter family of smooth bodies

$$\mathfrak{C} = \{\mathfrak{B}_\alpha\}_{\alpha\in\mathscr{I}}.$$

Here $\mathscr{I}$ is a set of indices that can be finite, countable or continual.

We introduce the notion of *total body* $\mathfrak{B}^*$ and *initial body* $\mathfrak{B}_*$ as follows

$$\mathfrak{B}^* = \bigcup_{\alpha\in\mathscr{I}} \mathfrak{B}_\alpha, \quad \mathfrak{B}_* = \bigcap_{\alpha\in\mathscr{I}} \mathfrak{B}_\alpha.$$

We shall say that the elements of $\mathfrak{C}$ corresponding to interior points of the interval $\mathscr{I}$ are the *intermediate* bodies.

We will distinguish discrete and continuous growth. In the case of a discrete growth the family $\mathfrak{C}$ is a finite sequence of nested sets:

$$\mathfrak{C} \ : \ \mathfrak{B}_1 \subset \mathfrak{B}_2 \subset \ldots \subset \mathfrak{B}_N. \tag{17.1}$$

If the growth is continuous then the family $\mathfrak{C}$ is represented by continuous family of bodies over the interval $\mathscr{I} = (\alpha, \beta) \subset \mathbb{R}$, that satisfy the following condition. There are two-dimensional smooth manifolds Ω_k and no more than a countable set of homeomorphisms such that

$$\begin{gathered}
\Psi_k \ : \ (\Omega_k, \alpha) \to \mathfrak{B}^*, \quad \alpha \in \mathbb{R}, \\
\forall \alpha < \beta \quad \mathfrak{B}_\alpha \subset \mathfrak{B}_\beta, \quad \forall \alpha \ \exists k \ \partial\mathfrak{B}_\alpha = \Psi_k(\Omega_k, \alpha), \\
\bigcup_k \Psi_k(\Omega_k \times \mathscr{I}_k) = \mathfrak{B}^*, \quad \bigcup_k \mathscr{I}_k = \mathscr{I}.
\end{gathered} \tag{17.2}$$

Obviously, the set $\mathfrak{C}$ has countable cardinality $|\mathfrak{C}| = \aleph_1$. Note that manifolds Ω_k represent the preimage of growth boundary. Relations (17.2) introduce on the manifold $\mathfrak{B}^*$ the structure of a smooth bundle [5, 18]. The interval $\mathscr{I}$ represents the base of the bundle while the manifolds Ω_k represent the fibers. If there is a single (universal) homeomorphism $\Psi = \Psi_1$ and a single manifold $\Omega = \Omega_1$, then the bundle becomes trivial. Otherwise topological structure of the growing boundary can vary. Changes in the topology of the preimage of growth boundaries correspond to the phenomenon of selfcontact of the image Ω_k, e.g. the transformation of a cylinder to a torus under the joining of the bases of the cylinder.

Arguing as above we see that growing body can be represented as a bundle over the total body $\mathfrak{B}^*$. Considering the fact that the base of the bundle is one-dimensional we denote the fiber by $\mathfrak{M}_\alpha$, where α is the base coordinate of the fiber. Structural properties of the bundle implies that the fibers are disjoint, and their union coincides with total body $\mathfrak{B}^*$, i.e. $\mathfrak{B}^* = \bigcup\limits_{\gamma \in \mathscr{I}} \mathfrak{M}_\gamma$.

In the process of growth the body $\mathfrak{B}_\alpha$ is presented by open subsets of total body $\mathfrak{B}^*$, whose boundary $\partial\mathfrak{B}_\alpha$ is a union of two separate fibers $\mathfrak{M}_{\alpha'}$ and $\mathfrak{M}_{\beta'}$, i.e. $\partial\mathfrak{B}_\alpha = \mathfrak{M}_{\alpha'} \cup \mathfrak{M}_{\beta'}$. In this case the body $\mathfrak{B}_\alpha$ can be presented as the union of fibers over an open interval $(\alpha', \beta') \subset \mathscr{I}$:

$$\mathfrak{B}_\alpha = \mathfrak{B}(\alpha', \beta') = \bigcup_{\gamma \in (\alpha', \beta')} \mathfrak{M}_\gamma. \tag{17.3}$$

Under the above mentioned assumptions we can define the growing body as the one-parameter family

$$\mathfrak{C} = \left\{ \mathfrak{B}_\alpha = \mathfrak{B}(\alpha_0, \alpha) \middle| \alpha \in \mathscr{I} \right\}.$$

Here α is a continuous parameter that characterizes the evolution of the growing body. As $\alpha \to \alpha_0$ the body degenerates into an infinitely thin film or a point. The obvious generalization of this definition is the following

$$\mathfrak{C} = \left\{ \mathfrak{B}_\gamma = \mathfrak{B}(\alpha_\gamma, \beta_\gamma) \middle| (\alpha_\gamma, \beta_\gamma) \subset \mathscr{I} \right\},$$

where $(\alpha_\gamma, \beta_\gamma)$ is a family of nested intervals.

According to the above definition the boundary of the growing body should be topologically equivalent to a typical fiber, which should be a smooth manifold. Hence the growing boundary must be topologically equivalent to a geometrically closed surface. If the growing boundary is topologically equivalent to a manifold with edge, then the growing body can be defined as follows

$$\mathfrak{C} = \left\{ \mathfrak{B}_\gamma = \mathfrak{B}_0 \cap \mathfrak{B}(\alpha_\gamma, \beta_\gamma) \middle| (\alpha_\gamma, \beta_\gamma) \subset \mathscr{I} \right\}.$$

Here $\mathfrak{B}_0$ is a fixed subset of the material manifold with smooth edge.

In present paper we consider the growth of a hollow circular cylinder of fixed height h. Suppose that the additional material is attached to the lateral surface of the growing cylinder. If the coordinate charts correspond to the placement coordinates in an actual configuration then the image of the set $\mathfrak{B}_0$ corresponds to a sufficiently large parallelepiped which height is equal to the height of the growing cylinder.

3 Stress-Strain State of the Body-Fiber

The stress-strain state of a growing body fundamentally differs from the stress-strain state of solids considered in classical solid mechanics. The most important is the fact that growing body has no natural (stress-free) configuration. Stress-strain state for growing bodies may be modelled in the framework of the theory of inhomogeneity developed in [13–15].

The representation of a body as a bundle of a smooth manifold allows one to use additional hypothesis concerning the properties of the fibers. In particular one can assume that each individual fiber has a natural configuration. Such hypothesis is adopted in present paper.

In order to describe stress-strain of a growing body it is necessary to determine the stress-strain state for a fiber as its structural element. In the case of discrete growth this structural element is a three-dimensional body $\mathfrak{B}_{n+1} \setminus \mathfrak{B}_n$, corresponding to the increment of the sequence (17.1). In the case of continuous growth the material surface $\mathfrak{M}_\gamma$, which corresponds to a fiber of a bundle (17.3), plays the role of structural element. Within the present work we assume that each separate body-fiber has a natural configuration immersed in Euclidean space. It is clear that the assembly of body-fibers have no such configuration.

In the case of discrete growth each assembly consists of a finite number of nested hollow cylindrical bodies. The second case is more abstract. It corresponds to the assembly of the continuum family of two-dimensional *material surfaces*.

We assume that the material of a body-fibers is hyperelastic and incompressible. Then the stress-strain state can be determined analytically by universal solutions of Rivlin–Ericksen type [19].

Let the image of stress-free (natural) configuration of the body-fiber $\mathfrak{B}_{n+1} \setminus \mathfrak{B}_n$ is embedded into physical (Euclidean) space $\mathscr{E}$. This embedding can be defined by the vector field of placements presented in Cartesian basis $\{\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3\}$ by decomposition $\mathbf{X} = X^m \mathbf{i}_m$. Here $\{X^1, X^2, X^3\}$ are Cartesian coordinates. Suppose that the deformation of the body-fiber is defined by the map $\mathbf{X} \mapsto \mathbf{x}$. We assume that this map has a symmetry relative to the axial axis of the cylindrical fibers and does not depend on coordinate X^3.

For a more compact formulation of the kinematic relations we use cylindrical coordinates $\{R, \Theta, Z\}$:

$$X^1 = R\cos\Theta, \quad X^2 = R\sin\Theta, \quad X^3 = Z.$$

The cylindrical coordinates define the local basis $\{\mathbf{e}_R, \mathbf{e}_\Theta, \mathbf{e}_Z\}$ and reciprocal basis $\{\mathbf{e}^R, \mathbf{e}^\Theta, \mathbf{e}^Z\}$. Elements of them can be presented by the decompositions

$$\mathbf{e}_R = \mathbf{e}^R = \mathbf{i}_1 \cos\Theta + \mathbf{i}_2 \sin\Theta,$$

$$\mathbf{e}_\Theta = -\mathbf{i}_1 R \sin\Theta + \mathbf{i}_2 R \cos\Theta,\ \mathbf{e}^\Theta = \frac{\mathbf{e}_\Theta}{R^2},\ \mathbf{e}^Z = \mathbf{e}_Z = \mathbf{i}_3.$$

The reference positions of material points in simplest form can be written as $\mathbf{X} = R\mathbf{e}_R + Z\mathbf{e}_Z$. Taking into account the central symmetry, the independence with respect to vertical coordinate Z, and the condition of incompressibility $|d\mathbf{x}/d\mathbf{X}| = 1$ we arrive at the following family of mapping (universal deformations belonging to the family 3 according to the classification given in [19])

$$\mathbf{x}(\mathbf{X}) = \mathbf{e}_R \sqrt{(\mathbf{e}^R \cdot \mathbf{X})^2 + a} + \mathbf{e}_Z \otimes \mathbf{e}^Z \cdot \mathbf{X}. \tag{17.4}$$

Here a is a deformation parameter that represents the change of the outer cylindrical surface radius. The deformation gradient $\boldsymbol{F}$ and left Cauchy–Green tensor $\boldsymbol{B} = \boldsymbol{F} \cdot \boldsymbol{F}^*$ (hereinafter the symbol $*$ denotes the transpose) are determined in terms of local basis corresponded to the reference position as follows

$$\boldsymbol{F} = \frac{R}{\sqrt{R^2 + a}} \mathbf{e}_R \otimes \mathbf{e}_R + \frac{\sqrt{R^2 + a}}{R^3} \mathbf{e}_\Theta \otimes \mathbf{e}_\Theta + \mathbf{e}_Z \otimes \mathbf{e}_Z.$$

The decomposition of tensor $\boldsymbol{B}$ and its inverse in the terms of the elements of the local basis corresponded to the actual position, i.e.

$$\mathbf{e}_R = \mathbf{e}_r, \quad \mathbf{e}_\Theta = \frac{\sqrt{r^2 - a}}{r} \mathbf{e}_\theta, \quad \mathbf{e}_Z = \mathbf{e}_z,$$

have the forms

$$\boldsymbol{B} = \frac{r^2 - a}{r^2} \mathbf{e}_r \otimes \mathbf{e}_r + \frac{1}{r^2 - a} \mathbf{e}_\theta \otimes \mathbf{e}_\theta + \mathbf{e}_z \otimes \mathbf{e}_z, \tag{17.5}$$

$$\boldsymbol{B}^{-1} = \frac{r^2}{r^2 - a} \mathbf{e}_r \otimes \mathbf{e}_r + \frac{r^2 - a}{r^4} \mathbf{e}_\theta \otimes \mathbf{e}_\theta + \mathbf{e}_z \otimes \mathbf{e}_z.$$

If the cylindrical body-fiber is produced from an incompressible material of Mooney–Rivlin type then the strain energy can be presented as a linear function of the first $I_1 = I_1(\boldsymbol{B})$ and second $I_2 = I_2(\boldsymbol{B})$ invariants of tensor $\boldsymbol{B}$, i.e.:

$$W(I_1, I_2) = C_1(I_1 - 3) + C_2(I_2 - 3),$$

$$I_1 = \mathrm{Tr}\boldsymbol{B} = 3 + \frac{a^2}{r^2(r^2 - a)}, \quad I_2 = I_1.$$

Here C_1, C_2 are material constants. Under the conditions of incompressibility we have the following decomposition of Cauchy stress tensor [22]

$$\boldsymbol{T} = -p\boldsymbol{I} + J_1\boldsymbol{B} + J_{-1}\boldsymbol{B}^{-1},$$

where p is hydrostatic pressure, $J_1 = 2\partial W/\partial I_1 = 2C_1$ and $J_{-1} = -2\partial W/\partial I_2 = -2C_2$ are coefficients of reaction, and $\boldsymbol{I}$ is a unit tensor. Note that constants C_1, C_2 can be defined by pair of engineering constants μ, β, i.e.:

$$C_1 = \mu(1+\beta)/4, \quad C_2 = \mu(1-\beta)/4.$$

Here μ corresponds to the shear modulus and β defines the additional parameter for nonlinear response. From thermodynamical restriction it follows that $-1 < \beta < 1$ [19].

After simple calculations we obtain the following:

$$\boldsymbol{T} = T^{rr}\mathbf{e}_r \otimes \mathbf{e}_r + T^{\theta\theta}\mathbf{e}_\theta \otimes \mathbf{e}_\theta + T^{zz}\mathbf{e}_z \otimes \mathbf{e}_z, \quad T^{rr} = -p + J_1\frac{r^2-a}{r^2} + J_{-1}\frac{r^2}{r^2-a},$$

$$T^{\theta\theta} = -\frac{p}{r^2} + J_1\frac{1}{r^2-a} + J_{-1}\frac{r^2-a}{r^4}, \quad T^{zz} = -p + J_1 + J_{-1}.$$

Hydrostatic stress component p can be determined by the equilibrium equation $\nabla \cdot \boldsymbol{T} = \mathbf{0}$. Integrating this equation with respect to r we get

$$T^{rr} = \frac{\mu}{2}\left(\ln\frac{r^2-a}{r^2} - \frac{a}{r^2}\right) + p_0, \quad T^{\theta\theta} = \frac{T^{rr}}{r^2} + \frac{\mu}{r^2}\left(\frac{r^2}{r^2-a} - \frac{r^2-a}{r^2}\right),$$

$$T^{zz} = T^{rr} + \mu\, a\,\frac{r^2 - (1+\beta)\,a/2}{r^2(r^2-a)}, \tag{17.6}$$

where p_0 is the constant of integration. Note that in the terms of physical basis $\mathbf{e}_{\langle r\rangle} = \mathbf{e}_r$, $\mathbf{e}_{\langle\theta\rangle} = \mathbf{e}_\theta/r$, $\mathbf{e}_{\langle z\rangle} = \mathbf{e}_z$ the stresses have the form

$$T_{\langle rr\rangle} = T^{rr}, \quad T_{\langle\theta\theta\rangle} = T^{\theta\theta}r^2, \quad T_{\langle zz\rangle} = T^{zz}.$$

Thus, the deformations and stresses can be defined up to the parameters a and p_0. This implies that the boundary conditions may be satisfied exactly only on the cylindrical surfaces if the constant hydrostatic load intensity p_i and p_e are given

$$\boldsymbol{T}\cdot\mathbf{e}^r\big|_{r=r_i} = p_i\mathbf{e}_r, \quad \boldsymbol{T}\cdot\mathbf{e}^r\big|_{r=r_e} = p_e\mathbf{e}_r, \tag{17.7}$$

Here r_i, r_e are the radii of the inner and outer cylindrical boundary surfaces.

Substituting expressions for the radial component of the stress (17.6) to the boundary conditions (17.7) and taking into account the kinematic relations (17.4) we obtain the system of equations

$$\begin{cases} \frac{\mu}{2}\left(\ln\frac{R_i^2}{R_i^2+a} - \frac{a}{R_i^2+a}\right) + p_0 = p_i, \\ \frac{\mu}{2}\left(\ln\frac{R_e^2}{R_e^2+a} - \frac{a}{R_e^2+a}\right) + p_0 = p_e, \end{cases} \tag{17.8}$$

where $R_i = \sqrt{r_i^2 - a}$, $R_e = \sqrt{r_e^2 - a}$ are reference values of the radii of the boundary surfaces. After eliminating of the parameter p_0 from the resulting system we obtain the equation with respect to the parameter a:

$$\ln\left(\frac{R_i^2}{R_e^2}\frac{R_e^2+a}{R_i^2+a}\right) = 2\,\frac{p_i - p_e}{\mu} + a\,\frac{R_e^2 - R_i^2}{(R_i^2+a)(R_e^2+a)}.$$

Let $x = a/R_e^2$ be a new variable that can be interpreted as a relative deformation parameter. Potentiating the left and right hand-sides of the resulting expression we obtain the equation with respect to x

$$F = 0, \quad F = \frac{1+x}{\gamma+x} - \frac{A}{\gamma}\,e^{x\frac{1-\gamma}{(1+x)(\gamma+x)}}. \tag{17.9}$$

Here $\gamma = R_i^2/R_e^2$, $A = \exp[2(p_i - p_e)/\mu]$. Because the internal radius in the reference configuration has always positive value then $0 < \gamma < 1$. Furthermore $x > -\gamma$. Limit relations

$$\lim_{x\to-\gamma} F = \infty, \quad \lim_{x\to\infty} F = 1 - \frac{A}{\gamma}$$

show that Eq. (17.9) has a solution only if $A > \gamma$, i.e. there is a limit for the difference of hydrostatic load intensities: $p_i - p_e > \mu/2\ln\gamma$.

If the value of x is determined then the absolute deformation parameter $a = R_e^1 x$ can be calculated and the corresponding value p_0 may be also determined. Thus, for given values of hydrostatic loads p_i, p_e and radii of the boundary surfaces R_i, R_e one can define the parameters a, p_0 and all components of strain tensors (17.5) and stresses (17.6) as well.

4 Discrete Growing

Consider a finite set of bodies. Let the elements of this set be the circular hollow cylinders of equal height h (in natural configuration). The motion (17.4) transform them to the hollow cylinders of the same height, but of another radii. Such deformation can be realized, e.g. by expanding the hollow cylinder which base lie on the smooth rigid slabs. We assume that the images of the actual configuration of the cylinders are pairwise disjoint and their union is a connected set. The final

composite body can be treated as a result of discrete growth because cylindrical parts cannot deform independently after joining.

Let N be the number of cylindrical parts. Assume that the following scenario of growth is realized. On the first step the joining of the first and second body-fibers is performed. A composite body appears which we call the first assembly. Then the third body is joint to the composite body, etc. On the internal $r = r^1_{i,n}$ and the outer boundary $r = r^n_{e,n}$ of this composite bodies the uniformly distributed pressure $p_{i,n}$ and $p_{e,n}$ are defined

$$\boldsymbol{T}\cdot\mathbf{e}^r\big|_{r=r^1_{i,n}} = p_{i,n}\mathbf{e}_r, \quad \boldsymbol{T}\cdot\mathbf{e}^r\big|_{r=r^n_{e,n}} = p_{e,n}\mathbf{e}_r. \tag{17.10}$$

Index n indicates the number of assembly. The indexing in the notation of intensity of hydrostatic loads $p_{i,n}$, $p_{e,n}$ shows that they may vary during the growing process. Suppose that the contact between body-fibers is ideal, i.e. inner surface of k-th fiber and the outer surface of $k+1$-th fiber in the actual configuration are the same and stresses on them are in equilibrium:

$$\boldsymbol{T}\cdot\mathbf{e}^r\big|_{r=r^k_{e,n}} = \boldsymbol{T}\cdot\mathbf{e}^r\big|_{r=r^{k+1}_{i,n}}, \quad r^k_{e,n} = r^{k+1}_{i,n},\ k = 1,2,\dots n-1. \tag{17.11}$$

The deformation parameters a^k_n and parameters $p^k_{0,n}$, $k = 1,2,\dots,n$ may be found from the system of $2n$ nonlinear equations (17.10) and (17.11). Taking into account (17.4) and (17.6) we get

$$\begin{aligned}
&\frac{\mu}{2}\left[\ln\frac{(R^1_i)^2}{(R^1_i)^2+a^1_n} - \frac{a^1_n}{(R^1_i)^2+a^1_n}\right] + p^1_{0,n} = p_{i,n},\\
&\frac{\mu}{2}\left[\ln\frac{(R^n_e)^2}{(R^n_e)^2+a^n_n} - \frac{a^n_n}{(R^n_e)^2+a^n_n}\right] + p^n_{0,n} = p_{e,n},\\
&\frac{\mu}{2}\left[\ln\frac{(R^k_e)^2}{(R^k_e)^2+a^k_n} - \frac{a^k_n}{(R^k_e)^2+a^k_n}\right] + p^k_{0,n} =\\
&\qquad = \frac{\mu}{2}\left[\ln\frac{(R^{k+1}_i)^2}{(R^{k+1}_i)^2+a^{k+1}_n} - \frac{a^{k+1}_n}{(R^{k+1}_i)^2+a^{k+1}_n}\right] + p^{k+1}_{0,n},\\
&(R^k_e)^2 + a^k_n = (R^{k+1}_i)^2 + a^{k+1}_n,\ k = 1,2,\dots,n-1.
\end{aligned} \tag{17.12}$$

The system of equation (17.12) may be transform to simpler form if one introduce the following variables and parameters:

$$\alpha_k = 1 + A_k/(R^k_e)^2, \beta_k = \gamma_k + A_k/(R^k_e)^2, \nu_k = (R^1_e)^2/(R^k_e)^2, x_n = a^1_n/(R^1_e)^2,$$

$$W_n = e^{2\frac{p_{i,n}-p_{e,n}}{\mu}}\ A_1 = 0,\ A_k = \sum_{p=2}^{k}\left((R^{p-1}_e)^2 - (R^p_i)^2\right),\ k = 2,3,\dots,n, \gamma_k = \left(\frac{R^k_i}{R^k_e}\right)^2.$$

Eliminating $p^k_{0,n}$ and potentiating of the left and right hand sides of above equation, we obtain

$$\prod_{k=1}^{n} \gamma_k \frac{\alpha_k + \nu_k x_n}{\beta_k + \nu_k x_n} = W_n \exp\left[\sum_{k=1}^{n} (1-\gamma_k) \frac{\nu_k x_n + \alpha_k - 1}{(\beta_k + \nu_k x_n)(\alpha_k + \nu_k x_n)}\right]. \quad (17.13)$$

We distinguish the following types of growth

1. Growth with a prescribed reference geometry. Here we suppose that the geometrical characteristics of body-fibers in the image of natural configuration are given, i.e. the reference radii of the unstrained body-fibers R^k_i and R^k_e are known.
2. Growth with a given actual geometry. The position of growing boundaries in the image of the actual configuration $\mathscr{R}_n$ and the thickness of body-fibers in the reference configuration are known, i.e. the values $\Delta^k = R^k_e - R^k_i$, $k = 1, \ldots, n$ are prescribed.

Let us consider these types of growth in detail.

Type 1. Growth with a given reference geometry. Using given values of the reference radius R^k_i, R^k_e, $k = 1, \ldots, n$ one can calculate values of α_k, γ_k, β_k, ν_k, and taking into account given values of hydrostatic load $p_{i,n}$, $p_{e,n}$ calculate the values of W_n. As a result one obtain a series of uncoupled non-linear equations (17.13). The solutions of this equations determines deformation parameters x_n independently. Thereafter one may calculate a^k_n, $k = 1, \ldots, n$ and define stresses by the relations (17.6).

Type 2. Growth with a given actual geometry. In this case the reference radii of the body-fibers are not known a priori, and Eqs. (17.13) have to be supplemented by additional equations that define the radius of growing boundary $\mathscr{R}_n$ in actual configurations

$$r^n_{e,n} = \sqrt{(R^n_e)^2 + a^n_n} = \mathscr{R}_n.$$

To analyze the system of equations firstly allocate in the left and right hand sides of Eq. (17.13) the terms corresponding to the n-th body-fiber, i.e.:

$$\gamma_n \frac{\alpha_n + \nu_n x_n}{\beta_n + \nu_n x_n} \prod_{k=1}^{n-1} \gamma_k \frac{\alpha_k + \nu_k x_n}{\beta_k + \nu_k x_n} =$$

$$= W_n \exp\left[\frac{(1-\gamma_n)(\nu_n x_n + \alpha_n - 1)}{(\beta_n + \nu_n x_n)(\alpha_n + \nu_n x_n)} + \sum_{k=1}^{n-1} \frac{(1-\gamma_k)(\nu_k x_n + \alpha_k - 1)}{(\beta_k + \nu_k x_n)(\alpha_k + \nu_k x_n)}\right]. \quad (17.14)$$

Unlike *type 1* the values of γ_n, α_n, β_n, ν_n, can't be defined a priori, because they depend on the dimensionless deformation parameter x_n. In fact, since

$$(R^n_e)^n = \mathscr{R}^2_n - a^n_n = \mathscr{R}^2_n - (R^1_e)^2 x_n - A_n, \quad R^n_i = R^n_e - \Delta^n,$$

the expression for A_n

$$A_n = A_{n-1} + (R_e^{n-1})^2 - (R_i^n)^2$$

is the algebraic equation whose solution determines A_n through the parameters with indices $m < n$, actual radius of the growing border $\mathscr{R}_n$ and the deformation parameter x_n. Substituting these expressions into Eq. (17.14) leads to the explicit form of non-linear equations:

$$\left(1 + \frac{2\,\xi_n}{\sigma_n - x_n}\right)^2 \frac{\zeta_n}{\phi_{n-1} + x_n} \prod_{k=1}^{n-1} \gamma_k \frac{\alpha_k + \nu_k x_n}{\beta_k + \nu_k x_n} =$$

$$= W_n \exp\left[\frac{\sigma_n - x_n - \xi_n}{\phi_{n-1} + x_n}\left(1 - \frac{(\sigma_n - x_n)^2}{4\zeta_n \xi_n}\right) + \sum_{k=1}^{n-1} \frac{(1-\gamma_k)\,(\nu_k x_n + \alpha_k - 1)}{(\beta_k + \nu_k x_n)(\alpha_k + \nu_k x_n)}\right], \tag{17.15}$$

where

$$\xi_n = \left(\frac{\Delta^n}{R_e^1}\right)^2, \quad \sigma_n = \frac{\mathscr{H}_n}{(R_e^1)^2}, \quad \zeta_n = \left(\frac{\mathscr{R}_n}{R_e^1}\right)^2,$$

$$\phi_{n-1} = \frac{A_{n-1} + (R_e^{n-1})^2}{(R_e^1)^2}, \quad \mathscr{H}_n = \mathscr{R}_n^2 - A_{n-1} - (R_e^{n-1})^2 + (\Delta_n)^2.$$

5 Continuous Growth

In the case of continuous growth it is convenient to introduce an intermediate configuration which image is not free from stresses but it may be immersed in Euclidean space. If such intermediate configurations perform additional conditions, i.e. the local configuration in the neighborhood of any interior point does not change during the growing process, then the total local deformation, which transforms the neighborhood of material point to the actual state, can be presented as a multiplicative decomposition

$$\boldsymbol{H} = \boldsymbol{F} \cdot \boldsymbol{K}, \quad \text{rot } F = \boldsymbol{0}, \quad \text{rot } K \neq \boldsymbol{0}, \quad \dot{\boldsymbol{K}} = \boldsymbol{0},$$

where $\dot{\boldsymbol{K}}$ denotes the derivative with respect to time, or to a time like parameter.

In general, the intermediate configuration is not compatible with actual external fields acting on the growing body. Thus one must attach a system of fictitious mass and surface forces which have the character of Eshelby forces [16].

Bearing in mind the idea of a bundle as a continual family of material surfaces, which separately has natural (stress-free) configuration in Euclidean space, we can

present the system of fictitious forces by continuous family of surface loads that hold the material surfaces in assembly.

Thus, with each material surface one can associate the deformation that transforms the surface from the intermediate configuration to the unstressed state

$$\mathbf{K} = \frac{R}{\sqrt{R^2+\alpha(R)}}\,\mathbf{e}_R \otimes \mathbf{e}^R + \frac{\sqrt{R^2+\alpha(R)}}{R}\,\mathbf{e}_\Theta \otimes \mathbf{e}^\Theta + \mathbf{e}_Z \otimes \mathbf{e}^Z.$$

Note that the union of these fields determines a single field of linear transformations (a three-dimensional field of second-rank tensors) which are not gradients of any vector field in Euclidean space.

The body deforms from the intermediate configuration to the actual in conventional sense. So it is subjected to the deformation $\mathbf{F}$ which has the form

$$\mathbf{F} = \frac{\tilde{r}}{\sqrt{\tilde{r}^2+A(t)}}\mathbf{e}_{\tilde{r}} \otimes \mathbf{e}^{\tilde{r}} + \frac{\sqrt{\tilde{r}^2+A(t)}}{\tilde{r}}\mathbf{e}_{\tilde{\theta}} \otimes \mathbf{e}^{\tilde{\theta}} + \mathbf{e}_{\tilde{z}} \otimes \mathbf{e}^{\tilde{z}}.$$

Here variable with tilde corresponds to the intermediate configuration. In this case the total distortion and corresponding strain are the following

$$\mathbf{H} = \mathbf{F}\cdot\mathbf{K} = \frac{\sqrt{\tilde{r}^2-\alpha(\tilde{r})}}{\sqrt{\tilde{r}^2+A(t)}}\mathbf{e}_{\tilde{r}} \otimes \mathbf{e}^{\tilde{r}} + \frac{\sqrt{\tilde{r}^2+A(t)}}{\sqrt{\tilde{r}^2-\alpha(\tilde{r})}}\mathbf{e}_{\tilde{\theta}} \otimes \mathbf{e}^{\tilde{\theta}} + \mathbf{e}_{\tilde{z}} \otimes \mathbf{e}^{\tilde{z}}.$$

Considering the general case we assume that growth starts on a non-empty initial body which is a hollow cylinder that is free of stresses at initial instant. Its inner and outer radii are $\tilde{r}_i$, $\tilde{r}_e$ respectively. Cylindrical material surfaces are attached to the outer surface of the body continuously increasing its external radius in the intermediate configuration. Let actual value of this radius is r_g. Suppose that on cylindrical surfaces of the growing body we have hydrostatic load p_e and p_i, i.e.

$$\boldsymbol{T}\cdot\mathbf{e}^{\tilde{r}}\big|_{\tilde{r}=\tilde{r}_i} = p_i\mathbf{e}_{\tilde{r}}, \quad \boldsymbol{T}\cdot\mathbf{e}^{\tilde{r}}\big|_{\tilde{r}=\tilde{r}_e} = p_e\mathbf{e}_{\tilde{r}}, \tag{17.16}$$

Then physical components of stresses can be presented by the formulas

$$T_{<\tilde{r}\tilde{r}>} = \begin{cases} I(\tilde{r},A) + p_i, \ \tilde{r}_0 \le \tilde{r} \le \tilde{r}_1 \\ I(\tilde{r}_1,A) + p_i + \int\limits_{\tilde{r}_1}^{\tilde{r}} \rho\Big(\frac{1}{\rho^2-\alpha} - \frac{\rho^2-\alpha}{(\rho^2+A)^2}\Big)d\rho, \ \tilde{r}_1 \le \tilde{r} \le \tilde{r}_g \end{cases}$$

$$T_{<\tilde{\theta}\tilde{\theta}>} = T_{<\tilde{r}\tilde{r}>} + \begin{cases} \frac{\tilde{r}^2+A}{\tilde{r}^2} - \frac{\tilde{r}^2}{\tilde{r}^2+A}, \ \tilde{r}_0 \le \tilde{r} \le \tilde{r}_1 \\ \frac{\tilde{r}^2+A}{\tilde{r}^2-\alpha} - \frac{\tilde{r}^2-\alpha}{\tilde{r}^2+A}, \ \tilde{r}_1 \le \tilde{r} \le \tilde{r}_g \end{cases} \tag{17.17}$$

$$T_{<\tilde{z}\tilde{z}>} = T_{<\tilde{r}\tilde{r}>} + \begin{cases} \frac{\tilde{r}^2+A-(1+\beta)A/2}{(\tilde{r}^2+A)\tilde{r}^2}A, & \tilde{r}_0 \le \tilde{r} \le \tilde{r}_1 \\ \frac{\tilde{r}^2+A-(1+\beta)(A+\alpha)/2}{(\tilde{r}^2+A)(\tilde{r}^2+\alpha)}(A+\alpha), & \tilde{r}_1 \le \tilde{r} \le \tilde{r}_g \end{cases}$$

$$I(\tilde{r}, A) = \int_{\tilde{r}_0}^{\tilde{r}} \rho\Big(\frac{1}{\rho^2} - \frac{\rho^2}{(\rho^2+A)^2}\Big)\, d\rho = \ln\frac{\tilde{r}\sqrt{A+\tilde{r}_0^2}}{r_0\sqrt{A+\tilde{r}^2}} - \frac{A}{2}\Big(\frac{1}{A+\tilde{r}_0^2} + \frac{1}{A+\tilde{r}^2}\Big). \tag{17.18}$$

Radial stresses $T_{<\tilde{r}\tilde{r}>}$ in the neighborhood of growing boundary are defined by the formula

$$T_{<\tilde{r}\tilde{r}>} = I(\tilde{r}_1, A) + p_i + \int_{\tilde{r}_1}^{\tilde{r}_g} \rho\Big(\frac{1}{\rho^2-\alpha} - \frac{\rho^2-\alpha}{(\rho^2+A)^2}\Big)d\rho. \tag{17.19}$$

Circumferential stress $T_{<\tilde{\theta}\tilde{\theta}>}$ in the neighborhood of growing boundary is

$$T_{<\tilde{\theta}\tilde{\theta}>|_{\tilde{r}=\tilde{r}_g}} = p_e + \frac{\tilde{r}^2+A}{\tilde{r}^2-\alpha(\tilde{r}_g)} - \frac{\tilde{r}^2-\alpha(\tilde{r}_g)}{\tilde{r}^2+A}. \tag{17.20}$$

Radius of the outer cylindrical boundary in the actual configuration is defined by

$$r_g = \sqrt{\tilde{r}_g^2 + A}. \tag{17.21}$$

The rate of change of the material composition of the body can be given a function $V(t)$, which determines the increasing of the volume of the growing body during the growing process. For incompressible material it is an invariant with respect to the change of configuration. Under the assumption that the inner radius of $\tilde{r}_i$ does not change, the outer one can be defined as follows

$$\tilde{r}_g = \sqrt{V(t)/(\pi h) + \tilde{r}_0^2}. \tag{17.22}$$

Consider the types of growth like in the case considered for discrete growth. We assume that in all cases the dependence of the volume $V(t)$ is known.

Type 1. Growth with given distortion. The distortion function $\alpha = \alpha(\tilde{r})$ is prescribed. To determine stress field one must find parameter $A(t)$ which is defined implicitly by the equation

$$F(A) = \Delta p, \quad F(A) = I(\tilde{r}_1, A) + \int_{\tilde{r}_1}^{\tilde{r}_g} \rho\Big(\frac{1}{\rho^2-\alpha} - \frac{\rho^2-\alpha}{(\rho^2+A)^2}\Big)d\rho.$$

Here $\Delta p = (p_e - p_i)/\mu$. Stresses that arise in the body in the process of growth can be determined by relation (17.17).

Type 2. Growth with a predefined displacements of growing boundary. The position of the image of growing boundary in the actual configuration is known, i.e. the function $z = r_g(t)$ is given. Since the radial coordinate of the growing boundary in intermediate configuration is given by (17.22), the parameter A can be determined from Eq. (17.21), i.e.

$$A = z^2 - \tilde{r}_g^2. \tag{17.23}$$

Substituting expressions (17.23) and (17.19) into the boundary conditions (17.16) we get integral equation with respect to $\alpha(\tilde{r})$

$$\int_{\tilde{r}_0}^{x} \rho\Big(\frac{1}{\rho^2 - \alpha} - \frac{\rho^2 - \alpha(\rho)}{(\rho^2 + z^2(x) - x^2)^2}\Big)d\rho =$$

$$= \Delta p - I(\tilde{r}, z^2(x) - x^2), \quad \Delta p = (p_e - p_i)/\mu.$$

As a result of change of variables $\xi = \rho^2$, $y = \rho^2 - \alpha(\rho)$, $\zeta = x^2$ we obtain the integral equation

$$\int_{a}^{\zeta} \Big(\frac{1}{y(\xi)} - \frac{y(\xi)}{(\xi - A(\zeta))^2}\Big)d\xi = Q(\xi) \tag{17.24}$$

with respect to function $y(\xi)$. Here the functions $A(\zeta) = z^2(\sqrt{\xi}) - \zeta$ $Q(\zeta) = 2(\Delta p(\sqrt{\xi}) - I(\tilde{r}_1, A(\zeta)))$ are prescribed.

The solution of this equation defines function $\alpha(\rho)$ and therefore all stresses (17.17).

6 Numerical Examples

In this section some results of numerical simulation of both discrete and continuous growth are given.

The rubbery material with the following mechanical characteristics $J_1 = 4.419\ 10^5$ Pa, $J_{-1} = -3.009\ 10^5$ Pa and $\mu = 7.928\ 10^5$ Pa, $\beta = 0.241$ is considered. The inner and outer diameters are $R_i^1 = 1.8h$, $R_e^1 = 2h$ in all calculations. In the discrete growth case it is studied three scenarios: 5, 10 and 25 layers joined to the initially body. The thicknesses of the layers were chosen from the following condition: the volume of final body was fixed. The pressure on inner and outer surface of initially body are taken as zero. The time of growth in the continuous case and the thickness of layers in the discrete one were chosen from the

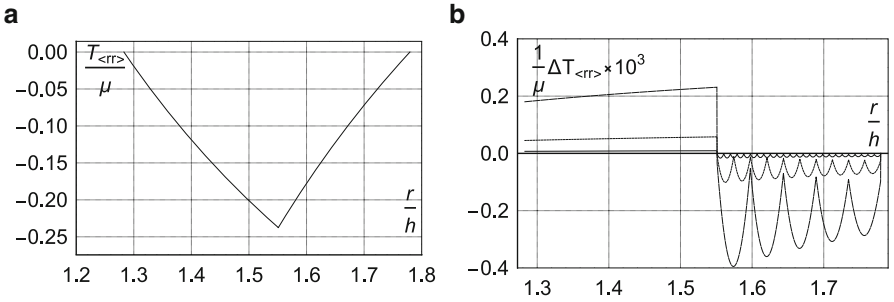


Fig. 17.1 (**a**) Relative radial stresses, (**b**) The difference between relative radial stresses for continues and discrete growth

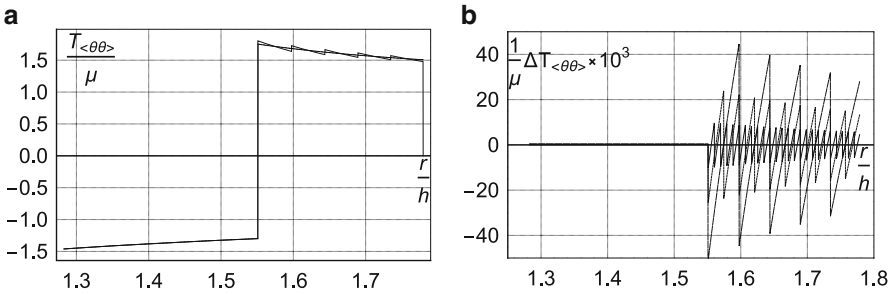


Fig. 17.2 (**a**) Relative circumferential stresses, (**b**) The difference between relative circumferential stresses for continues and discrete growth

condition that volume of initially body is doubled. The graphics of relative radial stress and circumferential stress distributions for discrete and continuous growth are shown in Figs. 17.1a and 17.2a. The graphics of the differences between relative stresses for discrete and continues growth (N = 5 – dashed, N = 10 – doted, N = 25 – solid line) are shown on Figs. 17.1b and 17.2b.

The computational examples show the convergence of solutions obtained for the discrete growth to corresponding solutions for continuous growth under the following conditions: the number of discrete body-fibers increases while their thickness decreases such that the final volume of growing solid is fixed.

Acknowledgements This work was supported by the Russian Science Foundation under Grant 14-19-01280.

References

1. Gibson, I., Rosen, D.W., Stucker, B.: Additive Manufacturing Technologies. Rapid Prototyping to Direct Digital Manufacturing. Springer, New York (2009)

2. Choy, K.L.: Chemical vapour deposition of coatings. Prog. Mater. Sci. **48**, 57–170 (2003)
3. Nastasi, M., Mayer, J.W.: Ion Implantation and Synthesis of Materials. Springer, Berlin/Heidelberg/New York (2006)
4. Aruytunyan, N.K., Drozdov, A.D., Naumov, V.E.: Mechanics of Growing Viscoelastoplastic Solids. Nauka, Moscow [in Russian] (1987)
5. Lychev, S.A., Manzhirov, A.V.: The mathematical theory of growing bodies. Finite deformations. J. Appl. Math. Mech. **77**(4), 421–432 (2013)
6. Lychev, S.A., Manzhirov, A.V.: Reference configurations of growing bodies. Mech. Solids. **48**(5), 553–560 (2013)
7. Lychev, S.A.: Universal deformations of growing solids. Mech. Solids. **46**(6), 863–876 (2011)
8. Lychev, S.A., Manzhirov A.V.: Discrete and Continuous Growth of Hollow Cylinder. Finite Deformations. Lecture Notes in Engineering and Computer Science: Proceedings of The Congress on Engineering 2014 (WCE 2014), London, 2–4 July 2014, pp. 1327–1332 (2014)
9. Klarbring, A., Olsson, T., Stalhand, J.: Theory of residual stresses with application to an arterial geometry. Arch. Mech. **59**, 341–364 (2001)
10. Epstein, M., Maugin, G.A.: Thermomechanics of volumetric growth in uniform bodies. Int. J. Plast. **16**(7), 951–978 (2000)
11. Yavari, A.: A geometric theory of growth mechanics. J. Nonlinear Sci. **20**(6), 781–830 (2010)
12. Manzhirov, A.V., Lychev, S.A.: The mathematical theory of growing solids: finite deformations. Dokl. Phys. **57**(4), 160–163 (2012)
13. Noll, W.: Materially uniform simple bodies with inhomogeneities. Arch. Ration. Mech. Anal. **27**(1), 1–32 (1967)
14. Wang, C.C.: On the geometric structure of simple bodies, or mathematical foundations for the theory of continuous distributions of dislocations. Arch. Ration. Mech. Anal. **27**(1), 33–94 (1967)
15. Epstein, M.: The Geometrical Language of Continuum Mechanics. Cambridge University Press, New York (2010)
16. Maugin, G.A.: Material Inhomogeneities in Elasticity. Chapman and Hall, London (1993)
17. Marsden, J.E., Hughes, T.J.R.: Mathematical Foundations of Elasticity. Dover, New York (1994)
18. Choquet-Bruhat, Y., Dewitt-Morette, C., Dillard-Bleick, M.: Analysis, Manifolds and Physics. Part 1. Basics. North-Holland, Amsterdam (1982)
19. Truesdell, C.: A First Course in Rational Continuum Mechanics: General Concepts. Academic, Boston (1991)
20. Manzhirov, A.V., Lychev, S.A.: Residual stresses in growing bodies. In: Topical Problems in Solid and Fluid Mechanics, pp. 66–79. Elite Publishing House, Delhi (2011)
21. Gurtin, M.E., Murdoch, A.I.: A continuum theory of elastic material surfaces. Arch. Ration. Mech. Anal. **27**, 291–323 (1975)
22. Lurie, A.I.: Nonlinear theory of elasticity. North-Holland, Amsterdam (1990)

Chapter 18
Design and Characterization of a Model Pilot Multi-tube for the Transfer of Heating the Fire Tube Boilers in Academic Laboratories

Austin Ikechukwu Gbasouzor

Abstract The aim of this research is to design and fabricate pilot multi-tube boiler using a diesel fired burner $(C_{13}H_{25})_9$ to generate 80 kg of steam hour. The boiler tank is made of pure mild steel. Mild steel is used to fabricate the fire tubes and other parts such as the furnace, smokestack and return chamber that make up the boiler. The heating surface area was increased for sake of efficiency and fast steam generation by reversing the direction of the gas through a second and third parallel tube (three pass). The boiler (which is fired by a diesel burner) generates dry saturated steam at a pressure of 1.5 bars and temperature of 111.4 °C. It can be used for domestic and industrial purposes.

Keywords Boiler cycle • Boiler efficiency • Dry saturated steam • Fuel firing rate • Gross caloric value • Heat absorbed • Steam flow rate • Specific heat • Strength of materials • Water enthalpy

1 Introduction

The word 'boiler', in everyday use, covers a wide range of equipment, from simple domestic hot water boilers to boilers housed within a power generation plant to convert fossil fuel to electricity. Generally, domestic hot water boilers do not produce steam and should operate at low pressure. While some combination boilers now operate at the pressure of the incoming cold water mains, this is still far below the normal operating pressure of steam-raising boilers.

The basic operation of steam turbines employs two concepts, which may be used either separately or together. In an impulse turbine the steam is expanded through nozzles so that it reaches a high velocity. The high-velocity, low-pressure jet of

A.I. Gbasouzor (✉)
Department of Mechanical Engineering, Anambra State University, P.M. B. 02, Uli, Nigeria
e-mail: unconditionaldivineventure@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_18

steam is then directed against the blades of a spinning wheel, where the steam's kinetic energy is extracted while performing work. Only low-velocity, low-pressure steam leaves the turbine.

In a reaction turbine the steam expands through a series of stages, each of which has a ring of curved stationary blades and a ring of curved rotating blades. In the rotating section the steam expands partially while providing a reactive force in the tangential direction to turn the turbine wheel. The stationary sections can allow for some expansion (and increase in kinetic energy) but are used mainly to redirect the steam for entry into the next rotating set of blades.

In most modern large steam turbines, the high-pressure steam is first expanded through a series of impulse stages sets of nozzles that immediately lower the high initial pressure so that the turbine casing does not have to withstand the high pressures produced in the boiler. This is then followed by many subsequent impulse or reaction stages (20 or more), in each of which the steam continues to expand.

The first reaction-type turbine was built by Hero of Alexandria in the first century AD. In his aeolipile, steam was fed into a sphere that rotated as steam expanded through two tangentially mounted nozzles. No useful work was produced by the aeolipile. Not until the nineteenth century were attempts made to utilize steam turbines for practical purposes. In 1837 a rotating steam chamber with exhaust nozzles was built to drive cotton gins and circular saws. A single-stage impulse turbine was designed by the Swedish engineer Carl Gustaf de Laval in 1882. A later American design had multiple impulse wheels mounted on the same shaft with nozzle sections located between each wheel. Subsequent advances in the design of steam turbines and boilers allowed for higher pressures and temperatures. These advances led to the huge and efficient modern machines, which are capable of converting more than 40 % of the energy available in the fuel into useful work.

Watt is credited as being the first inventor to separate the steam engine, and the boiler, into two separate units in the latter part of the eighteenth Century. In these early times, the primary use of the boiler was to generate steam for steam driven engines.

Steam driven engines replaced the horse as a means of motive power, it followed that steam driven engines were rated in 'Horsepower'.

Boiler design progressed from what was essentially a kettle to a relatively large-diameter flue pipe submerged in water – thus the first fire-tube boiler.

As power and pressure requirements increased, boilers became larger and the single-flue pipe became a larger number of smaller diameter flue tubes combined with an external, or internal, furnace for the combustion of the fuel. The modern-day 'modified Scotch Marine' boiler, generally comprising horizontal steel furnace combustion chambers) and/or fire-tube convective pass(es), in 'dry-back' or 'water-back' configurations, owes its heritage to these early multi-tube boilers and their application in ships constructed on Scotland's River Clyde.

The primary application of the boiler was still motive power; whether for pumping water from mines, driving machinery in mills, propelling steam locomotives or ships. Therefore, boiler ratings were based on the size of the steam engine that they

were capable of driving. The quantity of steam required to operate a 1 horsepower steam engine became known as 1 Boiler Horsepower. (Note that the water-tube boiler was not prevalent until after the first water-tube boiler design patent of 1867; thus, the term Boiler Horsepower (Bhp) has been associated with fire-tube boilers from the earliest days of boiler development).

During this period, variations in steam engine efficiency made it difficult to assign a qualified rating (i.e. Pounds per Hour (PPH)) to the amount of steam required to drive a 1 horsepower steam engine. Tests, conducted in 1876, determined approximately 30 pounds of steam per hour was required to produce 1 horsepower of mechanical work. In 1889, the American Society of Mechanical Engineers (ASME) standardized the term "Boiler Horsepower" as being based on a conventional steam engine evaporation rate of 30 pounds of steam per hour (PPH), at 70 PSIG pressure, and a feed water temperature of 100 °F. This definition was subsequently modified to: Boiler Horsepower – the unit of capacity expressed as the equivalent evaporation of 34.5 pounds of water per hour, from and at 212 °F (33,475 Btu/h.).

Also, it was determined that for the steel fire-tube boilers of the day, which utilized brick set bases incorporating large amounts of refractory, and generally coal fired, 10 square feet of fireside heating surface was necessary for a steam engine to generate I mechanical horsepower. As a result it became an industry standard practice to rale lire-tube boilers in Boiler Horsepower, and to base this rating on fire-side heating surface (1 Bhp per 10 square feet of fire-side heating surface).

Competition between fire-tube boiler manufacturers eventually forced improvements in boiler design and fuel burning equipment. This, together with a broad shift towards liquid and gaseous fuel utilization, resulted in cleaner and more reliable combustion and improved heat transfer within the boiler. Progressive reduction in the fire-side heating surface required, per Boiler Horsepower, was therefore consistent with these advancements. By the I960s, the 10 square feet of fire-side heating surface per Boiler Horsepower criterion decreased to 5 square feel of fireside heating surface per Boiler Horsepower – an axiom which is commonly cited today, particularly in the United States.

During the latter part of the twentieth Century, many manufacturers of fire-tube boilers designed and marketed boilers with greater input capacity burners, these boilers are equipped with requisite steam nozzle and safety valve(s) characteristics which therefore permitted operation at less than 5 square feet of fire-side heating surface per boiler horsepower with acceptable reliability and efficiency.

It was during this same period that certain jurisdictional mandates were promulgated that required licensed boiler operators for boilers in excess of a certain defined fire-side heating surface criterion. These limitations generally afforded opportunity to employ less than 5 square feet of fire-side heating surface per Bhp. This furthered development of fire-tube boiler designs specifically for particular jurisdictional requirements, again with acceptable reliability and efficiency.

Manufacturers of other boiler types, such as vertical, water-tube and cast iron sectional boilers, have used the Bhp output rating as a means of comparison with fire-tube boilers. They have not, typically, related Bhp output to a certain

square feet of fire-side heating surface criterion, opting • generally to rate by net output generated by the boiler; steam boilers expressed in measurements of PPM, 34.5 pounds per Boiler Horsepower from and at 212 °F, or millions of Btu/Hr (MBH), and hot water boilers expressed in measurements of 33,475 Btu/h per Boiler Horsepower, or millions of Btu/h (MBH).

1.1 Basic Design Manufacturing Requirements

- Compliance with the ASME Boiler and Pressure Vessel Code.
- Compliance with required safety and installation Codes.
- The ability to meet the required efficiency and other performance standards.
- The ability to meet the required level of pollutant emissions,
- Compliance with the requirements of the National Board of Boiler and Pressure Vessel Inspectors through local jurisdictions having authority (JHA),
- The ability to meet the perceived needs of the customer in terms of operational performance, reliability and maintenance costs.
- The ability to produce a competitively priced product.

Fire-tube boiler manufacturers have established over the years that these criteria can be satisfied with varying fire-side heating surface specifications. Thus, the nominal 5 square feet of fire-side heating surface per boiler horsepower axiom is less important as a critical design consideration.

1.2 High-Pressure

While the advances of eighteenth century, the call was for higher pressures; this was strongly resisted by Watt who used the monopoly his patent gave him to prevent others from building high-pressure engines and using them in vehicles. He mistrusted the materials' resistance and the boiler technology of the day.

1.2.1 The Important Advantages of High Pressure

- They could be made much smaller than previously for a given power output. There was thus the potential for steam engines to be developed that were small and powerful enough to propel themselves and other objects. As a result, steam power for transportation now became a practicality in the form of ships and land vehicles, which revolutionised cargo businesses, travel, military strategy, and essentially every aspect of society.
- Because of their smaller size, they were much less expensive.

- They did not require the significant quantities of condenser cooling water needed by atmospheric engines.
- They could be designed to run at higher speeds, making them more suitable for powering machinery.

1.2.2 The Disadvantages

- In the low pressure range they were less efficient than condensing engines, especially if steam was not used expansively.
- They were more susceptible to boiler explosions.

The main difference between how high-pressure and low-pressure steam engines work is the source of the force that moves the piston. In Newcomen's and Watt's engines, it is the condensation of the steam that creates most of the pressure difference, causing atmospheric pressure (Newcomen) or low-pressure steam. (Watt) to push the piston; the internal pressures never greatly exceed atmospheric pressure. In a high-pressure engine, most of the pressure difference is provided by the high pressure steam from the boiler; the low pressure side of the piston may be at atmospheric pressure or, if it is connected to a condenser, this only provides a small proportion of the pressure difference.

1.3 Considerations for the Future

American Boiler Manufacturers Association (ABMA) Stated in issue 1 of 2004 that It is apparent that the powerful analytical tools available today will continue to be applied to improve fire-tube boiler design, as well as that of the fuel burning equipment. This will inevitably result in smaller, higher efficiency fire-tube boilers with lower pollutant emissions.

Combustion technology, with specific regard to NOx reduction, is rapidly approaching the point of diminishing returns. Should further emission reductions develop, one method of achieving compliance will be a significant increase in boiler efficiency, and thus reducing fuel input consumption. Combined with an obvious general tendency by end users, architects and engineers, to minimize the boiler room footprint in building construction, the resultant more compact and higher efficiency boilers, requiring lower heating surface per Boiler Horsepower, will be furthered. Albeit, the resulting improvements will remain predicated upon the particular fire-tube boiler system application requirements, materials and manufacturing limitations.

Manufacturers of fire-tube boilers, recognizing the benefit of enhanced design and materials technology, currently apply these resources to advance product design,

performance and manufacturing competitiveness. Individually, and together with the US Department of Energy, programs have been initiated to further development of these, and other technologies, to meet with the ever increasing expectation of fire-tube boiler capabilities.

1.4 Design Objectives

The research work provides guidance on how to design a simple model fire-lube boiler for domestic and industrial use. This design guideline can assist upcoming engineers to understand the basic design of boiler and a suitable size, material and heat of combustion, The choice of fire-tube boiler and its design is crucial to give the best performance of boiler; good performance of boiler is influenced by maximum heat absorbed and minimum heat loss. The design of boiler may be influenced by factors including process requirements, economics, and safety. All the important parameters used in the guideline are explained in the definition section which helps the reader to understand the meaning of the parameters or term used.

2 The Organ of Boiler

A boiler is a closed vessel in which steam is produced from water by combustion of fuel. Steam boilers is made up of two major parts, that is, the combustion chamber, which provides heat by the combustion of fuel, and the heat exchanger which transforms water into steam through heat exchange in the medium (Saidur et al. 2010). Figure 18.1 shows a schematic diagram of a steam boiler.

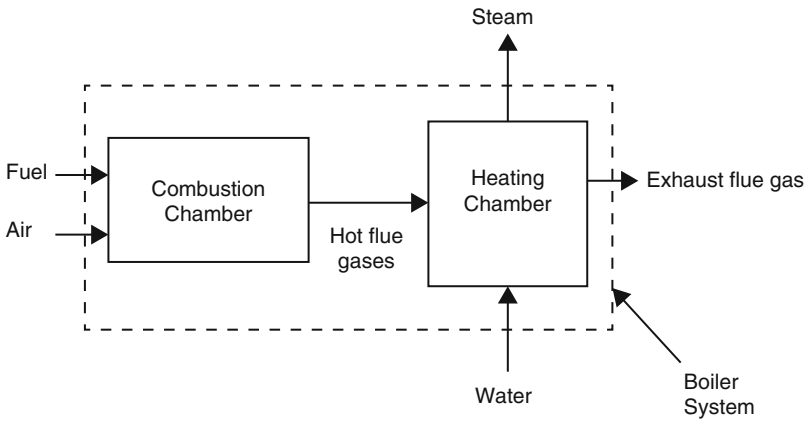


Fig. 18.1 A schematic diagram of a steam boiler

- Boiler systems are classified in variety of ways. They can be classified;
- According to the end use, such as foe heating, power generation or process requirements.
- According to pressure, materials of construction, size tube contents (for example waterside or fireside), firing, heat source or circulation (for example, oil-fired, gas-fired, coal-fired, or solid fuel-fired).
- According to their method of fabrication.
- And boiler can be pack aged or field erected.

2.1 Types of Boilers

2.1.1 Fire Tube Boilers

Fire tube boiler consists of boiler shell which is filled with water and perforated with tubes, there are different configurations of these tubes but horizontal configuration is the common one. Water is partially filled in the water tank and volume is left inside the tank to accommodate the steam. Long horizontal tubes arc called flues and these carry the hot combustion gases through the water tank and heating the water. The furnace is situated at one end of a fire tube which elongates the path of the hot gases, thus expanding the heating surface which can be further increased by making the gases reverse direction through a second tube or bundle of multiple tubes.

The water and steam in fire-tube boilers contained within a large diameter drum or shell, and such unit arc often referred to as 'shell type boiler'. Heat from the products of combustion is transferred to the boiler water by tubes and it goes out from the smokestack. Fire-tube boilers are approximated to 360 psi of steam pressure. In case of fire-tube boiler the whole lank is under pressure so if tanks burst it creates a major explosion and if one need to increase the steam pressure of fire-tube boiler then it is necessary to increase the thickness of the shell and material of tube sheet.

The water is confined by the outer shell of boiler. To avoid the need for a thick outer shell fire-tube boilers are used for lower pressure applications. Generally, he heat input capacities for fire-tube boilers are to 50 mbtu per hour or less, but in recent years the size of fire-tube boilers has increased.

Fire-tube boilers are subdivided into three groups; horizontal return tubular (HRT) boilers typically have horizontal self-contained fire-tubes with a separate combustion chamber. Scotch, scotch marine, or shell boilers have the fire-tubes and combustion chamber housed within the same shell. Firebox boilers have a water-jacketed firebox and employ at most three passes of combustion gases.

Most modern fire-tube boilers have cylindrical outer shells with a small round combustion chamber located inside the bottom of the shell. Depending on the constructions details, these boilers have tubes configured in one, two, three or four pass arrangements, because the design of fire-tube boilers is simple, they are easy to

construct in a shop and can be shipped fully assembled as a package unit. Fire-tube boilers typically have a lower initial cost, are more fuel efficient and are easier to operate.

2.1.2 Advantages of Fire-Tube Boilers

- Low cost.
- Fluctuations of steam demand can be met easily.
- It is compact in size.

2.2 Water-Tube Boilers

In water-tube boilers the rack of tubes are positioned vertically in the firebox and through these tubes water flows which gets heated upon flowing 'through these tubes, these vertical pipes are called riser and these extends from the water drum which is at the bottom of the boiler to the stem headers which are at the top of the boilers. These typically surround the firebox in many layers like a dense forest. As steam bubbles form, they rise to the steam drum where the steam exist through the header.

Water-tube in furnace can be arranged in many different configurations and they are often used to connect water drums at the bottom and water, and steam drum at the top, there exist a mono-tube boiler pump used to circulate water through a succession of coils, this type of boiler has very fast production of steam but has very less storage capacity. Water-tube boilers are preferred for high pressure application because of the high pressure steam/water is contained in smaller diameter pipes which withstand the high pressure.

Water tube boilers require less weight of metal for a given size, are less liable to explosion, produce higher pressure, are accessible and can respond quickly to change in steam demand. Tubes and drums of water-tube boilers are smaller than that of fire-tube boilers and due to smaller size of drum, higher pressure can be used easily in water-tube.

2.2.1 Advantages of Water-Tube Boilers

- High pressure of the order of 140 kg/cm^2 can be obtained.
- Heating surface is large; therefore steam can be generated easily,
- Large heating surface can be obtained by use of large number of tubes,
- As a reason of high movement of water in the tubes the rate of heat transfer becomes large resulting into a greater efficiency.

3 Methodology

3.1 Design Procedure of the Boiler

In the design of multi-tube fire boiler in this research work, many things were considered when analyzing these systems:

- Design specification.
- Design consideration.
- Detailed design.
- Technological details.

3.1.1 Design Specification

The fire tube boiler consists of various components and it will be great importance to have a detailed specification before the design. The arrangement of the fire-tube boiler is illustrated in Fig. 18.2.

The diesel burner used to heat up the furnace of the fire-tube has the following specifications:

- Mass firing rate = 2.5–5 kg/h
- Orifice diameter for exit (d) = Ø0.0005 m
- Motor rating = 0.5 horse power.

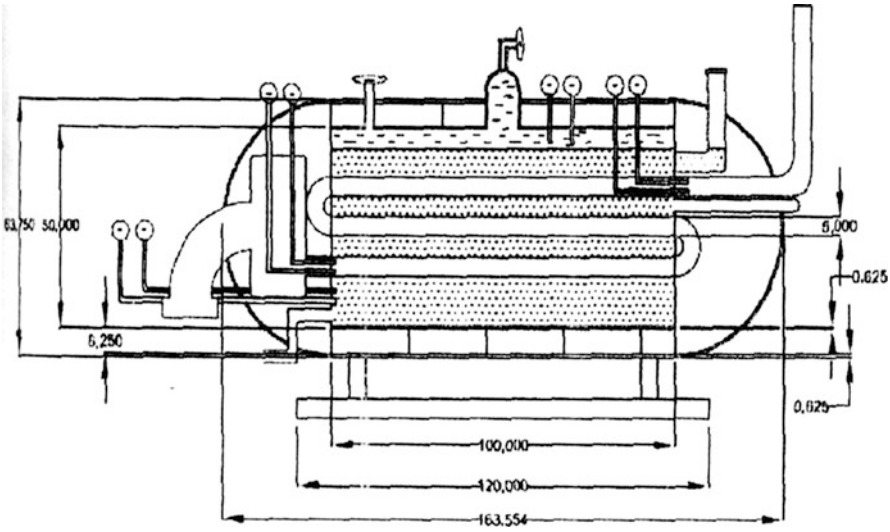


Fig. 18.2 Dimensioned diagrams of the three pass fire-tube boiler

The burner is connected to the furnace by the means of both external and internal circular flange (a projecting collar, rim, or rib on an object for fixing it to another object, holding it in place or strengthening it. Flanges are often found on pipes and shafts) of both the burner and furnace respectively. The flange specifications are given as follows;

- Outer diameter of circular flange
- $(d_0) = Ø0.017 \times 2$ m
- Inner diameter of circular flange
- $(dj) = Ø0.013 \times 2$ m
- Number of opening for bolts and nuts of flange = 4 openings
- Diameter of the bolts and nuts used $(d_b) = Ø00.014$ m

The furnace which is located inside the boiler pressure vessel (shell) and situated at one end of 5 section of longitudinal fire-tubes connected to it serially which elongates the path of the hot gases, thus expanding the heating surface. The idea of placing the furnace inside the boiler shell is to maximize the heat of the boiler rather than losing it to the surrounding. The furnace serves as a pre-heater in this case as it raises the temperature of the water.

The fire-tubes extend to a compartment known as the return chamber situated at another end in the boiler vessel (shell). The return chamber itself which is serving as an intermediary for hot gases transfer has another set of 5 fire-tubes connected to it in the same manner as that at the furnace. This was done to further increase the heating surface area by making the gases reverse direction through a second 5 sets of parallel tubes. The heat emitted by this other set of 5 longitudinal fire-tubes at the return chamber goes out from a smoke stack. The following are the specifications of the inner components in the boiler vessel (shell):

- Total of 15 pieces of fire-tubes.
- A furnace.
- Two return chamber.
- Smoke stack.

3.1.2 Design Consideration for Material Selection

For an intelligent design to be done, the knowledge of the materials available as well as the properties they posses are very important For the selection of the proper material to be used for the design of the fire-tube boiler, we shall consider the factors which affect the choice of material selected and used for design and there reasons.

3.1.3 Factors Considered Are

- Suitability of the material for the working conditions in service, considering characteristics such as; appearance, thermal conductivity, rate of emissivity, strength, stiffness, creep, etc.

- Availability of the material: the ease at which the materials are seen or purchased in the market.
- Workability of the material: considering possible methods of processing material selected into desired shape such as; weldability, machinability, formability, and workability.
- Expected load or force as well as adequate strength in conformity so as to function satisfactorily without failure.
- Cost of the material (economic consideration).

3.1.4 Choice of Material

Based on the above considerations, the materials used for the design of the fire-tube boiler were thus selected and tabulated below in Table 18.1.

3.2 Detailed Design

Having completed the material selection for the fire-tube boiler, the design of the various parts of the boiler shown in Fig. 18.3 is typified by the following features;

- The volumetric boiler pressure vessel (tank or shell).
- The furnace.
- The fire-tube.
- The return chamber.
- The smoke stack.

Table 18.1 Materials used and reasons

Parts	Material formally used or preferred and reasons	Material used and reasons
Furnace	Aluminum; good conductivity, high corrosion resistance but high melting point	Mild steel; affordable, available, weldable, malleable, strength, high conductivity, and corrosion resistance
Fire-tubes	Copper; high thermal conductivity, better formability	Mild steel; affordable, available, weldable, malleable, strength, high conductivity, and corrosion resistance
Return chamber	Aluminum; good conductivity, high corrosion resistance but high melting point	Mild steel; affordable, available, weldable, malleable, strength, high conductivity, and corrosion resistance
Pressure vessel	Wrought iron; toughness, malleable, and ductile	Steel; low cost of fabrication, stronger, quick weldability, cheaper and less labor
Smoke sack	Copper; high thermal conductivity, better formability	Mild steel; affordable, available, weldable, malleable, strength, high conductivity, and corrosion resistance

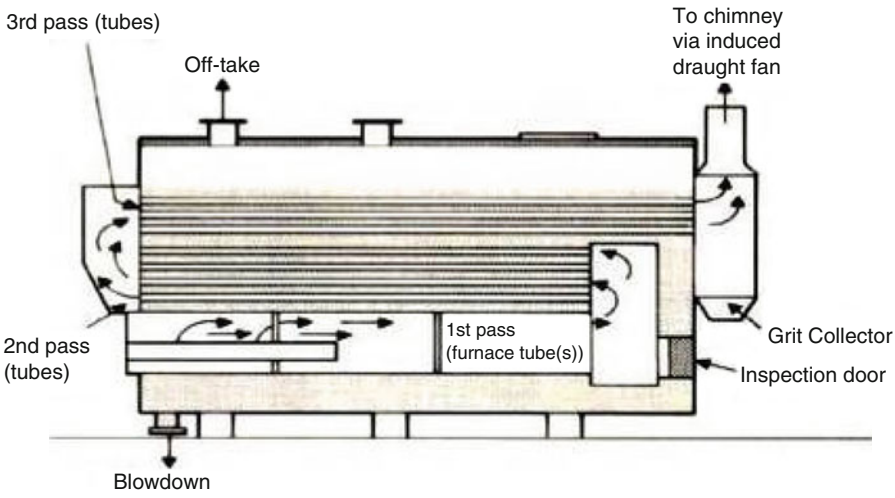


Fig. 18.3 Schematic diagrams of the three pass fire-tube boiler

- Actual volumetric capacity of the boiler.
- Pressure gauge.
- Temperature gauge.
- Safety valve.
- Thermal stresses and creep analysis (Fig. 18.3).

3.3 Design of the Volumetric Boiler Pressure Vessel

The boiler volumetric tank measures the quantity of water delivered to it at a given time. It has a capacity of 292 liters and it is made of steel metal of 0.006 m thickness for pressure resistance, Ø0.62 m and length of 0.76 m three holes were bored on the surface of the tank for both steam outlet (Ø0.50 m), turbine outlet (Ø0.50 m) and exhaust or smokestack outlet (Ø0.05 m). A hole of Ø0.178 m is also provided at one end of the longitudinal section of the tank for the cylindrical furnace placed inside it. Other dimensions are as follows;

Volumetric capacity of the drum (tank) = volume of cylinder + volume of hemisphere.

$$= \pi r^2 \times L + \frac{2}{3}\pi r^3$$

$$\text{Recall : r} = \frac{\text{d}}{2} = \frac{0.62}{2} = 0.31 \text{ m} \tag{18.1}$$

Placing value of r into Eq. 18.1

$$= \pi \times (3.31)^2 \times 0.76 + {}^{2}/_{3}\,\pi(0.31)^3$$
$$= 0.229449 + 0.062394$$
$$= 0.2918431\ \text{m}^3 = 292\ \text{litres}$$

To know the maximum pressure and temperature of the boiler, using the hoop law;

We know that,

Tensile stress of a mild steel = 60 Mpa Ultimate tensile stress of a mild steel = 410 Mpa

Hoop stress of mild steel = 140 Mpa Pressure of steam at 111.4 °C (p) = 1.5 bar = 0.15 Mpa

Thickness of pressure vessel (t) = 6 mm Diameter of vessel (d) = 620 mm

The hoop of stress of the steam

$$\sigma h = \frac{p \times d}{2t}$$
$$= \frac{0.15 \times 620}{2 \times 6}$$
$$= 7.75\ \text{Mpa} \tag{18.2}$$

The estimated maximum pressure of the vessel

$$P_{\max} = \frac{\sigma_{\max} \times 2t}{d}$$
$$= \frac{140 \times 2 \times 6}{620}$$
$$= 2.71\ \text{Mpa}$$
$$= 27.1\ \text{bar} \tag{18.3}$$

From steam table, temperature at 27.1 bar = 228°c

3.4 Design of the Furnace

The furnace made of mild steel located at one end of the boiler connected to a heat supply (diesel burner located outside the boiler vessel) in this case by means of a flange with specifications stated above, serves as the central system for heat (hot gases) distribution to the fire tubes. The furnace has a length of 0.40 m, thickness 10 mm and a diameter of 0.170 m. Other dimensions are given below;

Volume of the furnace

$$= \frac{\pi d^2}{4} \times L = \frac{\pi^{0.17^2}}{4} \times 0.40$$
$$= 0.0091 \ \text{m}^3 \tag{18.4}$$

3.5 Design of the Fire-Tubes

The fire-tubes made of mild steel is a total 15 in numbers and is sub-divided into three sections namely

- Furnace section = 5 fire-tubes of length 0.30 m and diameter 0.030 m each.
- Return chamber section = 5 fire-tubes of length 0.30 m and diameter 0.030 m each.
- The third pass section (section to the smokestack) – 5 fire-tubes of length 0.50 m and diameter 0.030 m each.

Other dimensions are as shown;
Volume of the fire-tubes in the furnace section

$$= \frac{\pi d^2}{4} \times L = \frac{\pi (0.03)^2}{4} \times 0.3$$
$$= 0.0002121 \ \text{m}^3 \tag{18.5}$$

Volume multiplied by the number of fire-tubes on this section

$$= 0.0002121 \times 5$$
$$= 0.001061 \ \text{m}^3$$

Volume of fire-tubes in the return chamber section

$$= \frac{\pi d^2}{4} \times L = \frac{\pi (0.03)^2}{4} \times 0.3$$
$$= 0.0002121 \ \text{m}^3 \tag{18.6}$$

Volume multiplied by the number of fire-tubes on this section

$$= 0.0002121 \times 5$$
$$= 0.001061 \ \text{m}^3$$

Volume of fire-tubes in the third passes section

$$= \frac{\pi d^2}{4} \times L = \frac{\pi(0.03)^2}{4} \times 0.5$$
$$= 0.00035343 \ \text{m}^3 \tag{18.7}$$

Volume multiplied by the number of fire-tubes on this section

$$= 0.00035343 \times 5$$
$$= 0.001767 \ \text{m}^3$$

3.6 Design of the Return Chamber

The return chamber made of mild steel which serves as an intermediary of heat transfer between the above mention sets of fire-tubes has a length of 0.15 m and diameter of 0.30 m. The volume of the return chamber

$$= \frac{\pi d^2}{4} \times L = \frac{\pi(0.3)^2}{4} \times 0.15$$
$$= 0.010603 \ \text{m}^3 \tag{18.8}$$

3.7 Design of the Smokestack (Exhaust)

The smokestack made of mild steel used to transport the flue out of the system has a length of 0.20 m and diameter of 0. 15 m, The volume of the smokestack

$$= \frac{\pi d^2}{4} \times L = \frac{\pi(0.15)^2}{4} \times 0.20$$
$$= 0.0035343 \ \text{m}^3 \tag{18.9}$$

3.8 Actual Volumetric Capacity of the Boiler

Actual capacity (i.e. volume) of the boiler = volume of the drum – total volume of the inner compartment of the boiler.

$$\text{Volumetric capacity of the dram} = 292 \text{ liters} \tag{18.10}$$

Total volumetric capacity of the inner compartments = volume of furnace + volume of fire-tubes on each sections + volume of the return chamber + volume of the smokestack = {0.0091 m + (0.001061 m^3 + 0.001061 m^3 4 0.001767 m^3) + 0.010603 m^3 + 0.0035343 m^3} = 0.0271263 m^3

Therefore,

Actual capacity of the boiler

$$= 0.2918431\ m^3 - 0.0271263\ m^3$$

$$= 0.2647168\ m^3 = 264.72\ \text{liters}$$

$$Q = -K_{12}\frac{T_2 - T_1}{X_{12}} = -K_{23}\frac{T_3 - T_2}{X_{32}} = -K_{34}\frac{T_4 - T_3}{X_{43}} \tag{18.11}$$

For the complete wall as a whole;

$$Q = -U\ (T_4 - T_1)$$

Where, U is the overall heat transfer coefficient of the wall

T is the temperatures of inner and outer wall

$$\frac{1}{U} = \frac{X_{12}}{K_{12}} + \frac{X_{23}}{K_{23}} + \frac{X_{34}}{K_{34}} \tag{18.12}$$

From the calculation Q = 22591.84w

$$\mathrm{Q} = -\mathrm{U}\,(T_4 - T_1)$$

$$\mathrm{U} = \frac{22591.84}{111.4 - 50} = 367.95\ \mathrm{W/m^2k}$$

Therefore,

$$\frac{1}{U} = \frac{X_{12}}{K_{12}} + \frac{X_{23}}{K_{23}} + \frac{X_{34}}{K_{34}} \tag{18.13}$$

K = thermal conductivity
K of aluminum = 210w/mk
K of fiber glass = 1.475w/mk
K of mild steel = 48w/mk
Thickness of mild steel (X_{12}) = 6 mm
Thickness of aluminum (X_{34}) = 1 mm

$$\frac{1}{36.795} = \frac{0.006}{48} + \frac{X23}{1.47} + \frac{0.001}{21.0}$$

$$0.0272 = 0.000125 + \frac{X23}{1.47} + 0.00000476$$

$$\frac{X23}{1.47} = 0.0272 - 0.000125 - 0.00000476$$

$$\frac{X23}{1.47} = 0.0259$$

$$X23 = \frac{0.0259}{1.47}$$

$$= 0399\text{m} \cong 40 \text{ mm}$$

4 Design Calculations and Mathematical Formulation

Order of calculation of the boiler is stated below;

- Combustion (Stoichiometric) calculation.
- Burner heat transfer rate per unit time.
- Thermal design calculation.
- Amount of steam generation.
- Velocity of steam.
- Pressure.

5 Combustion (Stoichiometric) Calculation Outline

Combustion (Stoichiometric) calculations provide much of the basic information necessary for the design of a boiler plant. They help find the amount of fuel to be fed for the required thermal output of the plant. The specifications of fans and blower are based on the air required for burning of atomizing that quantity of fuel. There is more to a burner than just blowing fire into a boiler or other heating device (furnace). The functions of our burner are;

- To provide heat to a boiler.
- To control the outlet temperature or pressure of a boiler.
- To provide a high turndown so that it does not shut off over the full range of the boiler load demands.
- To burn the fuel in the most efficient way possible to keep fuel consumption low.

Excess Air: The extra amount of air added to the burner above that which is required to completely burn the fuel.

Turndown: The ratio of the burner's maximum BTUH firing capability to the burner's minimum BTUH firing capability.

The fuel is diesel oil (gas oil) in the case of this project. The chemical formula of diesel is $C_{13}H_{25}$

For complete combustion

$$C_{13}H_{25} + O_2 - - - - - - CO_2 + H_2O$$

Balancing the equation

$$4C_{13}H_{25} + 77O_2 - - - - - - 52CO_2 + 50H_2O$$

Computer for the fuel bass ($4C_{13}H_{25}$)

$$\text{Molecular weight} : C = 12, H = 1, O = 16$$

i.e. $4C_{13}H_{25} = 4\,(12 \times 13 + 1 \times 25) = 724\ \text{kg}$

Computer for the mass of ($77O_2$)

$$77O_2 = 77 \times 16 \times 2 = 2{,}464\ \text{kg}$$

Computer for the mass of ($52CO_2$)

$$52CO_2 = 52\,(12 + 16 \times 2) = 2{,}288\ \text{kg}$$

Computer for the mass of ($50H_2O$)

$$50H_2O = 50\,(1 \times 2 + 16) = 900\ \text{kg}$$

The air-fuel ratio (A/F)

$$= \frac{\text{mass of O2} \times \text{percentage of O2 by mass in air}}{\text{Mass of fuel (diesel)}}$$

$$= \frac{2464 \times 100}{724 \times 23.3}$$

Stoichiometric air/fuel ratio (A/F) = 14.61

6 Excess Air Addition

Because gas flow rates are usually measured in cubic feet and fans are volumetric devices used to deliver the combustion air, conversions are made from pounds to cubic feet. As a "rule of thumb", 1 ft^3 (0.0283 m^3) of natural gas theoretically

Table 18.2 Excess levels for different fuels

Fuel	Type of furnace or burners	Excess air (%by wt)
Pulverized coal	Completely water-cooled furnace for slag-tap or dry-ash removal	15–20
	Partially water-cooled furnace for dry-ash removal	15–40
Coal	Spender stoker	30–60
	Water-cooler vibrating-grate stokers	30–60
	Chain-grate and travelling-stokers	15–50
	Underfeed stoker	20–50
Fuel oil	oil burners, register type	15–20
	Multi-fuel burners and flat-flame	20–30
Natural gas	High pressure burner	5–7
Wood	Dutch over (10–23 %through grates) and Hoff type	20–25

requires 10 ft3 (0.28 m^3) of air for complete combustion at 60^0f and 14.7 psi under perfect conditions. Of course this never happens, therefore burner combustion air fans may be sized for 25 % excess air in which case the fan delivers 12.5 ft^3 (0.354 m^3) of air per 1 ft^3. But in the course of this project 15–20 % excess air is prescribed for oil fuels, which is added for efficiency. The equation for combustion change to; (Table 18.2)

$$4C_{13}H_{25} + 1.2\,[77O_2 + (79/21)\,77N_2] - 52Co_2 + 50H_2O + (0.2)77O_2 + 1.2\,(79/21)\,77N_2$$

7 Modes of Heat Transfer in Fire Tube Boiler

The importance of heat transfer in boiler design is self-evident. Zeljko Warga stated that in conventional 3 and 4 pass fire tube boilers, only a smaller portion of total heat is transferred in the furnace as it has as much as 90 % and more heat transfer surfaces in the tubes. The radiation in the tubes is almost non existent compared to convection, while in the furnace, the radiation can be even smaller than convection, as the test boilers demonstrated. This is in total contradiction to water-tube boilers where convection represents less than 20 % of total amount of heat exchange in furnace as the test boilers demonstrated to the total heat exchange in the furnace can be as high as 80 %. Hot water test boilers showed additionally that the percentage of surfaces in tubes could be close to that found in the furnace. The industrial-sized steam test boiler has as low as 2.3 times more area in the tubes then in the rest of the boiler. Furthermore it was proved that the convection in the furnace of fire-tube boilers can be made even higher than the radiation.

The number of boiler tubes is limited by burner fan capability to overcome internal pressure loss. By that fact, the general direction in designing fire-tube boilers is given; namely to install only as many tubes as are necessary. This requires the exact analytical assessment of heat transfer in particular boiler sections to which

this research work was devoted. Thus, by proper design of the boiler (for which the in-deep knowledge of heat transfer is of primary importance), as demonstrated in this work, a sizable intensification of heat transfer and noticeable savings in boiler manufacturing cost can be achieved. By so doing lets identify three modes of heat transfer;

1. Conduction
2. Convection
3. Radiation

However, they have part to play in the boiler. Though conduction is not considered in simple boiler calculation.

8 Heat Transfer by Convection in Boilers

Convection in boilers takes place simultaneously with radiation. In tubes of fire-tube boilers more than 90 % of heat exchange takes place by convection, while in the furnaces the radiant part is greater than in tubes. Calculation of convection is conducted by standard equations for flows in straight tubes and channels it is also true for the boiler furnaces, whether they are circular or rectangular in cross-section. This picture totally changes when gas flow directly hits the surface involved in convection, such as in the case of the rear of the furnace. The rate of convection is much higher in these cases and cannot be assessed by classical equations for straight flow in tubes and channels. Tests showed much lower gas exit temperatures from the furnace than had been calculated which was found to be attributable to lack of taking into account the heat transfer from impinging jet of the flue-gases. Also test shows an overall improvement in heat transfer in boilers with use of cylinder of high temperature and corrosion resistant material (to have less scaling of surfaces by unburned fuel sulfur, less soot, while also affecting radiation and convection).

Heat transfer by convection relies on the bulk movement of a heated fluid in relation to a surface. It features strongly in the design of plant items such as super-heaters and re-heaters, where there is gas-to-metal and metal -to-steam convective heat transfers as shown in Table 18.3.

$$Q_{conv} = h\,(T_h - T_c)\,A_c \tag{18.14}$$

Table 18.3 Combustion temperatures, specific heat capacity and flue-gas enthalpy at that temperature

Combustion temperature of diesel oil	Enthalpy (h) diesel fuel	Density of diesel	Specific heat capacity (c_p) of diesel	Specific heat capacity (c_p) of air
540–650 °C	44,500 kj/kg	840 kg/m^3	1.75kj/kg °C	1.006kj/kg °C

9 Burner Heat Transfer Rate per Unit Time

A burner that should be able to fire a boiler to generate 80 kg/h, assuming that the water will be heated from ambient condition that is temperature of 30 °C to a temperature of 111.4 °C at 1.5 bar.

Quantity of heat energy needed to generate this quantity of steam is given by

$$Q = M_s h_f g$$

Where;

M_s = mass of steam produced per second

Hf_g = Enthalpy of vaporization of water

$M_s = 80/3{,}600 = 0.0222$ kg|s

From steam table hf_g of water at 111.4 °C and 1.5 bar is equal to 222×10^3j/kg
Therefore; Q = MS hf_g

10 Fuel Consumption Rate of the Burner

Power output = fuel consumption x calorific value of fuel

$$P_b = M_f \times CV \tag{18.15}$$

Where;

P_0 = power output

M_f = fuel consumption or mass flow rate

Cv = calorific value of fuel

Given that calorific value of diesel is 45,500kj/kg
∴ Fuel consumption or mass flow rate of fuel (M_f)
Mass flow rate = Volumetric flow rate x Density

$$Mf = Q_f \times P_f$$

$$\therefore Q_f = \frac{M_f}{P_f} \tag{18.16}$$

Where;

$$P_f = \text{density of diesel} = 840 \ \text{kg}|\text{m}^3$$

11 Thermal Design Calculation

The thermal design calculation involves the heat transfer from all heat sources located in the boiler as outlined:

11.1 Fire Tubes Heat Transfer Calculation

- Furnace heat transfer calculation.
- Fire-tubes heat transfer calculation.
- Return chamber heat transfer calculation.
- Smokestack calculation.

11.2 Furnace Calculation

$$\text{The sensible heat loss of flue gas at furnace exit} = m \times C_p \times \Delta T \tag{18.17}$$

Where;

$$m = \text{mass of flue gas (kg)}$$

$$C_p = \text{specific heat of flue gas}$$

$$T = (\text{flue gas temperature} - \text{ambient temperature}) \ \text{in}^0\text{C}$$

Theoretical air required from air fuel ratio
Mass of flue gas $[m_g\ (_P)] = m_a + m_f$
Heat loss $= M_p \text{ x } C_p \text{ x } \Delta T$

11.3 Radiation Heat Transfer from Furnace

$$Q_{rad} = \varepsilon\sigma \left(T_h{}^4 - T_c{}^4\right) A_c \tag{18.18}$$

Where;

q_{rad} = heat transfer per unit time (W)

σ = Stefan Boltzmann const = 5.6703×10^{-8} $\left(w/m^2k^4\right)$

ε = emissivity of material (mild-steel) used = 0.32

T_h = hot body absolute temperature (K) = temperature of water

= 565°C = 838°k

T_c = cold surroundings absolute temperature = temperature of furnace

= 30°C = 303°k

A_c = area of the object $\left(m^2\right)$

d = 170mm : $A_c = 0.02271 m^2$

11.4 Convective Heat Transfer from Furnace

$$Q_{conv} = h_c \times A \times (\Delta T) \tag{18.19}$$

Where; h_c – convective heat transfer coefficient $(w/m^2k) = 250 w/m^2k$

d = heat transfer diameter (m) = 0.17 m,

$$A = \frac{\pi d^2}{4} = 0.02271 \ m^2$$

ΔT = (temperature of furnace – temperature of water) ^{0} k

L = length of furnace = 40 cm = 0.4 m

11.5 Conduction Heat Transfer from Furnace

$$Q_{cond} = \frac{kc \times A \times (\Delta T)}{l} \tag{18.20}$$

Where;

K_c = thermal conductivity coefficient (w/mk) = 59w/mk

L = length of furnace

T_h − hot body absolute temperature

(K) = temperature of water

T_c = cold surroundings absolute temperature = temperature of furnace

A_c = area of the object $\left(m^2\right)$

Efficiency of furnace

Thermal efficiency of the furnace by direct method;

$$\text{Thermal efficiency of the furnace} = \frac{\text{Heat output from the burner}}{\text{Heat in the fuel consumed (heat input)}} \times 100$$

Heat output from burner = 55kw

$$\text{Heat in the fuel consumed} = \frac{\text{GCV of diesel } \left(\frac{\text{kj}}{\text{kg}}\right)}{\text{fuel consumption rate } \left(\frac{\text{kg}}{\text{h}}\right)} \tag{18.21}$$

Fire-tube heat transfer at the following sections:

(a) Furnace section to return chamber.
(b) Return chamber to 2nd pass tubes
(c) 2nd pass tubes to smokestack.

12 Radiative Heat Transfer in the Furnace to the Return Chamber Section

$$Q_{rad} = \varepsilon\sigma \left(T_h^4 - T_c^4\right) A_c \tag{18.22}$$

Where;

q_{rad} = heat transfer per unit time (W)

12.1 Convective Heat Transfer

$$Q_{conv} = h_c \times A \times (\Delta T) \tag{18.23}$$

Where;

$$h_c = \text{convective heat transfer coefficient } \left(w/m^2k\right)$$

12.2 Conduction Heat Transfer

$$Q_{cond} = \frac{kc \times A \times (\Delta T)}{l} \tag{18.24}$$

Where;

K_c = thermal conductivity coefficient $(w/mk) = 59$ w/mk

d = heat transfer diameter (m) 0.03=d×5=0.15 m, A=$^{\pi d2}/_4$=0.017671 m^2

ΔT = (temperature of furnace-temperature of water)° k

L = length of furnace $0.3 = L \times 5 = 1.5$ m

$$Q_{cond} = \frac{59 \times 0.0177 \times (535)}{1.5} = 37\text{I}.87 \text{ w}$$

13 Return Chamber Heat Transfer Calculation

13.1 Radiative Heat Transfer

$$Q_{rad} = \varepsilon\sigma \left(T_h^4 - T_c^4\right) A_c \tag{18.25}$$

Where;

q_{rad} = heat transfer per unit time (W)

13.2 Convective Heat Transfer

$$Q_{conv} = h_c \times A \times (\Delta T) \tag{18.26}$$

Where;

$$h_c = \text{convective heat transfer coefficient } \left(w/m^2k\right)$$

13.3 Conduction Heat Transfer

$$Q_{cond} = \frac{kc \times A \times (\Delta T)}{l} \tag{18.27}$$

Where;

$$K_c = \text{thermal conductivity coefficient } (w/mk)$$

14 Heat Transfer by the Fire Tubes from Return Chamber 1(2nd Pass) to Return Chamber 2 (3rd Pass)

14.1 Radiative Heat Transfer

$$Q_{rad} = \varepsilon\sigma \left(T_h^4 - T_c^4\right) A_c \tag{18.28}$$

Where;

$$q_{rad} = \text{heat transfer per unit time } (W)$$

14.2 Convective Heat Transfer

$$Q_{conv} = h_c \times A \times (\Delta T) \tag{18.29}$$

Where;

$$h_c = \text{convective heat transfer coefficient } \left(w/m^2k\right)$$

14.3 Conduction Heat Transfer

$$Q_{cond} = \frac{kc \times A \times (\Delta T)}{l} \tag{18.30}$$

Where;

$$K_c = \text{thermal conductivity coefficient } (w/mk)$$

15 Heat Transfer by the Fire Tubes from Return Chamber 2 (3rd Pass) to the Smoke Stack

15.1 Radiative Heat Transfer

$$Q_{rad} = \varepsilon\sigma \left(T_h^4 - T_c^4\right) A_c \tag{18.31}$$

Where;

$$q_{rad} = \text{heat transfer per unit time } (W)$$

Convective heat transfer

$$Q_{conv} = h_c \times A \times (\Delta T) \tag{18.32}$$

Where;

$$h_c = \text{convective heat transfer coefficient } \left(w/m^2k\right)$$

15.2 Conduction Heat Transfer

$$Q_{cond} = \frac{kc \times A \times (\Delta T)}{l} \tag{18.33}$$

Where;

$$K_c = \text{thermal conductivity coefficient } (w/mk)$$

16 Smoke Stack Heat Loss Calculation

16.1 Radiation Heat Loss

$$Q_{rad} = \varepsilon\sigma \left(T_h^4 - T_c^4\right) A_c \tag{18.34}$$

Where;

$$q_{rad} = \text{heat transfer per unit time (W)}$$

16.2 Convective Heat Loss

$$Q_{conv} = h_c \times A \times (\Delta T) \tag{18.35}$$

Where;

$$h_c = \text{convective heat transfer coefficient } \left(w/m^2k\right)$$

16.3 Conduction Heat Loss

$$Q_{cond} = \frac{kc \times A \times (\Delta T)}{l} \tag{18.36}$$

Where;

$$K_c = \text{thermal conductivity coefficient (w/mk)}$$

16.4 Amount of Steam Generated

The amount of steam generated is calculated from the formula below;

$$M_s = \frac{qt}{he} \tag{18.37}$$

Where;

m_s = mass of steam (kg/h)

q_t = calculated total heat transfer (kw)

h_e = evaporation energy of steam (kj/kg)

From steam table; under saturated steam of P_{sat} at 111.4°C

17 Boiler Efficiency

Basically boiler efficiency can be tested by the following methods;

- **The direct method;** where the energy gain of the working fluid (water and steam) is compared with the content of the boiler fuel.
- **The indirect method;** where the efficiency is the difference between the losses and the input.

In working out the efficiency of our boiler we will be applying the direct method.

18 Direct Method

This is also known as 'input – output method' due to the fact that it needs only the useful output (steam) and the heat input (i.e. fuel) for evaluating the efficiency. This efficiency can be evaluated using the formula mentioned in Fig. 18.4;

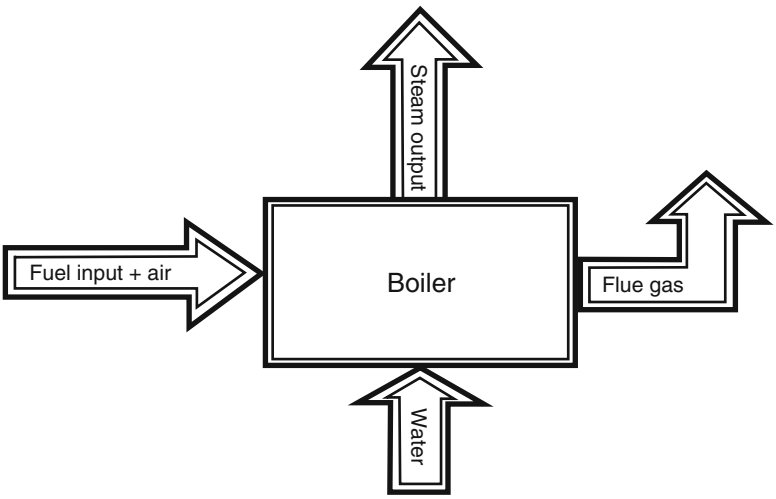


Fig. 18.4 Input and output

$$\text{Boiler efficiency} = \frac{\text{heat output} \times 100}{\text{heat input}}$$

$$\text{Boiler Efficiency} = \frac{37.93 \times 100}{55}$$

$$= 68.96\ \%$$

$$\sim = 69\ \%$$

19 Test and Result

The fabricated fire tube boiler was tested to evaluate its performances, efficiency and determine its evaporation ratio.

The purpose of the performance test is to determine the actual performance and efficiency of the oiler and compare it with design values or norms. It is an indicator for tracking day-to-day and season-to-season variations in boiler efficiency and energy efficiency improvements.

19.1 Test

When the burner is turned on and ignition occurs which produces the required fire in the furnace inside the furnace, a hot flue gas is produced which is forced through the fire-tubes (by the he p of blower in the burner) and heat is thus transferred into the water which in turn results in production of the required steam that may be used for industrial purposes.

19.2 Result

The result in this case is a torque produced at a steam pressure of 1.5 bar and a steam temperature of 111.4 °C also raising the temperature of the water from 30 °C to a generated steam quantity of 61.34 kg/h, with a diesel quantity of 5.2Htres/h. The efficiency of the burner after getting an adequate combustion air/fuel ratio and heat delivery from the burner resulted into 64.3 %. The efficiency of the boiler was also calculated to be 69 %.

20 Conclusion

The following conclusions can be drawn from the data and research that has been mentioned in this paper. Thermodynamics, heat transfer and strength of materials analysis subjected to temperature and pressure variations were conducted in the theoretical framework of the laboratory fire-tube steam boiler. Dimensions of

major and secondary parts were estimated from computations from the theoretical framework and 3D modelling process for the steam boiler was then carried out to present various working drawings of the steam boiler for possible construction.

Good boiler design practices must take into account the operation of the boiler and not simply the heat transfer, parameters that a good boiler design addresses include;

- Ample furnace volume must be included to absorb a significant portion of Radiative heat transfer and allow the low NO_x burner designs to function.
- Optimized pressure drop across the boiler convective passes, the pressure drop determines the fan size required for the boiler application.
- Ample steam storage and steam height. The volume of steam and distance from the steam nozzle to the normal water level determine to a very large extent the steam quality and the amount of water that will be carried over into the system. Boiler design and optimization programs have been written to determine the performance of fire-tube boilers. These programs can be applied to analyze a wide variety of the boiler scenarios for many different boiler applications extending from simple gas fired systems to complex waste heat applications.

In-flame gas temperature data for lire-tube boilers has been obtained. The data follows expected trends and has been very useful in the validation of predictive optimization models. This data is compared to predicted results from computational fluid dynamic combustion models and good agreement has been found. This data is used to optimize furnace and heat transfer surfaces for typical fire-tube boilers. Gas temperatures measured at the entrance to the convective tube surfaces provided excellent data that validated the heat transfer sub-models augmented surface tubes have proven to be a valuable resource in the design of fire-tube boilers for many special applications. The advantages of thee augmented tube are that it allows the designer to include larger steam storage and steam height resulting in higher steam quality and rapid load swing handling ability. Using the augmented tube also allows the designer to have a lower overall pressure drop with a boiler efficiency that is still over 81 %. The augmented tube boiler may be used to reduce the boiler shell diameter and still maintain standard steam volumes, steam heights, and boiler efficiency.

Conclusively, a simple laboratory lire-tube steam boiler is herein presented for fabrication, testing and further improvement. Production of a simple steam boiler of this sort will enable the availability of portable and affordable steam boilers for steam generation processes, especially in school laboratories. The availability of steam boilers in school laboratories will enhance students learning process, especially in the area of thermodynamics, heat transfer and energy studies.

21 Recommendation

I am recommending this research work to any other research institute, students and public to serve as a basis for further research and improvements on fire-tube boiler design. We also recommend that other per-heater devices and feed water pump should be applied for higher efficiency.

Appendix 1: The Technical Futures of the Fire-Tube Boiler

Fig. 18.5 The fabricated fire-tube boiler

Fig. 18.6 Side view of the boiler

Fig. 18.7 Plan of the boiler

Appendix 2: Difference Technical Views of the Fire-Tube Boiler

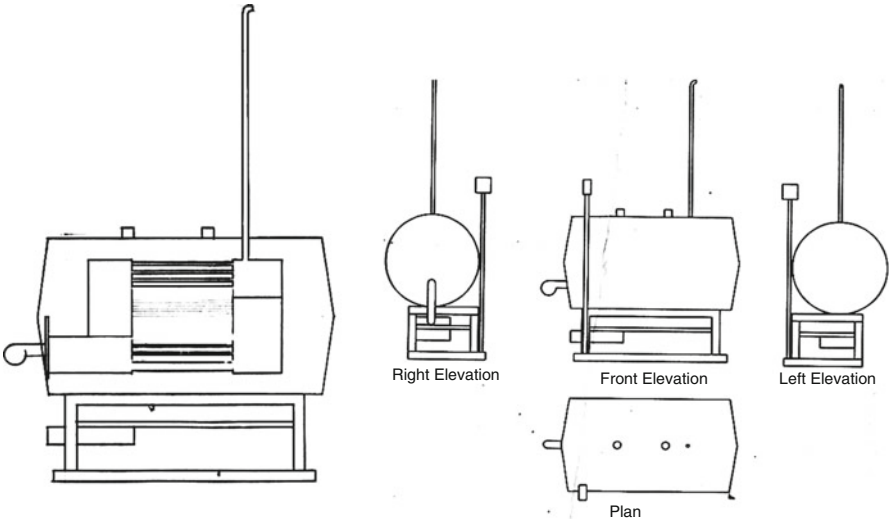


Fig. 18.8 Difference technical views of the boiler

References

1. American Society of Mechanical Engineers (ASME): Boiler and Pressure Vessel Code (BPVCK www.go.asme.org/bpvc13) (2010)
2. American Boiler Manufacturers Association (ABMA) White Paper: Historical Advancement of Fire-side Heating surface in fire-tube Boilers. PDF (2004)
3. Gbasouzor, A.I.: Fabrication of Pilot Multi-Tube Fire-Tube Boiler Designed for Teaching and Learning Purpose in Mechanical Laboratory, Lecture Notes in Engineering and Computer Science, in Proceedings of the World Congress on Engineering 2014, WCE2014, pp. 1478–1486. London. 2–4 July 2014
4. Boiler and Heaters, Improving Energy Efficiency, Canadian Industry Program -for Energy Conservation (August 2001) https://oee.nrcan.gc.ca/sites/oee.nrcan.gc.ca/files/pdf/publications/infosource/pub/cipec/boilersheaters.pdf (2006)
5. Cengel, Y.A., Boles, M.A.: "Thermodynamics: An Engineering" Approach, 5th edn. McGraw Hill, Companies Inc., Boston (2006)
6. Department of Energy (DOE): Fundamentals Handbook. Thermodynamics, heat transfer, and fluid flow. U.S. Department of Energy, Washington, DC [Online] (2014)
7. Landis, F.: Steam Engine Source: Compton's interactive Encyclopedia http://www.fem.unicamp.br/~em313/paginas/consulte/steame.htm (2014)
8. Kitto, J.B., Stuitz, S.C.: Steam: Its Generation and Use, 41st edn. The Babcock and Wilcox Company, Barberon (2005)
9. Saidur, R., Ahamed, J.U., Masjuki, H.H.: Energy, exergy and economic analysis of industrial boilers. Energy Policy 38, 2188–2197 (2010)

Chapter 19
Experimental and Computational Studies on Effects of Scale-Up of Riser on Heat Transfer Chatacterisitcs of Circulating Fluidized Bed

Ranjit S. Patil, Pinakeswar Mahanta, and Manmohan Pandey

Abstract In the present paper, CFD simulations using ANSYS-Fluent 14.5 were accomplished to study the effect of bed or sand inventory and effect of scale-up of riser on heat transfer characteristics like temperature and heat transfer coefficient. To accomplish the scale-up study, 3D CFD simulations were performed on the Circulating Fluidized Bed (CFB) risers of cross section 0.15×0.15 m, 0.30×0.30 m, and each of height 2.85 m. CFD simulations to predict heat transfer characteristics were accomplished under same operating conditions on heated portion (heater) of both risers. The walls of heater were maintained at the constant heat flux $q'' = 1{,}000$ (W/m^2). Modeling and meshing were done using ProE and ANSYS ICEM CFD software, respectively. RNG k-ε model was used for turbulence modeling. Eulerian model with Gidaspow phase interaction scheme was used to simulate the two phase flow (air + sand mixture flow). Computational (CFD simulation) data was compared with experimental data for the validation purpose. After validation, further simulations were conducted on riser of cross section 0.30×0.30 m. Based on scale-up study, empirical correlation has been developed to predict the heat transfer coefficient.

Keywords Ansys • Fluent • CFB • CFD • Heat transfer characteristics • Riser • Scale-up

R.S. Patil (✉)
Assistant Professor, Department of Mechanical Engineering, BITS Pilani – K K Birla Goa Campus, Sancoale, Goa 403726, India
e-mail: ranjitp@goa.bits-pilani.ac.in

P. Mahanta • M. Pandey
Professor, Department of Mechanical Engineering, IIT Guwahati, Guwahati 781039, India
e-mail: pinak@iitg.ernet.in; manmohan@iitg.ernet.in

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_19

1 Introduction

Present paper in a broad way is associated with energy sector – power plant engineering. In this domain, it is concerned with Circulating Fluidized Bed (CFB). CFB has several applications and can be used as a boiler (heat exchanger), dryer, gasifier, combustor etc. A large number of CFB units are installed for power generation throughout the world. The increase in diversity in CFB applications demand the need for the development of more efficient experimental techniques, realistic simulations, and other research and design tools.

Versatile tool like CFD and related software's may be therefore used to accomplish the research with accuracy and also to overcome the limitations of experimental aspects.

Reference [1] reported that CFD simulation is a versatile tool to simulate two phase problems to predict heat transfer characteristics such as temperature, heat transfer coefficient, and hydrodynamic characteristics such as pressure, velocity, volume fraction etc.

References [2–4] reported the Eulerian–Eulerian model with kinetic theory of granular flow, which is currently the most applicable approach to compute gas–solid flow in a CFB.

References [5–7] reported on different drag models, which are used to predict the most representative gas–solid interphase exchange coefficient.

Reference [2] reported on CFD simulation of gas/particle flow behavior in the riser section of a CFB, which was simulated using Fluent to predict velocity, volume fraction, pressure, and turbulence parameters for each phase.

Reference [8] reported a simulation study using Fluent- CFD software using multifluid Eulerian model integrating the kinetic theory for solid particles using, which was capable of predicting the gas–solid behavior of a fluidized bed. Time-average bed pressure drop and local voidage profiles showed similarities between the model predictions and experimental results.

Reference [9] reported that solid-phase wall boundary condition had minor effect on axial voidage profile when the Gidaspow drag model was used.

Literature review reveals that many researchers have reported CFD simulations based on Eulerian model to predict only hydrodynamic characteristics for two phase flow in the CFB riser.

Reference [10–12] reported that CFB risers of square and rectangular cross-sections are now widely used in boilers, biomass dryer etc. It is also reported that, there is a high demand to study the effects of scale-up on heat transfer characteristics of square cross-section risers. Experimental scale-up study was performed at IIT Guwahati to predict the effect of scale-up on heat transfer characteristics using three CFB units as in Fig. 19.1, each of height 2.85 m with riser cross sections of 0.15×0.15 m, 0.20×0.20 m, and 0.25×0.25 m, respectively. However, because of limitations like requirement of extra fund and extra space in laboratory to launch new experimental CFB set up having riser of cross section 0.30×0.30 m, present paper approached to accomplish the further scale-up study through CFD simulations, which were performed under same operating conditions on upper

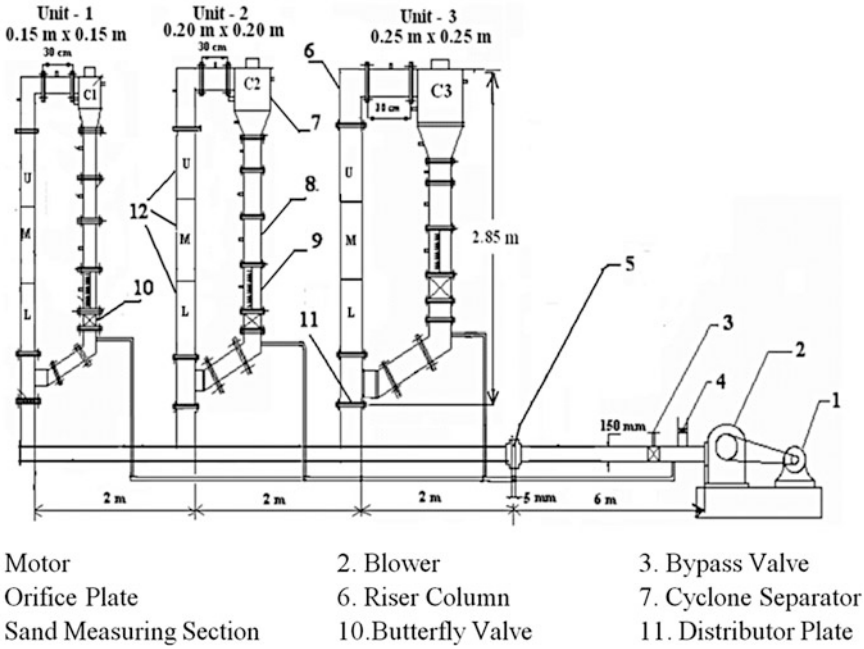


Fig. 19.1 CFB setup at IIT Guwahati

dilute region of risers of different cross sectional area to predict heat transfer characteristics. Hence present paper describes computational methods to predict the effect of change in geometrical parameter like change in cross sectional area of riser on heat transfer characteristics at upper dilute region of CFB risers. Effect of sand inventory on heat transfer characteristics also predicted at lower denser region. Empirical correlation has been developed to predict the heat transfer characteristics.

2 Experimental Setup

Figure 19.1 presents the three CFB setups having risers of cross section area 0.15×0.15 m, 0.20×0.20 m, and 0.25×0.25 m with each riser height of 2.85 m designed and fabricated at IIT Guwahati [13]. Riser of CFB setup was made up of plexiglass to facilitate flow visualization. A positive displacement type blower powered by a 20 HP motor supplies air. Heater section's positions are *U* (upper dilute or splash region of particles), *M* (middle dense region of particles) and *L* (lower splash or dense region). Experiments were conducted on the CFB unit with sand as the bed material and air as the fluidizing medium. Heat transfer characteristics along the riser were studied with incorporation of heater section in

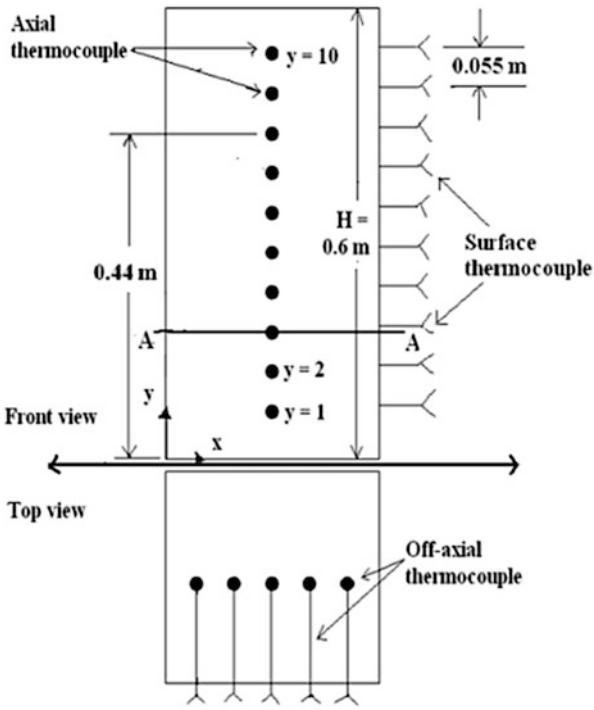


Fig. 19.2 Position of the thermocouples

the upper, middle and lower splash regions; having the same cross sectional area as that of the riser (0.15×0.15 m) and height of 0.6 m. The heater section was fabricated with MS sheet of 2 mm thickness with a height of 0.6 m. Nichrome wire or heater coil of 2 kW capacities was wound over the mica sheet of 1.5 mm thickness which covers the MS wall of the heater section. Another mica sheet, which acts as an electric insulator, was wrapped over the Nichrome wire. To avoid the heat losses by radiation, ceramic wool and asbestos sheets were wrapped over the assembly. Heat was supplied to the heater section with electrical supply through an auto transformer. To measure the temperature of the surface of the heater section and the bed, calibrated *T*- type thermocouples were installed at the same height on the wall as well as inside the heater section respectively as in Fig. 19.2.

Ten set of thermocouples with equal spacing of 5.5 cm along the height of the heater section were used to measure the bed temperature and surface temperature of the heater section, as in Fig. 19.2. A section *AA* was taken in the lateral direction at 0.16 m above the inlet of the heater and another one section was taken in the lateral direction 0.44 m above the inlet of the heater. Five thermocouples were placed along the horizontal direction in these sections with equal spacing at the non-dimensional distance [*X/b*] of 0.1, 0.3, 0.5, 0.7, and 0.9, respectively as in Fig. 19.2. Here the non-dimensional distance [*X/b*] is the distance *X* measured from the left hand side wall of the heater to the thermocouple end, normalized with respect to the width *b* of the heater.

3 Experimental Heat Transfer Study

Steady state experiments were conducted on the CFB setups under similar operating conditions to examine the effect of change in riser cross section, effect of bed or sand inventory on heat transfer characteristics. Experiments were carried out at five superficial velocities varying from 2.3 m/s to 3.0 m/s or at non-dimensional velocity ratios U^* (ratio of superficial velocity and minimum fluidizing velocity) of 5–8, particle size 460 μm and heat flux 1,000 W/m^2, with initial sand inventory above the distributor as 1,750 N/m^2. Further, experiments were conducted with five superficial velocities varying from 2.5 m/s ($U^* = 5$) to 4.0 m/s ($U^* = 8$) for a sand inventory so that weight per unit area of distributor plate P of each CFB setup was 3,050 N/m^2 maintained in each CFB unit. Preliminary experiments reveal that below the lower limit of weight of sand (1,750 N/m^2) fast fluidization was not achieved. Similarly, beyond the upper limit (3,050 N/m^2) of weight of sand fluidization was not achievable due to the high load on the blower. Superficial velocity of air U was considered in the range of $U^* = 5$ to $U^* = 8$. This is because, at $U^* < 5$, fast fluidization was not achievable and $U^* > 8$ was limited by the capacity of the blower.

The local heat transfer coefficient h is calculated by

$$h = Q / [A_S \cdot (T_S - T_B)] \quad W.m^{-2}.K^{-1} \tag{19.1}$$

where Q is rate of heat supplied to the heater, measured using a Wattmeter. A_s is the surface area of heater, $q'' Q/A_s$ is the heat flux, T type calibrated thermocouples and data acquisition system with Dasy Lab software version 8.0 was used to measure the surface temperature T_S and bulk mean bed temperature T_B. The local heat transfer coefficient is measured at 10 locations (y = 1 to 10 as in Fig. 19.2) along the height of heater. Average heat transfer coefficient (h_{avg}) along the heater section at its any particular location above the distributor plate is calculated by

$$h_{avg} = \frac{1}{H}\int_0^H h_y.dy \quad \mathrm{W.m^{-2}.K^{-1}} \tag{19.2}$$

where H is the height of the heater 0.6 m, h_y is the local heat transfer coefficient. Local heat transfer coefficient h_y is calculated at 10 different points (y = 1, 2 . . . . . .10 as shown in Fig. 19.2) along the height of heater section.

Uncertainty analysis was carried out for the heat transfer coefficient. Uncertainty is depending upon connections of thermocouples, accuracy of T type thermocouple ±0.5 °C, wattmeter accuracy ±5 W, accuracy in length measurement ±1 mm etc. Uncertainty analysis, using the method of Kline and McClintok [14], showed that the heat transfer coefficients estimated in the present study were within ±4 %. Please give a label to each figure/table. Each label should be numbered and with a brief description.

4 CFD Modeling and Simulation

In Ansys – Fluent [15] the governing equations are discretized using the finite volume technique. The discretized equations, along with the boundary conditions, are solved to obtain a numerical solution. Eulerian multiphase model was used for the simulation of air-sand flow. 3-D CFD simulations were done on heater section (portion U as in Fig. 19.1 each of height 0.6 m). While meshing the heater, cell type selected was tetrahedral/hybrid with T grid meshing scheme. Selected boundary conditions as in Fig. 19.3 were air velocity at inlet, bottom of riser = varying from 2.30 to 4 m/s, and volume fraction of sand = 0; volume fraction of sand at inlet at right hand side wall of riser = 1 and sand velocity = varying from 0.00468 m/s to 0.0075 m/s; outlet boundary condition was pressure outlet at top of the riser = 0 Pa gauge pressure of air-sand mixture. While simulating the riser in Ansys, 3D steady state solver was used. Other parameters selected were density of sand = 2,600 kg/m^3, mean diameter of sand = 460 μm, density of air = 1.225 kg/m^3, turbulence model used = RNG k-ε model, numerical method used for pressure velocity coupling = phase coupled SIMPLE, discretization scheme = 1st order upwind, under relaxation parameters for pressure = 0.1, density = 0.1, body forces = 0.1, momentum = 0.1, volume fraction = 0.2, energy = 0.1; convergence criteria = 0.001, solution initialization = from all zones,

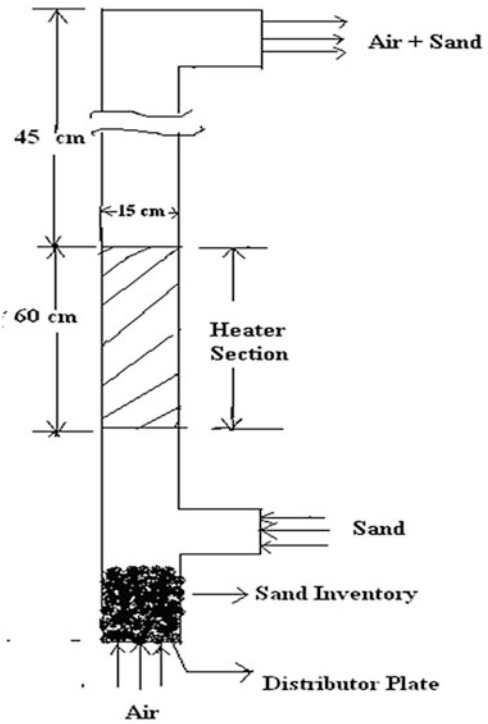


Fig. 19.3 CFB riser (figure drawn not to the scale)

height of the sand inventory above the distributor plate in the different CFB risers was maintained to obtain the static pressure on distributor plate as 1,750 N/m^2 or 3,050 N/m^2.

In the Eulerian multiphase (gas-solid, two fluids) model, conservation equations of mass and momentum for both phases are developed and solved simultaneously. The link between the two phases is through the drag force in the momentum equations.

Continuity (kth phase)

$$\frac{\partial}{\partial t}\left(\varepsilon_k \rho_k\right) + \nabla \cdot \left(\varepsilon_k \rho_k \vec{u}_k\right) = \sum_{p=1}^{n} \dot{m}_{pk} \tag{19.3}$$

where

$$k = f \;\; for \;\; fluid$$
$$k = s \;\; for \;\; solids.$$

Momentum (fluid phase)

$$\begin{aligned} &\tfrac{\partial}{\partial t}\left(\varepsilon_f \rho_f \vec{u}_f\right) + \nabla \cdot \left(\varepsilon_f \rho_f \vec{u}_f \otimes \vec{u}_f\right) \\ &\quad = -\varepsilon_f \nabla p + \nabla \cdot \bar{\bar{\tau}}_f + \varepsilon_f \rho_f \vec{g} + K_{sf}\left(\vec{u}_s - \vec{u}_f\right) + \vec{F}_f \end{aligned} \tag{19.4}$$

Momentum (solids phase)

$$\begin{aligned} &\tfrac{\partial}{\partial t}\left(\varepsilon_s \rho_s \vec{u}_s\right) + \nabla \cdot \left(\varepsilon_s \rho_s \vec{u}_s \otimes \vec{u}_s\right) \\ &\quad = -\varepsilon_s \nabla p - \nabla p_s + \nabla \cdot \bar{\bar{\tau}}_s + \varepsilon_s \rho_s \vec{g} + K_{sf}\left(\vec{u}_f - \vec{u}_s\right) + \vec{F}_s \end{aligned} \tag{19.5}$$

Total volume fraction conservation

$$\varepsilon_s + \varepsilon_f = 1 \tag{19.6}$$

Equation (19.3) represents the mass balance of each phase with temporal and spatial gradients on the left hand side and mass creation $\dot{m}$ of the p_{th} species (in this case, zero) by reaction or phase change. Equations 19.4 and 19.5 are momentum conservation equations for the fluid (air in this case) and solid phase, respectively. The left side represents temporal and spatial transport terms whereas the right hand side has terms for the various interaction forces involved. Note that the hydrodynamic pressure is shared by both phases and hence, the gradient of pressure p is premultiplied by the respective volume fractions ε in both equations. ρ, $\vec{u}$ and $\vec{g}$ represent to density, velocity and acceleration due to gravity respectively. The stress term τ_f represents the shear stress in gas phase in (19.4). Equation 19.5 represents the solids phase equation, where τ_s represents the shear stress term due to collision among particles.

Terms $K_{sf}\left(\overrightarrow{u}_f - \overrightarrow{u}_s\right)$ and $K_{sf}\left(\overrightarrow{u}_s - \overrightarrow{u}_f\right)$ represent the momentum exchange or drag between the two phases in (19.4) and (19.5). These terms are equal in magnitude and opposite in sign and account for the friction at the interface between the phases. The terms $\overrightarrow{F}_f$ in (19.4) and $\overrightarrow{F}_s$ in (19.5) represent all other forces that may affect the flow, such as electrical, magnetic and other effects.

The drag is an effective way of representing the surface integral of all the forces that exist at the interface between the phases. Interphase momentum exchange factor of Gidaspow's drag closure as in (19.7).

$$\begin{aligned} &For \\ &\varepsilon_g > 0.8, \\ &K_{s,f} = \tfrac{3\varepsilon_s \varepsilon_g \rho_g}{4 d_s} C_{d,s}\left(\varepsilon_g{}^{-2.65}\right)\left|\overline{u_s} - \overline{u_g}\right| \\ &C_{d,s} = \tfrac{24}{\varepsilon_g.\mathrm{Re}_s}.\left[1 + 0.15.\left(\varepsilon_g.\mathrm{Re}_s\right)^{0.687}\right] \\ &For \\ &\varepsilon_g \le 0.8, \\ &K_{s,f} = 150.\tfrac{\varepsilon_s{}^2.\mu_g}{\varepsilon_g.d_s{}^2} + 1.75.\tfrac{\varepsilon_s.\rho_g}{d_s}.\left|\overline{u_s} - \overline{u_g}\right| \end{aligned} \tag{19.7}$$

where Re_s and d_s are the Reynolds number and diameter of solid particles respectively and other symbols have their standard meaning which already are defined.

Entire CFB riser of 0.15×0.15 m was modeled and meshed with 230,982 tetrahedral cells and for riser 0.30×0.30 m, 512,018 tetrahedral cells which were selected for its simulation after grid independence test. Energy Eq. 19.8 was applied during heat transfer 3-D simulations for heater section as in Fig. 19.4.

$$\frac{\partial}{\partial t}(\rho E) + \nabla.(\overline{v}(\rho E + p)) = \nabla.\left(k_{eff}\nabla T - \sum_j \dot{h_j} j_j + \left(\overline{\overline{\tau}}_{eff}.\overline{v}\right) + S_{hj}\right) \tag{19.8}$$

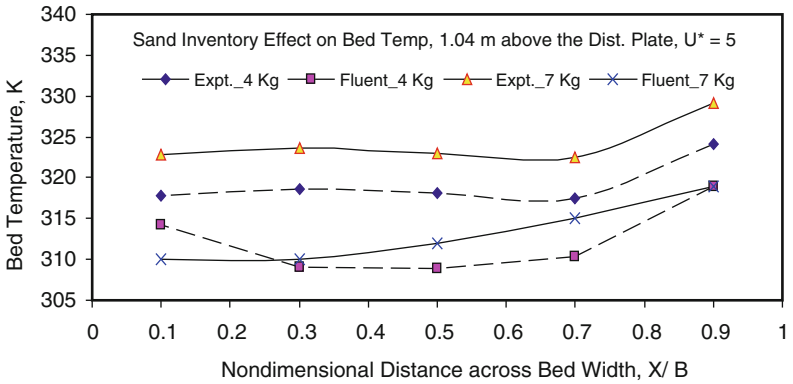


Fig. 19.4 Bed temperature distribution

The four terms of the right hand side in (19.8) represents energy transfer due to conduction, species diffusion, viscous dissipation and volumetric heat sources S_{hj}.

Now, heat transfer simulations (by enabling energy equation) for the heater section were carried out using both multiphase models to obtain the bulk mean bed temperature T_b and wall temperature T_s. Simulations were carried at constant heat flux $q'' = 1{,}000$ W/m^2 for sand inventory 1,750 and 3,050 N/m^2 and for a superficial velocity of air varying from 2.3 to 4 m/s. Local convective heat transfer coefficient h is calculated by using (19.1). Simulated results were compared with experimental results.

5 Results and Discussion

5.1 Studies on Effect of Bed Inventory on Heat Transfer Characteristics

Figures 19.4 and 19.5 show the effect of bed inventory on the distribution of bed temperature and heat transfer coefficient. Eulerian multiphase model with Gidaspow phase interaction scheme was used to simulate the two phase flow. Inventory in the bed was varied from 4 kg i.e. 1,750 N/m^2 to 7 kg i.e. 3,050 N/m^2. Parameters velocity $U^* = 5$, heat flux 1,000 W/m^2, and particle size 460 μm were maintained same for the simulations conducted for the different inventories. It is observed that increase in the inventory of sand in the riser increases the bed temperature and heat transfer coefficient. This is because sand particles concentration i.e. suspension density increases with increase inventory of the bed. Consequently, more quantity of

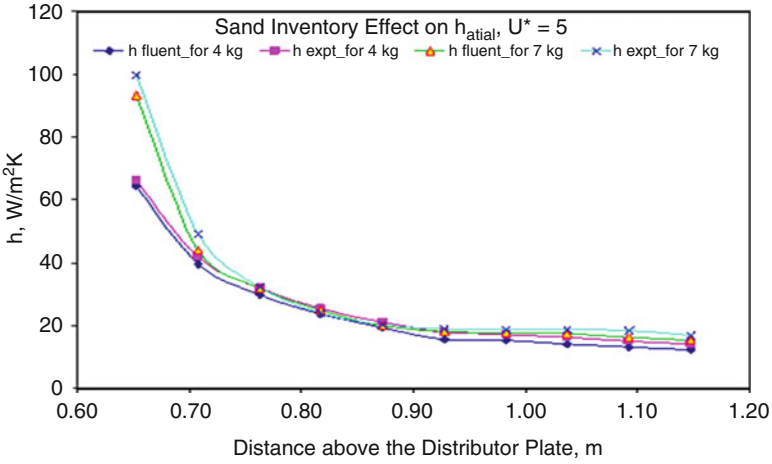


Fig. 19.5 Axial distribution of heat transfer coefficient

particles in the lower splash region promotes more heat transfer through conduction, because of which bulk temperature of bed across the sections taken at 1.04 m above the distributor plate was observed to be higher than that for the lower inventory.

5.2 *Studies on Effect of Scale-Up on Heat Transfer Characteristics*

Effects of riser cross section on heat transfer characteristics were investigated as shown in Figs. 19.6, 19.7, 19.8, and 19.9. Riser of bed cross section 0.30×0.30 m was modeled and simulated using Ansys-Fluent. Results obtained on bed temperature distribution and heat transfer coefficient were compared with results obtained

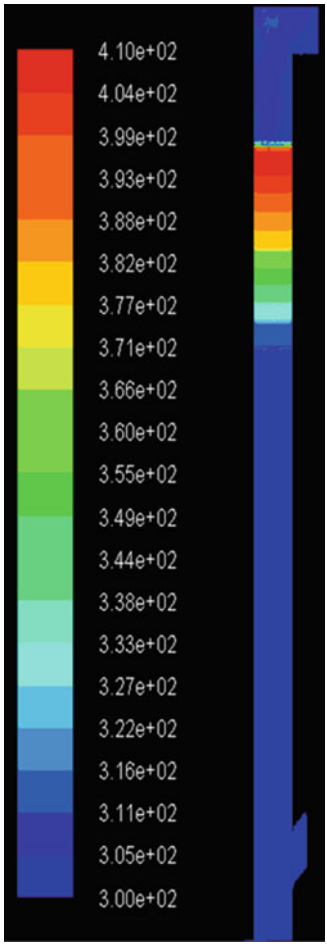


Fig. 19.6 Temperature counter (K) – upper splash region of 0.15×0.15 m at $U = 4$ m/s and $P = 3{,}050$ N/m^2

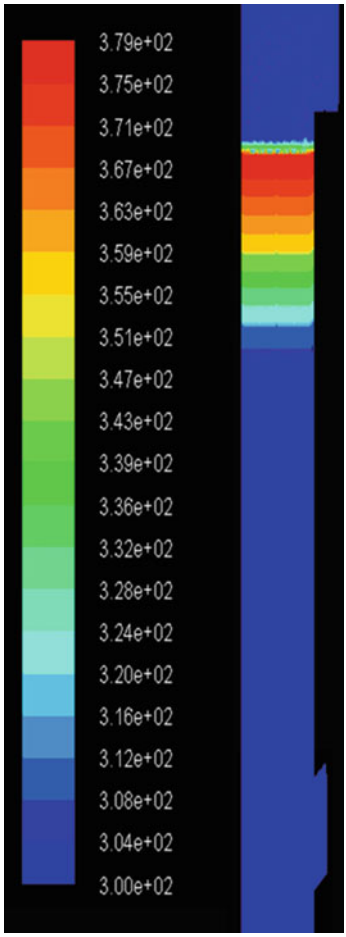


Fig. 19.7 Temperature counter (K) – upper splash region of 0.30 × 0.30 m at U = 4 m/s and P = 3,050 N/m^2

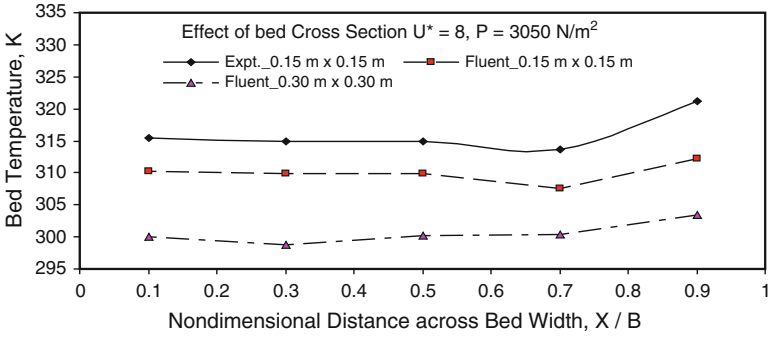


Fig. 19.8 Bed temperature distribution

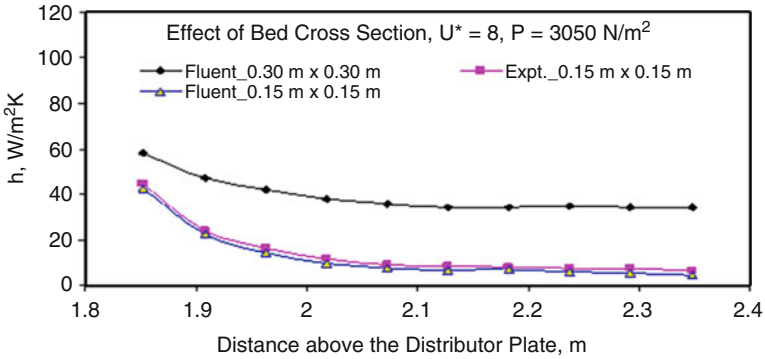


Fig. 19.9 Axial distribution of heat transfer coefficient

0.15×0.15 m CFB bed. Simulations were completed for the upper splash region for both the risers with superficial velocity of air as 4 m/s, and maintaining the pressure 3,050 N/m^2 due to weight of sand inventory per unit area of the distributor plate. Eulerian multiphase model with Gidaspow phase interaction scheme was used to simulate the two phase flow. Particle size used was 460 μm and heat flux of 1,000 W/m^2 was applied at the walls of heater. Sand velocity at inlet at right hand side wall of risers was 0.0075 (m/s) at $U^* = 8$ for CFB unit 1 (0.15×0.15 m) and CFB unit of cross section 0.30×0.30 m. Remaining details are mentioned in Sect. 4 of present paper. Temperature counters are presented in Figs. 19.6 and 19.7.

Figure 19.8 represent the bed temperature distribution across a bed cross section at 2.24 m above the distributor plate and axial variation of local heat transfer coefficient, respectively. Average bed temperature at lateral section at 2.24 m above the distributor plate was more in smaller cross section heater than larger cross section heater. This is expected because sand inventory in larger cross section CFB setup was kept proportionately more than the smaller size CFB setup so as to maintain the same weight of sand per unit area of the distributor plate. Therefore, weight of sand particles per unit surface area of the larger cross section heater was more than the smaller heater. Therefore at same heat flux applied at heater wall of each CFB setup, distribution of amount heat extracted due to conduction from wall of the heater took place into large number particles, which were comparatively more in larger cross section CFB setup, hence average bed temperature was less for lager size heater than smaller heater. As shown in Fig. 19.9, it is observed that in the wall-to-bed heat transfer study, heat transfer coefficient increases with increase in bed size. This is because the driving temperature difference $(T_S - T_B)$ in the larger size bed was lesser than the smaller size beds, while the heat flux was held fixed. This is because of the higher concentration of solid particles near the wall of the larger bed, and consequently, lower thermal resistance from the bed-to-wall causing better heat conduction.

5.3 Correlation

A dimensional analysis was done using Rayleigh's method. Nondimensional numbers were Nusselt number, Reynolds number, non-dimensional density ratio $[\rho_{sus}/\rho_g]$ (ratio of suspension density to the density of gas, i.e. air), and non-dimensional geometrical parameter $[H/B]$ (ratio of height of the heater to the hydraulic diameter of the bed). A best-fit equation involving these nondimensional numbers was obtained using Findfit function of Mathematica [16]. The best-fit equation is as follows.

$$\mathrm{Nu_B} = 42.98[\mathrm{Re_B}]^{0.238}\left[\rho_{\mathrm{sus}}/\rho_g\right]^{0.25}[\mathrm{H/B}]^{-1.571} \tag{19.9}$$

Variation of Prandtl number Pr was not significant enough to cause significant variation in the predicted value of heat transfer coefficient. Therefore it has not been included in the empirical correlation. The correlation Eq. 19.9 is valid in the following ranges:

$$20,629 < Re_B < 84270, 6 < \left[\rho_{sus}/\rho_g\right] < 270, 0.7 < [H/B] < 4$$

Figure 19.10 shows the comparison of the experimental data of present work plus data obtained from literature with the prediction of the correlation Eq. 19.9 showing

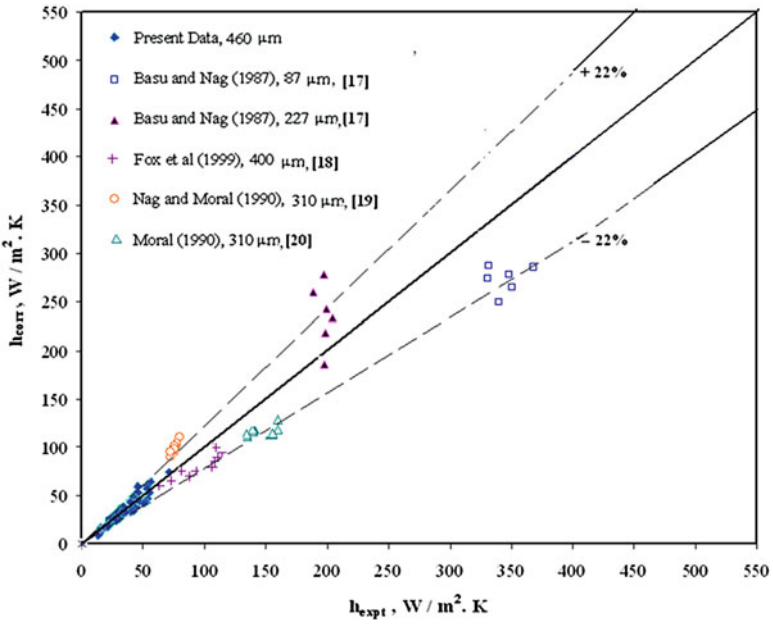


Fig. 19.10 Comparison of experimental data with proposed correlation

Table 19.1 Experimental data on various parameters

Author	U (m/s)	ρ_{sus} (kg/m^3)	Hr (m)	H/B	d_p (μm)
Basu and Nag [17]	3–5	22–96.79	5.5	0.98	227
Fox et al. [18]	3.5–8.3	30–170	5.966	0.69	400
Basu and Nag [17]	3.7–5	21.50–58.63	5.5	0.98	87
Nag and moral [19]	7.2–12.5	25–62	5.15	2.55	310
Moral [20]	7.2–12.5	25–62	5.15	1.7	310
Present work	2.5–4	7.2–288	2.85	2–4	460

rms deviation of ± 22 %. Experiments of other researchers were carried out over a wide range of velocity, bed density particle size, and *[H/B]* ratio. Table 19.1 represents the operating range of various operating parameters and non-dimensional numbers of different researchers. In order to facilitate the easy comparison of predicted values with experimental values, all data were plotted in Fig. 19.10, with the measured heat transfer coefficient and the theoretical prediction as the coordinates. The middle line (line at 45°) is the line of prefect agreement, and other two extreme lines show the boundary of ± 22 %.

6 Conclusion

Heat transfer experiments were conducted on in-house fabricated CFB setup – at the lower denser, middle and upper dilute region of sand particles via heater, and 3-D numerical simulations using Fluent. Numerical and experimental results were in good agreement for the riser of cross sectional area 0.15×0.15 m. To accomplish the scale-up study, further CFD simulations were done on riser having cross sectional area 0.30×0.30 m. Bed temperature decreases and heat transfer coefficient increases with increase in riser cross sectional area. Effect of sand inventory on heat transfer characteristics was also studied. Bed temperature and heat transfer coefficient increases with increase in bed inventory. Based on experimental and computational study, new correlation was developed for bed Nusselt number as a function of Reynolds number, non-dimensional density parameter, and non-dimensional geometrical parameter.

References

1. Patil, R.S., Pandey, M., Mahanta, P.: CFD simulations for heat transfer and hydrodynamic characteristics of circulating fluidized bed riser. In: 1st International Conference on Energy Engineering and Eco- Balance (IC-EEE), MITR. Pune, 16–18 Feb 2009
2. Benyahia, S., Arastoopour, H., Knowlton, T., Massah, H.: Simulation of particles and gas flow behavior in the riser section of a circulating fluidized bed using the kinetic theory approach for the particulate phase. Powder Technol. **112**, 24–33 (2000)

3. Zheng, Y., Wan, X., Qian, Z., Wei, F., Jin, Y.: Numerical simulation of the gas–particle turbulent flow in risers reactor based on k–ε–k_p–ε_p–Θ two fluid model. Chem. Eng. Sci **56**, 6813–6822 (2001)
4. Chan, C., Guo, Y., Lau, K.: Numerical modeling of gas–particle flow using a comprehensive kinetic theory with turbulent modulation. Powder Technol. **150**, 42–55 (2005)
5. Syamlal, M., Brien, T.O.: Derivation of a drag coefficient from velocity-voidage correlation, U.S. Department of Energy, Office of Fossil Energy, National Energy Technology Laboratory, Morgantown, April 1987
6. Gidaspow, D., Bezburuah, R., Ding, J.: Hydrodynamics of circulating fluidized beds, kinetic theory approach. In: Fluidization VII, Proceedings of the 7th Engineering Foundation Conference on Fluidization, Gold Coast (Australia). pp. 75–82 (1992)
7. Arastoopour, H., Pakdel, P., Adewumi, M.: Hydrodynamic analysis of dilute gas–solids flow in a vertical pipe. Powder Technol. **62**, 163–170 (1990)
8. Taghipour, F., Ellis, N., Wong, C.: Experimental and computational study of gas–solid fluidized bed hydrodynamics. Chem. Eng. Sc. **60**, 6857–6867 (2005)
9. Zhou, X., Gao, J., Xu, C., Lan, X.: Effect of wall boundary condition on CFD simulation of CFB risers. Particuology **11**, 556–565 (2013)
10. Patil, R.S., Pandey, M., Mahanta, P.: Study on the heat transfer characteristics in the lower splash region of circulating fluidized bed riser. In: 7th World Conference on Experimental Heat Transfer, Fluid Mechanics and Thermodynamics (ExHFT-7), pp. 555–561. Krakow (2009)
11. Patil, R.S., Pandey, M., Mahanta, P.: Parametric studies and effect of scale-up on wall-to-bed heat transfer characteristics of circulating fluidized bed risers. Exp. Thermal Fluid Sci. **35**, 485–494 (2011)
12. Patil, R.S., Pandey, M., Mahanta, P.: Effect of superficial velocity of air and riser cross sectional area on heat transfer coefficient of circulating fluidized bed. In: Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 1425–1430. London 2–4 July 2014
13. Pandey, M., Mahanta, P., Patil, R.S., Gawali, M.V.: Development of an experimental facility for the study of scale effects in circulating fluidized beds. In: National Conference on Recent Advances in Mechanical Engineering (NCRAME), pp. 170–175. Silchar (2008)
14. Kline, S.J., Mcclintok, F.A.: Describing uncertainties in single – sample experiments. Mech. Eng. **75**, 3–8 (1953)
15. Ansys Software, Version 13, Documentation
16. Mathematica Software, Find Fit Function, Version 10
17. Basu, P., Nag, P.K.: An investigation into heat transfer in circulating fluidized beds. Int. J. Heat Mass Transf. **30**(11), 2399–2409 (1987)
18. Fox, W.B., Grewal, N.S., Moen, D.A.: Wall-to-bed heat transfer in circulating fluidized beds. Int. Commun. Heat Mass Transf. **26**(4), 499–508 (1999)
19. Nag, P.K., Moral, M.N.A.: Effect of probe size on heat transfer at the wall in circulating fluidized beds. Int. J. Energy Res. **14**, 965–974 (1990)
20. Ali Moral, M.N.: Some studies on heat transfer in circulating fluidized beds, Ph.D. Thesis, Indian Institute of Technology Kharagpur, December 1990

Chapter 20
Predicting the Influence of the Machining Parameters on the Tool Tip Temperature

S.A. Chowdhury, M.N. Islam, and B. Boswell

Abstract This research has indicated that, by using a computer model, a numerical method can be effective in predicting the tool tip temperature. This allows the effects of changes made to the machining parameters to be simulated. Temperatures through the tool tip can therefore be reduced by using the optimum cooling method for particular machining parameters.

Keywords Air cooling • CFD • Cutting fluids • MQL • Tool-chip interface temperature • Vortex tube

1 Introduction

Measuring a tool tip interface temperature has never been an easy task to achieve using thermal couples due to the hostile environment in which thermal couples have to work. Limited success has been achieved from the use of infra-red pyrometers as the chips are in direct line of sight of the cutting zone. Calculated values (for example, using Cook's equation) are limited in their usefulness when the generated heat is being continually cooled by coolant. Recent development of computers with higher computational capability has allowed numerical methods to investigate the influence of cutting fluids on the tool chip interface temperatures.

The main functions of a cutting fluid are cooling and lubricating the cutting tool. Jayal et al. [1] expanded this to include chip removal from the cutting zone, an increase from two to three functions of cutting fluids. In addition to this coolant provides temporary protection against oxidation and corrosion, which can be classed as an important secondary function. For these reasons it is apparent that coolant has an important role to play in metal cutting. Any alternative cooling method will

S.A. Chowdhury • M.N. Islam • B. Boswell (✉)
Curtin University, Perth, WA, Australia
e-mail: sayem.chowdhury@postgrad.curtin.edu.au; m.n.islam@curtin.edu.au; b.boswell@curtin.edu.au

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_20

need to provide the same properties. The challenge of this research is to prove that flood coolant can be replaced in many machining applications. Traditionally the only way to show that coolant could be replaced by alternative cooling methods was by carrying out robust testing of the machining operation. CFD is used in this analysis as research has shown that it is an easier and quicker method to apply and an appropriate computer analysis to determine the best cooling method. This chapter is a revised work of a conference paper on the World Congress on Engineering [2].

The mechanical energy consumed in a machining operation is converted into thermal energy [3] mainly by shearing the material, with a secondary heat being generated as a chip slides over the rake face of the tool tip. For this reason it is obvious that for whatever cooling method is used it is the effectiveness of dissipating the heat from the cutting zone that determines its usefulness. According to the National Institute for Occupational Safety and Health [4], metalworking fluids can be classified as the following types:

- straight oils
- soluble oils
- semi-synthetics and
- synthetics

Metalworking fluids often contain a mixture of other substances including biocides, corrosion inhibitors, metal fines, tramp oils, and biological contaminants [4]. Literature review identifies several health, safety, and environmental issues that should be addressed with regard to the application of cutting fluid: (a) toxicity of the fluid components, (b) flammability of the fluid, and (c) fluid disposal [5]. According to the Occupational Safety and Health Administration (OSHA) in the USA [5], metalworking cutting fluids can cause adverse health effects through skin contact with contaminated materials, spray or mist and through inhalation from breathing a mist or aerosol. These effects include skin disorders, respiratory diseases, and cancer. In recent years, the application of cutting fluids in machining operations has attracted also immense scrutiny due to its adverse effects on the environment. Consequently, a vast amount of research has been undertaken to reduce the application of cutting fluids [6]. Despite these efforts the problem remains unsolved because the research has shown that the effectiveness of the cutting fluid fluctuates on a case-by-case basis. According to Astakhov et al. [7], experimental studies in metal cutting are expensive and time consuming. In addition, the results are valid only for the specific experimental conditions used. The results also depend on the accuracy of the calibration of the experimental equipment. However, it is believed that the problem can be solved by developing a numerical model to help select the amount of cutting fluid in a scientific way, which minimises the amount of coolant needed for effective cooling of the tool tip.

Computational Fluid Dynamics (CFD) was adopted to carry out the numerical analysis (this procedure is now relatively common place). Figure 20.1. shows a simple CFD model of an oblique cutting operation using flood cooling. Numerical

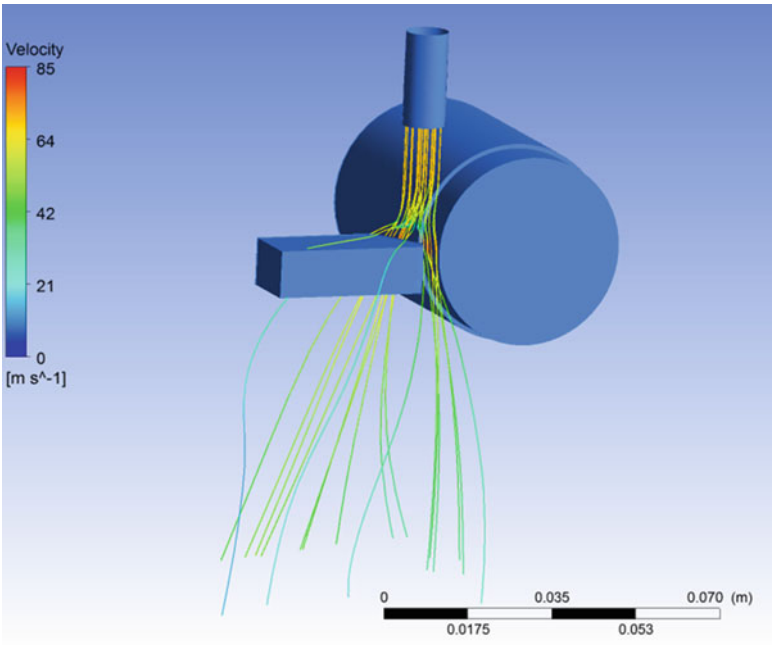


Fig. 20.1 Cutting fluid flow simulation in orthogonal metal cutting process

methods have been applied for modelling machining operations. They can be applied in calculating the temperature distribution and thermal deformation in the tool, chip, and workpiece. These methods are including but not limited to:

- Finite Difference Method (FDM)
- Finite Element Method (FEM)
- Boundary Element Method (BEM)

The FEM is used to simulate the temperature distribution the tool insert, work piece etc. FEM is also used to simulate the thermal deformation [3].

Numerical methods are an alternative approach to experimental studies. There are several numerical methods, but FEM is the most popular for metal cutting studies. FEM can predict the various outputs and characteristics of the metal cutting process such as the cutting force, stresses, temperatures, and chip geometry [7].

Umbrello et al. [8] implemented empirical tool wear models using numerical codes in order to validate capabilities of the models. Bruni et al. [9] concluded that analytical and artificial neural network models are capable of predicting the surface roughness under different machining conditions. Similarly, Khidhir and Mohamed [10] found that the experimental measurement of tool temperature at the cutting tool and workpiece interface is problematic, but the finite element simulation can be

used for quick and accurate prediction of the tool temperature under various cutting conditions. Knowles [11] concluded that finite element analysis can be viewed as a technique for solving equations, and as such, it has found application in many other areas including fluid flow, lubrication, seepage, electromagnetic fields, heat transfer, biomechanics, and others. FEM is used with some models to predict the heat distribution of an object. In this research, thermal analysis is performed using the Finite Element Modelling Software ANSYS CFX, which is capable of simulating and analysing turning objects [12]. Shen developed a thermal model using ANSYS. Koyama [13] also used Computation Fluid Dynamics (CFD) simulation with the CFX; which is general purpose CFD software for heat transfer and fluid flow analysis by ANSYS Inc. Mostofa et al. [14] used ANSYS CFX for a CFD simulation to investigate water jet air and abrasive velocity.

2 Cutting Tests and Set-Up

The heat dissipating effects of the air was improved by using a "Ranque-Hilsch Vortex Tube", a device that can produce cold air and provide a cold jet of air to the tool interface. The Vortex Tube (VT) was set at 90 psi compressed air giving a volumetric flow rate of 11 SCFM and the lowest temperature recorded for these conditions was -11C. The MQL was delivered from a Uni-max cutting tool lubrication system which distributed atomised coolube lubricant to the cutting zone. This system operates on the same principle as a Serv-O-Spray allowing the lubricant to be sprayed from a single air source, which allows adjustment to the amount of lubricant delivered to the cutting zone. A traditional emulsified cutting fluid (Cocol ultra cut) was used for the flood test machining as this is suitable for a range of materials. The cutting tool selected for all of the tests was a Sandvik (PDJNR2525M15) tool with a coated tungsten carbide insert (DNMG432-PF4215). All cooling nozzles used during the tests were kept at approximately 25 mm from the tool during all tests. Cutting tests were carried out using three conditions; dry, flood, and combined cooled air with MQL. Typical machining practices were used in machining the 40 mm diameter workpiece using the parameters as given in Table 20.1.

Table 20.1 Machining parameters

Cooling methods	Cutting speed (m/min)	Depth of cut (mm)	Feedrate (mm/rev)
Dry			
Flood	150	1	0.25
Cooled air combined with MQL			

3 CFD Analysis

The steps for CFD simulation are shown in Fig. 20.2, and is Pre-processing, which can be further divided into:

- defining the geometry
- meshing
- defining the physical model i.e. the equations of motion, conservation etc.
- defining the boundary conditions

The first step of the model is to develop the geometry of the metal cutting process. This was generated by using Ansys Design Modeler. It is also possible to import tool geometry data from other CAD software. The next step produced the mesh for the geometry by using the Ansys Meshing function. The number of elements should be reasonably high to get a best output from the simulations. If the number of elements is low then the accuracy of the analysis will not be high. Therefore it is important to find a reasonable balance between accuracy and the run time.

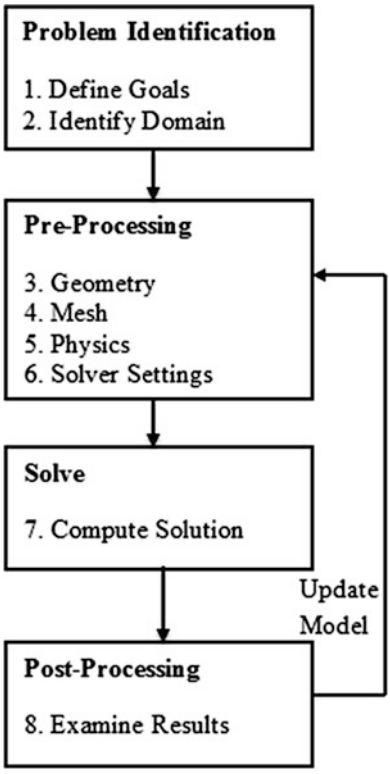


Fig. 20.2 CFD modelling steps shown in the figure [15]

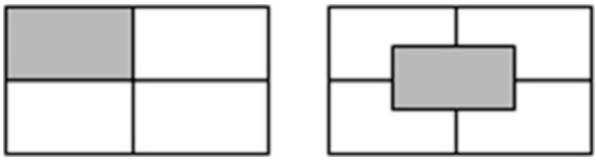


Fig. 20.3 The cell centered control volumes of fluent (*left*) and node-centered control volumes of CFX (*right*) [16]

In CFX, the simulation execution equations can be selected to be solved iteratively, and the post processer is used to analysis the results of the solution. One of the differences between Ansys CFX and Fluent is that the Fluent control volumes are cell-centered (i.e. they correspond directly with the mesh) while CFX control volumes are node-centered as shown in Fig. 20.3.

Super computers play an important role in the field of computational science, as they are capable of handling large volume of data at high speed. One of the examples of the use of super computers is weather forecasting. This is why the iVEC Super computer is ideal for the necessary computations in predicting the tool tip temperature. Ansys CFX software was used for the simulation of the CFD analysis of cutting fluid in the orthogonal metal cutting process. Ansys CFD solvers are based on the finite volume method, which is discretised into a finite set of control volumes.

The methodology used for the heat generated factors, and subsequent cooling process during machining for the CFD model to predict is shown in the block diagram Fig. 20.4. In this model, the generating heat considerations are depth of cut, feedrate, cutting speed and material properties workpiece and tool. The cooling considerations during cutting are different fluid quantity, velocity and properties.

The machining process has many variable parameters that influence the amount of heat generated, that then need to be dissipated by the cutting fluid. In this research, the following parameters are considered: cutting fluid properties, work material, tool geometry and material removal rate. Conventional wisdom has shown that any change in the machining parameters can have a substantial effect on the temperature generated at the tool tip and changing the rate of wear of the tool tip.

4 Results and Discussion

The FE model presented estimates the tool chip interface temperature distribution and also cooling processes for oblique turning operations having a variety of machining parameters. The model was validated with three experimental readings taken from the machining of a steel workpiece with different material removal rates.

The agreement obtained between the experimental and numerical results indicate that the proposed FEM model is capable of predicting the tool tip temperature for turning operations under different cooling processes. This enables the cost of manufacturing to be reduced by refining or using alternative cheaper cooling methods. Alternative cooling methods need to be tested against numerous machining parameters. According to some literature the cooling cost in metal cutting can contribute up to 17 % of the cost of the total machining cost.

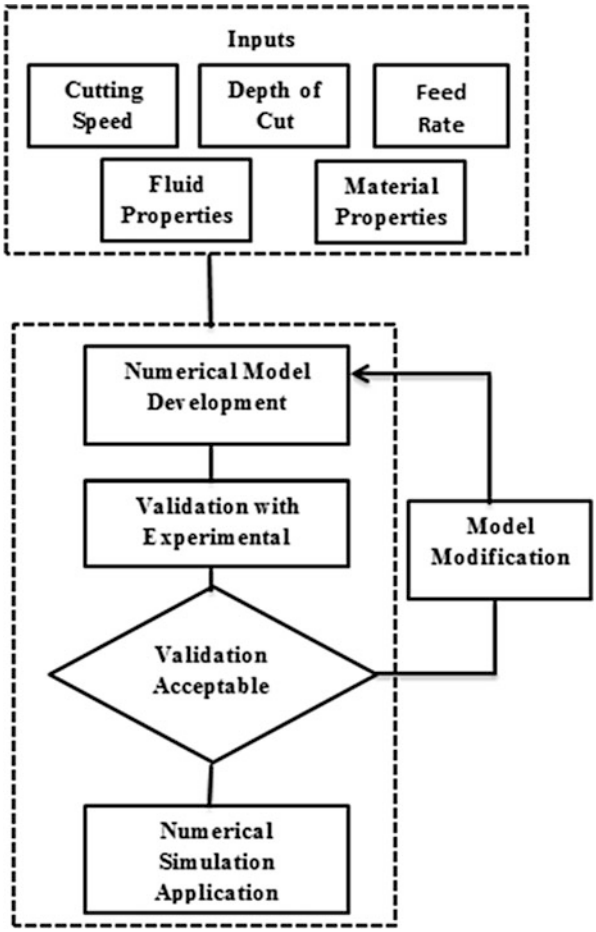


Fig. 20.4 The methodology used for developing the numerical model

There are several advantages if cold compressed air is used as a coolant in machining. Air is readily available, safe to use, inexpensive, environmentally suitable, and most workshops have compressed air available near the machine. The limitation of using air as a coolant is that the tool and workpiece start to increase in temperature after a number of machining operations. Therefore, there is a need for the tool and workpiece to have additional cooling added. This problem does not exist with liquid cooling as liquid coolants cover the tool and workpiece and reduce the heat buildup during machining Fig. 20.5.

Tool life is one of the considerations to achieve sustainable machining. Tool life depends on the tool interface temperature and can be severely shortened at high temperature. Therefore, the temperature distribution simulation of the tool insert is useful in predicting the tool life for the three different cases i.e. dry machining, flood cooling and a combination of cold air and Minimum Quantity Liquid (MQL) cooling. In all three cases, the same heat generating parameters are considered for the simulations.

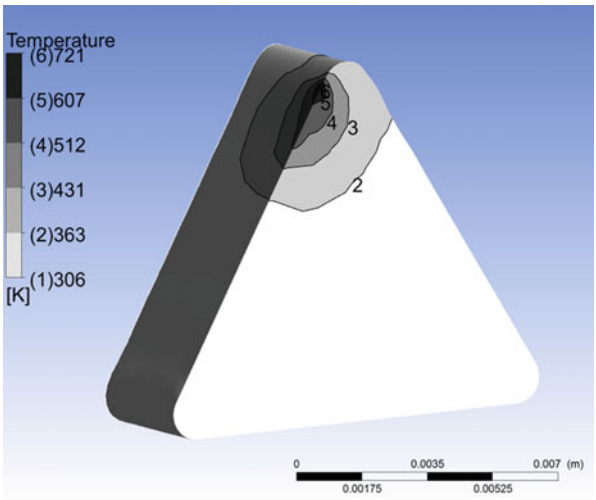


Fig. 20.5 Tool chip interface temperature simulation by flood cooling according to the cutting parameters of Table 20.1

It was assumed that the mechanical work done during the metal cutting process was converted into the equivalent amount of heat. Total heat generation P_{total} was calculated using the following Eq. 20.1.

$$P_{total} = K_s \times DOC \times Feedrate \times Cutting\ Speed \tag{20.1}$$

The Specific Cutting Force K_s, primarily depends on the work material and the feedrate. The value of K_s was obtained from the tool manufacturer's data handbook [17]. The Specific Cutting Force, K_s, for the above condition is approximately 3,645 N/mm^2. In all three cases, the same heat generating parameters are considered in the simulations. Equation 20.1 was applied to calculate the heat inputs, using the machining parameters. It was taken that 3.3 % of the total calculated heat entered into the tool tip insert as Fleisher et al. [18] he concluded that in general 92.7 % of the heat is conducted away by the chips, 4 % of the heat goes into the workpiece, and 3.3 % of the heat enters into the tool tip. The CFD model was able to capture the cooling effect of the chips, removing the heat from the cutting zone, as shown by the tool interface Fig. 20.6. The tool interface temperature is shown to be less for MQL and cold air simulation Fig. 20.7. Some of the boundary conditions or inputs for this CFD analysis were as following:

- air flow was 10 SCFM
- air temp was −10 °C
- liquid fluid temperature was 25 °C

According to Fig. 20.5 CFD simulation the chip tool interface temperature is found to be the lowest, when liquid cutting fluid was used as coolant. The initial coolant temperature was considered to be 25 °C, and the flow rate was 24 l/min.

By using the CFD model it is now possible to predict the tool interface temperature for different cooling methods. For example, when the air mist temperature

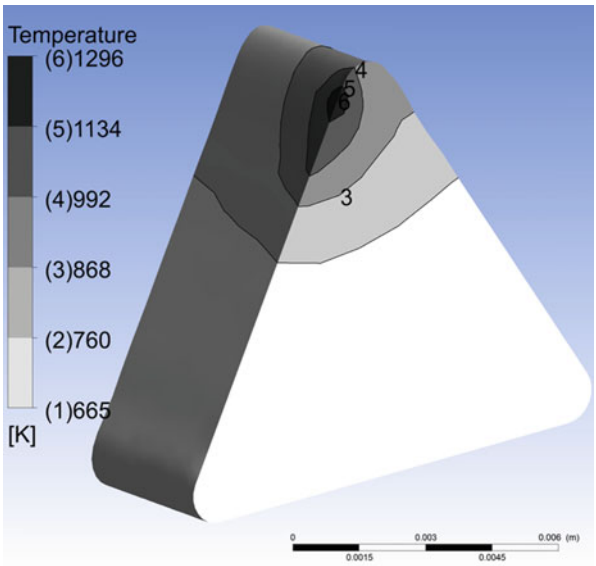


Fig. 20.6 Tool chip interface temperature simulation in dry machining according to the cutting parameters of Table 20.1

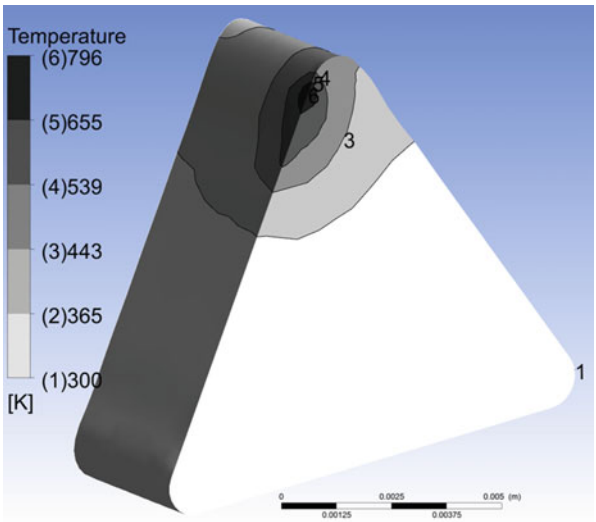


Fig. 20.7 Tool chip interface temperature simulation by combined −9 °C air and minimum liquid cooling according to the cutting parameters of Table 20.1

is reduced further as shown in Fig. 20.8 there is a reduction of tool tip interface temperature. Temperature distribution of the tool insert is further reduced, as that shown in the Fig. 20.9, when the cold air temperature is reduced to −50 °C.

Obtaining the tool life is crucial in testing and detecting the effectiveness of the coolants. As a result of the simulation, the maximum interface temperature for the dry machining is 1,023 °C (1,296 K) as shown in Fig. 20.6. Similarly the maximum interface temperature for cold air with MQL cooling is 523 °C (796 K). On the other

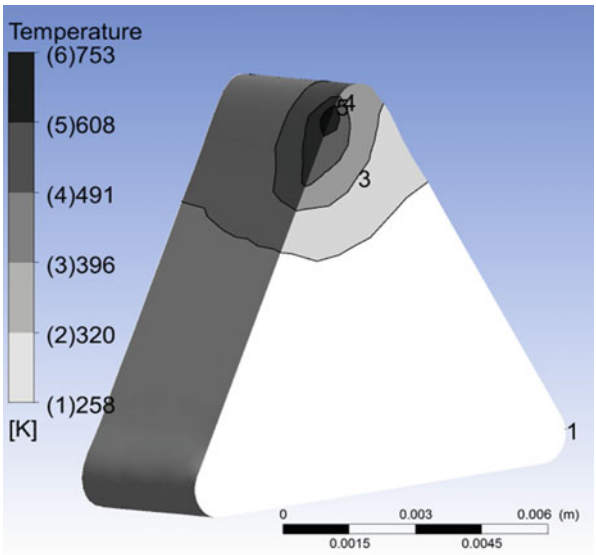


Fig. 20.8 Tool chip interface temperature simulation by combined −50 °C air and minimum liquid cooling according to the cutting parameters of Table 20.1

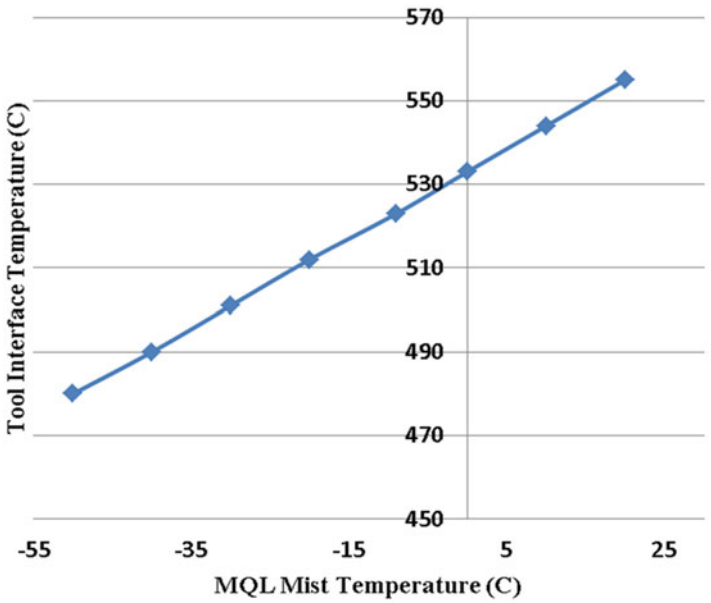


Fig. 20.9 The effect of mist air temperature on the tool chip interface temperature in combined cold air with MQL according to the cutting parameters of Table 20.1

hand, the maximum interface temperature for the flood cooling is 448 °C (721 K). This research has shown that the cold air with MQL can operate at a temperature that is effective for metal cutting. The temperature difference between the combination

of cold air with MQL and flood cooling is only 75 degrees or 10 %. This research has also shown that flood cooling may provide longer tool life according to the simulated result but the tool interface temperature difference between the combined of cold air with MQL and flood cooling is not enormous. Therefore, the possibility of tool failure in combined cold air with MQL will not be subsequently high when comparing that of flood coolant as the tool interface temperature difference is not very significant.

5 Conclusion

This research showed that it was feasible to use CFD to determine the effectiveness of the cooling method, before cutting takes place. Virtual prototyping with ANSYS or similar soft should be used increasingly by manufacturing engineers around the world to help deliver their products to the market; faster, cheaper and more environmentally friendly. Even though the virtual model produced a slightly higher tool interface temperature for MQL, and cold air the resulting data was still good enough to identify the best cooling method. This was shown to be achieved by significantly reducing the amount of physical prototyping or machining testing required in determining the most effective tool cooling method. In addition to the cooling method CFD was an effective method of determining the best cutting parameters.

In conclusion, the combination of cold air with MQL can be an effective alternative to flood cooling, proving that environmental benefits are possible by adopting MQL and cold air cooling. Finally, by virtual modelling the different complex machining scenarios, allows additional cutting zone knowledge to be obtained, which would be otherwise difficult to achieve by traditional testing methods.

References

1. Jayal, A.D., et al.: Machining performance and health effects of cutting fluid application in drilling of A390. 0 cast aluminum alloy. J. Manuf. Process. **9**(2), 137–146 (2007)
2. Chowdhury, S.A., Islam, M.N., Boswell, B.: Effectiveness of using CFD for comparing tool cooling methods. In: Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp 1076–1081. London 2–4 July 2014
3. Sutherland, J.W.: Introduction to Heat in Machining. Available http://www.mfg.mtu.edu/marc/primers/heat/heat.html
4. NIOSH.: Health hazard evaluation and technical assistance report: HETA. National Institute for Occupational Safety and Health 005-0227-3049, Diamond Chain Company, Indianapolis (2007)
5. OSHA: Metalworking Fluids: Safety and Health Best Practices Manual. U.S. Department of Labor, Occupational Safety and Health Administration, Salt Lake City (1999)

6. Islam, M.N., Rafai, N.H., Heng, B.C.: Effect of cutting fluid supply strategies on surface finish of turned parts. Adv. Mater. Res. **383–390**, 4576–4584 (2012)
7. Astakhov, V.P., Outeiro, J.C.: Metal cutting mechanics, finite element modelling. In: Machining, pp. 1–27. Springer, London (2008)
8. Umbrello, D., et al.: Prediction of tool wear progress in machining of carbon steel using different tool wear mechanisms. Int. J. Mater. Form. **1**(1), 571–574 (2008)
9. Bruni, C., et al.: Surface roughness modelling in finish face milling under MQL and dry cutting conditions. Int. J. Mater. Form. **1**(1), 503–506 (2008)
10. Khidhir, B.A., Bashir, M.: Study of cutting speed on surface roughness and chip formation when machining nickel-based alloy. J. Mech. Sci. Technol. **24**(5), 1053–1059 (2010)
11. Knowles, N.C.: Finite element analysis. Comput. Aided Des. **16**(3), 134–140 (1984)
12. ANSYS CFX Technical Specifications. ANSYS, Canonsburg (2010)
13. Koyama, K.: CFD simulation on LNG storage tank to improve safety and reduce cost. Syst. Model. Simul. Part 1, 39–43 (2007)
14. Mostofa, M.G., Kil, K.Y., Hwan, A.J.: Computational fluid analysis of abrasive waterjet cutting head. J. Mech. Sci. Technol. **24**(1), 249–252 (2010)
15. Introduction to CFD, Ansys CFX training materials. ANSYS, Canonsburg (2010)
16. Difference between ANSYS CFX and Fluent. Available www.cfd-online.com/Forums/cfx/110640-difference-between-ansys-cfx-fluent.html
17. D. T. I. a. W. C.: Tool and Manufacturing Engineering Handbook, 4th ed. vol. 1. Society of Manufacturing Engineers, Dearborn (1983)
18. Fleischer, J., Pabst, R., Kelemen, S.: Heat flow simulation for dry machining of power train castings. CIRP Ann. Manuf. Technol. **56**(1), 117–122 (2007)

Chapter 21
Production Planning for Make-to-Order Flow Shop System Under Hierarchical Workforce Environment

Sujan Piya and Nasr Al-Hinai

Abstract In make-to-order system orders are scheduled for production based on the due date agreed with customer and the strategy of company. Production planning of such system includes scheduling of orders to the production periods and allocation of workers at different work centers. Complexity in the system arises when the operation to perform next and its processing time is skilled dependent where a higher qualified worker type can substitute a lower qualified one, but not vice-versa. Under such working environment, efficient scheduling of orders and allocation of workers at different work center play major role to improve system performance. This paper develops a mathematical model for make-to-order flow shop system under hierarchical workforce environment. The model helps identify optimum schedule of orders and allocation of workers with an objective of minimizing the weighted average earliness and tardiness. A heuristic method is also proposed to overcome the complexity of mathematical model and solve the problem efficiently. Numerical analysis indicates that proposed heuristic is capable of finding optimal or near optimal solution in a considerably reduced amount of computational timing.

Keywords Flow shop • Hierarchical workforce • Make-to-order • Production planning • Scheduling • Worker allocation

1 Introduction

Due to fierce competition in the market there is an increasing trend of company moving from Make-to-Stock (MTS) model of production system towards Make-to-Order (MTO) system. The main reason for the popularity of MTO system is that the firm can offer a product which is unique and customized in nature [1]. Also, due to the success of Dell Corporation, many firms seek to offer greater product

S. Piya (✉) • N. Al-Hinai
Department of Mechanical and Industrial Engineering, College of Engineering, Sultan Qaboos University, AL-Khod 123, Muscat, Sultanate of Oman
e-mail: sujan@squ.edu.om; nhinai@squ.edu.om

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_21

variety at a low cost by eliminating finished goods inventory using MTO production model [2]. In MTO system, company starts production only after getting confirmed orders from the customers and each order will be for a unique product [3]. The scheduling of orders for production depends on the due date attached to the order and the strategy of company. The complexity in the production of orders occur when the operations to be performed depends on the level of skill that the workers posses, where a higher qualified worker type can substitute for a lower qualified one but not vice-versa [4]. This hierarchical nature of workforce will affect the time at which order once scheduled will be finished. It will have direct implication on when to start production and who should be allocated to produce it. Producing earlier than the due date is not desirable to the manufacturer as it will incur holding cost of inventory. Nor they can produce order late as it will incur tardiness cost or loss of company goodwill.

We can find abundant literatures dedicated to production scheduling (i.e., scheduling of job to the machines for production) and still it is growing rapidly. To name few of them are [5–17]. These literatures have developed new approaches or heuristics to solve scheduling problem either in a flow shop or job shop settings under stochastic or deterministic processing time with different objective functions such as minimizing makespan, mean flow time, weighted tardiness, weighted earliness, total ideal time and so on. In these literatures, scheduling is done by considering the number and capacity of available machines. However, if the industry is labor intensive, then only considering the number and capacity of machines is not enough. In such situation, the capability of worker for performing a specific job, rather than the capacity of available machine, will determine the duration on which the job can be completed. Since each worker will have their own capability, forming a hierarchical nature of workforce, scheduling order based on the capability of workers is another challenging task in such system. Therefore, together with scheduling and sequencing of orders to the machines, proper allocation of worker is also important to optimize available resources. According to [18] there are very few published papers that have dealt on hierarchical workforce problem and more research is needed in this area.

Research on hierarchical workforce using combinatorial method can be traced back to [19]. The paper developed an algorithm for single shift workforce scheduling that generates feasible schedule under the assumption that each worker must have 2 off-days per week. In the same line of research [20] discussed necessary and sufficient condition for a labor mix to be feasible and presented one-pass method that gives the least labor cost. Reference [21] developed a mathematical model that solves the problem by determining the most economical size of various hierarchies of workers, assigning them working and off days in each week. Reference [22] illustrated that integer programming approach is well suited for solving the problem studied by Ref. [20]. Reference [23] extended single shift workforce scheduling problem by presenting an optimal algorithm for multiple shifts scheduling of employees on the basis of 4 days or 3 days workweeks. Reference [24] developed mixed-integer programming model for the problem under multi-shift scheduling based on which a specialized scheduling heuristic is subsequently developed.

Reference [18] introduced the concept of compressed workweeks in the model of [22]. The work of [18] is further extended by Ref. [25] by introducing the concept of suitability of task assignment to individual employees. Recently, [26] extended the model of [18] and [22] by developing mathematical models under the assumption that the work is divisible. The models proposed in these literatures give the optimum number of workforces with different hierarchies needed to satisfy customer order. However, in MTO system each order will have certain due date attached to it. Hence, it is necessary for the company to identify the sequence on which order to produce first and then which next together with identifying optimum numbers of workers with different hierarchies to minimizing the total cost of production. Furthermore, reviewed literatures on hierarchical workforce neglect the precedence relationships that exist between the operations in the order, which is a critical issue when job needs to be processed at different work centers.

From the literature review it can be seen that there are many papers published on production scheduling and few on hierarchical workforce planning. However, these two issues have been researched separately. Nevertheless, many manufacturing companies have to deal with both of these issues simultaneously, especially when the company is labor intensive. This research aims at developing a mathematical model by integrating hierarchical workforce planning with order scheduling problem with an objective of minimizing weighted average earliness and tardiness, which as far as we know is the first attempt in this area of research.

The rest of the paper is structured as follows. Section 2 discusses the problem that will be addresses in this research. Section 3 explains the proposed mathematical model in detail. Section 4 discusses a heuristic method developed to overcome the complexity related to mathematical model. Section 5 is dedicated to numerical analysis. Finally, concluding remarks and future research directions are highlighted in Sect. 6.

2 Problem Statement

Scheduling of orders and allocation of workers to the available resources are some of the most important issues related to production planning and control. These issues will have direct impact on the productivity and cost of production. This paper has considered the integration of aforementioned issues in a periodic setting where all the orders that arrived on a given period will be produced in the next succeeding period. Once the order arrives with certain due date attached to it, order remains in a pre-shop pool till the next period. Company will accumulate all the orders of a given period plus any backlog of past period to make schedule for next period depending on which orders will go for production. To minimize the weighted average earliness and tardiness of orders, which is our objective function, it is necessary during scheduling to consider the hierarchical nature of workforce. Three different levels of worker hierarchy are considered, namely level 1, level 2 and level 3. Workers in these levels differ in terms of their capability to operate various machines and in the speed with which they can perform certain operation. Level 1 represents the

most efficient group of workers who can do all types of tasks that are necessary to be performed in the order or it is the group of workers who can operate all the machines. Plus, this group of workers can perform the operation faster than any other groups. On the other hand, level 3 represents the group of workers who can work only on fewer machines and takes more time to finish job. Capability of workers that fall on level 2 lies in between level 1 and level 3 in terms of number of machines they can operate and the speed at which they can perform the assigned operation. Here, the problem is to identify the required number of workers of each level, allocate these workers to different work centers and schedule the orders on the basis of capability of identified numbers of workers of various levels.

The paper assumes this problem in a flow shop environment. The shop basically consists of various work centers. An order consists of batch of one product type that will enter from the initial work center and exit through the last work center. Each order involves operations at all the work centers. The objective is to develop order scheduling model by taking into account the hierarchical nature of work force to minimize the weighted average earliness and tardiness. Following assumptions are used for developing mathematical model:

- Same order cannot be processed on more than one machine at a time.
- The processing time of operation of an order on a given machine by the given level of worker is predefined.
- Preemption of order is not allowed.
- Each machine can process at most one operation at a time.
- Due date of orders are known and fixed.
- Allocation of worker in the work center for the given order remains fixed once the work is started.
- Allocation of worker in the work center for different order may be different.
- Set up time is considered negligible.
- All orders are available for processing at the beginning of period.

3 Mathematical Model

This section discusses the proposed mathematical model developed to solve the problem described in Sect. 2. All indexes, parameters, and decision variables used in the model are listed below.

Index	
i	Work center, $i = \{1, \ldots .., m\}$
o	Customer order, $o = \{1, \ldots \ldots, O\}$
k	Worker level, $k = \{1, \ldots ..,K\}$
f	Unit in the order, $f = \{1, \ldots ..,F\}$
v	Position of order in the sequence, $v = \{1, \ldots \ldots, V)$

Parameter	
q_o	Total number of units in order o
d_o	Due date of order o
t_o	Time at which production of order o can be started
Δ_{oi}	Waiting time for operation in work center i of order o
p_{iok}	Processing time worker level k takes to process one unit of order o in work center i
Cap	Capacity of work center in a period
CT_{iok}	Completion time of a batch of order o in work center i by worker level k
ST_{iof}	Starting time of f unit of order o at work center i
w'	Earliness penalty of order o for each time unit of earliness
w''	Tardiness penalty of order o for each time unit of tardiness

Decision variable	
CT_{mo}	Completion time of order o at work center m
E_o	Earliness time of order o
T_o	Tardiness time of order o
X_{iov}	1 if at machine i order o is assigned to v^{th} position in the sequence; 0 otherwise
Y_{iok}	1 if in work center i order o is processed by worker level k; 0 otherwise

The Model:

$$\min z = \sum_{o=1}^{O} \left(w'_o E_o + w''_o T_o \right) \tag{21.1}$$

Subject to:

$$E_o \geq d_o - CT_{mo} \tag{21.2}$$

$$T_o \geq CT_{mo} - d_o \tag{21.3}$$

$$CT_{mok} = CT_{(m-1)ok} + p_{mok} \quad o = 1, \ldots\ldots\ldots, O;\ k = 1, \ldots\ldots\ldots\ldots, K \tag{21.4}$$

$$CT_{iok} = ST_{iof} + p_{iok} \quad i = 1, \ldots\ldots\ldots, m-1;\ o = 1, \ldots\ldots\ldots, O;\ k = 1, \ldots\ldots\ldots, K;\ f = 1 \tag{21.5}$$

$$\sum_{k=1}^{K} Y_{iok} = 1 \quad i = 1, \ldots\ldots\ldots\ldots, m; \quad o = 1, \ldots\ldots\ldots, O \tag{21.6}$$

$$\sum_{o=1}^{O} X_{iov} = 1 \quad i = 1, \ldots\ldots\ldots\ldots, m \tag{21.7}$$

$$Cap_i \geq \sum_{o=1}^{O} q_o p_{iok} Y_{iok} \quad i = 1, \ldots\ldots, m; \forall k \tag{21.8}$$

$$ST_{i(o+1)f} \geq CT_{iok} \quad o = 1, \ldots\ldots\ldots.., O; k = 1, \ldots\ldots\ldots\ldots., K; f = 1 \tag{21.9}$$

$$ST_{(i+1)of} \geq CT_{iok} \quad o = 1, \ldots\ldots\ldots.., O; k = 1, \ldots\ldots\ldots\ldots., K; f = 1 \tag{21.10}$$

$$\begin{aligned} &CT_{mo}, E_o, T_o \geq 0 \quad o = 1, \ldots\ldots\ldots.., O; \\ &Y_{iok}, X_{ior} = \{0, 1\} \quad i = 1, \ldots\ldots.., m; o = 1, \ldots\ldots\ldots, O; \end{aligned} \tag{21.11}$$

$$k = 1, \ldots\ldots\ldots., K \tag{21.12}$$

The objective function in Eq. (21.1) minimizes the weighted average earliness and tardiness. While developing the model it is assumed that all the orders that arrive in previous period can be completed within next period with the regular production time of workers. Equation (21.2) calculates the earliness of an order o which is greater than or equal to the difference between its due date and completion time at the last work center m. Similarly, Eq. (21.3) calculates the tardiness of an order. Completion time of an order at the last work center will determine the earliness and tardiness of an order. This can be calculated by Eq. (21.4). However, completion time of order at the last work center depends on its completion time at the preceding work center, which is calculated by Eq. (21.5). It indicates completion time of an order in a given work center by worker level k. Equation (21.6) is a binary variable to ensures that in each work center i order o will be processed by one worker level k. The value of Y_{iok} equals 1 means that worker level k is assigned to work center i to process order o. Otherwise, the value of it will be zero. Similarly, Eq. (21.7) is also a binary variable to ensure that each order is assigned only once in the given machine. The capacity constraint is represented by Eq. (21.8) and indicates that the total processing time for orders with a given worker type must be less than or equal to the available capacity within a period. Constraint that a machine can operate only one operation of an order at a time is indicated by Eq. (21.9). Equations (21.10), (21.11, and (21.12) are precedence constraint, non-negative constraint and binary variable respectively.

The starting time in the above equations can be calculated by using iterative method as follows.

$$\frac{\text{When } i = 1 \text{ and } f = 1}{ST_{iof} = \max\left[t_o, \Delta_{io}\right]} \tag{21.13}$$

$$\frac{\text{When } i = 1 \text{ and } f = 2, 3, \ldots\ldots, F}{ST_{iof} = ST_{io(f-1)} + p_{iok}} \tag{21.14}$$

$$\frac{\text{When } i = 2, 3, \ldots., m \text{ and } f = 1}{ST_{iof} = \max\left[ST_{(i-1)of} + p_{(i-1)ok}, \Delta_{io}\right]} \tag{21.15}$$

$$\frac{\text{When } i = 2, 3, \ldots., m \text{ and } f = 2, 3, \ldots\ldots, F}{ST_{iof} = \max\left[ST_{io(f-1)} + p_{iok}, ST_{(i-1)of} + p_{(i-1)ok}\right]} \tag{21.16}$$

4 Proposed Heuristic

Flow shop scheduling problem with the objective function considered in this paper is NP-hard [27]. The complexity of the problem is exacerbated by hierarchical nature of workforce that we have considered in this paper. Therefore, to solve industrial size problem in a reasonable timing, a heuristic method is proposed by assuming that at least one level of worker (Level 1) can work on all the machines and they are the most experienced group of workers thus resulting into less working time as well in all the work centers. Let *So* be the set of orders to be produced in any given period. Proposed heuristic works as follow:

Step 1: Generate sub-sets of order by selecting any two orders from the set (*So*). The total number of sub-sets (z) will be equal to $\sum_{a=1}^{n-1}(n-a)$ where, n is the total orders in the set *So*.

For eg.: Let, *So* = (o1, ...,o6) and assume that these orders have to be processed in six different machines (*m1*, . . . , *m6*). Also, let's assume that there are three levels of workers (*l*1, *l*2, *l*3) out of which level 1 can work on all the machines, level 2 on machines *m3*, *m4*, *m5* and *m6*. However, level 3 can work only on machines *m5* and *m6*. For this example, the sub-sets are (o1, o2), (o1, o3), . . . , (o5, o6).

Step 2: Select any sub-set and define the possible sequence. Here, since each sub-set will have two orders, the total number of possible sequences will be equal to $2z$. For sub-set (o2, o6) the possible sequences will be (o2-o6) and (o6-o2).

Step 3: For any possible sequence of orders allocate Level 1 worker on all the machines as it has been assumed that worker that falls in this level can work on the entire machines. Then,

3(a): Identify the number of different levels of workers who can work on same sets of machines and allocate the worker on those machines accordingly. This results into $3y^r$ different possibilities on worker allocation where, y is the

number of different levels of workers that can work on same sets of machines and r is the number of machines on which y can work.

For our example, in *m1* and *m2* only worker with level 1 can work. So, for any sequence of order in these machines, assign worker with level 1. Then, for machines *m3* and *m4*, workers with level 1 and 2 can work. So, one possible allocations for order sequence (o2-o6) on these machines will be [o2(*m3*/*l1*);o2(*m4*/*l1*)-o6(*m3*/*l1*;o6(*m4*/*l2*)]. The total possibilities for worker allocations here will be equal to 12.

3(b): Repeat Step 3(a) for the remaining sequence. This results into $6y^r$ different possibilities on worker allocation for orders (o2, o6).

3(c): For each possibility, calculate the objective function and then select the one that gives the minimum value of objective function. Here, it should be noted that while calculating the objective function, for remaining machines *m5* and *m6*, still worker belonging to level 1 will be allocated.

Let the best possibilities for our example be [o2(*m3*/*l1*);o2(*m4*/*l2*)-o6(*m3*/*l2*;o6(*m4*/*l1*)].

3(d): Similarly, repeat the steps from step 3(a) for remaining machines by assigning worker level on preceding machines based on the result of 3(c).

For the example, remaining machines for worker allocation are *m5* and *m6* and all level of workers can work on it. The total possible solution for worker allocations here will be equal to 27. These solutions will be enumerated by assigning worker level 1 in *m1*, *m2* for o2 and o6. Also, based on the result from 3(c), worker level 1 will be assigned at *m3* and *m4* for o2 and o6 respectively; and at *m4* and *m3* of o2 and o6 respectively worker with level 2 will be assigned.

Step 3 results into $\sum_{\forall s \in S} 3y_s^{r_s}$ different possibilities on order sequencing and worker allocation for a set of order where $s = (1, \ldots, S)$ represents set of worker level who can work on specific set of machine. Two sets of workers for the example are (*l1*, *l2*) and (*l1*, *l2*, *l3*) who can work on machine sets (*m3*, *m4*), (*m5*, *m6*) respectively.

The outcome of Step 3 will be the best sequencing of orders and allocation of workers at all the machines for orders (o2, o6).

Step 4: Repeat steps from *step 2* for all the remaining sub-set of orders and select the sub-set with the sequence and allocation that gives minimum objective function. This results into $\sum_{\forall s \in S} 6y_s^{r_s}(z)$ possibilities of order sequencing and worker allocations for all the sub-sets generated in step 1.

The outcome of Step 4 will be the best sequencing of orders and allocation of workers at all the machines for the pair of order that must be scheduled first for production.

Step 5: Next, repeat steps from step 1 for the remaining orders in the set (*So*) until the set remains empty. While repeating the steps it should be noted that the order pair that has been selected in Step 4 with their sequence and worker allocation

should be fixed and the sequencing of new order pair will start after the selected order pair.

Proposed heuristic results into $\sum_{\forall s \in S} 3y_s^{r_s} \left[\sum_{a=o}^{x} (n-2a) \ \{n-(2a+1)\} \right]$ possible solutions on order sequencing and worker allocations for all the orders in the set S_o instead of $\prod_{\forall s \in S} 3y_s^{r_s} n!$ by mathematical model.

5 Numerical Analysis

Numerical analysis is conducted to check the effect of various parameters on objective function and also to compare the effectiveness of proposed heuristic method with that of optimal solution generate by mathematical model. The effectiveness is measured in terms of objective function and the computational time i.e., the time to solve the problem.

5.1 Experimental Set Up

To carry out numerical analysis, number of test problems are generated by varying the size of orders, level of workers and number of work centers i.e., (o, k, i). Increase in the value of these parameters increases the level of complexity, which represents the number of possible solution for the given test problem. Size of orders are varied from (3, 7), level of workers from (1, 3) and work centers from (2, 6). This results into 75 problem instances and five replication of each problem instance is used in the numerical analysis. Time necessary to carry out processing for each order on given work center by the given level of worker is uniformly distributed as shown in Table 21.1. From the table it is clear that the worker who falls on level 1 need least processing time as compared to level 2 and level 3. The due date on each order is then generated by adding some allowances (α) in the average processing time $(P_o)_{Avg}$, which is the average value of total processing time of order o on all the

Table 21.1 Processing time for different level of worker

Worker (k)	Processing time (p_o)	Due date (d_o)
Level 1 ($l1$)	$U(4,7)$	$d_o = (P_o)_{Avg}(1+\alpha)$ where, $(P_o)_{Avg} = \frac{\sum_{k=1}^{K} \left(q_o \sum_{i=1}^{m} p_{iok} \right)}{K}$
Level 2 ($l2$)	$U(7,9)$	
Level 3 ($l3$)	$U(10,13)$	

machines by all three level of workers. The total unit (q_o) in the order is uniformly distributed between (2, 4) and the weights on earliness (w') and tardiness (w'') is assumed to be 0.5.

For simplicity, in the numerical analysis, it is assumed that all three different levels of worker can work on all the machines. However, the level is distinguishable based on the processing time that each level of worker takes to carry out task on the given machine. All test codes are executed using Matlab program of version 7.9.0 R2009b on an Intel® Core™ 2 Quad CPU @ 3.00 GHz with 4.00 GB RAM.

5.2 Result Discussion

Table 21.2 shows the result of numerical analysis for the optimal solution and proposed heuristic. In the result, the objective function (*WET*) and computational time (*CT*) for each test problem represents an average value for the five replications generated with the set up as mentioned in the experimental setup section. Worker level in the second column of Table 21.2 represents number of level of workers. $k = 1$ in the column means that only one level of worker i.e., $l1$ is available for production. Similarly, $k = 2$ means that two level of workers i.e., $l1$ and $l2$ are available and $k = 3$ means that all three level of workers are available for production. Out of 75 test problems, Table 21.2 shows the result of only 25 test problems which were selected randomly. From the table it is obvious that all three parameters (o, k, i) has influence on both the measure of effectiveness, especially *CT*.

The effect of increase in the size of order and number of work center on *WET* is inconsistent. Most of the time increase in the value of these parameters increases *WET*. However, there are multiple instances when *WET* decreases with the increase in these values; for eg. instances (3, 1, 4) and (5, 1, 4). It mainly depends on the due date assigned to the order. At a given due date, increase in the number of machines at which processing of an order has to be carried out leads to increase in the completion time of an order. This in turn will affect earliness or tardiness of an order. Similar effect can be observed when number of order increases. Increase in the number of orders affects the position of order in a given schedule which dictates the completion time of an order. Therefore, at a given due date, change in completion time affects earliness or tardiness of an order.

At a given value of o and i, *WET* decreases with increase in the worker level. When the level of worker is only 1 then objective function is not as good as when it is 3. This is due to the reason that with only one level of worker ($l1$), the earliness of an order may be high as this level of workers can work faster than any other level. Therefore, at a constant due date, the result will be better with multiple level of workers as it provides more flexibility on selecting the level of workers needed to improve objective function.

As discussed before, increase in the value of above three parameters increases the level of complexity, which leads to increase in the *CT* to solve the problem.

Table 21.2 Comparison between optimal solution and proposed heuristic

Order (o)	Worker level (k)	Machine (i)	Optimal solution		Heuristic method	
			WET	*CT*	*WET*	*CT*
3	1	4	51	00:00:01	51	00:00:01
3	2	4	46	00:00:01	46	00:00:01
3	2	6	55	00:00:02	55	00:00:02
3	3	5	47	00:00:01	47	00:00:01
3	3	6	50	00:00:02	53	00:00:01
4	1	2	46	00:00:06	46	00:00:01
4	1	5	61	00:00:09	61	00:00:03
4	2	3	47	00:00:20	47	00:00:01
4	3	2	44	00:00:59	45	00:00:01
4	3	6	55	00:01:11	55	00:00:01
5	1	4	49	00:08:33	55	00:00:01
5	2	2	40	00:08:18	44	00:00:01
5	2	5	53	00:08:53	58	00:00:01
5	3	2	40	00:13:08	40	00:04:04
5	3	6	58	00:11:45	58	00:05:39
6	1	3	57	00:38:18	59	00:04:58
6	1	5	69	00:40:24	74	00:04:25
6	2	2	50	00:52:09	50	00:05:40
6	3	4	58	01:03:40	58	00:07:06
6	3	5	66	01:01:23	79	00:08:09
7	1	5	71	02:18:36	78	00:11:16
7	2	5	68	02:32:15	68	00:11:25
7	2	6	72	02:38:08	79	00:13:38
7	3	4	71	02:56:52	88	00:13:43
7	3	6	72	03:23:28	76	00:14:26

From the result for optimal solution in Table 21.2 it is obvious that increase in the size of order has the highest effect followed by number of worker level on increasing the level of complexity rather than by the increase in the number of work center.

From Table 21.2 we can observe that the performance of proposed heuristic is as good as that of mathematical model in terms of *WET*. For most of the test problem *WET* is similar or very much closer to the optimal solution. To check the closeness of objective function obtained by proposed heuristic with respect to that of optimal solution, we calculated the average relative error, the result of which is shown in Table 21.3. Relative error (*RE*) in Eq. (21.17) indicates the percentage difference between the objective function (*WET*) obtained by heuristic method (*heu*) and the one obtained from optimal solution (*opt*) for the particular size of order.

Table 21.3 Error in the objective function by proposed heuristic

Test instance	Relative error (*RE*)	Average relative error
(3,*k*,*i*)	0.716	2.606
(4,*k*,*i*)	0.127	
(5,*k*,*i*)	2.550	
(6,*k*,*i*)	3.896	
(7,*k*,*i*)	5.743	

$$RE = \frac{\sum_{k=1}^{K}\sum_{i=1}^{m} WET_{(o,k,i)}^{heu} - \sum_{k=1}^{K}\sum_{i=1}^{m} WET_{(o,k,i)}^{opt}}{\sum_{k=1}^{K}\sum_{i=1}^{m} WET_{(o,k,i)}^{opt}} * 100 \tag{21.17}$$

As shown in Table 21.3 the average relative error on objective function for the proposed heuristic is 2.606 %. It indicates that proposed heuristic is working satisfactorily. Also, the table reveals that relative error increases with increase in the problem size.

From computational point of view proposed heuristic is much better. It is evident from Table 21.2 that for optimal solution computational time increases exponentially with the increase in the size of the problem. Therefore, even for medium size industrial problem mathematical model cannot generate result in a reasonable timing. The result shows that proposed heuristic is capable of finding optimal or near optimal solution in a considerably reduced amount of computational timing.

6 Conclusion and Future Work

This paper developed a mathematical model to allocate hierarchical nature of workforce to different orders at various work centers in a flow shop environment. This allocation problem was integrated with order scheduling problem. The problem that the paper dealt with is strongly *NP*-hard. Therefore, a heuristic method was also proposed to solve industrial size problem efficiently. The proposed heuristic reduces the number of possible combinations of order sequence and worker allocation drastically. Numerical analysis was conducted to show the effectiveness of proposed heuristic method with that of optimal solution generated by mathematical model.

The developed model helps a company to identify required number of workers with different level of hierarchies, order processing sequence, and allocation of different workers to the available work centers. As far as we know this is the first attempt to integrate order schedule problem with worker allocation under the situation where different levels of workers exists.

From numerical analysis it is evident that for the problem size of (7, 3, 6) the computational timing of proposed heuristic is around 14 min. It implies that even with the proposed heuristic, if the size of order is very big, with many levels of worker hierarchy and many work centers, then it may be difficult to solve the problem in a reasonable timing. Therefore, introduction of intelligent method such as GA, PSO, TS can be considered as an extension of present research. The next avenue for the research expansion may be to consider the problem by putting constraint on working hours of workers or number of availability of various levels of workers. Also, exploring the problem with different cases of due dates (i.e., tight due date, slack due date) may provide valuable insights into this area of research.

Acknowledgment The research was supported by the internal research grant (IG/ ENG/MIED/ 14/05) from Sultan Qaboos University. This support is gratefully acknowledged.

References

1. Gallien, J., Tallec, Y.L., Schoenmery, T.: A Model for Make to Order Revenue Management. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA (2004) (Working Paper)
2. Kreamer, K.L., Dedrick, J., Yamashiro, S.: Refining and extending the business model with information technology: Dell computer corporation. Inf. Soc. **16**, 5–21 (2000)
3. Piya, S., Katsuhiko, T., Katsumi, M.: Introducing quotation and negotiation structure in the order acceptance decision model. Int. J. Oper. Quant. Manag. **15**(4), 293–318 (2009)
4. Piya, S., Al-Hinai, N. Production planning under hierarchical workforce environment. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, WCE 2014, London, 2–4 July 2014, pp. 1029–1033
5. Applegate, D., Cook, W.: A computational study of the job shop scheduling problem. ORSA J. Comput. **02**, 149–157 (1991)
6. Brandimarte, P.: Routing and scheduling in a flexible job shop by tabu search. Ann. Oper. Res. **41**, 157–183 (1993)
7. Lee, I., Shaw, M.J.: A neural net approach to real time flow shop sequencing. Comput. Ind. Eng. **38**, 125–147 (2000)
8. Kacem, I., Hammadi, S., Brone, P.: Approach by localization and multi-objective evolutionary optimization for flexible job-shop scheduling problems. IEEE Trans. Syst. Man Cybern. **32**(1), 1–13 (2002)
9. Toktas, B., Azizoglu, M., Koksalan, S.K.: Two machine flow shop scheduling with two criteria: maximum earliness and makespan. Eur. J. Oper. Res. **64**, 278–285 (2004)
10. Tang, L., Liu, J., Liu, W.: A neural network model and algorithm for the hybrid flow shop scheduling problem in a dynamic environment. J. Intell. Manuf. **16**, 361–370 (2005)
11. Liu, L., Gu, H., Xi, Y.: Robust and stable scheduling of a single machine with random machine breakdowns. Int. J. Adv. Manuf. Technol. **31**, 645–654 (2007)
12. Yagamahan, B., Yenise, M.: Ant colony optimization for multi-objective flow shop scheduling problem. Comput. Ind. Eng. **54**(3), 411–420 (2008)
13. Pezzella, F., Morganti, G., Ciaschetti, G.: A genetic algorithm for the flexible job-shop scheduling problem. Comput. Oper. Res. **35**, 3202–3212 (2008)
14. Xia, Y., Chen, B., Yue, J.: Job sequencing and due date assignment in a single machines shop with uncertain processing times. Eur. J. Oper. Res. **184**, 63–75 (2008)

15. Manikas, A., Chang, Y.L.: Multi-criteria sequence-dependent job shop scheduling using genetic algorithms. Comput. Ind. Eng. **56**(1), 179–185 (2009)
16. Sha, D.Y., Lin, H.H.: A multi-objective PSO for job-shop scheduling problems. Expert Syst. Appl. **37**(2), 1065–1070 (2010)
17. Al-Hinai, N., El Mekkawy, T.Y.: Robust and stable flexible job shop scheduling with random machine breakdowns using a hybrid genetic algorithm. Int. J. Prod. Econ. **132**, 279–291 (2011)
18. Seckiner, S.U., Gokcen, H., Kurt, M.: An integer programming model for hierarchical workforce scheduling problem. Eur. J. Oper. Res. **183**, 694–699 (2007)
19. Emmons, H., Burns, R.N.: Off-day scheduling with hierarchical worker categories. Oper. Res. **39**, 484–495 (1991)
20. Hung, R.: Single-shift off-day scheduling of a hierarchical workforce with variable demands. Eur. J. Oper. Res. **78**, 49–57 (1994)
21. Narasihman, R.: An algorithm for single shift scheduling of hierarchical workforce. Eur. J. Oper. Res. **96**, 113–121 (1996)
22. Billonnet, A.: Integer programming to schedule a hierarchical workforce with variable demands. Eur. J. Oper. Res. **114**, 105–114 (1999)
23. Narashiman, R.: An algorithm for multiple shift scheduling of hierarchical workforce on four-day or three-day workweeks. INFOR **38**, 14–32 (2000)
24. Al-Yaqoob, S.M., Sherali, H.D.: Multiple shift scheduling of hierarchical workforce with multiple work centers. J. Inf. **18**(3), 325–342 (2007)
25. Pastor, R., Corominas, A.: A bicriteria integer programming model for the hierarchical workforce scheduling problem. J. Model. Manag. **5**, 54–62 (2010)
26. Ozguven, C., Sungur, B.: Integer programming models for hierarchical workforce scheduling problems including excess off-days and idle labor times. Appl. Math. Model. **37**, 9117–9131 (2013)
27. Koulamas, C.: The total tardiness problem: review and extensions. Oper. Res. **42**(6), 1025–1041 (1994)

Chapter 22
A Conceptual Framework for Analysing Supply Chain Performance of Oil Palm Refinery in Malaysia

Fitra Lestari, Kamariah Ismail, Abu Bakar Abdul Hamid, Eko Supriyanto, and Wahyudi Sutopo

Abstract The purpose of this research is the development of a supply chain framework in order to provide the process reference model based on the relationship between entities. This research chooses multiple cases of oil palm refinery in Malaysian because there is a barrier in implementing supply chain strategy in the form of inadequate measurement performance system. The methodology developed in this research comprised three phases involving business process reengineering, supply chain relationship model and performance indicator. There are 35 refineries in the peninsular Malaysia. However, measurement of supply chain performance will be selected in 6 oil palm refineries. The findings show that the conceptual framework can provide an instrument to analyze supply chain performance and evaluate the existing supply chain strategy.

Keywords Business process reengineering • Oil palm refinery • Performance indicator • SCOR model • Supply chain performance • Supply chain relationship model

F. Lestari • K. Ismail (✉) • A.B. Abdul Hamid
Faculty Management, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: fitra_lestari@yahoo.com; m-maria@utm.my; m-abakar@utm.my

E. Supriyanto
Faculty Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: eko@utm.my

W. Sutopo
Industrial Engineering, Sebelas Maret University, Surakarta, Indonesia
e-mail: sutopo@uns.ac.id

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_22

1 Introduction

The oil palm industry has several entities of processing industries involving Milling, Crushing Refinery and Oleochemical in order to win the market through strengthening business networks into supply chain strategies [1]. To conduct of the supply chain strategy, most of researchers divide the oil palm business into upstream and downstream [2, 3]. According Arip et al. [4] downstream sector has the biggest added value of products rather than upstream sector. Therefore, it has the potential to enhance the comparative advantage in order to develop industrial products of oil palm feedstock. In addition, there is one of processing industries in the downstream sector that connect with overall entities which it is the oil palm refinery. This industry produces many type derivative products of oil palm nevertheless they are processed through the different technology of processing industry [5].

The business process of the oil palm industry requires to improve the performance in supply chain strategies. Most of the research finds there is a barrier in the application of supply chain strategy in the form of the measurement system is inadequate [6, 7]. Omain et al. [8], conduct the research on the implementation of supply chain strategy in the oil palm industry in Malaysia. The finding shows one of the barriers in implementing supply chain strategy is in term of relationship between members of the supply chain that cause the measurement system inadequate. Therefore, supply chain strategies were used for different constructs for different country [9].

The purpose of this study is to conduct the method of the supply chain strategy in order to develop a conceptual framework for analyzing supply chain performance of the oil palm refinery in Malaysia. The method will be constructed using the Supply Chain Operations Reference (SCOR) model which this method can be used for modeling, benchmarking and performance measurement [10]. In this case, SCOR model is used to conduct relationships between entities in the oil palm downstream industry. Based on the gap of this research, it requires a tool to analyze supply chain strategy in the oil palm refinery that provide the reference process model based on the relationship between entities. To achieve objectives of the study, the oil palm refinery will be conducted based on the activity business process in Malaysia.

2 Measuring Supply Chain Performance

Measurement of supply chain performance aims to determine the extent of supply chain strategy that has been set in running business processes working properly. Akyuz and Erkan [11] conducted a literature review on supply chain performance measurement. They revealed that this concept is useful for controlling and managing the system. Thus, the several stages of improvements that are planned obtain optimal results. Supply chain performance measurement is very necessary to assist in evaluating supply chain strategy [12]. Many supply chain performance measurement

methods that have been studied. However, application of these methods must be adapted to the case studies that are conducted. Kurien and Qureshi [13] suggest that the selection of the method of supply chain performance measurement must be in accordance with a system that will be improved. Indeed, it is necessary to answer the question of how measure supply chain performance of oil palm downstream industry in Malaysia.

Several methods are used by researchers in measuring supply chain performance. Balanced Scorecard (BSC) was developed by Kaplan and Norton in 1992 which it purpose to measure performance based on perspectives, financial, customer, internal business and innovation and learning. It methods focus on performance measurement based on the vision and mission of the organization [14, 15]. Furthermore, Activity-Based Costing (ABC) is a technique to measure performance in supply chain strategy based on data of financial flow involving financial value and variable cost [16] nevertheless it can not used for non-financial measurement. In addition, Ramos et al. [17], measurement of the supply chain performance using Total Quality Management (TQM) to analyze and improve the goal of the business process in form products, services, and processes by creating value for customers. This approach focuses achieving the target and customer satisfaction. Moreover, Supply Chain Operations Reference (SCOR) model is used to evaluate the relationship between supplier, manufacturing and customer in order to measuring the performance more detail into activity of business process [18]. These methods are used to measure supply chain performance and support problem solving of supply chain cases, however the advantages of the methods depend on the scope of the problem.

This research develops a framework of performance measurement in supply chain strategy using the SCOR model which it was built by the Supply Chain Council (SCC). This method is used to investigate the relationship between entities in business processes [19]. Therefore, distribution of physical products from the supplier to the customer has a pattern that can be standardized into operational procedures within business processes [20]. In addition, the relationship between entities in supply chain strategy require developing the modeling tool which this method can adopt business process reengineering based on the stages that have standardized involving Plan, Make, Source, Deliver and Return [21]. This approach also is used as a set of key performance indicators [22].

3 Oil Palm Refinery in Malaysia

The upstream industry produces raw materials from oil palm business. These industries involve the plantation that produces Fruit Fresh Bunch (FFB) and the milling that extracts the raw material in the form of oils [23]. Thus, these industries commonly are an upstream stage in its production process. Moreover, downstream industry involves processing the materials collected from the upstream industry into finished products which it consists of crushing, refinery and oleochemical [24, 25]. The type of customer will vary depending on market destination such

as international market and local market because the downstream industry has direct contact with customers through the finished products. Furthermore, oil palm downstream industries in Malaysia have the high competition in the global market because industries of oil palm derivative products have been clustered into the specific finished product. These products will be processed by the developed countries to retail businesses such as food products, chemical industry, personal care products, animal feed, cleaning products and several sources of energy [26].

The oil palm refinery has the relationship with the overall downstream industries involving milling, crushing, oleochemical, delivery service providers and customer to process the oil palm derivative products [27]. Therefore, the oil palm refinery requires raw materials from the milling industry in the form of crude palm oil (CPO) and the crushing industry in the form of crude palm kernel oil (CPKO). The finished products are RBD palm oil, RBD palm kernel oil and palm fatty acid distillate. Finished products that consist of various types of products will be distributed in warehouses or direct shipping to await orders from the customer. In addition, the final products will deliver to export through port and local industry in the form of business to business (B2B) or business to customer (B2C). The integrations of oil palm refinery with other entities within the oil palm downstream industry show there is several configurations in the supply chain strategy. Thus, stakeholders can optimize the integration in the oil palm refinery with measuring supply chain performance based on the relationship between entities involving supplier, manufacturer, distributor and customer [28]. Figure 22.1 show supply chain oil palm refinery in Malaysia.

Furthermore, Malaysia has 55 refinery processes including 35 refineries in the Peninsular Malaysia, 6 refineries in Sabah and 14 refineries in Serawak [29]. The cases of oil palm refinery are chosen to be part of the sample with a specific

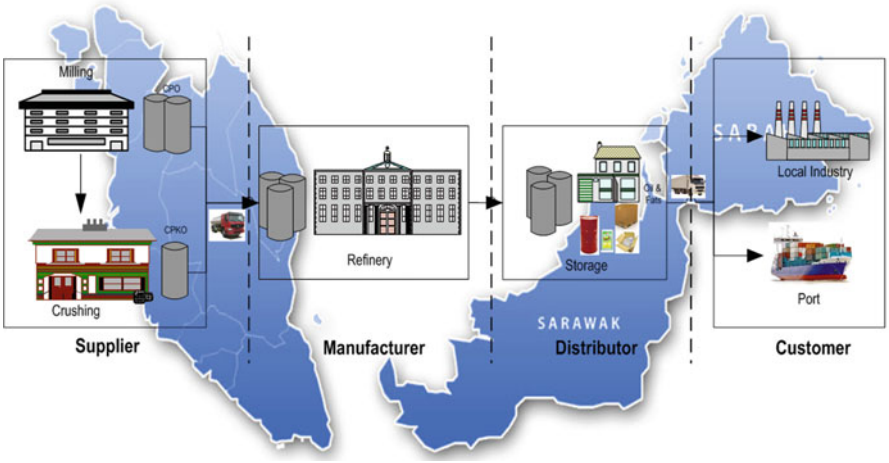


Fig. 22.1 Supply chain oil palm refinery in Malaysia

purpose of the research. In addition, the overall refineries in Malaysia cannot directly distribute their finished products to international market through the port and local industry such as oleochemical industry. Thus, the samplings of the case study will select certain refineries that deliver the finished product both of them which the oleochemical industry is only in Peninsular Malaysia. It technique adopts sampling based on purposive personal judgment [30]. Indeed, this research can choose multiple cases in Malaysian oil palm refinery. There are 35 refineries in the peninsular Malaysia. However, measurement of supply chain performance will be selected in 6 oil palm refineries.

4 Development Supply Chain Framework

The following phases of frameworks for analyzing supply chain performance:

- Phase one: Reengineering of business process

  The SCOR model serves to improve activities in the supply chain strategy [31]. The model contains five different processes involving Plan, Source, Make, Deliver and Return. Furthermore, it is breakdown into several levels of SCOR model which the model can be seen in Fig. 22.2. The plan represents the planning activities in achieving corporate targets. Source represents suppliers that supply raw material to the industry. Make represent process transformation of raw materials into finished products. Deliver is a process to distribute and deliver finished products to the customer. Return represents customer complaints about product of rejecting, defect, and repair. The five categories will be selected based on activity in oil palm refinery in order to understanding the business process. Furthermore, the data collection will be developed into business processes using software which it can develop business processes based on SCOR model.
- Phase two: Determining supply chain relationship model

  Most of the manufacturers have similarity the core business; nevertheless, they have the different operational strategy in production [32]. This research will explore some of operational strategies in the oil palm refinery, which become several alternatives of supply chain configurations in the oil palm downstream industry. This strategy considers the relationship between supplier and buyer, which are based on the frame thinking of five force competitive models into by Michael Porter at 1980. It conducts the condition whenever supplier or buyer has power to affect activity in the business process. In addition, many oil palm industries can be categorized into a holding company which it means supplier and buyer allow to cooperate in order to obtain the optimal size of production. The categories of relationship model between supplier and buyer can be seen in Table 22.1.
- Phase three: Selecting performance indicator

  The numbers of alternative relationship models appear into several scenarios in supply chain strategies in order to measure performance indicator. SCOR

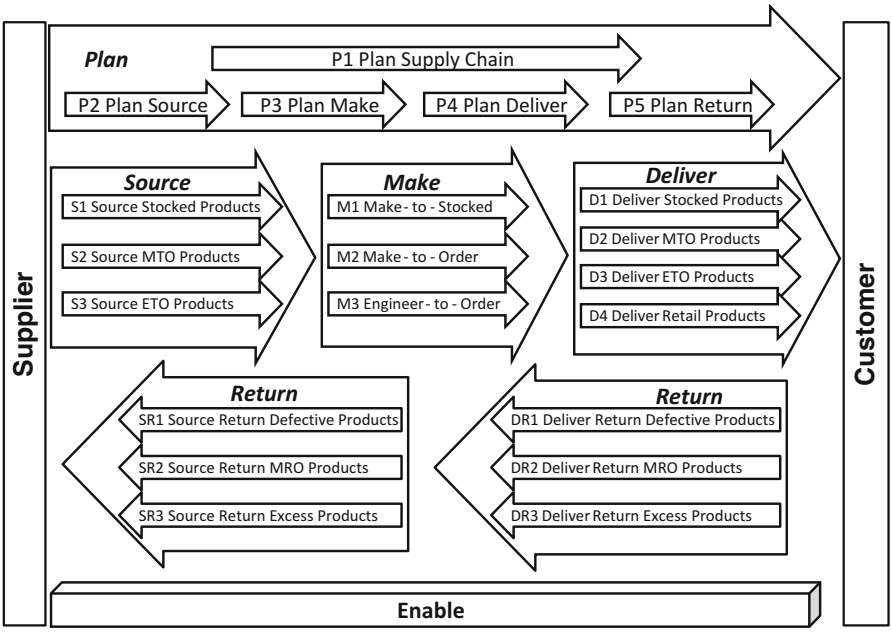


Fig. 22.2 The reference process modeling using SCOR model

Table 22.1 Type of supply chain relationship model

No	Supplier – buyer relationship	Description
1	Supplier dominance	The supplier has power to affect the buyer and uses their forecasting to push products to the buyer
2	Buyer dominance	The activity of business process is controlled by the buyer. Thus, suppliers produce a number of products from information of the buyer
3	Cooperate optimal size	Buyer and supplier conduct the optimal size of demand and supply because they are a holding company that allow deciding the optimal size of production
4	No collaboration	Production based on orders because there is not information sharing. Most of this type considers the high of variant product

model will describe performance indicators into five categories involving Plan, Source, Make, Deliver and Return. Then, it will be grouped from list of indicators to performance indicator which this research adopt qualitative content analysis into inductive category [33, 34]. Finally, the stakeholder will determine the appropriate performance indicator based on the relationship between the supplier and buyer. The performance indicator includes cost, quality, time, asset, reliability, responsiveness, utilization, and political stability.

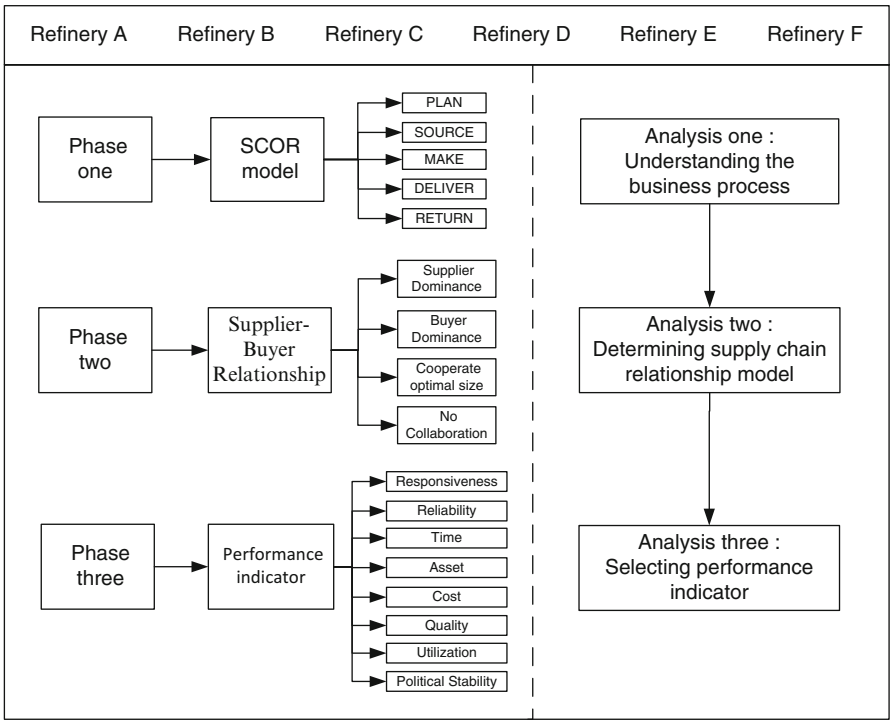


Fig. 22.3 Framework for analyzing supply chain performance in oil palm refinery

The expected results of this research will be achieved into a model to analyze the supply chain performance based on the relationship between entities in the oil palm downstream industry. It is obtained from several cases selection of the oil palm refinery in Malaysia. Number of cases required to collect the data until no significant new findings that are revealed. Thus, the framework is developed more concerned in reaching data saturation. The detail framework of supply chain performance in oil palm refinery can be seen in Fig. 22.3.

5 Conclusion and Future Work

This research has shown the development of a supply chain framework in order to analyze supply chain performance in oil palm refineries. Therefore, oil palm downstream industry involves several processing industries which they are represented by supply chain of business processes. From the framework, it has identified some potential areas for further research. First, developing of the reference process model adopted SCOR model. Second, determining supply chain relationship model based on supplier and buyer profiles. Third, selecting the appropriate performance

indicator of supply chain strategy is done into five categories of SCOR model. Finally, the model provides guidance for improving the business process in the oil palm refinery. Thus, the linkage of business process shows an illustration of the relationship between suppliers, manufacturer, distributor and customers. To support the above model in further research, the data analysis is supported by the Process wizard software in order to develop the reference process model.

Acknowledgment The authors thank to the Amanah Pelalawan Foundation for giving PhD Degree Scholarship Program.

References

1. Suksa-ard, C., Raweewan, M.: Optimization of supply and demand balance in a palm oil supply chain. Thammasat Int. J. Sci. Technol. **18**(2), 14–31 (2013)
2. Lestari, F., et al.: Supply chain configuration using hybrid SCOR model and discrete event simulation. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, WCE 2014, London, 2–4 July 2014, pp. 967–972
3. Duijn, V.G.: Traceability of the palm oil supply chain. Lipid Technol. **25**(1), 15–18 (2013)
4. Arip, M.A., Yee, L.S., Feng, T.S.: Assessing the competitiveness of Malaysia and Indonesia palm oil related industry. World Rev. Bus. Res. **3**(4), 138–145 (2013)
5. Gibon, V., et al.: Future prospects for palm oil refining and modifications. OCL – Oilseeds and fats, Crops and Lipids **16**(4), 193–200 (2009)
6. Fawcett, S.E., Magnan, G.M., McCarter, M.W.: Benefits, barriers, and bridges to effective supply chain management. Supply Chain Manag. Int. J. **13**(1), 35–48 (2008)
7. Richey, R.G., et al.: The moderating role of barriers on the relationship between drivers to supply chain integration and firm performance. Int. J. Phys. Distrib. Logist. Manag. **39**(10), 826–840 (2009)
8. Omain, S.Z., et al.: Supply chain management practices in Malaysia palm oil industry. The 11th Asia Pacific Industrial Engineering and Management Systems Conference, Melaka (2010)
9. Mollenkopf, D., Dapiran, G.P.: World-class logistics: Australia and New Zealand. Int. J. Phys. Distrib. Logist. Manag. **35**(1), 63–74 (2005)
10. Bolstorff, P., Rosenbaum, R.: Supply chain excellence: a handbook for dramatic improvement using the SCOR model, 3rd edn. Amacom, New York, USA (2007)
11. Akyuz, G.A., Erkan, T.E.: Supply chain performance measurement: a literature review. Int. J. Prod. Res. **48**(17), 5137–5155 (2010)
12. Hervani, A.A., Helms, M.M., Sarkis, J.: Performance measurement for green supply chain management. Benchmark. Int. J. **12**(4), 330–353 (2005)
13. Kurien, G.P., Qureshi, M.N.: Study of performance measurement practices in supply chain management. Int. J. Bus. Manag. Soc. Sci. **2**(4), 19–34 (2011)
14. Bigliardi, B., Bottani, E.: Performance measurement in the food supply chain: a balanced scorecard approach. J. Facil. Emerald **28**(5/6), 249–260 (2010)
15. Jie, F., Parton, K.A.: Balanced scorecard for Australian cattle producers: an application balanced scorecard for Australian cattle producers: an application. Australas. Farm Bus. Manag. J. **6**, 27–39 (2009)
16. Surowiec, A.: Costing methods for supply chain management. 1st Annual International Interdisciplinary Conference, pp. 24–26. Azores (2013)
17. Ramos, J.C., Asan, Ş.S., Majetic, J.: Benefits of applying quality management techniques to support supply chain management. International Logistics and Supply Chain Congress, Istanbul, Turkey, pp. 1–9 (2007)

18. Taghizadeh, H., Hafezi, E.: The investigation of supply chain's reliability measure: a case study. J. Ind. Eng. Int. **8**(1), 22–30 (2012)
19. Wang, W.Y.C., Chan, H.K., Pauleen, D.J.: Aligning business process reengineering in implementing global supply chain systems by the SCOR model. Int. J. Prod. Res. **48**(19), 5647–5669 (2010)
20. Soosay, C.A., Hyland, P.W., Ferrer, M.: Supply chain collaboration: capabilities for continuous innovation. Supply Chain Manag. Int. J. **13**(2), 160–169 (2008)
21. Erkan, T.E., Bac, U.: Supply chain performance measurement: a case study about applicability of SCOR model in a manufacturing industry firm. Int. J. Bus. Manag. Stud. **3**(1), 381–390 (2011)
22. Singh, R., et al.: Modeling supply chain performance of organized. Int. J. Sci. Res. **3**(3), 1–10 (2013)
23. Jamian, R., et al.: A conceptual approach of 5S to improving quality and environmental performance of Malaysian oil palm dealers. J. Teknol. **70**(1), 65–73 (2014)
24. Zwart, R.: Sustainability in the cultivation of oil palm –issues and prospects for the industry. J. Oil Palm Environ. **4**(5), 41–62 (2013)
25. Hai, T.C.: The Palm Oil Industry in Malaysia from Seed to Frying Pan. Petaling Jaya, Plantation Agriculture, WWF Malaysia (2002)
26. Sarmidi, M.R., Enshasy, H.A.E., Hamid, M.A.: Oil palm: the rich mine for pharma, food, feed and fuel. American-Eurasian J. Agric. Environ. Sci. **5**(6), 767–776 (2009)
27. Panapanaan, V.: Sustainability of Palm Oil Production and Opportunities for Finnish Technology and Know-How Transfer. Lappeenranta University of Technology, Lappeenranta (2009)
28. Li, S., et al.: The impact of supply chain management practices on competitive advantage and organizational performance. Omega **34**(2), 107–124 (2006)
29. Economics and Industry Development Division of Malaysian Palm Oil Board: Overview of the Malaysian Oil Palm Industry 2013, Malaysia. http://bepi.mpob.gov.my/images/overview/Overview_of_Industry_2013.pdf. Accessed 16 June 2014
30. Bigliardi, B., Bottani, E., Galati, F.: Open innovation and supply chain management in food machinery supply chain: a case study. Int. J. Eng. Sci. Technol. **2**(6), 244–255 (2010)
31. Huan, S.H., Sheoran, S.K., Wang, G.: A review and analysis of supply chain operations reference (SCOR) model. Supply Chain Manag. Int. J. **9**(1), 23–29 (2004)
32. Teece, D.J.: Business models, business strategy and innovation. Long Range Plann. **43**(2–3), 172–194 (2010)
33. Wang, W., Sedera, D., Tan, F.: Measuring CRM and SCM benefits: a preliminary measurement model. Pacific Asia Conference on Information Systems, Hyderabad, India (2009)
34. Elo, S., et al.: Qualitative content analysis: a focus on trustworthiness. SAGE Open **4**(1), 1–10 (2014)

Chapter 23
Investigations on Abrasive Electrochemical Grinding Process (AECG)

Jerzy Kozak and Grzegorz Skrabalak

Abstract This paper reports the development of the mathematical modeling of an abrasive electrochemical grinding (AECG) process. In AECG a metal-bonded grinding wheel is used instead of a graphite or metal wheel-electrode used in electrochemical grinding (ECG). Therefore, the mechanical abrasion is combined with the electrochemical dissolution. Significant improvements in material removal rates have been observed with AECG of advanced materials, such as superalloys, sintered carbides and metal matrix composites. The interaction of the abrasion and electrochemical dissolution in AECG is analyzed on the basis of computer simulation of the material removal process. The effects of main machining parameters such as feed rate, voltage and grit protrusion is studied.

Keywords Anodic dissolution • Grinding wheel • Hybrid machining • Micro-cutting • Process modeling • Simulation

1 Introduction

The technological improvement of machining processes can be achieved by combining different physico-chemical action on the material being treated. In particular a mechanical action, which is used in conventional material removal processes can be combined with respective interactions applied in unconventional manufacturing processes such as electrical discharge machining (EDM), electrochemical machining (ECM), laser beam machining (LBM) etc.

J. Kozak (✉)
Institute of Advanced Manufacturing Technology, ul. Wroclawska 37a, Krakow 30-011, Poland

Institute of Aviation, al. Krakowska 110/114, Warsaw 02-256, Poland
e-mail: jkozak64@wp.pl

G. Skrabalak
Institute of Advanced Manufacturing Technology, ul. Wroclawska 37a, Krakow 30-011, Poland
e-mail: grzegorz.skrabalak@ios.krakow.pl

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_23

The reasons for developing a hybrid process are to make use of the combined or mutually enhanced advantages, and to avoid or reduce some adverse effects the constituent processes produce when they are individually applied [1].

The most numerous group make Abrasive Hybrid Machining (AHM) processes, and they are the most commonly used in industry. There can be distinguished three main subgroups: Abrasive Electrical Machining (AEDM), Abrasive Electrochemical Machining (AECM), and Abrasive Electro-Chemical-Discharge Machining (AECDM) [2].

Essential conditions, under which any machining process is performed, are type of the tool and movements of the tool in relation to the workpiece. Main tools include: metallic electrodes containing abrasive grains, i.e. grinding wheels or abrasive sticks with metallic bond, metallic electrodes and free abrasive grit, and tools composed of abrasive segments and segmented metallic electrodes.

Depending on type of tool and working movements, that are used in particular process, there are many various methods, schematic diagrams of which are shown in Fig. 23.1. Among the others, there are: – methods making use of abrasive tool with metallic bond: Abrasive Electrochemical Grinding (AECG), Abrasive Discharge Grinding (AEDG), and Electrochemical Honing (ECH); – methods using free abrasive grains: Abrasive Electrochemical Finishing (AECF), Abrasive Electrical Discharge Finishing (AEDF), and Ultrasonic Electrochemical Machining (USECM).

Interconnections between machining mechanisms, which take part in metal removal process, cause the productivity of hybrid machining to be much greater than summed up results of the individual processes (assuming that the same parameters, characteristic for relevant processes, are kept). Similar positive effects have been achieved in respect to energy consumption and surface finish. In order to present the phenomena and factors that occur during given electro-mechanical machining processes, and which influence on these positive results, we will consider abrasive electro-chemical machining process, such as Abrasive Electrochemical Grinding (AECG) (Fig. 23.2.)

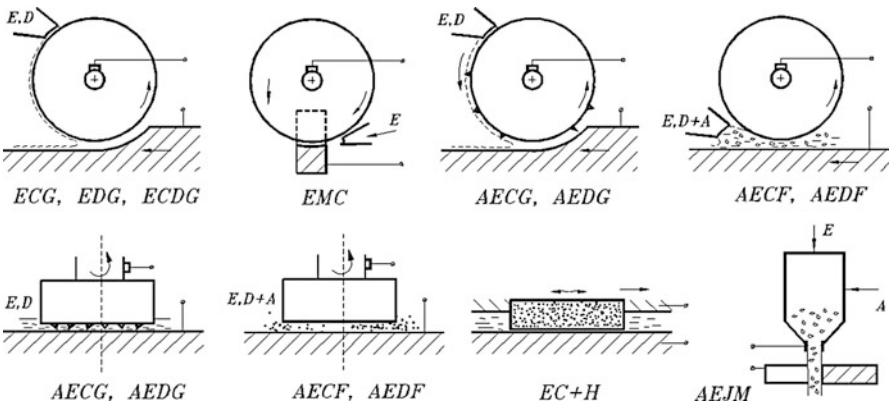


Fig. 23.1 Schematic diagram of selected methods of abrasive electrical machining (*A* abrasive, *D* dielectric, *E* electrolyte) [3]

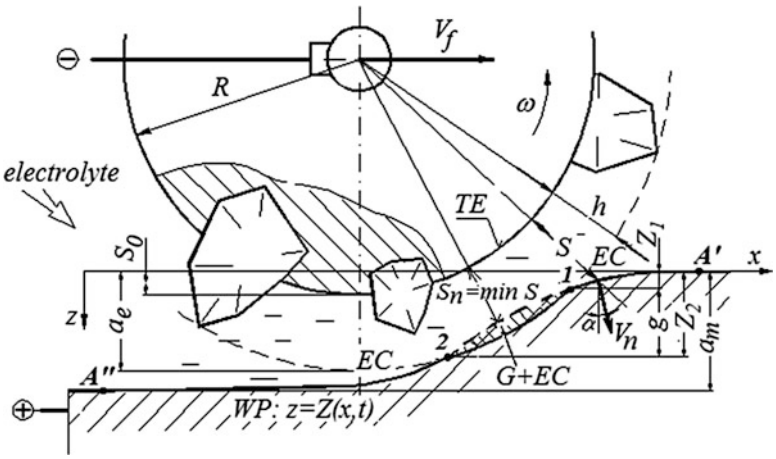


Fig. 23.2 Schematic diagram of abrasive electrochemical grinding (AECG)

Electrochemical grinding with metal bonded abrasive tool (AECG), consists in combination mechanical and electrochemical processes, acting on the workpiece, what considerable changes performance indexes of the machining process. Process productivity are increased many times, surface layer properties are improved, while tool wear and energy consumption decrease. Particularly effective is AECG process for machining parts made from difficult to cut materials, such as sintered carbides, creep resisting alloys (eg. Inconel, Nimonic), titanium alloys, metallic composites (eg. PCD-Co, Al-SiC, Al-Al_2O_3) [1]. Increase in performance indexes of hybrid machining processes, such as AECG process, results from interconnections between microcutting, electrochemical dissolving, changes in surface layer properties of material in machining area, and processes that take place in active layer of the grinding wheel surface.

Schematic diagram of AECG process in the case of plane surface grinding with circumferential face of the wheel is shown in Fig. 23.2. The following zones can be distinguished within the machining area: *EC* zone, where anodic dissolving prevails, and $G + EC$ zone, where microcutting predominates. The participation of either of these two mechanisms in material removal process, and in surface layer creation, changes as AECG process parameters are changed.

For example, when changing the feed rate V_f, while keeping the other parameters of the process constant, and V_f reaches value which is smaller than certain critical value *Vg*, then microcutting vanishes, because at these conditions the smallest gap *S*, between working surface of the electrode *TE* (grinding wheel bond surface) and workpiece (WP) surface, is greater than height *h* of the protruding fragment of abrasive grains. Depth of allowance to be removed, that is the real depth of cut a_m, can be considerable greater than set value a_e, and it depends on anodic dissolving. Action of the abrasive grains is limited to their influence on conditions existing in the gap, and in particular on electric field, transport of electrolyte, and hydrodynamic effect on near-anode boundary layers.

Increase in lengthwise-feed rate V_f results in gap size S. decrease, and, after value V_g has been exceeded, gap width s. along the definite length *1–2* (Fig. 23.2), become smaller than value h, and in consequence microcutting contribution to metal removal process increases. Further increase in feed rate V_f causes gap width under axis of the wheel to become also smaller than h. At this moment, depth of cut is approaching value a_e, because beyond the grinding wheel anodic dissolving contribution to metal removal process is relatively small. However, it should be noted, that influence of anodic dissolving on ground surface layer condition in area beyond the wheel may be significant.

For determination of basic characteristics of the process, namely the output data considering between depth of microcutting, g, the real depth of cut, a_m, material removal rate *MRR*, as a function of chosen input data, such as setting depth of cut, a_e, feed rate, V_f, working voltage, U, angular velocity, ω, the mathematical modeling and computer simulation have been developed.

2 Mathematical Modeling of the AECG Process

During the first stage of AECG process modeling, it is essential to derive the distribution of interelectrode gap (*S)*, which results from anodic dissolution. At this stage, the microcutting process is not taken into account for estimating material removal rate, however it is included in the electrochemical machinability coefficients (k_v) and total overpotential (E). When the distribution of interelectrode gap *(S)* is derived, it is possible to determine the $G + EC$ zone with microcutting, as the zone where abrasive grains sticking out (height of grain – h in the Fig. 23.1, $h > S_n$) of the bonding material of grinding wheel (Fig. 23.2 – hatched area between points 1 and 2) cross with the surface of machined material. In this way it is possible to derive dimensions of machining zones EC and $G + EC$, and the depth (g) of layer machined by microcutting, as well as ratio g/a_r.

Deriving of interelectrode gap distribution (S) in case of AECG is similar to ECG (often described as electrochemical machining using rotating electrode – *ECM-RE*), however in this case the working electrode (ER) is metal surface (with $\boldsymbol{R}$ radius), which is the averaged contour of metal bonding of grinding wheel. In the literature [4, 5, and others], shape of the anode surface was derived basing on the steady state conditions of ECM-RE. It led to determining gap distribution (S) basing on the solution of nonlinear differential equation of first order. The drawback of this method, resulting in relatively big mistakes and ambiguous solutions due to lack of a priori initial conditions.

The model presented below incorporates the electrochemical dissolution process of the anode surface (workpiece) from start to end of machining process. Moment of ending machining process is defined by machining time of position of the electrode tool.

Modeling was performer using following main assumptions:

- linear distribution of electric potential along the gap size, S,
- electrical conductivity of electrolyte, κ in the gap is constant,
- uniform distribution of abrasive grains on the surface of grinding wheel (β - ratio of metallic uncovered by grains surface of grinding wheel vs active side cylindrical surface of grinding wheel is known),
- the electrochemical reactions will be accounted for by introducing the total overpotential $\boldsymbol{E = E_a - E_c}$, where $\boldsymbol{E_a}$ and $\boldsymbol{E_c}$ are the overpotential potential of anode and cathode, respectively,

Basing on the theory of electrochemical shaping, the evolution of machined surface shape $\boldsymbol{F(x,y,z,t) = 0,}$ is described by the equation [5]:

$$\frac{\partial F}{\partial t} + k_v i \, |\nabla F| = 0 \tag{23.1}$$

The coefficient of electrochemical machinability k_v is equal to the volume of material dissolved per unit electrical charge – in general it is the function of, among others, current density.

Going forward to the coordinate system (x, z) – Fig. 23.2, change of machining surface geometry $\boldsymbol{z = Z(x, t)}$ during AECG process is described by:

$$\frac{\partial Z}{\partial t} = k_v i \sqrt{1 + \left(\frac{\partial Z}{\partial x}\right)^2} \tag{23.2}$$

with initial condition $\boldsymbol{z = Z_0(x)}$ describing surface before machining.

The current density $\boldsymbol{i}$ is depends on medium conductivity in the gap and on the voltage $\boldsymbol{U}$ according to Ohm's law, which is extrapolated to the whole gap size.

According to the assumptions concerning potential distributions, current density may be described as:

$$i = \frac{\beta \kappa \, (U - E)}{S} \tag{23.3}$$

where: $\boldsymbol{i}$ – mean anode current density including the presence of abrasive grains (which are insulators) by β ratio

During AECG process, the gap is equal to:

$$S = \sqrt{(x - x_0)^2 + (Z - z_0)^2} - R \tag{23.4}$$

where: $\boldsymbol{x_0 = x_0(t),\ z_0 = z_0(t)}$ – equations describing movement of grinding wheel central point (TE).

Presented above Eqs. 23.2, 23.3, and 23.4 describe the process of electrochemical shaping. The system of Eqs. 23.2, 23.3, and 23.4 has been solved numerically

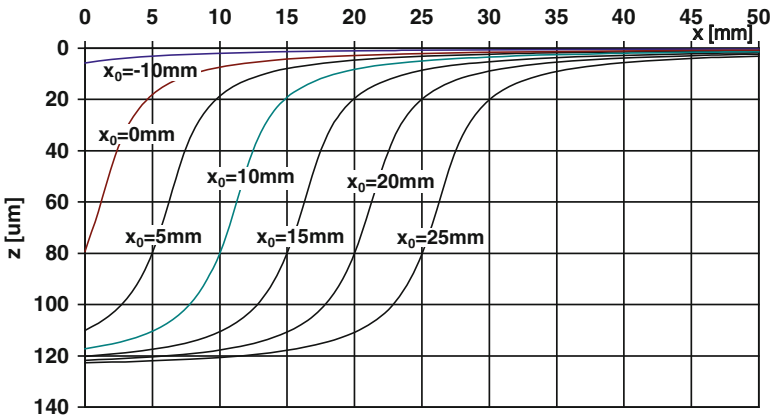


Fig. 23.3 Evolution of anodic dissolved surface geometry

using the Finite Difference Method. The simulation software basing on the same mathematical model were described in Ref. [6].

Figure 23.3 shows the evolution of anode shape achieved during simulation of AECG process for various positions of grinding wheel ($\boldsymbol{x_0}$). While analyzing presented graphs, it may be noticed that shape of the anode surface settles up with time and process becomes stationary.

3 Analysis of Computer Simulation Results

For the needs of AECG process analysis, simulations were performed for various machining parameters and various characteristics of $k_v(i)$, conductivity κ, β – ratio and dimensions R i h. Below are presented selected results for $\boldsymbol{S_0} = 0$, $R = 50$ mm, $k_v = 1$ mm^3/Amin, $\kappa = 0.016$ A/Vmm, $\beta = 0.3$ and for following investigated parameters $U\text{-}E = 6, 9, 12$ V; $V_f = 60, 120, 240, 480$ mm/min; $h = 25, 50, 70$ μm.

Figures 23.4 and 23.5 show the steady state surfaces of the anode and distribution of gap (S) thickness for various machining parameters. There is also sketched the working electrode (ER) together with the abrasive grains sticking out of the wheel up to $h = 70$ μm. Basing on the presented curves it may be stated that for parameters 1 and 2, the depth of grinding depends on the anodic dissolution (when $U\text{-}E = 12$ V and $V_f = 60$ mm/min) and real/achieved depth of machining $a_m = 117$ μm is 1.64 times deeper than pre-set depth of $a_e = h = 70$ μm. Lowering interelectrode voltage to $U\text{-}E = 9$ V results in increase of microcutting process in material removal and depth of grinding drops to $a_m = 75$ μm, the same reaching the level of pre-set a_e. When the interelectrode voltage is set at $U\text{-}E = 6$ V, similar effect may be achieved by increasing linear speed V_f. For example, doubling the speed V_f ($V_f = 120$ mm/min) caused lowering of maximal depth of dissolution to

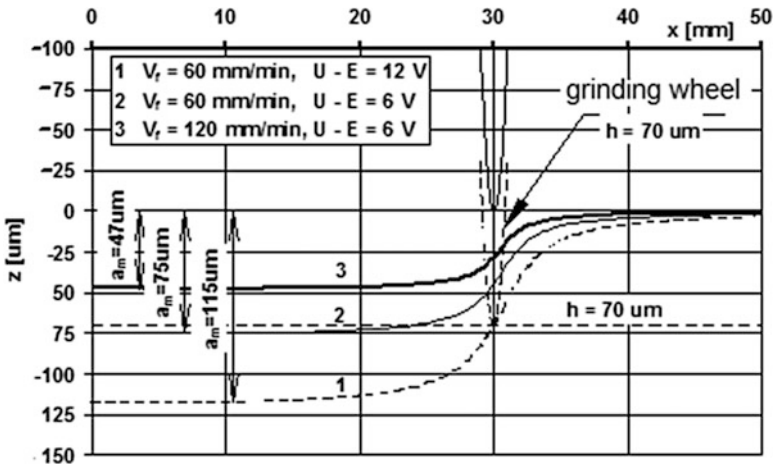


Fig. 23.4 Geometry of steady state anodic surfaces

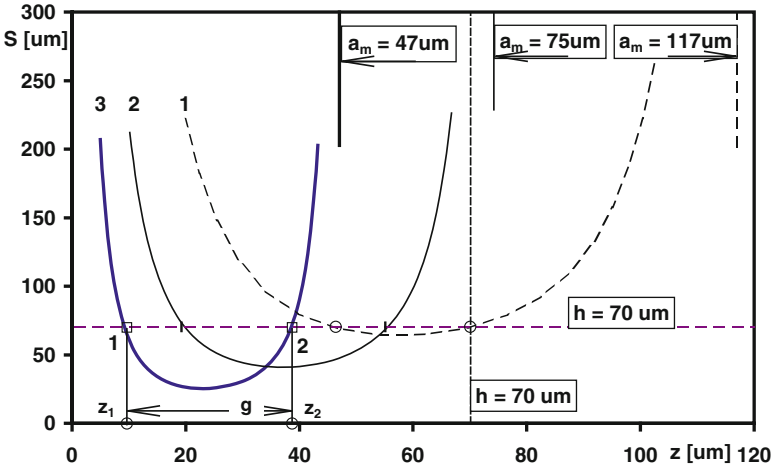


Fig. 23.5 Distribution of the gap size along z axis

$a_m = 47$ μm. In this case it may be stated that the depth of machining is mainly dependant on the pre-set depth $a_e = 70$ μm, and that during AECG process, the biggest share in material removal comes from microcutting process. The influence of electrochemical processes influences mainly mechanical properties of machined surface layer and reducing surface integrity.

Important characteristics of the AECG process concern dependencies between gap width (S_n), maximal dissolution depth a_m and process parameters. These two factors, as well as, height of abrasive grains (h) influence the run of the AECG process. Dependencies of $S_n(V_f)$ and $a_m(V_f)$ for constant interelectrode voltage

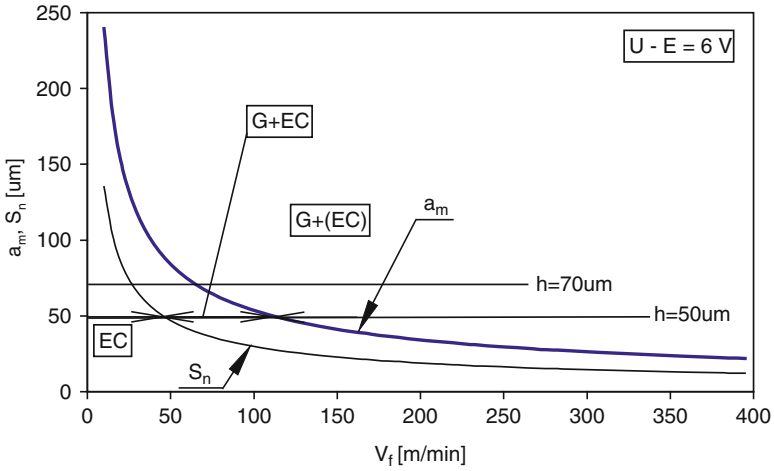


Fig. 23.6 Interelectrode gap $S_n(V_f)$ and depth of cut $a_m(V_f)$

U-$E = 6$ V, is shown in Fig. 23.6. In Fig. 23.6 are also marked lines for two heights of abrasive grains: $h = 50$ i 70 μm, what enables specifying ranges of V_f, for which different material removal processes are dominant. For example, for $h = 50$ μm:

- when V_f, < 45.6 mm/min – dominant process is electrochemical dissolution (EC);
- 45.6 mm/min $< V_f$, < 111 mm/min – height of grain above bonding becomes bigger than S_n and the same microcutting appears (its share grows together with V_f);
- when V_f, > 111.4 mm/min – microcutting is dominant (marked as G + (EC) in Fig. 23.6) and depth of cut is almost equal to setting value a_e;

Above listed boundary values of V_f, depend on abrasive grain height *(h)* and interelectrode voltage (U). Basing on the results of simulations, regression equations for shares of processes have been derived:

- border between *EC* and *G + EC*:

$$\mathbf{189}[(\mathbf{U\text{-}E})\,/\mathbf{V_f}]^{\mathbf{0.66}} = \mathbf{h}$$

- border between *G + EC* and *G+(EC):*

$$\mathbf{334}[(\mathbf{U\text{-}E}/\mathbf{V_f})]^{\mathbf{0.65}} = \mathbf{h}$$

Figure 23.7 presents limits for V_f and U-E when $h = 25$ i 50 μm. Point P (marked on the graph) shows the conditions, when probability of sparking drastically increases (in case of voltage increase).

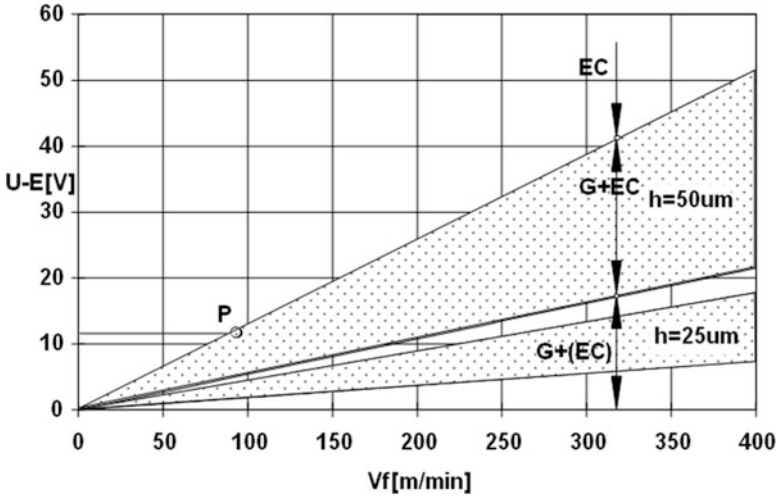


Fig. 23.7 Relation between V_f and *U-E* for boundary conditions

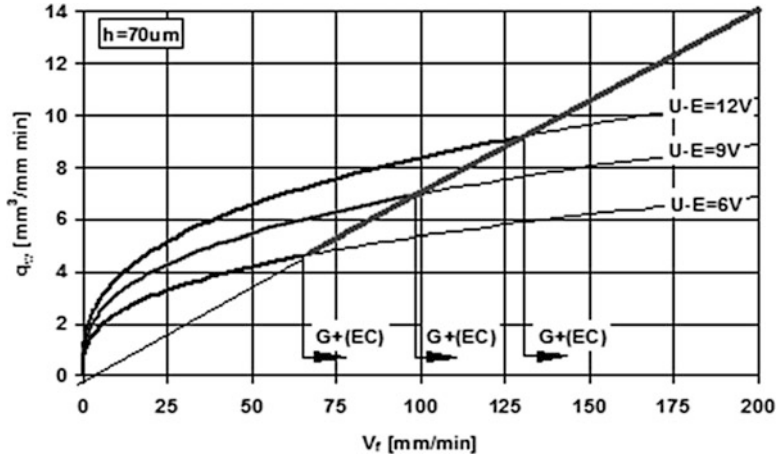


Fig. 23.8 Efficiency of AECG as the function of linear feed rate

Basing on the real/experimental depth of grinding, it is possible to derive AECG process efficiency – $\boldsymbol{q_w}$ in relation to grinding wheel width. It is expressed as $q_w = a_m V_f$. Achieved relationships between process efficiency and linear feed rate, for $h = 70$ μm and various *U-E*, are presented in Fig. 23.8. Radical drops of characteristics result from changing conditions and process shares from $G + EC$ to dominant role of microcutting $G + (EC)$.

Important influence on the cutting forces, during grinding process, has the depth of layer which is microcut – g. In case of AECG process it is limited with the dimensions of G + EC i $G + (EC)$ zones. Resulting from simulations values of

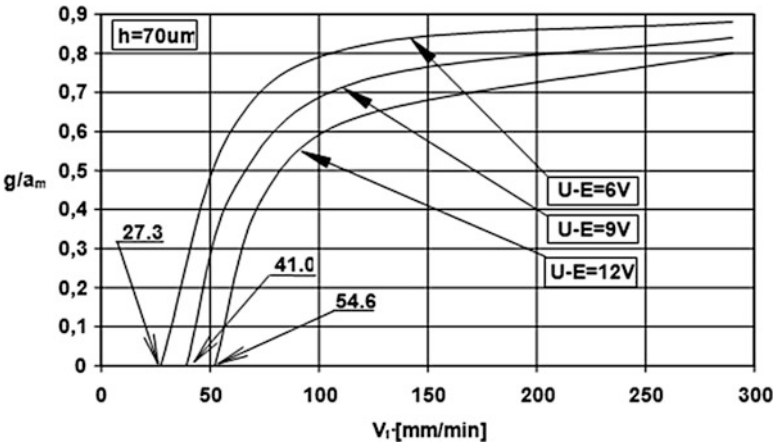


Fig. 23.9 Ratio g/a_m as the function of linear feed rate V_f

ratio g/a_m are presented in Fig. 23.9. For the feed rates (V_f) slower than those equal to the limits of $G + EC$ and $G + (EC)$, the share of g layer in grinding dept his significantly reduced. It leads to significant reduction of cutting forces, even to 0, when there appears pure electrochemical dissolution process. The boundary values of V_f, at which microcutting disappears, and which are marked in Fig. 23.9, depend on electrical parameters, electrochemical machinability of the material in a given electrolyte, and grinding wheel features, especially grain size and height h.

Dependencies achieved during simulation of the AECG process are similar to the results of machining.

4 Experimental Investigations of AECG Process

Experiments were conducted for machining slots of width equal to 0.3 mm in 0.2 mm thick stainless steel plate using brass bonded diamond grinding wheel (particle size 76 μm with concentration of 30%) (Figs. 23.10 and 23.11).

Electrolyte was 3% water solution of $NaNO_3$. Tests were performed for constant interelectrode voltage of 6, 8 and 10 VDC with constant feed rate $V_f = 1.2$ mm/min in each case. Results are presented in Figs. 23.12, 23.13, and 23.14.

Analyzing the presented SEM pictures of machined slots, there can be evaluated share of material removal processes. In case of machining with interelectrode voltage U-E = 6VDC (Fig. 23.12), material was removed mainly by microcutting/grinding process – phase G+(EC) from Fig. 23.6. There are visible traces of abrasive grains on the machine surface (bottom and side walls of the slot). When voltage was increased to 8VDC, also the share of electrochemical dissolution increased significantly (Figs. 23.13 and 23.14). In this case there is less traces of

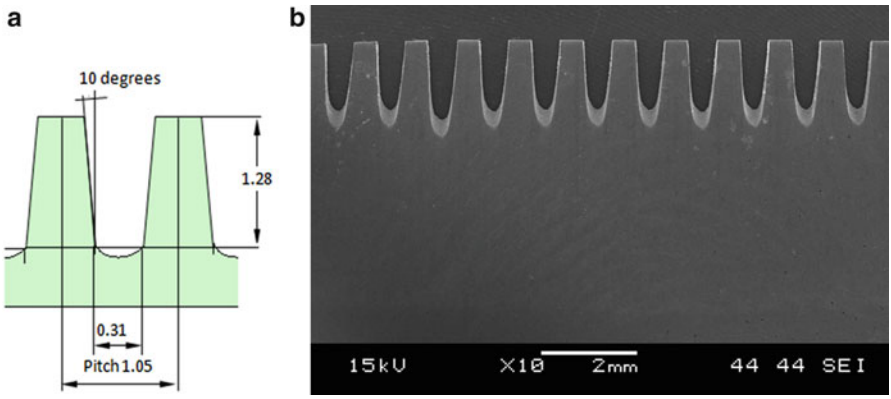


Fig. 23.10 (**a**) Schematic overview of the dimensions of the slot and (**b**) view of slots created during machining with 6 V, 1,000 rev/min, feed rate 0.9 mm/min, angle 60°

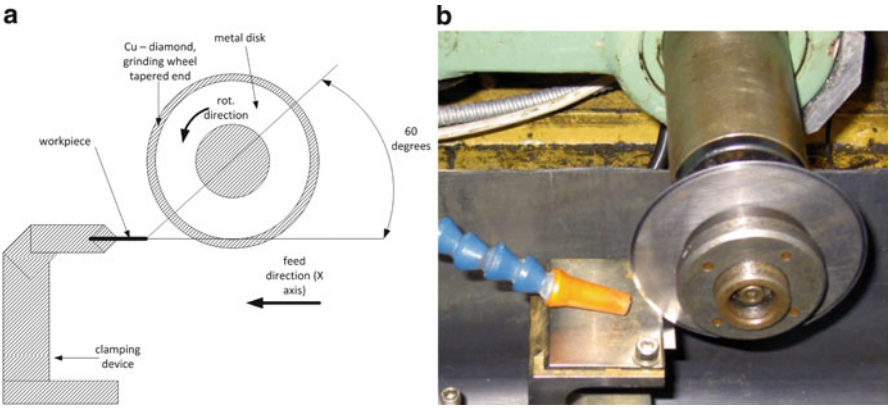


Fig. 23.11 (**a**) The machining scheme with 60° between the workpiece and disk and (**b**) test stand

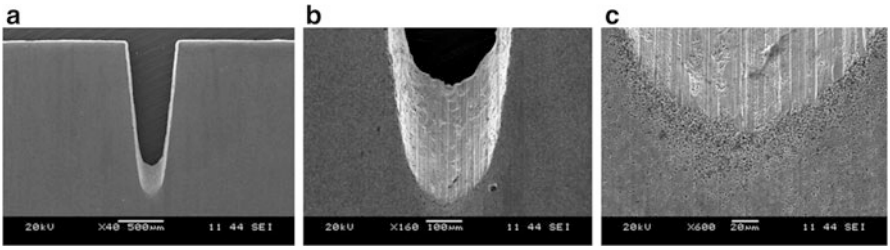


Fig. 23.12 Slots machined with AECG process with U-E = 6 VDC ((**a**) – magnification 40×, (**b**) – magn. 160×, (**c**) – magn. 600×) [7]

microcutting both in the bottom of the slot and its side walls – phase G + EC from Fig. 23.7. After increasing voltage to 10VDC, the share of microcutting decreased

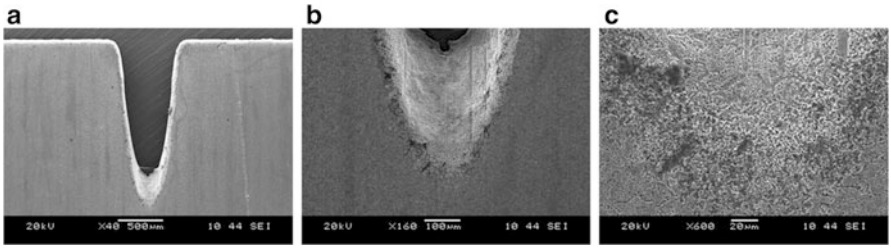


Fig. 23.13 Slots machined with AECG process with U-E = 8 VDC ((**a**) – magnification 40×, (**b**) – magn. 160×, (**c**) – magn. 600×) [7]

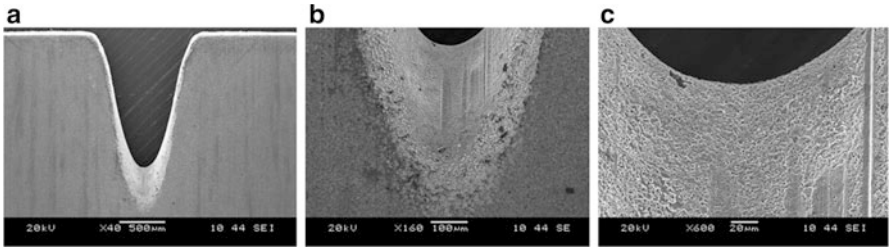


Fig. 23.14 Slots machined with AECG process with U-E = 10 VDC ((**a**) – magnification 40×, (**b**) – magn. 160×, (**c**) – magn. 600×) [7]

and the dominant removal process was electrochemical dissolution – phase EC +(G) from Fig. 23.7. High rate of electrochemical dissolution process is visible both on the geometry of the slot and on the surface at the bottom. However there are traces of abrasive grains on the machined surface, the grinding process has minor share in material removal.

Another example of practical application of the AECG process is shaping of composite, difficult to cut materials. Application of AECG process enabled significant improvement of Material Removal Rate while shaping cutting tips made of DDCC (Diamond Dispersed Cemented Carbide). Comparing to conventional grinding process, the MRR was increased by 16 times: from 0.14 mm^3/min (in case of conventional grinding) to 2.3 mm^3/min (in case of AECG process). In Fig. 23.15 there are presented geometries of used grinding wheel and machined groove, whilst Figs. 23.16 and 23.17 show the surface of machined groove and influence of the process on the surface layer. It can be observed on the mentioned SEM photographs of machined surfaces that AECG process does not influence the surface in the negative way. Although there are visible some cracks on the surface after machining (Fig. 23.16), they are of very small depth (not exceeding 20 µm) – as can be observed in Fig. 23.17.

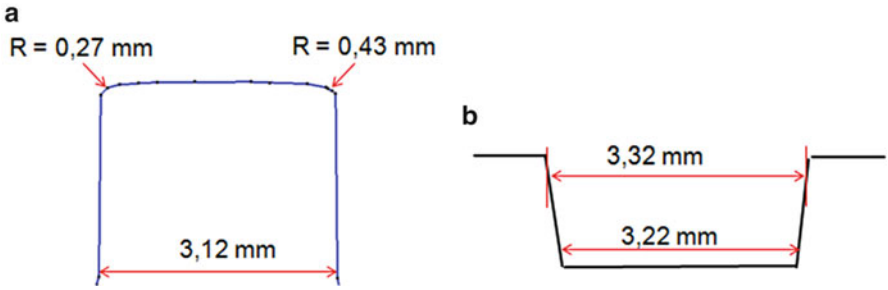


Fig. 23.15 (**a**) Geometry of grinding wheel used for AECG machining of DDCC composite; (**b**) geometry of grove machined in DDCC material (U-E = 30 VDC, V_f = 1 mm/min)

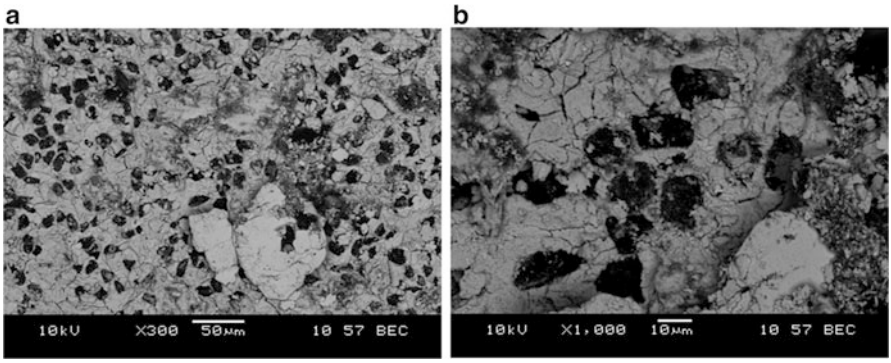


Fig. 23.16 Surface of machined groove in DDCC composite ((**a**) – magnification 300× and (**b**) – magnification 1,000×)

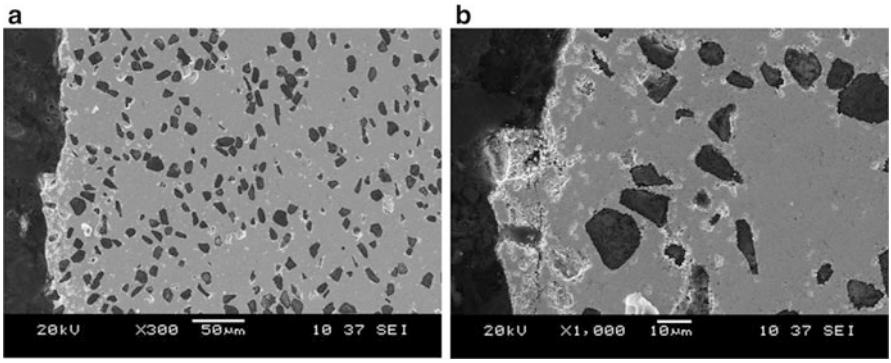


Fig. 23.17 Cross section of surface of machined groove in DDCC composite ((**a**) – magnification 300× and (**b**) – magnification 1,000×)

5 Conclusion

Described simulation of AECG process allows to derive approximate characteristics of the material removal process including machining conditions. It is worth to notice that performed mathematical modeling, especially in the aspect of evaluating geometry of interelectrode area, bases on the assumptions idealizing the machining process. In this aspect, results of simulations shall be treated as approximation. Described model is good starting point for further development of the detailed model of AECG process, which would incorporate changes of the electrolyte properties in the interelectrode gap resulting from products of dissolution and microcutting.

AECG process due to its characteristics is very useful in practical application. As shown in the presented example it may be applicable to simple-step shaping and sharpening of cutting blades. Analyzed process is also used for machining of difficult to cut, composite materials, like DDCC – Diamond Dispersed Cemented Carbide.

Acknowledgment This work was supported in part by the National Centre for Research and Development in Poland under Applied Research Programme, Grant No.: PBS1/A5/7/2012

References

1. Rajurkar, K.P., Zhu, D., McGeough, J.A., Kozak, J., De Silva, A.: New developments in electrochemical machining. Ann. CIRP **48**(2), 569–579 (1999)
2. Kozak, J., Oczos, K.: Selected problems of abrasive hybrid machining. J. Mater. Process. Technol. **109**(3), 360–366 (2001)
3. Kozak, J.: Selected problems of hybrid electro-mechanical machining. Arch. Mech. Technol. Automat. **16**, 59–65 (1996) (in Polish)
4. Noble, C.F.: Electro-mechanical action in peripheral electrochemical grinding. Ann. CIRP **32**(1), 123–127 (1983)
5. Davydov, A., Kozak, J.: High Rate of Electrochemical Shaping. Nauka, Moscow (1990)
6. Kozak, J., Rajurkar, K.P., Ruszaj, A., Sławiński, R.J.: Sculptured surface finishing by NC-electrochemical machining with ball-end electrode (ECM-CNC). Adv. Technol. Mach. Equip. **22**(1), 51–74 (1998)
7. Kozak, J., Skrabalak, G.: Analysis of abrasive electrochemical grinding process (AECG). Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, WCE 2014, London, 2–4 July 2014, pp. 1147–1152

Chapter 24
Pre-cooling Effects on the Resulting Grain Size in Friction Stir Processing of AZ31B

Ali H. Ammouri, Ghassan T. Kridli, George Y. Ayoub, and Ramsey F. Hamade

Abstract The effects of in-process cryogenic LN cooling on the resulting grain size of the friction stir processing (FSP) of twin roll cast (TRC) magnesium alloy AZ31B. Sheets (3 mm-thick) of TRC AZ31B were friction stir processed using a wide range of processing parameters (mostly tool feed and spindle speed). The tool rotational speed was varied between 600 RPM and 2,000 RPM while the tool feed rate varied between 75 mm/min and 900 mm/min. Thrust force and torque values were experimentally measured using a 4-component dynamometer. Temperature measurements were monitored during the different tests using Infrared sensors and thermocouples. The microstructure of processed samples was observed using optical microscopy. It was found that thrust force and torque values of the pre-cooled samples were 5% higher than those of the room temperature samples due to the material hardening induced by the cooling effect. Finer and more homogenous microstructure was observed for the pre-cooled samples when compared with samples processed at room temperature. The average grain size of pre-cooled samples was predicted using a relation -previously introduced by the authors- that relate grain size and the Zener-Hollomon parameter for TRC AZ31B. This equation was found to correctly predict grain diameter for in-line cooled FSP AZ31B samples at temperatures lower than room temperature.

Keywords AZ31B • Cryogenic cooling • Friction stir processing • Forces • Microstructure • Zener-Hollomon

A.H. Ammouri (✉)
Industrial and Mechanical Engineering Department, Lebanese American University, Byblos, Lebanon
e-mail: ali.ammouri@gmail.com

G.T. Kridli
Department of Industrial and Manufacturing Systems Engineering, University of Michigan Dearborn, Dearborn, MI 48128, USA
e-mail: gkridli@umich.edu

G.Y. Ayoub • R.F. Hamade
Department of Mechanical Engineering, American University of Beirut, Beirut 11-0236, Lebanon
e-mail: ga80@aub.edu.lb; rh13@aub.edu.lb

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_24

1 Introduction

Friction Stir Processing (FSP) is a solid state hot shear stirring process that has been widely used for grain size refinement. Cooling in manufacturing processes is known to improve tool life and surface finish. When it comes to friction stir processing and friction stir welding (FSW), cooling has some other advantages mainly related to microstructure enhancements.

Reduction in the width of the heat affected zone was observed during cryogenic CO_2 cooling of friction stir welded Aluminum alloy AA7010-T7651 [1]. Liquid nitrogen cooling resulted in Sub-micrometer grain size during FSP of AZ31B [2]. Water cooling during FSW of AA2219-T62 aluminum alloy was found to increase the strength and ductility of the resulting samples under tensile testing [3]. Improvement to joint strength of friction stir welded AA7075-T6 butt joints [4] and pure copper joints [5] was achieved by rapid water cooling.

The cooling techniques that were just described depend on continuously flooding the processed area with the cooling fluid. Pre-cooling that is introduced in this work depend on cooling the friction stir processing setup prior to engaging the processing tool into the workpiece. The effect of starting from a temperature lower than the normal operating temperature on thrust force, torque, process temperature, and microstructure is investigated in this work.

2 Experimental Setup

Friction stir processing of 3 mm thick TRC AZ31B sheets was performed on a HAAS-VF6 vertical machining center that was retrofit to perform friction stir processes. The sheets were firmly mounted on a 3 cm thick C30 steel backing plate using two holders. The workpiece holder was mounted to the machine's trunion using a KURT vice as shown in Fig. 24.1.

The FSP tool was made from SVERKER 21 (aka AISI-D2) tool steel manufactured by Uddeholms AB (SE-683 85 Hagfors, Sweden). The tool had a shoulder diameter of 19 mm, with a 6.4 mm pin diameter and a 2.7 mm pin height. After machining the tool to its final dimensions, it was hardened as per the datasheet provided by Uddeholms.

Force measurements are collected using the Kistler's Rotary 4-Component (Fx, Fy, Fz, and Torque) Dynamometer (Type 9123C) shown in Fig. 24.1. The Kistler 5223B charge amplifier acquires and amplifies the signal emanating from the dynamometer which is then collected by custom LabVIEW software through four analog input channels of a USB-6251 DAQ.

Two different temperature probes are utilized for monitoring temperatures at different locations of the setup. Thermocouples were placed at the middle of the workpiece on both sides of the top surface to log the temperature history throughout

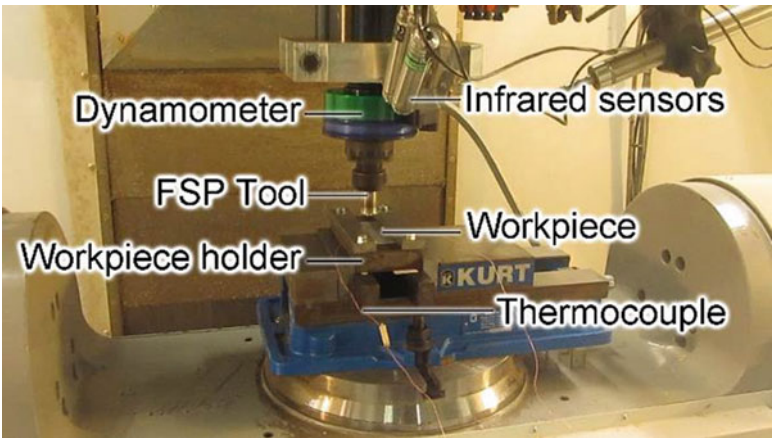


Fig. 24.1 The FSP experimental setup mounted on the trunion of the HAAS-VF6 vertical machining center

the friction stir process. Teflon pads were used as insulators between the fixture clamps and the workpiece to prevent heat loss through conduction with the clamps. The thermocouple signals were acquired through a NI USB-9213 DAQ which has built in signal conditioning for thermocouple signals.

Two OS151-USB compact non-contact infrared temperature sensors from Omega were used to monitor the FSP tool and workpiece temperatures. A laser sighting tool was used to align the sensors to the final position shown in Fig. 24.1. The 4–20 mA output signal from the sensors was acquired by the National Instruments USB-6251 DAQ device through a CB-68LP screw connector. The signal was calibrated and logged for all the test cases considered through the developed LabVIEW software.

Pre-cooling was achieved by spraying cryogenic liquid nitrogen on the FSP fixture with the mounted workpiece. The cryogenic cooling system comprises a 180 L liquid nitrogen tank with a delivery copper pipe and a control valve. Figure 24.2 shows the pre-cooling cryogenic setup positioned on HAAS vertical machining center. The nozzle can be seen at the end of the piping system that connects an accumulator to the workpiece.

The temperature measurements collected from the sensors are used for computing the amount of heat generated as:

$$Q = h \times A \times (T_s - T_\infty) \tag{24.1}$$

where h, A, T_s, T_∞ are convective heat transfer coefficient of air, cross-sectional area of the work piece, surface temperature of the work piece, and free stream temperature.

Fig. 24.2 The pre-cooling cryogenic setup: shown is the nozzle at the end of the piping system from accumulator to the workpiece fixtured on HAAS vertical machining center

The calculated values of Q from Eq. 24.1 are equal to the heat that should be removed from the work piece according to:

$$Q = \dot{m} \times h_{fg} \tag{24.2}$$

where $\dot{m}$ is mass flow rate of LN2 and h_{fg} is latent heat of vaporization of LN2.

The methodology combining Eqs. 24.1 and 24.2 was operated on a range of surface temperatures likely to be encountered in FSP processes. The mass flow rate values corresponding to these temperatures were calculated. The analysis give mass flow rates for surface temperature at the work-piece as shown in Fig. 24.3. The figure allows for estimating the mass flow rate needed to absorb the heat created at the tool-work interface at different temperatures. The desired temperature that we want the work piece to reach is 25 °C, however this will change for each temp set by user.

The pre-cooling cryogenic setup is shown in Fig. 24.4 while spraying liquid nitrogen as shown. Liquid nitrogen is sprayed over the area to be cooled until the temperature of the setup reached the target value of −10 °C. This temperature drop would guarantee that the starting temperature of the friction stir process is around 0 °C by the time the process starts. Temperature of the setup during the pre-cooling phase was monitored by the thermocouples.

The simulations were run according to the test matrix shown in Table 24.1 with and without pre-cooling. The tool rotational speed was varied from 600 RPM to 2,000 RPM and the traverse feed from 75 mm/min to 900 mm/min. Three feeds were considered for each tool rotational speed to cover the applicable operating range following recommendations by [6].

Upon the successful completion of a friction stir process the samples were prepared for optical microscopy imagery. Samples were cut, ground, and polished

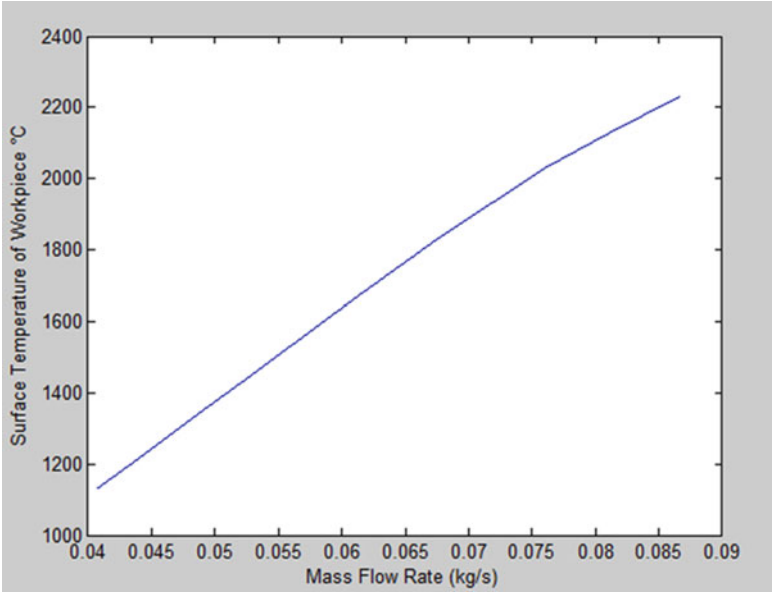


Fig. 24.3 Surface temperature vs Mass flow rate

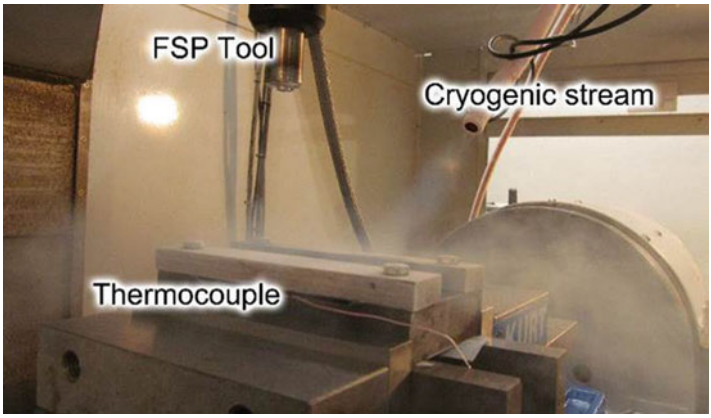


Fig. 24.4 The pre-cooling cryogenic setup: shown spraying liquid nitrogen

on the BUEHLER Metaserv 2000 with autopol I & II. Acetic-picral solution was used to etch the samples before optical imagery on a B41 Olympus microscope with the motorized X-Y stage and Olympus XC50 digital camera. Figure 24.5 shows the non-uniform grain size distribution of the as received TRC AZ31B at 50× magnification. The large variation of grain size is noted with grains ranging from 2 to 100 μm (average grain size of 13 μm).

Table 24.1 The experimental test matrix

	Tool rotational speed, RPM							
	600	800	1,000	1,200	1,400	1,600	1,800	2,000
Feed rate, mm/min	75 (A1)	100 (B1)	100 (C1)	150 (D1)	300 (E1)	350 (F1)	400 (G1)	500 (H1)
	100 (A2)	125 (B2)	150 (C2)	250 (D2)	500 (E2)	550 (F2)	600 (G2)	700 (H2)
	125 (A3)	150 (B3)	200 (C3)	350 (D3)	700 (E3)	750 (F3)	800 (G3)	900 (H3)

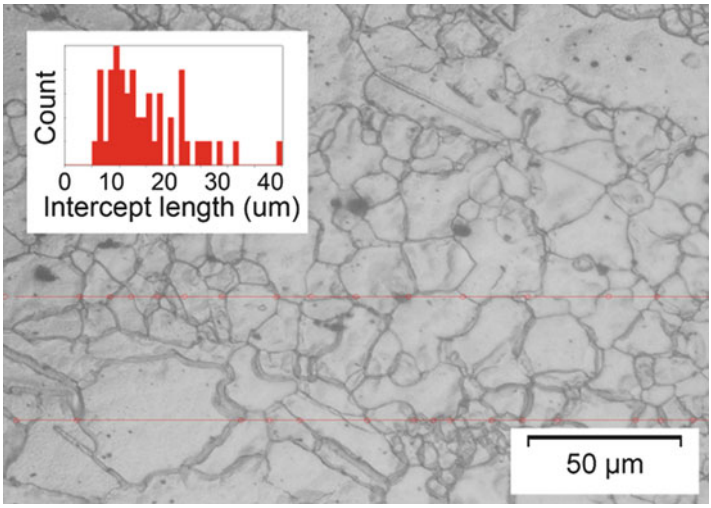


Fig. 24.5 Micrograph of the as-received TRC AZ31B at 50× magnification

3 Results

Samples of the friction stir processed specimen for both the room temperature and pre-cooled test cases are shown in Fig. 24.6a–d. Test cases F2 and G3 of Table 24.1 which correspond to (1,600 RPM – 550 mm/min) and (1,800 RPM – 800 mm/min) respectively, are shown at room temperature (Fig. 24.6a, c) and with pre-cooling (Fig. 24.6b, d). The pre-cooled samples are suffixed by a "C" to differentiate it from the room temperature samples.

The first obvious difference between the room temperature samples and the pre-cooled ones is the improvement in the surface finish of the pre-cooled samples. Without cooling, the material at the top surface of the workpiece is heavily plasticized and thus resulting in the surface roughness shown in Fig. 24.6a, c. On the other hand, pre-cooling the samples resulted in decreased ductility which improved the surface quality of these samples.

Forces, torques, temperature profiles, and microstructure of the pre-cooled samples were compared to the ambient temperature samples. The effect of pre-cooling on the temperature profile during FSP can be clearly noticed from the thermocouple and IR sensors that generated the signals shown in Fig. 24.7a–d for

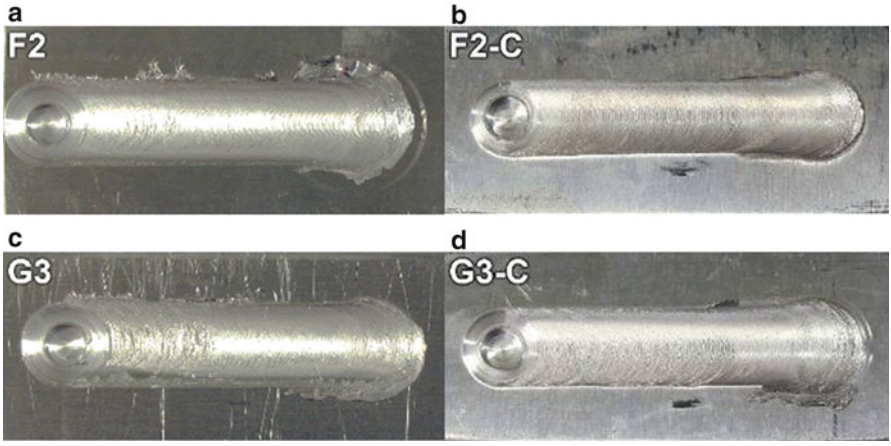


Fig. 24.6 FSP processed samples for test cases F2 and G3 of Table 24.1 shown at (**a**)-(**c**) room temperature and (**b**)-(**d**) with pre-cooling, respectively

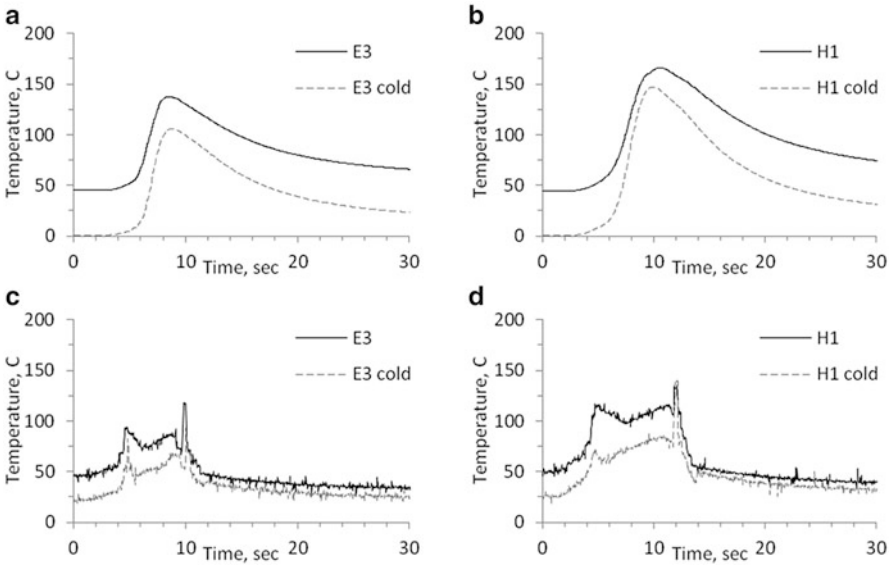


Fig. 24.7 Comparison between precooled and room temperature thermocouple readings of test case (**a**) E3 and (**b**) H1, and IR readings for test case (**c**) E3 and (**d**) H1

test cases E2 and H1. The 50 °C difference in temperature shifted both pre-cooled temperature profiles (Fig. 24.7a, b) downward at the beginning. The IR signals shown in Fig. 24.7c, d are capped at the sensors' minimum reading range (20 °C). As the tool traversed through the workpiece the temperature difference decreased to around 30 °C.

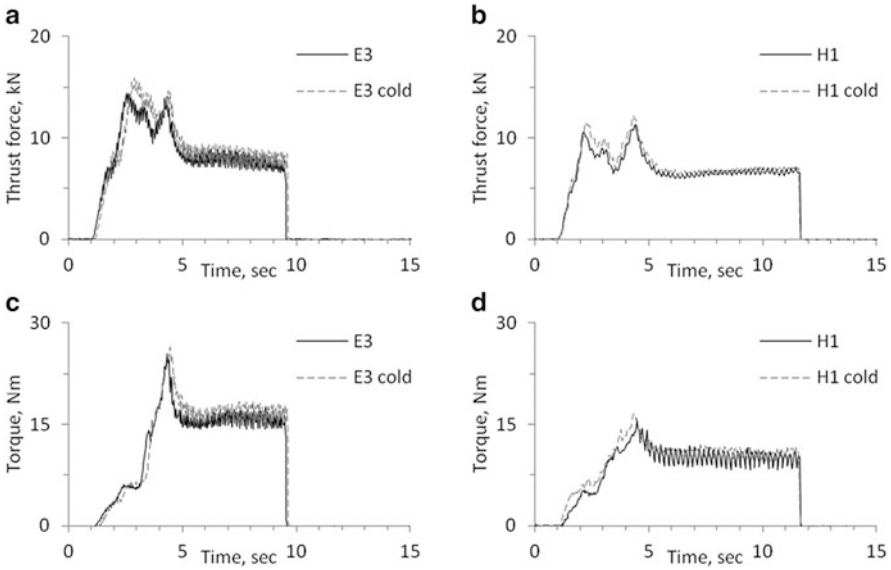


Fig. 24.8 Comparison between precooled and room temperature thrust forces of test case (**a**) E3 and (**b**) H1, and torques for test case (**c**) E3 and (**d**) H1

Compared with the room temperature test cases, the pre-cooled samples had higher thrust forces and torques which is due to the material hardening induced by the cooling effect. Both the maximum plunging and steady state traverse thrust forces as well as the maximum plunging and steady state torques were slightly increased by an average ratio of 5%. Figure 24.8a–d shows the thrust force and torque comparison for test cases E3 and H1. The effect is more obvious during the plunging phase since the workpiece is still cold. As the tool traverses along the processing line, the workpiece builds up heat and the pre-cooling effect is reduced.

The effect of pre-cooling on the final microstructure of the friction stir processed samples can be seen in Fig. 24.9. For the room temperature samples F2 and G3 shown in Fig. 24.9a, c, the average grain size was 8.0 μm and 7.7 μm, respectively. The average grain size of the pre-cooled test cases on the other hand had finer grains with an average diameter of 6.5 μm and 6 μm, respectively. The pre-cooled samples had more uniform grains size with less spread when compared to the room temperature samples. This can be noticed by examining the histograms that are also shown in Fig. 24.9a–d.

In order to explain the refinement caused by pre-cooling the samples, Eqs. 24.3 and 24.4 which relate the average grain size (d) of magnesium alloys during dynamic recrystallization to the Zener-Hollomon parameter (Z) [7] were used.

$$\ln d = 8.79 - 0.23 \ln Z \tag{24.3}$$

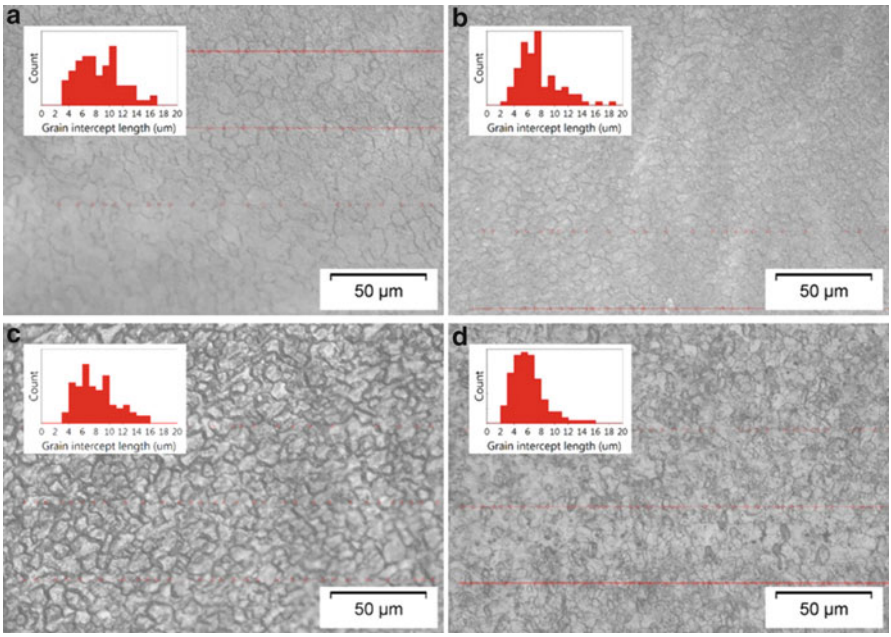


Fig. 24.9 Microstructure comparison of test cases (**a**) room temperature F2, (**b**) precooled F2, (**c**) room temperature G3, and (**d**) precooled G3

$$Z = \dot{\varepsilon} \exp\left(QR^{-1}T^{-1}\right) \tag{24.4}$$

The Zener-Hollomon parameter is described as a temperature compensated strain rate where Q is the activation energy of AZ31 (130 kJ mol^{-1}) and R the universal gas constant (8.314 J mol^{-1} K^{-1}). To describe the linear relation between ln Z and ln d for AZ31B, the coefficients in Eq. 24.3 are as proposed by Ammouri et al. [8]. Equations 24.3 and 24.4 are only valid upon the onset of DRX which occurs at temperature and strain rate values exceeding 227 °C and 0.5 s^{-1}, respectively [9].

The average temperatures and strain rates in the stir zone of the cases F2 and G3 were calculated to be (470 °C and 466 °C) and (98 s^{-1} and 104 s^{-1}), respectively. These values were found from the results of simulations of FSP using an experimentally validated FE model [10]. Given the 30 °C temperature offset of the pre-cooled samples during the traverse phase of FSP shown in Fig. 24.7, and using the strain rate values of test cases F2 and G3 in Eqs. 24.3 and 24.4, the predicted grain size of the pre-cooled samples would be 6.3 µm and 5.9 µm, respectively. This decrease in grain size is caused by the increase in the Zener-Hollomon parameter described in Eq. 24.4 due to the temperature decrease. The difference of 0.2 µm and 0.1 µm between the predictions of Eq. 24.3 and the experimental results for test cases F2 and G3 validates the performance of this equation in predicting the average grain size during FSP.

4 Conclusions

The effect of cryogenic pre-cooling on the friction stir processing of 3 mm thick sheets of TRC AZ31B magnesium alloy was investigated in this work. For a wide range of process parameters ranging between 600 RPM and 2,000 RPM for the tool rotational speed and between 75 and 900 mm/min for the tool traverse feed, the effect of pre-cooling was examined.

It was found that pre-cooling by exposure to cryogenic fluid had the following effects on the friction stir processing of TRC AZ31B:

1. The 50 °C temperature difference between the pre-cooled samples and the room temperature sample was maintained during the plunging phase of FSP and decreased to around 30 °C during the traverse feed. This observation was validated via thermocouple and IR temperature sensor measurements.
2. Decreasing the operating temperature of the friction stir process resulted in a slight increase of 5% in the thrust force and torque of the pre-cooled samples. This is attributed to the increased hardness of the material caused by the temperature reduction.
3. The microstructure of both room-temperature and pre-cooled samples was finer and more homogenous than the as received material. When comparing the room temperature samples to the pre-cooled samples, the later had finer and more homogenous microstructure.
4. The linear relation between the Zener-Hollomon parameter and the average grain size of dynamic recrystallization during FSP that has the form $\ln d = 8.79 - 0.23 \ln Z$ was validated by the experimentally measured average grain size at temperatures lower than room temperature.

The above findings indicate that pre-cooling in friction stir processing can result in desirable improvements in grain structure due to in-line cooling process. These improvements are reflected via finer grain size and better surface finish when compared to samples processed at room temperature.

Acknowledgments This publication was made possible by the National Priorities Research Program (NPRP) Grant # 09-611-2-236 from the Qatar National Research Fund (a member of The Qatar Foundation). The authors like to acknowledge the financial support of the university research board (URB) of the American University of Beirut (AUB). The statements made herein are solely the responsibility of the authors. Also, special thanks to Chehab Medical Gases (Beirut, Lebanon) for their help in providing components that are compatible with Liquid Nitrogen and for sharing their knowledge and expertise.

References

1. Jariyaboon, M., et al.: Effect of cryogenic cooling on corrosion of friction stir welded AA7010-T7651. Anti Corros. Methods Mater. **57**(2), 83–89 (2010)

2. Anwaruddin, M.: Achieving ultrafine grains in Mg AZ31B-O alloy by cryogenic friction stir processing and machining. MS Theses, University of Kentucky (2011)
3. Xu, W.F., et al.: Improvements of strength and ductility in aluminum alloy joints via rapid cooling during friction stir welding. Mater. Sci. Eng. A **548**, 89–98 (2012)
4. Fratini, L., Buffa, G., Shivpuri, R.: Mechanical and metallurgical effects of in process cooling during friction stir welding of AA7075-T6 butt joints. Acta Mater. **58**(6), 2056–2067 (2010)
5. Xue, P., et al.: Achieving friction stir welded pure copper joints with nearly equal strength to the parent metal via additional rapid cooling. Scr. Mater. **64**(11), 1051–1054 (2011)
6. Ammouri, A.H., Kheireddine, A.H., Hamade, R.F.: FEM optimization of process parameters and in-process cooling in the friction stir processing of magnesium alloy AZ31B. Proceedings of ASME 2013 International Mechanical Engineering Congress & Exposition, IMECE2013 (2013)
7. Wang, Y., Lee, C., Huang, C., Lin, H., Huang, J.: Influence from extrusion parameters on high strain rate and low temperature superplasticity of AZ Series Mg-based alloys. Mater. Sci. Forum **426**, 2655–2660 (2003)
8. Ammouri, A.H., Kridli, G., Ayoub, G., Hamade, R.F.: Relating grain size to the Zener-Hollomon parameter for twin-roll-cast AZ31B alloy refined by friction stir processing. J Mater. Process. Technol. (in press)
9. Darras, B.: Integrated thermo-mechanical investigations of friction stir processing of light weight alloys. PhD Theses, University of Kentucky (2008)
10. Ammouri, A.H., Kridli, G., Ayoub, G., Hamade, R.F.: Investigating the effect of cryogenic pre-cooling on the friction stir processing of AZ31B. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 1137–1140. WCE, London, 2–4 July 2014

Chapter 25
Survey on Simulation Methods in Multi-axis Machining

Khadidja Bouhadja and Mohamed Bey

Abstract In industry, the evolution of productivity and quality of mechanical manufacture of complex shape parts (mold, automobile, form . . .) is marked by the development of several machining simulation techniques for modeling and predicting the manufacturing process to represent the most realistic cut phenomenon. There exist several machining simulation techniques and touch various levels. Thus, this chapter summarizes the literature review and presents the techniques in a simplified scheme for the rapid exploration in this area, and in order to direct the reader to select an appropriate approach linked to a geometric or physical problem at a given scales in Part-Tool-Machine system. Particular attention is given to geometric simulation methods of the macroscopic scale; completed by brief comparison between models of workpiece representation for material removal process (Dexel, Voxel, Triple-Dexel).

Keywords Geometric modeling • Machining simulation • Multi-axis machining • NC verification • Swept volume • Virtual workpiece

1 Introduction

Mechanical parts with free form surfaces used in various industries (molds, aerospace, etc . . .) are machined on multi-axis CNC milling machines because of their highly complex geometric shapes [1]. These parts must meet functional and/or style requirements, which need special attention in their production phase. Before real machining, it is essential to simulate virtually the machining to verify the tool path, to detect collisions, to predict cutting forces and to evaluate roughness of the finished surface . . . etc. For this, there are various methods in the machining simulation, which touch various levels of complexity. In order to clarify and separate the difficulties, this chapter summarizes the literature review on the various methods of multi-axis machining simulation. These methods are identified and presented

K. Bouhadja (✉) • M. Bey
Centre de Développement des Technologies Avancées, Baba Hassen BP.17, Algiers 16303, Algeria
e-mail: bouhadjakhd@yahoo.fr; bey_mohamed@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_25

in a simplified structure (Fig. 25.1) permitting a rapid survey, by answering the key questions: why (objective: geometric or physical simulation), at what scale (human, macroscopic, microscopic), and how (geometric models, dynamic models). In the simulation of material removal process in multi-axis machining, 03 geometric model representations are considered: workpiece model, tool model and tool swept volume model. Particular attention is given to geometric representation methods at macroscopic scale, concluded with a comparison between the three popular models of workpiece representation: Dexel, Voxel and Triple Dexel.

2 Simulation Categories

The machining simulation is divided into geometric and physical simulations (Fig. 25.2).

2.1 Geometric Simulation

The geometric simulation is used for verifying graphically the absence of interferences and collisions and the respect of tolerances imposed by the designer. In addition, it can provide geometric information necessary to the physical simulation.

2.2 Physical Simulation

The physical simulation of a machining process aims to reveal the physical aspects of a machining process such as cutting forces, vibrations, surface roughness, machining temperature and tool wear. It is based on the geometric simulation and on the choice of the cutting tool material [2].

3 Simulation Scales

The study of the machining is often dealt with by using multi-scale approach to separate difficulties by limiting the number of phenomenon to be considered and the size of the model at a given scale. Three levels of analysis can be distinguished: human, macroscopic and microscopic.

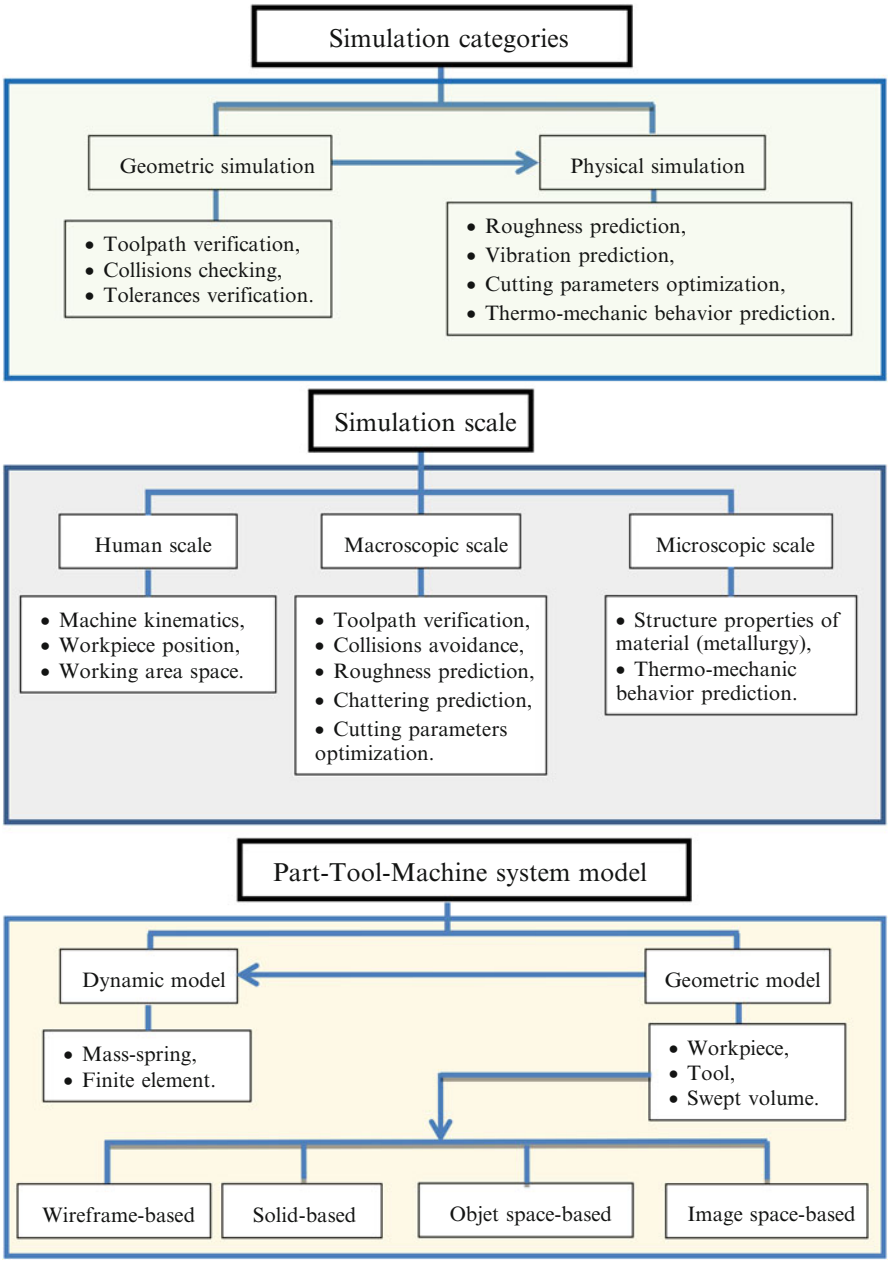


Fig. 25.1 Classification of simulation methods [1]

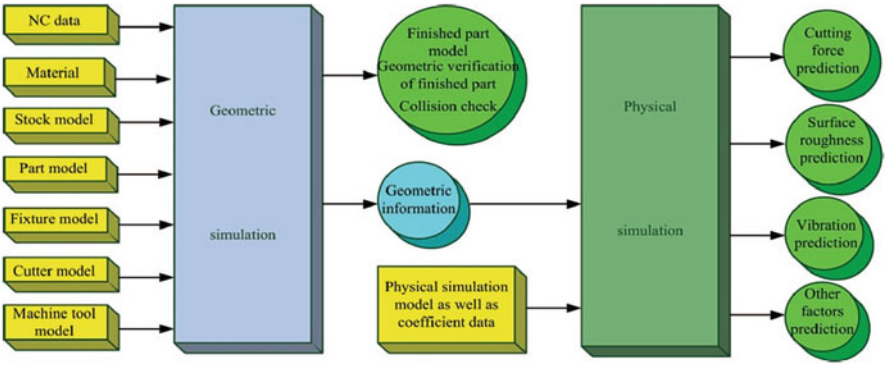


Fig. 25.2 General architecture of machining simulation [2]

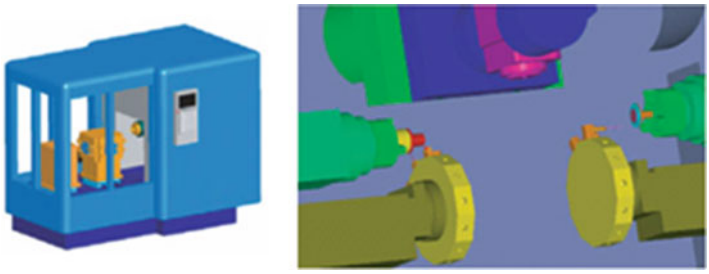


Fig. 25.3 Machine tool simulation [3]

3.1 Human Scale

It is a global simulation of the machining environment where the objective is to predict the behavior of production means to prepare the machining process by considering axes movements, workpiece position on the table and space of the working area (Fig. 25.3). This step is necessary when the means of production are complex and the movements of the workpiece relative to the tool are difficult to anticipate (multi-axis machine, machining robot, etc.). It allows the detection of possible collisions during machining.

3.2 Macroscopic Scale

In an industrial approach, it is very important to look closely to the part in order to visualize the removal of the material. The purpose of the simulation is to determine the volume of the material removed for each tool movement during part machining (Fig. 25.4).

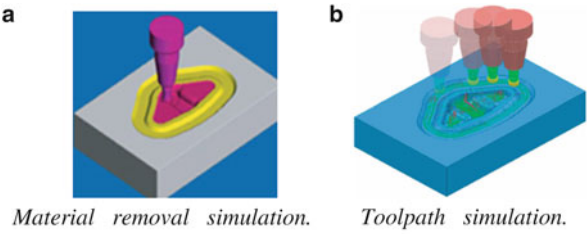


Fig. 25.4 Macroscopic scale machining simulation [3]

At this scale, simulation techniques allow to visualize and to anticipate surface defects totally related to the programmed strategy or to the machine kinematics. In the literature referenced, different kinds of works are cited. Some considered the representation of the workpiece [4, 5]. Other works, considered the generation of the tool swept volume [6–8]. The difficulty at this level is related to the kinematics of the 05-axis machine where the tool translates and rotates simultaneously. For a higher precision, other works used the theory of multi-body systems kinematics [9–11]. Other mechanical phenomena have an impact on the surface quality such as the chattering phenomenon where the appearance conditions are difficult to predict. This phenomenon degrades the machined surface quality and accelerates the wear of certain sensitive parts of the production means such as cutting tools and spindle. Discontinuous and periodic nature of the cut in milling is also the cause of systematic vibrations of the system constituted by Part-Tool-Machine. Other investigations are related to the prediction of the cutting forces to optimize the cutting conditions [12]. Artificial intelligence is used to avoid collisions for multi-axis machines [13].

3.3 Microscopic Scale

The simulation in this scale is related to the study of materials. It deals with the deduction of some properties from the material structure. Among these properties is the behavior law of the used material. The Mesoscopic scale is found at a larger scale than the microscopic scale. At this scale, the chip formation is studied. Based on a thermo-mechanical description involving physical and metallurgical phenomena, but in a scale of the continuum mechanics, the simulation is often dealt by the finite element method [14].

4 Model System (Part-Tool-Machine)

To simulate the current phenomena in the Part-Tool-Machine system dynamics at the macroscale, dynamic model and geometric model are introduced.

4.1 Dynamic Model

The used dynamic model of Part-Tool-Machine system can be a simple spring-mass model or a complex finite element model. The finite element modeling allows a much finer and a more flexible spatial discretization. It allows also the obtaining of more realistic vibration modes and to address the case where the workpiece and/or the tool are deformable in the working area.

4.2 Geometric Model

The used geometric models can range from the simplest one, a series of points [15], to the most complex one, facetised surface description or representation using Z-buffer or Dexel [16].

4.2.1 Geometric Model of the Workpiece

Three families of geometric representations are distinguished:

- The boundary of the volume can be represented by a list of points projected on a plane.
- The boundary of the volume can be represented by surfaces (B-Rep model) [17].
- The geometric model can also be a solid model using Voxels [18], Dexels [19] (Fig. 25.5) or Triple-Nailboard (Fig. 25.6) [20].

4.2.2 Geometric Model of the Tool

The geometric modeling of the tool permits to generate the machined surface and to calculate the geometric properties used in the cutting law. Several studies have been

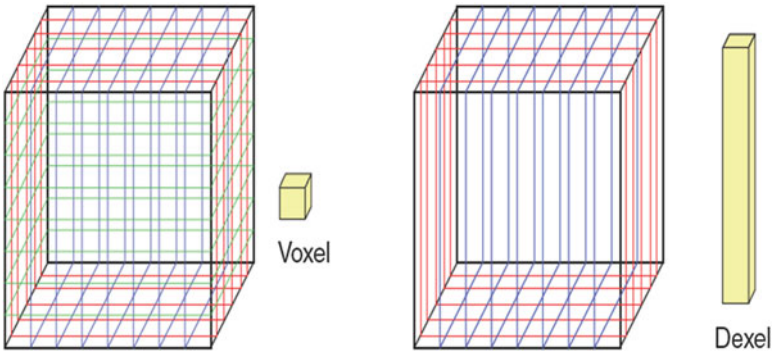


Fig. 25.5 Voxels and Dexels models [14]

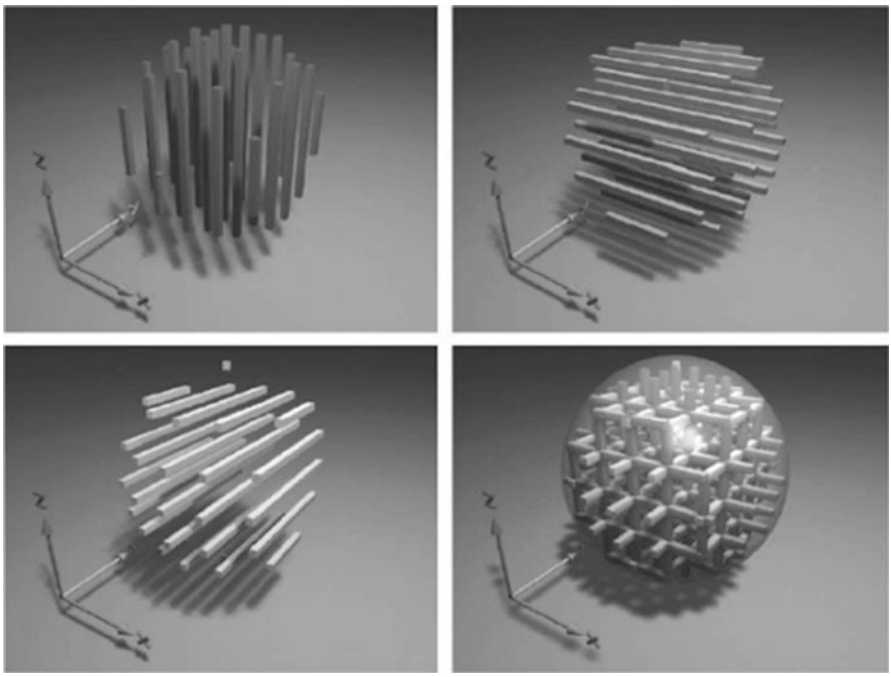


Fig. 25.6 Three-Nailboard model [20]

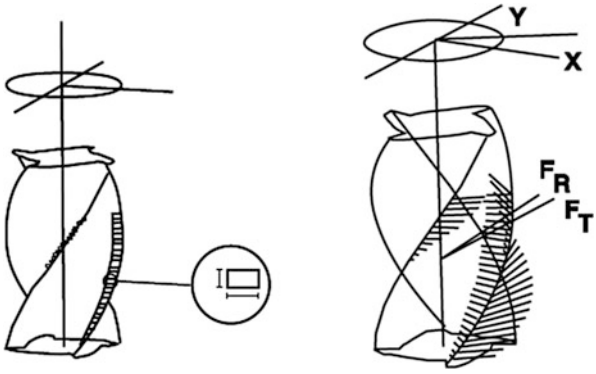


Fig. 25.7 Cutting tool decomposition [14]

performed to improve the modeling starting from the consideration of the complex tool geometry. This problematic is complex because modeling requires a law model for cutting forces and a geometric description of the tool. The response was found in the modifications developed in the calculus of the static cutting forces. Simulations dedicated to milling profile operations were inspired from cutting forces model dealing with complex tools geometries as a sum of basic tools (Fig. 25.7) [21].

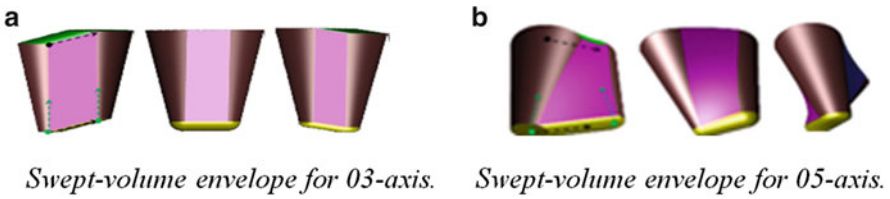


Fig. 25.8 Swept-volume envelope of a conical tool [4]

4.2.3 Geometric Model of Swept Volume

The modeling of the tool swept volume is based on the CSG representation (Constructive Solid Geometry) of the solid envelope of the tool path for 03-axis machining. The recent works are focused on the generation of the tool swept volume for 05-axis machining [6–8] where the difficulty is increased by the kinematics of the machine since the tool translates and rotates simultaneously (Fig. 25.8).

5 Geometric Simulation at the Macroscale

The literature shows that there are different ways for classifying the geometric representation in simulation at the macroscale. In Ref. [2], the used methods are classified as follows: wireframe-based, solid based, object space-based, image space-based and web-based simulation system. In this study, the web-based simulation is not considered.

5.1 Wireframe-Based

In wireframe-based simulation, the trajectory and the shape of the machined workpiece are displayed under shape of wire. This model has a simple and fast data structure. It has been applied extensively in the beginning of the machining simulation. This model remains applicable to parts of simple geometry.

5.2 Solid-Based

The solid-based simulation is a 3D volume representation. It is used for the geometric and the physical simulations. This model permits a very accurate geometric representation but expensive [2]. The two existing models for this case are CSG-based and B-Rep-based.

5.2.1 CSG-Based

It defines the constructive form of a 3D model using primitive volumes such as cylinders, spheres ... etc. Although, the Boolean operations and the consistency check are simple, visualization or data analysis may require a transformation into B-Rep model. The approximate cost of the simulation using CSG is $O(n^4)$ where n is the number of tool movements [22]. So, the simulation for machining freeform surfaces becomes intractable [23].

5.2.2 B-Rep-Based

This model is suitable for viewing. Unlike the CSG model, the B-Rep model explicitly defines the volume by a list of surfaces, edges and vertices. The computational cost is high in terms of time, storage of data and complexity. For n tool movements, the cost of the simulation is estimated to $O(n^{1.5})$ [24].

5.3 Object Space-Base

In a machining simulation as object based space, the parts are represented by a set of discrete points with vectors or surfaces with vectors or some volume elements. There are three main decomposition methods for machining simulation patterns for object based space model: Z-map method, vector method and Octree-based method.

5.3.1 Z-Map Method

It consists in decomposing the model of the part in several 3D vectors (Fig. 25.9). Each vector begins with the value of the height of the raw part. During the simulation process, 3D vectors heights are updated for each tool movement. In this case, the Boolean operations have only one dimension and therefore the simulation is very

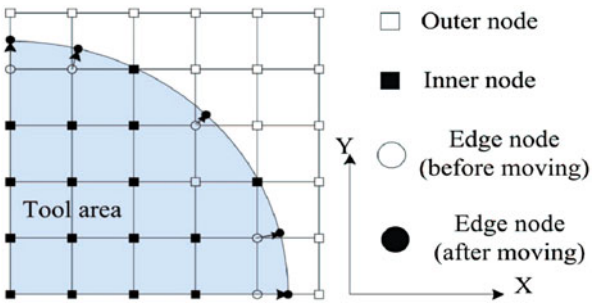


Fig. 25.9 Z-map method [30]

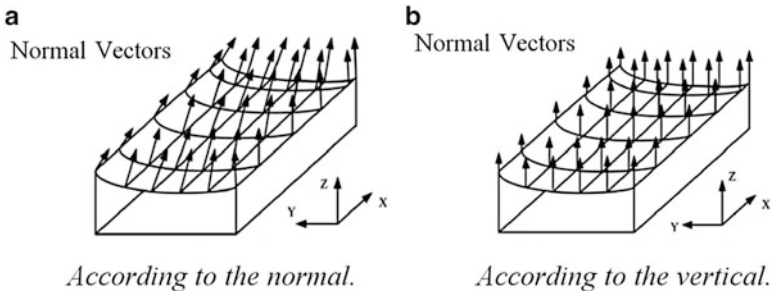


Fig. 25.10 Vectors orientations [2]

fast. In Ref. [25], this method was used in the collision detection algorithm for 03-axis CNC milling machine. This method is not usable for 04-axis and 05-axis machining since the tool axis is not vertical. Later, many researchers have used different approaches to improve the Z-map model [26–29].

5.3.2 Vector Method

This method involves discretization of the surface according to specific methods to obtain a set of points. For each point, a vector is associated with limits between the nominal surface and the raw part. Its vectors can be oriented in two ways (Fig. 25.10):

1. According to the surface normal (accurate): in this case, each vector is linearly independent from the other vectors.
2. According to the Z-axis of the tool (simplified): in this case, all vectors are parallel to Z-axis. it is adapted to 03-axis machining.

To simulate machining operations, the intersection of the vectors with the envelope of the tool swept volume must be calculated for each the tool displacement. The length and direction vectors are changed for each elementary tool movement. To detect non-machined areas, just check the direction and length of the vectors:

– Positive direction: no machined area;
– Negative direction: machining under the nominal surface;
– Length of vectors: if they are not in the machining tolerances, a correction is necessary.

5.3.3 Octree-Based Method

This method represents the workpiece in a tree structure (Fig. 25.11). Each node is recursively subdivided into eight disjoint child nodes until satisfying the required accuracy. This representation on a hierarchical Octree provides to the NC machining

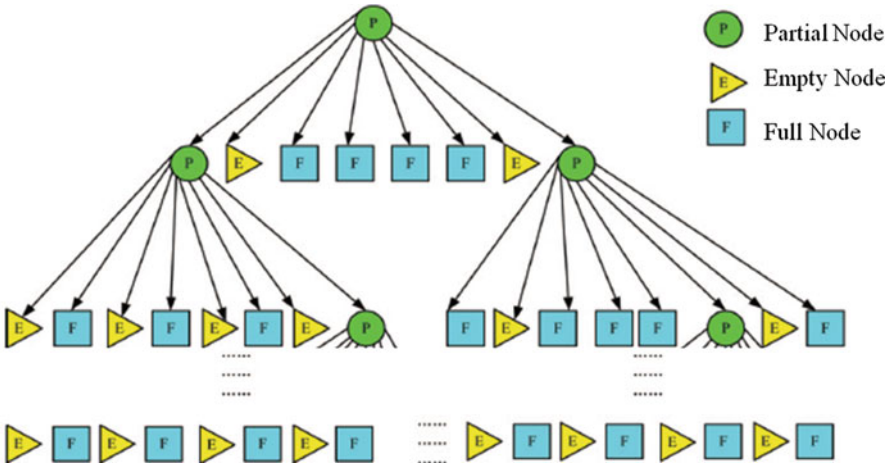


Fig. 25.11 Octree model [2]

simulation simplicity of Boolean operations calculation even when the local cutting area is complex. In Ref. [31], a machining simulation system is developed in which the part was represented by a traditional Octree for the creation and the modification of the model. Subsequently, it was represented in B–Rep model to animate the display, to verify and to optimize. The authors present the decomposition algorithm of the Octree model into three Quadtree models which store the geometry along the three main directions. Subsequently, this system was extended to the physical simulation for the prediction of the cutting forces based on the material removal rates [24]. For the optimization of the cutting parameters, Karunakaran et al. [12] found good results compared to the experimental. In Refs. [32, 33], geometric and physical simulation were integrated to predict the cutting forces. Kawashima et al. [34] developed an extended Octree called Graftree to represent more faithfully 3D objects in the geometric simulation (Fig. 25.12). For this case, each boundary cell has been described in the form of CSG with some restrictions. Kim et al. [35, 36] used the super-sampling method to enhance the Octree model.

5.4 Image Space-Based

In this model, parts are represented by the depths of the pixels (Dexels). It is an extension of the Z-buffer. The basis of the method is the projection of a grid (or a screen) in a given direction on a surface according to a selected view (Fig. 25.13). It fits well for 03-axis machining simulation with the projection direction is the tool axis (Z-axis). The construction of the surface is obtained by the intersection between a set of straight parallel lines to the Z-axis and the swept volume envelope. The bijective surfaces are the most suitable ones. For a set of surfaces, it is not always

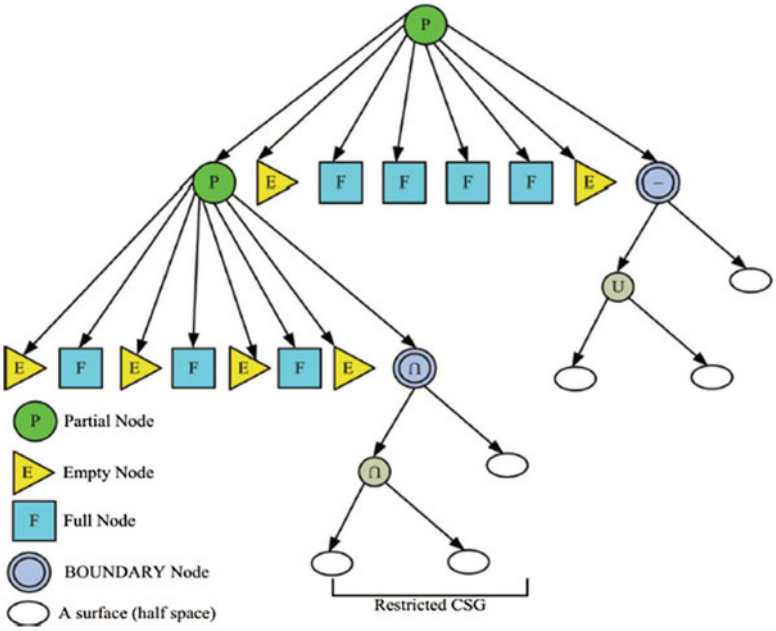


Fig. 25.12 Graftree model [34]

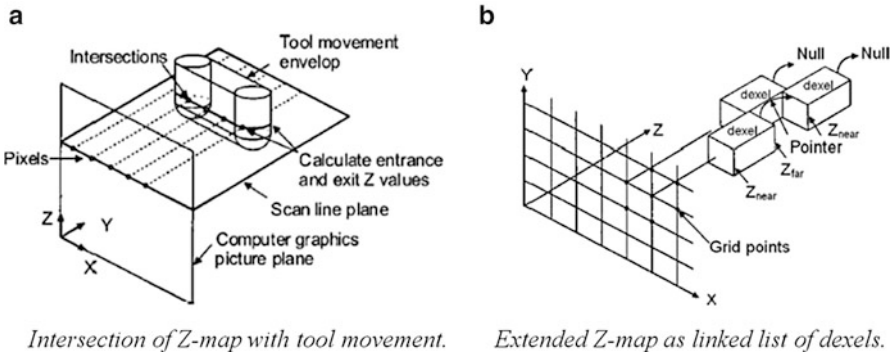


Fig. 25.13 Image space-based simulation [12]

possible to machine all surfaces. For each line, all intersections with all surfaces are calculated and the highest intersection belonging to the skin of the part is retained thereby allowing the machining the outer envelope of the part [2]. In Refs. [37, 38], the model in Dexels for milling with ball end mill tool is used with the integration of geometric and physical simulations to predict the cutting forces for 03-axis and 05-axis machining.

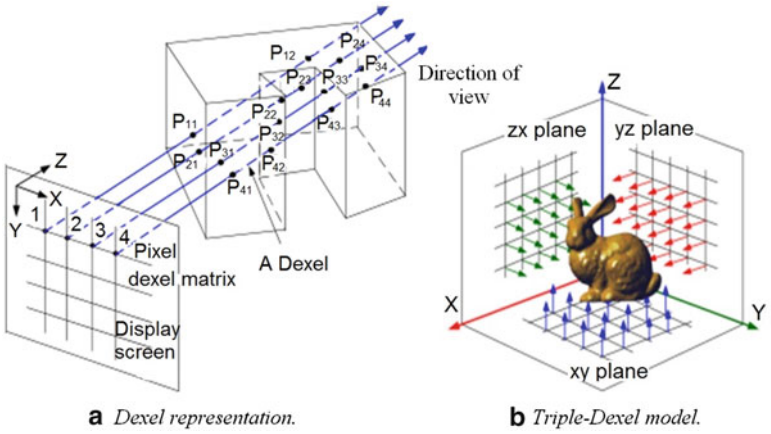


Fig. 25.14 (**a**) Single Dexel representation, (**b**) triple-Dexel model [39]

6 Geometric Model Workpiece Representation for Material Removal

The Dexel model introduced in Sect. 5.4 is limited in regions where the surfaces normal are nearly perpendicular to the ray direction (Fig. 25.14a) [39]. To overcome this problem, a Triple-Dexel modeling, which is an extension of single Dexel modeling, is constructed by casting rays in three orthogonal directions (X, Y and Z) to discretize the model (Fig. 25.14b). The authors of [40] show clearly that the generated surface from the Triple-Dexel data (Fig. 25.15b, d) is more accurate than the reconstructed surface from the single Dexel data presented in Fig. 25.15a, c when using the same ray resolution.

Triple-Dexel modeling is highly suitable for real-time graphics-based simulation applications such as numerical control (NC) machining verification and virtual sculpting. In [41], the authors present a novel surface reconstruction method from Triple-Dexel data by first converting the Triple-Dexel data into contours on three sets of orthogonal slices and then generating the solid's boundary surface in triangular facets from these contours. The developed method is faster than the Voxel-based method, and the reconstructed surface model is more accurate than the surface reconstructed from Voxel representation using the marching cube algorithm [41].

7 Conclusion and Future Work

The machining simulation is a technique used to check the tool path, to detect collisions, to predict the surface roughness, to predict cutting forces for optimizing the cutting parameters. These objectives require an accurate modeling of the

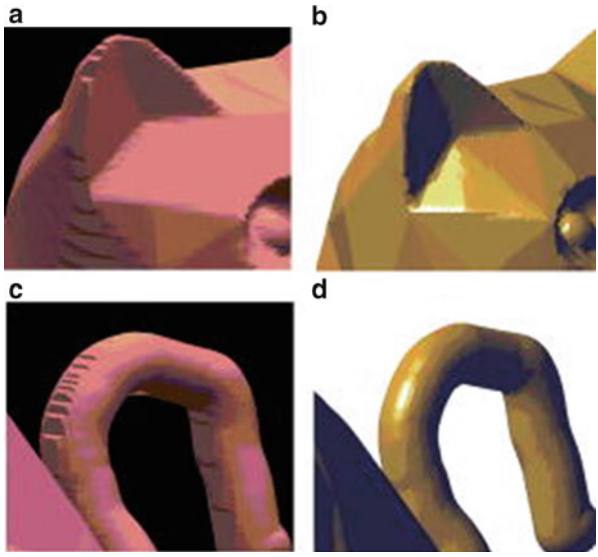


Fig. 25.15 Comparison between single-Dexel data and triple-Dexel data [40]

machining environment. This synthesis was carried out to clarify and to separate the difficulties due to the complexity and the difficulty of this technique. In perspective, the selection and the adoption of one or more techniques for geometric and physical simulation model representation for material removal in 05-axis machining will be considered.

References

1. Bouhadja, K., Bey, M.: Classification of simulation methods in machining on multi-axis machines. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 992–997. WCE, London, 2–4 July 2014
2. Zhang, Y., Xu, X., Liu, Y.: Numerical control machining simulation: a comprehensive survey. Int. J. Comput. Integr. Manuf. **24**(7), 593–609 (2011)
3. Kalay, F.: Simulation numérique de l'usinage-Application à l'aluminium AU4G (A2024-T351). Technique De L'ingénieur, l'expertise technique de référence, bm7002-2 (2010)
4. Lee, S.W., Nestler, A.: Virtual workpiece: workpiece representation for material removal process. Manuf. Technol. **58**, 443–463 (2012)
5. Aras, E., Feng, H.-Y.: Vector model-based workpiece update in multi-axis milling by moving surface of revolution. Manuf. Technol. **52**(913–927), 2011 (2011)
6. Lee, S.W., Nestler, A.: Complete swept volume generation, part I: swept volume of a piecewise C1-continuous cutter at five-axis milling via Gauss map. Comput. Aided Des. **43**, 427–441 (2011)
7. Lee, S.W., Nestler, A.: Complete swept volume generation – part II: NC simulation of self-penetration via comprehensive analysis of envelope profiles. Comput. Aided Des. **43**, 442–456 (2011)
8. Mann, S., Bedi, S., Israeli, G., Zhou, X.: Machine models and tool motions for simulating five-axis machining. Comput. Aided Des. **42**, 231–237 (2010)

9. Ibarakia, S., Sawadaa, M., Matsubaraa, A., Matsushitab, T.: Machining tests to identify kinematic errors on five-axis machine tools. Precis. Eng. **34**, 387–398 (2010)
10. Kong, L.B., Cheung, C.F.: Prediction of surface generation in ultra-precision raster milling of optical freeform surfaces using an integrated kinematics error model. Adv. Eng. Softw. **45**, 124–136 (2012)
11. Hong, C., Ibaraki, S., Matsubara, A.: Influence of position-dependent geometric errors of rotary axes on a machining test of cone frustum by five-axis machine tools. Precis. Eng. **35**, 1–11 (2011)
12. Karunakaran, K.P., Shringi, R., Ramamurthi, D., Hariharan, C.: Octree-based NC simulation system for optimization of feed rate in milling using instantaneous force model. Manuf. Technol. **46**, 465–490 (2010)
13. Ahmad, R., Tichadou, S., Hascoet, J.Y.: 3D safe and intelligent trajectory generation for multi-axis machine tools using machine vision. Int. J. Comput. Integr. Manuf. **26**(4), 365–385 (2013)
14. Assouline, S.C.: Simulation numérique de l'usinage à l'échelle macroscopique: prise en compte d'une pièce déformable. Thèse doctorat, Ecole Nationale Supérieure d'Arts et Métiers – CER de Paris (2005)
15. Montgomery, D., Altintas, Y.: Mechanism of cutting force and surface generation in dynamic milling. ASME Trans. J. Eng. Ind. **113**, 160–168 (1991)
16. Mizugaki, Y., Hao, M., Kikkawa, K.: Geometric generating mechanism of machined surface by ball-nosed end milling. Ann. CIRP **50**, 69–72 (2001)
17. Marty, A.: Simulation numérique de l'usinage par outil coupant à l'échelle Macroscopique: contribution à la définition géométrique de la surface usinée. PhD thesis, Ecole Nationale Supérieure d'Arts et Métiers – CER de Paris (2003)
18. Ratchev, S., Liu, S., Huang, W., Becker, A.A.: Milling error prediction and compensation in machining of low-rigidity parts. Int. J. Mach. Tool Manuf. **44**, 1629–1641 (2004)
19. Hook, T.V.: Real-time shaded NC milling display. SIGGRAPH'86 **20**, 15–35 (1986)
20. Weinert, K., Zabel, A.: Simulation based tool wear prediction in milling of sculptured surfaces. Ann. CIRP **53**, 217–223 (2004)
21. DeVor, R.E., Kline, W.A., Zdeblick, W.J.: A mechanistic model for the force system in end milling with application to machining airframe. In: 8th North American Manufacturing Research Conference, pp. 297–303 Dearborn, USA, (1980)
22. Voelcker, H.B, Hunt, W.A.: The role of solid modeling in machining process modeling and NC verification. SAE Technical Paper 810195 (1981)
23. Jerard, R.B., Drysdale, R.L., Hauck, K.E., Schaudt, B., Magewick, J.: Methods for detecting errors in numerically controlled machining of sculptured surfaces. IEEE Comput. Gr. Appl. **9**(1), 26–39 (1989)
24. Karunakaran, K.P., Shringi, R.: A solid modelbased off-line adaptive controller for feed rate scheduling for milling process. J. Mater. Process. Technol. **204**(1–3), 384–396 (2008)
25. Anderson, R.O.: Detecting and eliminating collision in NC machining. Comput. Aided Des. **10**(4), 231–237 (1978)
26. Hsu, P.L., Yang, W.T.: Real-time 3D simulation of 3-axis milling using isometric projection. Comput. Aided Des. **25**(4), 215–224 (1993)
27. Lee, S.K., Ko, S.L.: Development of simulation system for machining process using enhanced Z-map model. J. Mater. Process. Technol. **130–131**, 608–617 (2002)
28. Kang, M.J., Lee, S.K., Ko, S.L.: Optimization of cutting conditions using enhanced z map model. CIRP Ann. Manuf. Technol. **51**(1), 429–432 (2002)
29. Lee, S.H., Lee, K.S.: Local mesh decimation for view-Independent three-axis NC milling simulation. Int. J. Adv. Manuf. Technol. **19**(8), 579–586 (2002)
30. Yun, W.S., Ko, J.H., Lee, H.U., Cho, D.W., Ehmann, K.F.: Development of a virtual machining system, part 3: cutting process simulation in transient cuts. Int. J. Mach. Tool Manuf. **42**(15), 1617–1626 (2002)
31. Karunakaran, K.P., Shringi, R.: Octree-to-BRep conversion for volumetric NC simulation. Int. J. Adv. Manuf. Technol. **32**(1–2), 116–131 (2007)

32. Ling, H.J., et al.: Method of determining integration limit for cutting force model of flat end milling process. J. Tool Technol. **38**(4), 11–13 (2004)
33. Li, J.G., Yao, Y.X., Xia, P.J., Liu, C.Q., Wu, C.G.: Extended Octree for cutting force prediction. Int. J. Adv. Manuf. Technol. **39**(9–10), 866–873 (2008)
34. Kawashima, Y., Kawashima, Y., Itoh, K., Ishida, T., Nonaka, S.: A flexible quantitative method for NC machining verification using a space-division based solid model. Vis. Comput. **7**(2–3), 149–157 (1991)
35. Kim, Y.H., Ko, S.L.: Development of a machining simulation system using the Octree algorithm. Lect. Notes Comput. Sci. **3482**(III), 1089–1098 (2005)
36. Kim, Y.H., Ko, S.L.: Improvement of cutting simulation using the Octree method. Int. J. Adv. Manuf. Technol. **28**(11–12), 1152–1160 (2008)
37. Fussell, B.K., Jerard, R.B., Hemmett, J.G.: Robust feedrate selection for 3-axis NC machining using discrete models. J. Manuf. Sci. Eng. **123**(2), 214–224 (2001)
38. Fussell, B.K., Jerard, R.B., Hemmett, J.G.: Modeling of cutting geometry and forces for 5-axis sculptured surface machining. Comput. Aided Des. **35**(4), 333–346 (2003)
39. Peng, X., Zhang, W.: A virtual sculpting system based on triple Dexel models with haptics. Comput. Aided Des. Appl. 01/2009 **6**(5), 645–659 (2009). doi:10.3722/cadaps
40. Zhang, W.: Virtual Prototyping with Surface Reconstruction and Freeform Geometric Modeling Using Level-Set Method. Missouri University of Science and Technology, pp. 98–99 (2008)
41. Zhang, W., Leu, M.C.: Surface reconstruction using Dexel data from three sets of orthogonal rays. Faculty Research & Creative Works. Paper 3678. http://mst.bepress.com/faculty_work/367 (2009)

Chapter 26
An Approach to Modelate Human Error on Biodiesel Plants: Analysis of Active Failures and Latent Conditions

R.D. Calvo Olivares, S.S. Rivera, and J.E. Núñez Mc Leod

Abstract During the last decade biodiesel production facilities have had a fast growth all over the world. Production levels and installed capacity have increased continuously to respond the demands of renewable oils. This growth has been accompanied by increasing accident rates. This fact raises the necessity of understand accidental causes in order to eliminate or diminish them. The present study applies the Reason' Swiss Cheese model of Human Error to a series of accidents that have taken place at biodiesel facilities in the period 2003 to January 2014. It allows identifying the unsafe acts and latent conditions that have conducted to accidents, and implementing tools to manage them.

Keywords Accidents • Active failures • Biodiesel • Human error • Human reliability • Incidents • Model • Latent conditions • Risks

1 Introduction

The world economic growth and the consequent increase of energy demand, the rise of petroleum price due to the geopolitical instability and armed conflicts, in addition to the exhaustion of oil reserves and the environmental problems caused by fossil fuels, have encouraged the research, promotion, development and utilization of alternative and renewable energy sources. Among these, particular importance has been given to biodiesel production due to its environmental advantages and use with little or no adequacy of engines.

R.D. Calvo Olivares (✉) • J.E. Núñez Mc Leod
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)/CEDIAC Institute, Eng. Faculty, Cuyo National University, Mendoza, Argentina
e-mail: rcalvo@cediac.uncu.edu.ar; jnmcleod@cediac.uncu.edu.ar

S.S. Rivera
CEDIAC Institute, Eng. Faculty, Cuyo National University, Mendoza, Argentina
e-mail: srivera@cediac.uncu.edu.ar

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_26

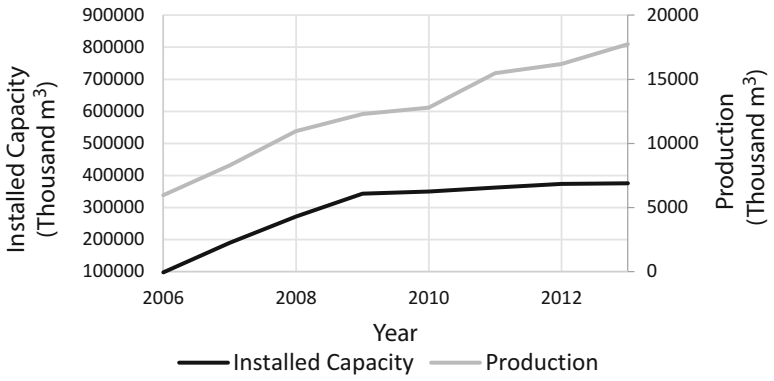


Fig. 26.1 Biodiesel installed capacity and production in thousand of m^3 versus year. Installed capacity is represented on left axis, and Production, on right axis

In the last decades, biodiesel installed capacity and production has increased significantly to respond to global demands according to Fig. 26.1. Data correspond to United States, Europe and Argentina, and were obtained from U.S. Energy Information Administration [1] and from Biodiesel magazine [2].

The implementation of public policies and the establishment of legal frameworks that limit carbon dioxide emissions, regulate the percentage of biodiesel to be blended with gasoil and, in some cases (e.g. Argentina), promote regional economies, small and medium scale enterprises, and local farmers, have been local factors that have also influenced in the increase of production.

The growth of biodiesel production has been accompanied by an increase of accident rates. The belief that the process to obtain biodiesel is simple added to the lack of expert operators, safe technologies and the scarcity of accident register are determinant factors for incident and accident occurrence [3–5].

Poor or inexistent records of accidents and incidents limit the understanding of the process and human- machine interfaces, the generation of lessons learned and the exchange of information with other companies to avoid accidents occurrence or diminish their consequences. During the last decade, data reveals that the same accident has occurred in different years and companies, and the same consequences (e.g. total loss of the plant) have affected different biodiesel production facilities [5]. This situation becomes more critical when safety is considered like a secondary cost that should be limited to the minimum when it cannot be eliminated (e.g. in Argentina). Although the process is simple, it is a chemical process and the facility is subject to the same hazards as any other chemical process facility [6]. Therefore, the production process requires reliable storing, transport and control mechanisms and infrastructure.

Accidents that are more relevant have been presented and some of them analyzed in detail by several authors [3, 7–10]. Recently, accidents and incidents occurring at biodiesel facilities have been gathered and organized in a database [5]. Availability

of information in a complete and unified way is crucial. However, the development of a framework or model is required for a better understanding of accident causation on these kind of facilities.

Transesterification of triglycerides of virgin oils or renewables sources (refined/edible oils) is the method most widely used to obtain biodiesel. It consists in the reaction of oil with an alcohol, such as methanol, in presence of a catalyst, generally a strong base such as sodium or potassium hydroxide. After the reaction, methyl esters and glycerol obtained are separated in two streams. The methyl esters stream goes to a neutralization step in which acid is added in order to neutralize any residual catalyst and to split any soap formed during the reaction. Then, methanol is removed. Next step is the washing of biodiesel to remove any remaining substance and finally, the drying step, to remove any water from the biodiesel. The glycerol stream is refined adding acid [11].

Main risks of this process are associated with handling and storing of methanol and involve fires, explosions and toxic spill or releases. The use of acid in the neutralization step of glycerol also involves risks of explosion, since the excess acid can cause overpressurization of the reactor [4, 5]. The rising plant capacity and complexity has also contributed to the increase of hazards due to the number and dimension of equipment required, and to the larger inventory of chemicals [8].

According to collected information, excluding those accidents under investigation or with no data about causes, the 20.5% of the cases was due to human error [5]. Human error is defined as *a failure to perform a given task that could cause damage to equipment and property or disruption of scheduled operations* [12]. Most common type of human errors founded at biodiesel facilities are chemical mixtures did in an improper ratio, valves left opened, welding operations on tanks that contain or contained vapors of inflammable substances, operations unattended and improperly disposal of oily rags. The inadequate performance at biodiesel facilities is related to a poor training or skill, a poor supervision, and scarcity of written equipment operating and maintenance procedures.

2 Human Error State of the Art

The study of human reliability and human error can be traced back to the late 1950s when the human aspect was included in a system reliability analysis by H.L. Williams [13, 14]. Human error classification systems and human error databases were developed during the decades of 1960s and 1970s, although main applications were in the military domain and in some early nuclear power plant developments. The Three Mile Island accident in 1979 raised the importance of human error in many industries. Some years later, the emergent method of Human Reliability Assessment (HRA) becomes compulsory in nuclear power plants all over the world [15]. HRA is the common name for an assortment of methods and models used to predict the occurrence of human errors [16]. In the 1980s, there was an important development of HRA techniques, which allowed a deeper

understanding of human error, including causes, manifestation and consequences. During the 1990s decade, HRA techniques became more mature and aspects such as organizational influences were considered in human error models [15]. Human error approach has changed over the years. In the beginnings, the focus was upon individual erroneous actions. In the last years, the emphasis has been on managerial factors [16]. However, much remain to research and improve in the field of human error, since human behavioral science and engineering are not enough integrated [17].

3 Human Error Models: A Review

Many models and techniques have been developed for the assessment of human error, especially in Nuclear and Aviation areas. However, specific methods to analyze and classify errors in biodiesel plants have not been founded yet, although the increasing rate of accidents raises the need of doing it. The absence of a human error model or taxonomy makes difficult incident analysis to identify more probable causes of accidents, prevent their recurrence and generate useful data. An adapted model or taxonomy for biodiesel facilities will allow identifying those errors that may affect system's goals and the inherent risk associated to them [18].

Existent models range from simple lists of error types to error classification systems based on operator performance. Some of them are useful to classify errors, like task-based taxonomies [19–21] but do not offer insight into underlying causes of error since they do not involve internal cognitive task components and psychological mechanisms associated. Others models take into account the errors that can occur during the communication process [22–26]. Elements such as the medium, the circumstances, the noise on the signal, the role of feedback, the skills and attitudes of the source and receiver, between others, are considered as possible sources of error.

Information Processing models are mainly based on Broadbent's work in 1958 [27]. In later models, the person is seen as a serial information processor, who receives an input of data that then is recoded, stored, processed to take a decision and the corresponding output. Early models omitted certain elements such as processes of thinking, problem solving, and the role of past person's experience or expectations [15]. In 1992, Wickens [28] proposed the most broadly known model of information processing. It is useful for researching the components of human performance and examining performance limitations although some authors such as Reason [29] and Hollnagel [30] have criticized it. They claim that this model does not represent human behavior adequately since any behavior is not an organized sequence.

Symbolic Processing Models consider that knowledge structures and 'mental models' are critical determinants of human performance, involving the form of knowledge and the content of the representation [15]. The central role in these

type of models is played by the representation, activation and utilization of different types of knowledge. Authors such as Rasmussen [31–33], Pew [34], Reason [29], Hollnagel [30] and Norman [35] develop models under these concepts.

Among these kind of models, the Skill-, Rule- and Knowledge-Based Model (SRK), and Generic Error-modelling System are founded. SRK is an empirical and qualitative model, based on reports containing information about human errors committed in plants. The model considers that human behavior can be classified into three levels associated with decreasing levels of expertise with the task: skill-, rule- and knowledge-based behavior. The first level involves actions made by operators without conscious attention or control. Their performance is controlled by stored patterns of behavior, and the reaction to stimuli involves little conscious effort or consideration. Therefore, tasks can be performed in parallel. Rule-based behavior involves actions in familiar settings, made according to stored or readily available rules. The response to stimuli is not automatic; it requires an initial appreciation of the necessity to apply such rules consciously. Stored rules or procedures are acquired from previous problem-solving, planning behavior, or learned from other people. Parallel tasks are possible. Errors at this level are often associated with the application of the wrong rule or procedure. Knowledge-based behavior involves actions that require high attentional control such as problem- solving, goal- selection and planning. It is generally used in novel situations. In a specific event, the operator behaves according to the knowledge of the system that he or she has. The model allows identifying the error that occurred, how it occurred and its causes. It also provides a useful framework to classify errors. It was developed as a retrospective tool but it can also be applied to predict errors [15, 18]. The framework has had an important influence, particularly at industrial installations, due to its robustness, validity and its easy understanding. However, it has received some critics by some authors such as Dougherty [36] and Hollnagel [30].

In 1990, Reason [29] made a classification of unsafe actions according to the prior intention to act. Non- intentional actions appear as *slips* (actions not as planned) and *lapses* ('errors resulting as a failure in the execution and/or storage stage of an action sequence, independently of the adequacy of the plan that guided it to achieve the objective'). Intentional actions are classified as *mistakes* (rules or procedures are not applied, or are unsuitable or do not exist) and violations (willful disregard of rules and regulations). The same author [29, 37] developed another conceptual framework joining these concepts to the Rasmussen's SRK framework. It resulted in three basic error types: ***skill-based slips and lapses***, ***rule-based mistakes*** and ***knowledge-based mistakes***. Each one of these types of error contains a number of 'failure modes'. Skill-based errors are related to monitoring failures (inattention or over-attention). Rule-based mistakes involve misapplication of right rules or application of wrong rules. Finally, knowledge-based mistakes originate from a limited conscious workspace to display the problem space and incorrect/incomplete knowledge. These ideas were then used by Reason to develop other frameworks among which we can find 'The Swiss Cheese Model'.

3.1 Swiss Cheese Model of Human Error

Reason [29] did another important contribution developing a 'model' of how accidents could be seen not simply as a consequence of human error, but as the result of the relationship between 'unsafe acts' (active failures) by operators and latent conditions of the system. The model, known later as the 'Swiss Cheese model', resulted very pedagogical and many accident analysts started to apply it in different industrial domains such as nuclear, aeronautic and health care [38]. It was evolving through the years, from the first one that consisted in five 'planes' with benign and pathological aspects (top level decision makers, line management, preconditions, productive activities and defenses) to the last one in which there were three elements: hazards, defenses and losses. The 'planes' were replaced by Swiss cheese slices that represent all the many barriers, defenses, safeguards and controls that a system should have (Fig. 26.2). According to the resulting model [39, 40] accident causes are due to the successive penetration of the defenses by either active failures or latent conditions. Active failures are defined as *unsafe acts* made by operators and they involve errors and violations. Latent conditions, the 'pathogen agents' include contributory factors that may dormant in the system until they contribute to an accident (e.g. organizational culture, management decisions, procedure design, or deficient training). According to Shappell and Wiegmann [41] although this model revolutionized common views of accident causation, it is a theory in which the 'holes in the cheese' are not defined clearly. For an accident investigator it is crucial to know what these system failures or 'holes' are, in order to be detected and eliminated to avoid accident occurrence. In previous work [42] several latent failures on biodiesel small-scale production plants were identified and the associated problems or consequences for the system. It is the objective of the present work to describe, in a similar way that was done for aviation, what the holes in the slices are (active and latent failures) according to

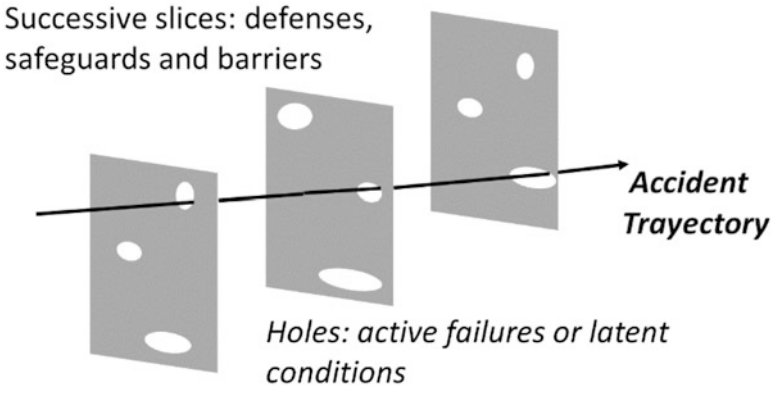


Fig. 26.2 Reason's Swiss Cheese model, 2000

a register of 86 accidents occurred at biodiesel plants between 2003 and January 2014. This will allow to develop a framework to understand the causes in a better way [43]. Data were obtained through the analysis and collection of information from different documental sources such as newspapers, magazines, technical reports and bibliography [5].

4 Modelling Errors on Biodiesel Plants

4.1 Unsafe Acts

Unsafe acts can be divided in errors and violations [29]. Reason defines error as an action not perform as intended and it can be reduced by improving the quality and availability of information processed by human. Violations are deviations from normal operating procedures, rules and regulations, and require motivational and organizational changes to be reduced. The same author states that these human actions are classified according to Rasmussen's SRK model.

Human errors founded in the 86 collected accidents were modeled using the classification previously explained. Accidents or incidents involving knowledge-based mistakes and violations were not founded [43]. Next, results are presented.

4.1.1 Skill-Based Errors

The study of information collected about accidents on biodiesel plants allows finding the following skill-based errors:

- *Attentional slips*: occur when the operator fail to monitor the progress of routine actions at some critical point [44]. An example for this type of error was the accident that occurred when a transfer of a processing- chemical mixture was left unattended, causing a small tank overflowed. The mixture ran across a driveway into a small inlet along a local river [45].
- *Memory lapses*: took place when the person forgets what earlier intended to do or when steps from a plan or action or procedure are omitted [44]. Examples for this type of error involves storage tanks and silos that were left opened, closing valves which were also left opened or process equipment that was left on.

4.1.2 Rule-Based Mistakes

An example for this type of error is the improper application of procedures. This occurred in two opportunities, when operators were mixing sulfuric acid with glycerin, during the neutralization step. They introduced 21 times more acid into the vessel than the process was designed for. It created an exothermic reaction that

Table 26.1 Examples of unsafe acts on biodiesel facilities

Skill-based errors
Inattention during chemical transfer process
Process equipment left on
Closing valve of storage silo left opened
Silo containing flammable material left opened to the air
Glycerin tank left opened
Rule-based errors
Failure to apply adequate procedure for doing a chemical mixture
Failure to dispose oily rags according to procedure
Failure to apply procedures for welding tasks on tanks and pipes

caused an explosion, followed by a spill in one of the cases and by a fire in the other [46, 47].

Other example is the improper disposal of oily rags that have created fires in presence of ignition sources. The plant usually disposed rags (used to clean biodiesel) and filters in a water tank, but that day operators put the rags in a drum not properly contained, they caught fire and damaged a filter press [48].

Table 26.1 summarizes the unsafe acts founded on biodiesel facilities.

4.2 Latent Conditions

It refers to the conditions that are present in the organization long before a specific incident occurs. Most of them are due to the organization itself, because of its design or of managerial decisions. The presence of latent conditions is universal in all organizations nevertheless of their incident and accident record [40]. Next, latent conditions found on biodiesel plants are presented according to the following classification: unsafe supervision, organizational influences and process design [43].

4.2.1 Unsafe Supervision

Information analyzed can be classified in two categories:

- **Inadequate supervision:** the role of any supervisor is not only to check the others work, but also to provide a guide, training opportunities and to motivate operators to do their work in an efficient and safe way. However, this is not always the case. Several accidents occurring at biodiesel plants were the result of a poor or inexistent supervision. As previously mentioned, in two cases [46, 47] operators mixed, during neutralization phase, sulfuric acid with glycerin in an improper ratio. This caused the explosion of the mix vessel as a result of

Table 26.2 Examples of unsafe supervision on biodiesel facilities

Inadequate supervision
Failed to provide adequate training
Failed to check an adequate qualification
Failed to instruct about facility's hazards
Failed to provide hot work permissions
Failed to provide minimal work conditions
Supervisory violations
Failed to enforce and/or fulfill normative and regulations

overpressure. If a supervisor had been checking the operation or he had provided the correct training, the accident could have been avoided.

Another situation that repeats are the tasks of welding on tanks that contain or have contained flammable substances, and on pipe parts that in general, are connected to equipment that contain a dangerous substance. In most analyzed cases, this kind of work was made by contractors during the scheduled maintenance of the biodiesel plant. Contractors are usually not employees of the dangerous establishment so the installations and their associated risks are not familiar to them [49]. Supervisor's function is to communicate working conditions, work environment hazards and to ensure that the contract worker is well trained and know procedures. Unfortunately, this not always occurs and at seven cases [50–56], contract operators began their welding tasks causing explosions and fires, with fatal consequences. The welding work generates sparks or an increment of temperature that are elemental keys to ignite flammable environments.

- **Supervisory violations:** occur when existing laws and regulations are willfully omitted by supervisors. There have been cases in which supervisors have not taking in care basic safety measures such as adequate ventilation, or have failing to follow local legislation (e.g. OSHA Acts, Georgia Environmental Acts, etc.).

Table 26.2 summarizes the *inadequate supervision* and *supervisory violations.*

4.2.2 Organizational Influences

Supervisory and operators practices are affected directly by decisions of upper-level management. Usually, if an investigator do not have a framework to consider and investigate organizational errors, they go unobserved [41]. However, they are an important part of the 'pathogens' that contributes to an accident. According to the accidents collected occurring on biodiesel plants, organizational influences are divided in two groups [43]:

- **Resource management:** it refers to how resources such as personnel, machinery, tools and equipment, and money are managed in order to achieve the company's goal of cost-effective operations and safety of the process. However, in times of financial resources austerity, one of the first costs to be cut is safety. For example,

until the year 2009 many biodiesel plants in the United States were based on tax credit, however it expired causing many plants to reduce installed capacity or shut down operations [57]. Cost-cutting may lead to the purchase of equipment (e.g. storing vessels) that is not optimal and adequate for the operation and the risks involved. Other practices include reducing of maintenance tasks not only for equipment but also for workplaces. The 30% of the studied accidents have been originated by equipment mechanical failures such as motor overheating, thermostat and safety valve failures, between others. These failures that finished in some cases with important consequences like the destruction of a building or equipment, could have being avoided with adequate maintenance.

- **Organizational Process:** involves decisions and rules that conduct tasks on the organization. It includes the use of operating procedures. It has been founded that the use of procedures is not common on biodiesel industry. This originated accidents, as mentioned previously, due to the lack of knowledge about the proper steps to follow in a chemical reaction (e.g. mixture of sulfuric acid and glycerin).

It is important to point out that when facilities apply not only working procedures but also safety plans and procedures, response of operators in case of accidents is faster. In addition, if the program includes communicating to local authorities (e.g. fire or police stations) about the facility hazards and substances stored, mitigation procedures are quicker and the magnitude of consequences diminishes.

Table 26.3 summarizes organizational influences.

4.2.3 Process Design

According to available technical reports of performed inspections at some biodiesel plants [58] organizations fail to have adequate information about the process technologies (e.g. safe upper and lower limits for temperature, pressures, flows and compositions), to perform periodic inspections on process equipment and machinery, and to correct deficiencies noted during equipment inspections.

Table 26.3 Examples of organizational influences

Resource management	Process design
Personnel	*Machinery and equipment*
Selection process	Use of inadequate engine for pumping methanol
Training	Use of storing tanks of inadequate material
Financial resources	Lack of control about ignition sources in explosive
Cost cutting	Environments
Organizational process	Lack of grounding of the tanks to avoid static loads
Procedures	Use of metallic tanks with unsafe locked
Instructions	
Documentation	
Safety programs	

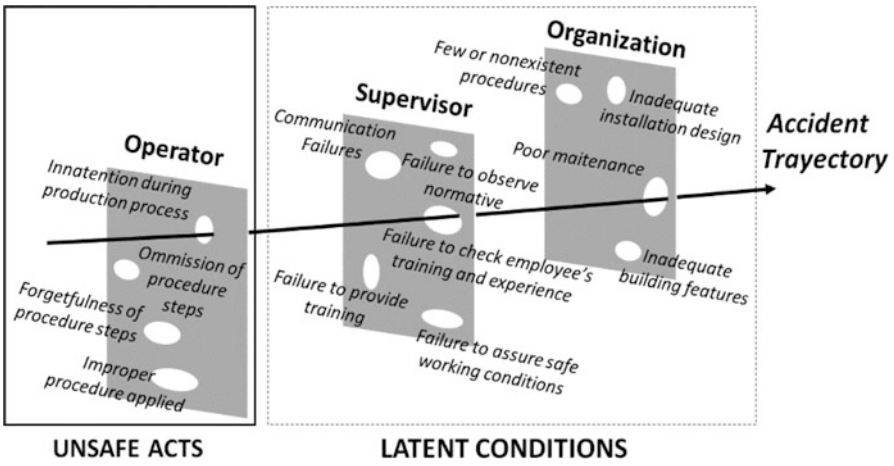


Fig. 26.3 Adaptation of Swiss cheese model to analyze accident causation on biodiesel facilities

Other common failure on process design is the lack of correct electrical classification, including proper grounding of storing tanks and equipment.

In other cases, choosing of inadequate material for biodiesel and raw material (e.g. sulfuric acid) storage vessels originated fires and spills of important magnitude [59, 60].

It has also been found the use of improper equipment such as for example, wrong electrical equipment in the control room [58], or the engine of a methanol pump that produced sparks, caused an injured person and the destruction of the plant [61].

Finally, fast spread of fire on most of the accidents shows the lack of fireproof materials, fire detection equipment (e.g. air handlers or smoke detectors) and suppressant systems (e.g. sprinklers).

Figure 26.3 summarizes the previous analysis. It is an adaptation of the Swiss cheese model to show accident causes on biodiesel plants.

5 Summary of Discussions

Analysing the exposed results, accidents on biodiesel plants have been the result of a combination of diverse factors among which the human error can be founded. So far, knowledge about causes of accidents that have occurred because of human error is limited due to the lack of complete information. This fact restricts the understanding of accidents since do not allow identifying other possible active and latent conditions existing in these complex systems. It has been founded that there is not information for the 37% of the accidents and for the 17%, causes are 'under investigation'. That means loss of information about causes in more than half of the accidents.

The present work is an attempt to contribute to general knowledge of accident and incident causation in biofuel industry. However, it is a limited study because of the scarcity of data. A great effort is necessary to begin reporting and recording useful and complete information about adverse events such as event sequence, mitigation measures, causes and consequences. These data will help accident researchers to identify other possible unsafe acts and latent conditions that have originated or can be the origin of accidents, or even cause the recurrence of one or more accidental events. As a result, a deeper understanding of the causes will be possible.

6 Conclusions and Recommendations

This is a first approach to model human error on biodiesel plants. Error models and taxonomies provide a useful framework to obtain information about the causes of an accident and understand its underlying error mechanisms. Once a type of error has been identified and associated to specific conditions or scenarios, it can be replicated through simulation. This makes possible to investigators obtaining a better knowledge about it and proposing the mechanisms to eliminate it. According to analyzed data, it is crucial to train personnel, providing safer working procedures, both for normal and for abnormal operating situations, and to ensure a good level of supervision on biodiesel facilities. When working with contractors, it is recommended that facilities implement the use of hot working permissions. Previously to do the maintenance tasks, contractors should be communicated about the installation hazards. It is also a safety practice that supervisors be sure about the experience and training of the contractor. It is also important to assure through equipment selection and installation, that components have the adequate reliability to reduce mechanical failures. Finally, the studying of the taxonomy allows identifying those accidents that are more likely to occur or that repeated with more frequency. This will help to focus on resources and obtain a more efficient solution in its mitigation or prevention.

References

1. International Energy Statistics: U.S. Information Administration. Available at www.eia.gov. Accessed Sept 2014
2. Argentina biodiesel: taxes, domestic mandates and production. Biodiesel magazine, 30 Apr 2014. Available at www.biodieselmagazine.com. Accessed Sept 2014
3. Rivera, S.S., Núñez Mc Leod, J.E.: Human error in biofuel plants. Lecture Notes in Engineering and Computer Science: Proceedings of Lecture Notes in Engineering and Computer Science: Proceedings of World Congress on Engineering 2008, pp. 1214–1219. London, 2–4 July 2008
4. Rivera, S.S., Núñez Mc Leod, J.E.: Recommendations generated about human reliability on biodiesel plants. Lecture Notes in Engineering and Computer Science: Proceedings of Lecture Notes in Engineering and Computer Science: Proceedings of World Congress on Engineering 2012, pp. 588–593. London, 4–6 July 2012

5. Calvo Olivares, R.D., Rivera, S.S., Núñez Mc Leod, J.E.: Database for accidents and incidents in the biodiesel industry. J. Loss Prev. Process Ind. **29**, 245–261 (2014)
6. Van Gerpen, J., Shanks, B., Pruszko, R., Clements, D., Knothe, G.: Biodiesel Production Technology Report, July 2004, NREL/SR-510-36244
7. Marlair, G., Rotureau, P., Breulet, H., Brohez, S.: Booming development of biofuels for transport: is fire safety of concern? Fire Mater. **33**(1), 1–19 (2009)
8. Salzano, E., Di Serio, M., Santacesaria, E.: Emerging risks in the biodiesel production by transesterification of virgin and renewable oils. Energy Fuel. **24**(11), 6103–6109 (2010)
9. Salzano, E., Di Serio, M., Santacesaria, E.: Emergency safety issues for biodiesel production plant, Florence, vol. 19 (2010)
10. Salzano, E., Di Serio, M., Santacesaria, E.: State-of-art and historical analysis for the safety of biodiesel production by trans-esterification of virgin and renewable oils, Prague (2010)
11. Gerpen, J.V.: Biodiesel processing and production. Fuel Process. Technol. **86**(10), 1097–1107 (2005)
12. Hagen, E.W. (ed.): Human reliability analysis. Nucl. Saf. **17**, 315–326 (1976)
13. Williams, H.L.: Reliability evaluation of the human component in man–machine system. Electrical Manufacturing, pp. 78–82, Apr 1958
14. Dhillon, B.S.: Human Reliability with Human Factors. Pergamon Press, New York (1986)
15. Isaac, A., Shorrock, S.T., Kennedy, R., Kirwan, B., Andersen, H., Bove, T.: Technical Review of Human Performance Models and Taxonomies of Human Error in ATM (HERA), European Organization for the Safety of Air Navigation, pp.6–7, 19–51 (2002)
16. Rivera, S.S., Núñez Mc Leod, J.E., Calvo Olivares, R.D.: IAENG Transactions on Engineering Sciences: Special Issue of the International MultiConference of Engineers and Computer Scientists 2013 and World Congress on Engineering 2013, Chapter: Causes of latent failures on biodiesel plants and accepted practices to diminish them. In: Sio-Iong Ao, Alan Hoi-Shou Chan, Hideki Katagiri, Li Xu (eds.), pp. 225–234. Taylor & Francis Group, London (2014)
17. Baziuk, P., et al.: Toward a unified human reliability model. Advances in Safety, Reliability and Risk Management, pp. 119–126. London (2012)
18. Kirwan, B.: A Guide To Practical Human Reliability Assessment, pp. 81–90. Taylor & Francis Group, London (1994)
19. Swain, A.D.: Modelling of response to nuclear power plant transients for probabilistic risk assessment. Proceedings of the 8th Congress of the International Ergonomics Association, Tokyo, Aug 1982
20. Swain, A., Guttmann, H.: A handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, USNRC, Washington (1983)
21. Spurgin, A.J., Lydell, B.D., Hannaman, G.W., Lukic, Y.: Human Reliability Assessment: A Systematic Approach. In Reliability '87, NEC, Birmingham (1987)
22. Lasswell, H.D.: The structure and function of communication in society. In: Bryson, L. (ed.) The Communication of Ideas. Harper and Row, New York (1948)
23. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
24. Dance, F.E.X.: Human Communication Theory. Holt, Rinehart and Winston, New York (1967)
25. Berlo, D.K.: The Process of Communication: An Introduction to Theory and Practice. Holt, Rinehart and Winston, New York (1960)
26. Andersch, E.G., Staats, L.C., Bostom, R.N.: Communication in Everyday Use. Holt, Rinehart and Winston, New York (1969)
27. Broadbent, D.E.: Perception and Communications. Pergamon, London (1958)
28. Wickens, C.: Engineering Psychology and Human Performance, 2nd edn. Harper-Collins, New York (1992)
29. Reason, J.: Human Error. Cambridge University Press, Cambridge, UK (1990)
30. Hollnagel, E.: Human Reliability Analysis: Context and Control. Academic, London (1993)
31. Rasmussen, J.: Human Errors. A Taxonomy for Describing Human Malfunction in Industrial Installations. Risø National Laboratory, DK-4000, Roskilde (1981)

32. Rasmussen, J.: Human errors: a taxonomy for describing human malfunction in industrial installations. J. Occup. Accid. **4**, 311–335 (1982)
33. Rasmussen, J.: Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. North Holland, New York (1986)
34. Pew, R., et al.: Evaluation of Proposed Control Room Improvements Through Analysis of Critical Operator Decisions. Electric Power Research Institute, Palo Alto (1982)
35. Norman, D.A.: Categorization of action slips. Psychol. Rev. **88**, 1–15 (1981)
36. Dougherty, E.M.: Human reliability analysis – where shouldn't thou turn? Reliab. Eng. Syst. Saf. **29**, 283–299 (1990)
37. Reason, J.T.: Generic error-modelling system: a cognitive framework for locating common human error forms. In: Rasmussen, J., Duncan, K.D., Leplat, J. (eds.) New Technology and Human Error. Wiley, Chichester (1987)
38. Reason, J., Hollnagel, E., Paries, J.: Revisiting the 'Swiss Cheese' model of accidents. Eurocontrol Experimental Center, EEC Note No. 13/06, Oct 2006
39. Reason, J.: Human error. West. J. Med. **172**(6), 393–396 (2000)
40. Reason, J.: Human error: models and management. BMJ **320**, 768–770 (2000)
41. Shappell, S.A., Wiegmann, D.A.: The Human Factors Analysis and Classification System-HFACS. FAA. US Department of Transportation (2000)
42. Rivera, S.S., Núñez Mc Leod, J.E., Calvo, D.R.: Latent Failures on Biodiesel Plants. International Association of Engineers; Lecture Notes in Engineering and Computer Science, London (2013)
43. Calvo Olivares, R.D., Rivera, S.S., Núñez Mc Leod, J.E.: Analysis of Active Failures and Latent Conditions on Biodiesel Production Facilities, Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 1013–1017. WCE, London, 2–4 July 2014
44. Leiden, K., Ronald Laughery, K., Keller, J., French, J., Warwick, W.: A review of Human Performance Models for the Prediction of Human Error. Ann Harbor, Sep 2001
45. Department of Ecology, State of Washington, USA: Ecology settles penalty with Seattle Biodiesel. Available: www.ecy.wa.gov (2013). Accessed May 2013
46. The Seattle Times: Biodiesel fuel plant between Aberdeen and Hoquiam explodes. Available: http://seattletimes.com (2013). Accessed May 2013
47. Chi Town Daily News: OSHA launches investigation at plant that exploded. Available: www.chitowndailynews.org (2013). Accessed May 2013
48. The Press- Enterprise: TEMECULA: fire damages biodiesel business. Available: www.pe.com (2013). Accessed May 2013
49. European Commission: Chemical Accident Prevention & Preparedness- Major accidents involving contractors. Number 2, Dec 2012
50. Pioneiro: Tanque explode e mata uma pessoa na Oleoplan em Veranópolis. Available: http://pioneiro.clicrbs.com.br/rs (2013). Accessed May 2013
51. Herald Sun: Man critical after factory explosion. Available: www.heraldsun.com.au (2013). Accessed May 2013
52. Folha de S. Paulo: Explosão em fábrica de biodiesel deixa três mortos em Formosa. Available: www.folha.uol.com.br. Accessed May 2013
53. WESM 91.3 FM: 1 dead after explosion at Princess Anne biodiesel plant. Available: http://wesm913.org/news (2013). Accessed May 2013
54. Canada.com: One dead as blast rips biodiesel plant in High River. Available: www.canada.com (2013). Accessed May 2013
55. News 12: EPD finds environmental threat at biodiesel plant. Available: www.wrdw.com (2013). Accessed May 2013
56. Casper Star-Tribune: Explosion, fire at biodiesel plant kills one. Available: http://trib.com/ (2013). Accessed May 2013
57. Biodiesel Magazine: Chemical Safety Matters. Available: www.biodieselmagazine.com (2010). Accessed 7 Nov 2010
58. Occupational Safety and Health Administration Website. Available: www.osha.gov

59. The Star-Ledger: Sulfuric acid explodes in Newark plant, injures employee. Available: www.nj.com (2013). Accessed May 2013
60. WFTV Channel 9 (abc): Explosions Rock Biodiesel Plant in Osceola County. Available: www.wftv.com (2013). Accessed May 2013
61. Newschannel 5 (WTVF-TV): Crews Battle Large Chemical Fire in Coffee County. Available: www.newschannel5.com (2013). Accessed May 2013

Chapter 27
Exploring Pareto Frontiers in the Response Surface Methodology

Nuno Ricardo Costa and João Alves Lourenço

Abstract Multiple response optimization problems have many optimal solutions that impact differently on process or product. Some of these solutions lead to operation conditions more hazardous, more costly or more difficult to implement and control. Therefore, it is useful for the decision-maker to use methods capable of capturing solutions evenly distributed along the Pareto frontier. Three examples were used to evaluate the ability of three methods built on different approaches for depicting the Pareto frontier. Limitations of a desirability-based method are illustrated whereas the consistent performance of an easy-to-use global criterion gives confidence to use it in real-life problems developed under the Response Surface Methodology framework, as alternative to the sophisticated physical programming method.

Keywords Bias • Desirability • Multiresponse • Optimization • Pareto • Programming

1 Introduction

High performance, low through-life cost and manufacturability are examples of process and product characteristics that must be satisfied simultaneously. However, finding compromise solutions for multiple objective or response problems is not a trivial issue. Many of the generated solutions may not be optimal and among the optimal (Pareto) solutions one can expect to have solutions that impact differently on process or product. In practice, some Pareto solutions will lead to operation

N.R. Costa (✉)
Instituto Politécnico de Setúbal -Escola Superior de Tecnologia de Setúbal, Campus do IPS, Estefanilha, Setúbal 2910-761, Portugal

UNIDEMI/DEMI, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica 2829-516, Portugal
e-mail: nuno.costa@estsetubal.ips.pt

J.A. Lourenço
Instituto Politécnico de Setúbal -Escola Superior de Tecnologia de Setúbal, Campus do IPS, Estefanilha, Setúbal 2910-761, Portugal

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_27

conditions more hazardous, more costly or more difficult to implement and control. Therefore, it is of paramount importance for the decision-maker (DM) to use an optimization method capable of providing the maximum information on the Pareto surface (or Pareto frontier). If the method fails to capture a representative set of Pareto solutions, solutions where any improvement in one response cannot be done without degrading the value of, at least, another response, the DM may have denied the possibility of finding a more favorable compromise solution. In addition, the methods ability to depict Pareto frontiers has been rarely evaluated in the Response Surface Methodology (RSM) framework, which difficult the practitioners task of choosing an effective method to solve multiresponse optimization problems.

This chapter evaluates the working abilities of three optimization methods, two easy-to-use and one mathematically more sophisticated, and contributes to help practitioners in making more informed decisions when they need to select an optimization method for solving real-life problems developed under the RSM framework. For details on RSM the reader is referred to [1].

The remainder of the chapter is organized as follows: Next section provides a literature overview; then optimization methods are reviewed; Sects. 4 and 5 include the examples and results discussion, respectively; Conclusions are presented in Sect. 6.

2 Literature Overview

Most real-life problems involve multiple and conflicting objectives so their analysis has been a widely research subject. As a result, a great quantity and variety of methods to generate solutions for multiresponse optimization (MO) problems are available in the literature. The two most popular criteria in the RSM framework are built on desirability and loss function approaches.

An extensive review on desirability-based methods is presented in [2]. From the popular desirability method introduced in [3], later modified in [4], to less known proposal presented in [5], 12 methods were reviewed. In [6] it is provided an extensive review on loss function-based methods and summarizes the relative merits of 12 multivariate loss function-based and desirability-based methods. In [7] the strengths of two popular loss function-based methods [8, 9] are combined. Other contributions introduced in the last decade are the mean squared error [10], weighted signal-to-noise ratio [11], PCA-based grey relational analysis [12], weighted principal component [13], capability index [14], patient rule induction [15], design envelopment analysis [16], compromise programming [17], goal programming [18], physical programming [19, 20], bayesian probability [21], weighted Tchebycheff formulations [22], modified ε-constraint method [23, 24]. Relationships and differences among several criteria are highlighted in [25]. This list is not exhaustive. Many other researchers have contributed to the growing wealth of knowledge in the field. However, little attention is paid to methods ability to depict Pareto surfaces. Exceptions are the works reported in [26–28].

3 Optimization Criteria

In the next subsections three methods built on different approaches are reviewed, namely a desirability-based method (DAM method), a global criterion-based method (GC method), and a physical programing-based method (PP method).

3.1 DAM Method

Desirability-based methods are easy to understand, and flexible for incorporating the decision-maker's preferences (priority to responses). The most popular of them, the so-called Derringer and Suich's method [3], or modifications of it [4], is available in many data analysis software packages. However, to use this method the analyst needs to assign values to four shape parameters ("weights"). This is not a simple task and impacts on the method performance. An alternative desirability-based method that requires minimum information from the user was proposed in [29]. The proposed method is easy to understand and implement in the readily available Microsoft Excel®-Solver tool and, in addition, requires less cognitive effort from the analyst or decision-maker. The user only has to assign values to one shape parameter, which is a relevant advantage over the extensively used Derringer and Suich's method.

In [29] the individual desirability functions are defined as

$$d = \frac{2\widehat{y} - (U + L)}{U - L} + 1 = \frac{2\widehat{y}}{U - L} + \frac{-2L}{U - L} = m\widehat{y} + c \tag{27.1}$$

where $\hat{y}$ represents the response's model, U and L are the upper and lower response bounds, respectively. The global desirability (composite) function is as follows

$$D = \left(\sum\nolimits_{i=1}^{p} \omega_i \left|d_i - d_i\left(\theta_i\right)\right|\right) / p \tag{27.2}$$

where $d_i(\theta_i)$ is the value of the i-th individual desirability function at the target value θ_i, ω_i is the weight (degree of importance or priority) assigned to i-th response, p is the number of responses, and $\sum_{i=1}^{p} \omega_i = 1$. The aim is to minimize D.

These authors illustrated the DAM method only for Nominal-The-Best (NTB- the value of the estimated response is expected to achieve a particular target value) response type. To evaluate the DAM method performance in problems where other responses type exist, namely the Larger-The-Better (LTB- the value of the estimated response is expected to be larger than a lower bound L; $\widehat{y} > L$) and Smaller-The-Better (STB – the value of the estimated response is expected to be smaller than an upper bound U; $\widehat{y} < U$), three examples with different responses type are considered here. Under the assumption that it is possible to establish the specification limits U and L for LTB-type and STB-type responses, as instance,

based on product and process knowledge or practical experience, $d_i(U_i)$ and $d_i(L_i)$ are used in Eq. 27.2 instead of $d_i(\theta_i)$. An alternative to U and L values is to use the maximum and minimum values yielded by model fitted to response.

3.2 GC Method

In [30] it is proposed the minimization of an arithmetic function defined as

$$\sum_{i=1}^{n} \left(\frac{|\widehat{y}_i - \theta_i|}{U_i - L_i} \right)^{e_i} \tag{27.3}$$

where e_i are user-specified parameters (shape or power factors, $e_i > 0$). To accommodate a STB-type response, the target value is set equal to L ($\theta = \mathrm{L}$); for a LTB-type response the target value is set equal to U ($\theta = \mathrm{U}$).

3.3 PP Method

Physical programing (PP) is an elegant approach to multiresponse optimization introduced in [19] that consists of converting MO problem into single-response problem by using class functions which capture the user's preferences in a mathematically consistent manner [31]. A four step procedure to implement is summarized as follows: (1) classify each response into (soft) class 1S (when response is of STB-type), 2S (when response is of LTB-type), 3S (when response is of NTB-type), or 4S (when response is of between-is-better type (BIB-type): the estimated response value is expected to take any value inside its specification limits); (2) define the ranges for each response; (3) form the class function; (4) form the aggregate function; (5) perform the computational optimization. In mathematical terms, the compromise solutions for a MO problem are achieved by solving the following optimization problem:

$$Min \log_{10} \left(\frac{1}{n} \sum_{i=1}^{n} g_i \left(f_i \left(\mathrm{X} \right) \right) \right) \tag{27.4}$$

subject to

$$\begin{aligned} f_i \left(\mathrm{X} \right) &\leq f_{i5} && \text{for class 1S} \\ f_i \left(\mathrm{X} \right) &\geq f_{i5} && \text{for class 2S} \\ f_{i5L} \leq f_i \left(\mathrm{X} \right) &\leq f_{i5R} && \text{for classes 3S and 4S} \end{aligned}$$

where $g_i(f_i(\mathrm{X}))$ are class functions that capture the designer's preferences, f_i is the i-th response, and $f_{i1}, \ldots, f_{i5}$ are physically meaningful values chosen by the decision-maker to specify the class function associated with the i-th response. Notice that aggregate function is formed based on the class functions and not on the responses. Class functions provide the means for the decision maker to express and specify ranges of differing levels of preferences for each response. Six ranges can be specified for classes 1S and 2S, ten ranges for class 3S, and eleven for class 4S. The class functions map responses into non-dimensional, strictly positive real numbers. Considering the class 1S as an illustration, the ranges in order of decreasing preference are as follows:

- Highly desirable ($f_i \leq f_{i1}$): an acceptable range over which the improvement that results from further reduction of the preference metric is desired, but is of minimal additional value.
- Desirable ($f_{i1} \leq f_i \leq f_{i2}$).
- Tolerable ($f_{i2} \leq f_i \leq f_{i3}$).
- Undesirable ($f_{i3} \leq f_i \leq f_{i4}$): a range that, while acceptable, is undesirable.
- Highly undesirable ($f_{i4} \leq f_i \leq f_{i5}$): a range that, while still acceptable, is highly undesirable.
- Unacceptable range ($f_i \geq f_{i5}$): the range of values that the objective function may not take.

Once the range parameters have been defined, the class function is constructed. Considering the case of class 1S, the class function $g_i(f_i(\mathrm{X}))$ are defined as

$$g_i^k = T_0\left(\xi_i^k\right) g_{i(k-1)} + T_1\left(\xi_i^k\right) g_{ik} + \overline{T}_0\left(\xi_i^k, \lambda_i^k\right) s_{i(k-1)} + \overline{T}_1\left(\xi_i^k, \lambda_i^k\right) s_{ik} \quad (27.5)$$

where k is the region ($k = 2,\ 3,\ 4,\ 5$), $\xi_i^k = \left(f_i - f_{i(k-1)}\right) / \left(f_{ik} - f_{i(k-1)}\right)$ with $0 \leq \xi_i^k \leq 1$, $\lambda_i^k = f_{ik} - f_{i(k-1)}$, and

$$T_0(\xi) = \frac{1}{2}\xi^4 - \frac{1}{2}(\xi - 1)^4 - 2\xi + \frac{3}{2} \quad (27.6)$$

$$T_1(\xi) = -\frac{1}{2}\xi^4 + \frac{1}{2}(\xi - 1)^4 + 2\xi - \frac{1}{2} \quad (27.7)$$

$$\overline{T}_0(\xi, \lambda) = \lambda\left[\frac{1}{8}\xi^4 - \frac{3}{8}(\xi - 1)^4 - \frac{1}{2}\xi + \frac{3}{8}\right] \quad (27.8)$$

$$\overline{T}_1(\xi, \lambda) = \lambda\left[\frac{3}{8}\xi^4 - \frac{1}{8}(\xi - 1)^4 - \frac{1}{2}\xi + \frac{1}{8}\right] \quad (27.9)$$

$$s_{ik} = \left.\frac{\partial g_i^k}{\partial f_i^k}\right|_{f_i^k = f_{ik}} \quad (27.10)$$

As an example, for the highly desirable region, the class function expression (g_i) can be represented by an exponential function as follows:

$$g_i = g_{i1} \exp\left[\frac{s_i^1}{g_{i1}}(f_i - f_{i1})\right] \tag{27.11}$$

The class 2S is the mirror image of class 1S, and the class 3S is a composition of classes 1S and 2S [19]. For details on the qualitative meaning and mathematical nature of class functions the reader is referred to [19, 31–34] and references therein. For alternative representation of class functions, see [35–38], as examples.

4 Examples

To better understand the working abilities of the DAM, GC, and PP methods, namely their ability to depict Pareto frontiers, three examples were selected from the literature. These examples only deal with the optimization of two responses so as to graphically represent the Pareto frontier. The first and second examples have appeared repeatedly in the literature [2, 27], and the third one was reported in [39]. The objective of this third example, like of the first one, is to optimize one mean response and one standard deviation response. To optimize two mean responses is the objective of the second example.

Example 1 The metal removal rate for a cutting machine was evaluated using a central composite design with three replicates. Cutting speed (x_1), cutting depth (x_2), and cutting feed (x_3) are the control variables. The models fitted to mean ($\widehat{\mu}$) and standard deviation ($\widehat{\sigma}$) responses are as follows [27]:

$$\begin{aligned}\widehat{\mu} &= 79.89 + 1.25x_1 - 0.15x_2 + 0.08x_3 - 1.47x_1x_2 + 0.75x_1x_3 + 0.87x_2x_3 \\ &\quad -2.07x_1^2 - 0.22x_2^2 - 0.49x_3^2 \\ \widehat{\sigma} &= 1.79 + 0.11x_1 + 0.35x_2 - 0.15x_3 + 0.64x_1x_2 - 0.18x_1x_3 + 0.97x_2x_3 \\ &\quad -0.26x_1^2 - 0.09x_2^2 + 0.04x_3^2\end{aligned}$$

The mean response is of NTB-type ($69 < \widehat{\mu} < 83$) with target value equal to 71 and $\widehat{\sigma}$ is of STB-type ($\widehat{\sigma} < 1.95$) with target value equal to 0. The constraints for the input variables are $-\sqrt{3} \le x_i \le \sqrt{3}$ (i = 1, 2, 3).

This example shows no significant difference in methods performance. All the criteria performed satisfactorily, generating an appropriate set of solutions along the Pareto frontier (see Figs. 27.1, 27.2, and 27.3). It is important to note that, if desirable, the number of solutions generated from PP method can be higher, by increasing the number of preference values.

Example 2 A central composite design with four center points was run to determine the settings for reaction time (x_1), reaction temperature (x_2), and amount

Fig. 27.1 DAM frontier

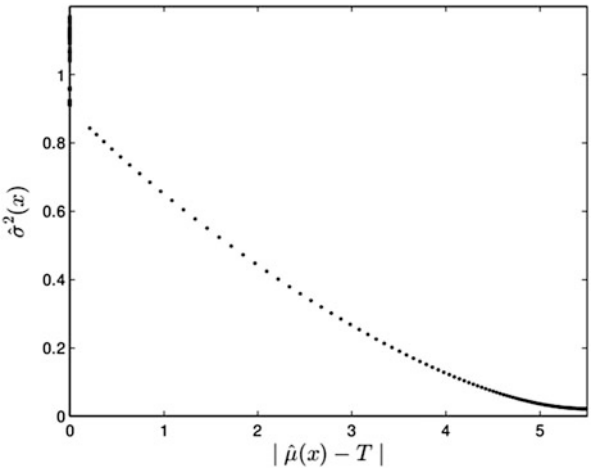
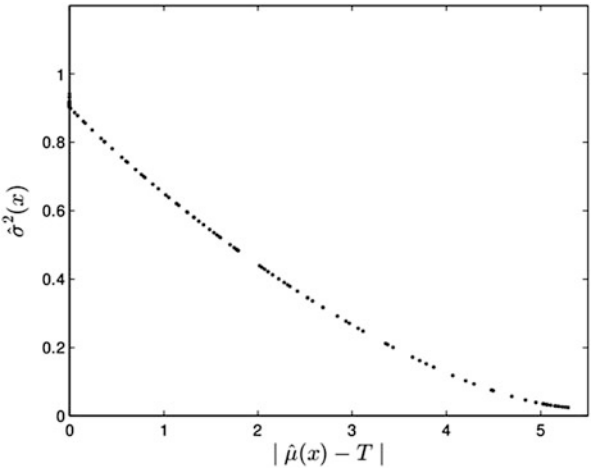


Fig. 27.2 GC frontier



of catalyst (x_3) to maximize the conversion (y_1) of a polymer and achieve a target value for the thermal activity (y_2). Estimated response models are

$$\begin{aligned}\widehat{y}_1 &= 81.0943 + 1.0290x_1 + 4.0426x_2 + 6.2060x_3 - 1.8377x_1^2 + 2.9455x_2^2 \\ &\quad - \ 5.2036x_3^2 + 2.1250x_1x_2 + 11.3750x_1x_3 - 3.8750x_2x_3 \\ \widehat{y}_2 &= 59.8505 + 3.5855x_1 + 0.2547x_2 + 2.2312x_3 + 0.8360x_1^2 \\ &\quad + \ 0.0742x_2^2 + 0.0565x_3^2 - 0.3875x_1x_2 - 0.0375x_1x_3 + 0.3125x_2x_3\end{aligned}$$

The ranges for y_1 and y_2 are [80, 100] and [55, 60], respectively. Assuming that y_1 is a LTB-type response, its target value is set equal to 100; y_2 is a NTB-type response and its target value is set equal to 57.5 [2]. The constraints for the input variables are $-1.682 \leq x_i \leq 1.682 \quad (i = 1,\ 2,\ 3)$.

Fig. 27.3 PP frontier

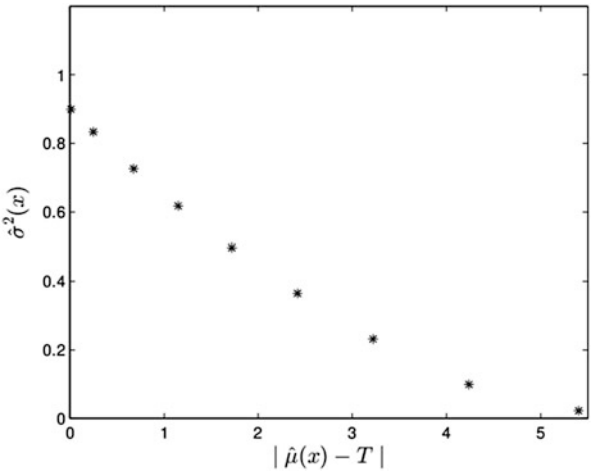


Table 27.1 DAM solutions

(ω_1, ω_2)	(x_1, x_2, x_3)	$(\hat{y}_1, \hat{y}_2)$	D
(0.22, 0.78)	(−0.5434, 1.682, −0.5984)	(95.21, 57.50)	0.053
(0.89, 0.11)	(0.0221, 1.682, −0.2019)	(96.13, 60.00)	0.227
(0.86, 0.14)	(−1.682, 1.682, −1.058)	(98.03, 55.00)	0.155

DAM method can't yield a satisfactory representation of Pareto frontier. This method only generated three alternative solutions (see Table 27.1 for details) whereas GC method generated a larger set of alternative solutions. Figure 27.4 shows that GC method yielded a satisfactory representation of the Pareto frontier, generating solutions with low bias values for $\hat{y}_2$, including solutions with $\hat{y}_2$ on target value, and solutions with $\hat{y}_1$ value close to the target ($\widehat{y}_1 = 98$). Figure 27.5 provides evidence that PP method can generate an evenly distributed set of alternative solutions along the Pareto frontier and, like GC method, can satisfy the decision-makers with different sensitivities to conversion and thermal activity responses, which are in conflict.

Example 3 In [39] is formulated a bi-objective problem with the estimated mean (μ) and standard deviation (σ) functions defined as

$$\mu = (x_1 - 4.0)^3 + (x_1 - 3.0)^4 + (x_2 - 5.0)^2 + 10.0$$
$$\sigma = \tfrac{1}{3}\sqrt{\left[3(x_1 - 4.0)^2 + 4(x_1 - 3.0)^3\right]^2 + 4(x_2 - 5.0)^2}$$

subject to

$$-x_1 - x_2 + 6.45 + 2 \leq 0$$
$$2 \leq x_i \leq 9, \quad i = 1, 2.$$

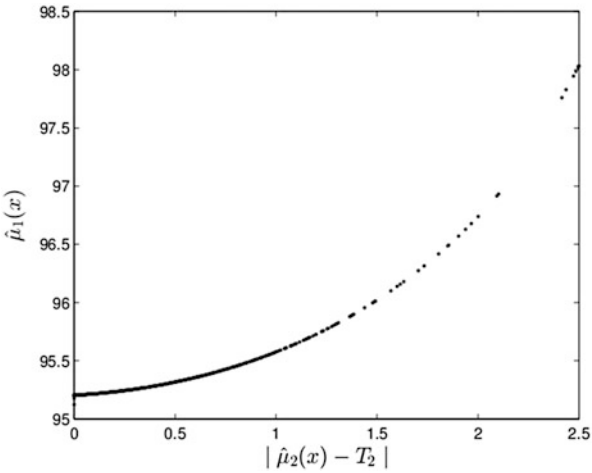


Fig. 27.4 GC frontier

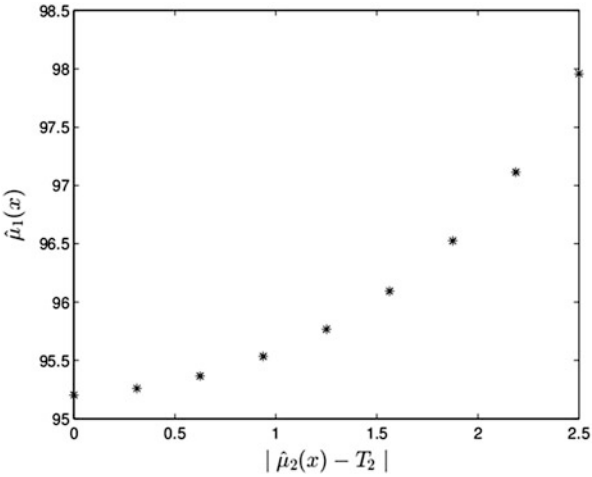


Fig. 27.5 PP frontier

We assume that μ is a STB-type response with U = 10, its target and lower specification limits are equal to 5.10, and σ is a STB-type response with U = 1.50 and both target and lower specification limits equal to 0.

This is not a standard dual response (mean and standard deviation responses optimization) problem developed under the *RSM* framework. Nevertheless, such as it is usual in bi-objective or dual response problems, the mean and standard deviation responses are in strong conflict (better values for μ are only possible degrading the σ value), which is appropriate to evaluate methods performance.

Figures 27.6 and 27.7 provide evidence of GC and PP methods ability to depict the Pareto frontier, which includes convex and nonconvex regions [39]. The limitations of DAM method are displayed in Fig. 27.8, where one can see that only a very small region of Pareto frontier is depicted.

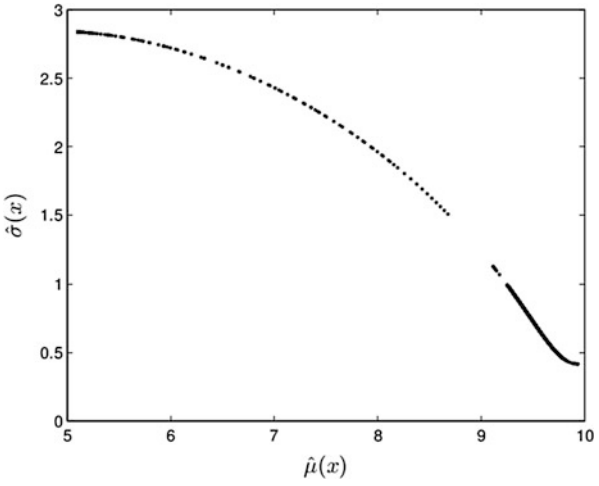


Fig. 27.6 GC frontier

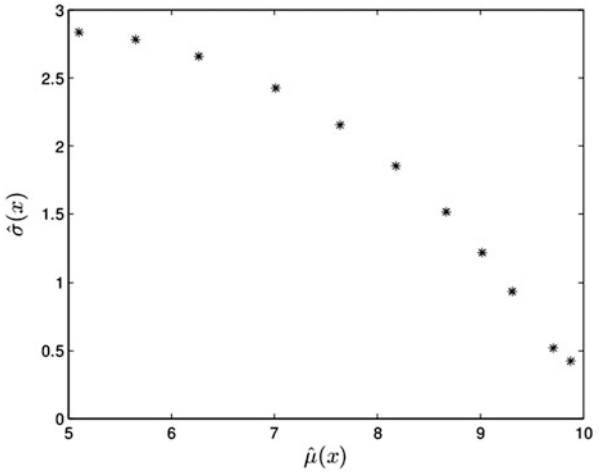


Fig. 27.7 PP frontier

5 Discussion

The lack of a generally agreed upon examples that must be used to evaluate optimization criteria performance does not contribute to a clear understanding of methods working ability. Results of some examples make a method look effective when, in fact, it has serious limitations. In addition, methods ability to depict Pareto surfaces has been rarely illustrated in the literature.

The exact determination of the Pareto frontier is unrealistic for real-world problems, as it is usually an infinite set. Effective optimization methods can yield, at least, a discrete representation of Pareto frontier. However, presented examples show that desirability-based (DAM) method does not perform always as desired. Example 1 does not show differences in methods performance, but Examples 2 and 3 show

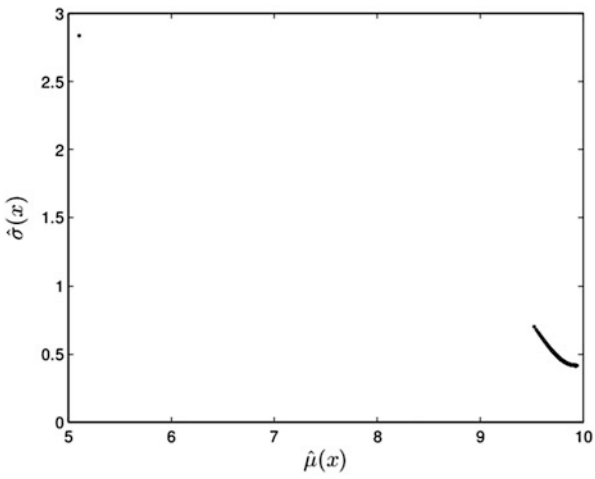


Fig. 27.8 DAM frontier

the DAM method inability to depict the Pareto frontiers. In fact, results yielded by DAM method do not give confidence to use it in real-life problems. This method is similar to a weighted sum method $\left(\text{of the form } \sum \omega \, \mathrm{F}(\mathrm{x})\right)$, whose limitations to depict highly convex and nonconvex regions of Pareto frontiers are well illustrated in the literature [40–42]. The poor performance of DAM method was expected, and contrast with that of GC and PP methods, which depicted the Pareto frontier in all the examples.

GC method is a weighted exponential sum function and presented examples provide evidences that it can capture solutions in convex and nonconvex regions of Pareto frontier. According to [27], shape factors $0.25 \leq e_i \leq 3$ are, in general, appropriate to GC method depict a representative set of optimal solutions for problems developed in the RSM framework. This is a relevant advantage over other methods available in the literature, which requires subjective information from analyst or decision-maker and do not offer guidelines about weights or shape factors range, giving confidence to use GC method in real-life problems. Nevertheless, higher e_i values may be necessary to obtain complete representations of some Pareto frontiers, namely for those where exist highly convex and nonconvex regions [43, 44]. In these cases, to use higher e_i values enables to better capture Pareto solutions, but non-optimal solutions may be also captured [44].

Physical programming method is among the most effective multicriteria mathematical programming techniques for the generation of Pareto solutions that belong to both convex and non-convex efficient frontiers [20]. Examples 1–3 confirm that it can yield solutions well distributed along the Pareto frontier for an evenly distributed set of preference parameters. However, simplicity and application easiness are not attributes of PP method. The class functions are calculated satisfying several mathematical properties so their construction is not ease [37, 45]. In addition, the PP algorithm includes a few free parameters whose uncertainty in their determination can't be fully removed, is highly sensitive to initial conditions, and the splines

(representations of class functions) obtained as well as their curvatures do not ensure the non-existence of local minima in the aggregate preference function [46]. According to [45], the PP problem model is, in general, a constrained non-linear multimodal optimization, therefore, to solve it, a global optimization technique has to be used as a substitute for the non-linear programming optimization originally used in the PP method [37]. These conditions make less appealing the PP method, namely for those who have limited mathematical and computational background.

6 Conclusion and Future Work

Determining the optimal factor settings that optimize multiple objectives or responses is critical for producing high quality products and high capability processes, and can have tremendous impact on reducing waste and costs. However, conflicting responses are usual in real-life problems and optimal variable settings for one or more responses may lead to degradation of, at least, another one, such as it is illustrated in Examples 1–3.

A large variety of alternative solutions can be found for multiple response optimization problems, and different impacts on process or product can also be expected from these alternative solutions. Therefore, to satisfy decision-makers with different sensitivity to each response, a method capable of capturing solutions evenly distributed along the Pareto frontier have to be used.

This article shows the working ability of global criterion-based and physical programming-based methods to generate solutions along the Pareto frontier in problems with conflicting objectives and their superiority over a desirability-based method. GC method is relatively easy to understand and apply, which are appealing advantages over other methods, including the PP method, and a stimulus to apply it in real-life problems. Nevertheless, further research is needed to better understand GC method working ability in problems developed under the RSM framework and set the range values for shape factors when the number of responses is large (four or more) and responses surface is highly nonconvex. The comparison with other effective approaches, namely in problems with a higher number of control variables, number of responses, and responses type must also be considered. To evaluate PP method in these problems, to confirm its effectiveness, and further research on construction of preference functions are also proposals for future research. Another suggestion is the development of a design space exploration approach that can be used to visualize n-dimensional Pareto surfaces and provide information about solutions in the neighborhood of optimal solutions for tradeoff analysis and decision making.

Acknowledgement The authors are grateful to Instituto Politécnico de Setúbal for its finantial support to this publication.

References

1. Myers, R., Montgomery, D., Anderson-Cook, C.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 3rd edn. Wiley, Hoboken (2009)
2. Costa, N., Lourenço, J., Pereira, Z.: Desirability function approach: a review and performance evaluation in adverse conditions. Chemom. Intell. Lab. Syst. **107**, 234–244 (2011)
3. Derringer, G., Suich, R.: Simultaneous optimization of several response variables. J. Qual. Technol. **12**, 214–218 (1980)
4. Derringer, G.: A balancing act: optimizing a product's properties. Qual. Prog. **27**, 51–58 (1994)
5. Das, P., Sengupta, S.: Composite desirability index in cases of negative and zero desirability. J. Manag. Resour. **10**, 25–38 (2010)
6. Murphy, T., Tsui, K., Allen, J.: A review of robust design methods for multiple responses. Res. Eng. Des. **15**, 201–215 (2005)
7. Ko, Y., Kim, K., Jun, C.: A new loss function-based method for multiresponse optimization. J. Qual. Technol. **37**, 50–59 (2005)
8. Pignatiello, J.: Strategies for robust multi-response quality engineering. IIE Trans. **25**, 5–15 (1993)
9. Vining, G.: A compromise approach to multiresponse optimization. J. Qual. Technol. **30**, 309–313 (1998)
10. Köksoy, O.: A nonlinear programming solution to robust multi-response quality problem. Appl. Math. Comput. **196**, 603–612 (2008)
11. Pal, S., Gauri, S.: Assessing effectiveness of the various performance metrics for multi-response optimization using multiple regression. Comput. Ind. Eng. **59**, 976–985 (2010)
12. Tong, L.-I., Wang, C.-H.: Multi-response optimisation using principal component analysis and grey relational analysis. Int. J. Ind. Eng. **9**, 343–350 (2002)
13. Liao, H.-C.: Multi-response optimization using weighted principal component. Int. J. Adv. Manuf. Technol. **27**, 720–725 (2006)
14. Awad, M., Kovach, J.: Multiresponse optimization using multivariate process capability index. Qual. Reliab. Eng. Int. **27**, 465–477 (2011)
15. Kwak, D.-S., Kim, K.-J., Lee, M.-S.: Multistage PRIM: patient rule induction method for optimisation of a multistage manufacturing process. Int. J. Prod. Res. **48**, 3461–3473 (2010)
16. Al-Refaie, A.: Optimizing performance with multiple responses using cross-evaluation and aggressive formulation in data envelopment analysis. IIE Trans. **44**, 262–276 (2012)
17. Govindaluri, S., Cho, B.: Robust design modeling with correlated quality characteristics using a multicriteria decision framework. Int. J. Adv. Manuf. Technol. **32**, 423–433 (2007)
18. Kazemzadeh, B., Bashiri, M., Atkinson, A., Noorossana, R.: A general framework for multiresponse optimization problems based on goal programming. Eur. J. Oper. Res. **189**, 421–429 (2008)
19. Messac, A.: Physical programming: effective optimization for computational design. AIAA J. **34**, 149–158 (1996)
20. Chen, W., Sahai, A., Messac, A., Sundararaj, G.: Exploring the effectiveness of physical programming in robust design. J. Mech. Des. **122**, 155–163 (2000)
21. Peterson, J., Miró-Quesada, G., Del Castillo, E.: A bayesian reliability approach to multiple response optimization with seemingly unrelated regression models. Qual. Technol. Quant. Manag. **6**, 353–369 (2009)
22. Dächert, K., Gorski, J., Klamroth, K.: An augmented weighted Tchebycheff method with adaptively chosen parameters for discrete bicriteria optimization problems. Comput. Oper. Res. **39**, 2929–2943 (2012)
23. Lee, D., Jeong, I., Kim, K.: A posterior preference articulation approach to dual response surface optimization. IIE Trans. **42**, 161–171 (2010)
24. Lee, D., Kim, K., Köksalan, M.: An interactive method to multiresponse surface optimization based on pairwise comparisons. IIE Trans. **44**, 13–26 (2012)

25. Ardakani, M., Wulff, S.: An overview of optimization formulations for multiresponse surface problems. Qual. Reliab. Eng. Int. **29**, 3–16 (2013)
26. Costa, N., Lourenço, J.: Optimization criteria ability to depict Pareto frontiers. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 958–961. London, UK (2014)
27. Costa, N., Lourenço, J., Pereira, Z.: Multiresponse optimization and Pareto frontiers. Qual. Reliab. Eng. Int. **28**, 701–712 (2011)
28. Köksoy, O., Doganaksoy, N.: Joint optimization of mean and standard deviation using response surface methods. J. Qual. Technol. **35**, 239–252 (2003)
29. Ch'ng, C., Quah, S., Low, H.: A new approach for multiple-response optimization. Qual. Eng. **17**, 621–626 (2005)
30. Costa, N., Pereira, Z.: Multiple response optimization: a global criterion-based method. J. Chemometr. **24**, 333–342 (2010)
31. Messac, A., Mattson, C.: Generating well-distributed sets of Pareto points for engineering design using physical programming. Optim. Eng. **3**, 431–450 (2002)
32. Lin, K.-P., Luo, Y.-Z., Tang, G.-J.: Multi-objective optimization of space station logistics strategies using physical programming. Eng Optim. (2014). doi:10.1080/0305215X.2014.954568
33. Messac, A., Melachrinoudis, E., Sukam, C.: Mathematical and pragmatic perspectives of physical programming. AIAA J. **39**, 885–893 (2001)
34. Yuan, Y., Ling, Z., Gao, C., Cao, J.: Formulation and application of weight-function-based physical programming. Eng. Optim. **46**, 1628–1650 (2014)
35. Martínez, M., Sanchis, J., Blasco, X.: Integrated multiobjective optimization and a priori preferences using genetic algorithms. Inform. Sci. **178**, 931–951 (2008)
36. Martínez, M., Sanchis, J., Blasco, X.: Multi-objective engineering design using preferences. Eng. Optim. **40**, 253–269 (2008)
37. Sanchis, J., Martínez, M., Blasco, X., Reynoso-Meza, G.: Modelling preferences in multi-objective engineering design. Eng. Appl. Artif. Intel. **23**, 1255–1264 (2010)
38. Utyuzhnikov, S., Fantini, P., Guenov, M.: A method for generating a well-distributed Pareto set in nonlinear multiobjective optimization. J. Comput. Appl. Math. **223**, 820–841 (2009)
39. Chen, W., Wiecek, M., Zhang, J.: Quality utility: a compromise programming approach to robust design. J. Mech. Des. **121**, 179–187 (1999)
40. Das, I., Dennis, J.: A closer look at drawbacks of minimizing weighted-sums of objectives for Pareto set generation in multicriteria optimization problems. Struct. Optim. **14**, 63–69 (1997)
41. Marler, R., Arora, J.: The weighted sum method for multi-objective optimization: some insights. Struct. Multidiscip. Optim. **41**, 853–862 (2010)
42. Athan, T., Papalambros, P.: A note on weighted criteria methods for compromise solutions in multi-objective optimization. Eng. Optim. **27**, 155–176 (1996)
43. Messac, A., Sundararaj, G., Tappeta, R., Renaud, J.: Ability of objective functions to generate points on non-convex Pareto frontiers. AIAA J. **38**, 1084–1091 (2000)
44. Marler, R., Arora, J.: Survey of multi-objective optimization methods for engineering. Struct. Multidiscip. Optim. **26**, 369–395 (2004)
45. Sanchis, J., Martínez, M., Blasco, X.: Multi-objective engineering design using preferences. Eng. Optim. **40**, 253–269 (2008)
46. Martínez, M., Sanchis, J., Blasco, X.: Multiobjective controller design handling human preferences. Eng. Appl. Artif. Intel. **19**, 927–938 (2006)

Chapter 28
The Implications of Wet and Dry Turning on the Surface Quality of EN8 Steel

Zulfiqar Ahmad Khan, Matthew Grover, and Mian Hammad Nazir

Abstract This paper, by experimental and investigation, examines the effects of dry and flood cutting conditions by comparing the rate of tool wear during metal turning and the produced surface roughness to determine if dry cutting can be a cost effective solution. For efficient manufacturing, the surface roughness of the turned parts should be dependent on their intended application, factors such as environment of operation or further manufacturing processes will determine this level of surface roughness required, as the performance and mechanical properties of the material can be affected. EN8 steel has been selected as the work material for its popularity and low hardness. The results show both wet and dry conditions have their benefits in relation to the intended application of the part, but mostly dry turning produces competitive surface roughness when finished by turning when compared to wet, and acceptable levels of tool wear while rough cutting. It would be recommended that in most circumstance for rough cutting, dry conditions should be employed with the knowledge of slight increased tool wear and possibly shorter life but with reduced manufacturing costs and environmental hazards.

Keywords Build-up-edge • Coolant • Dry turning • Manufacturing • Surface roughness • Tool wear

1 Introduction

Research has been conducted on the tool wear and surface quality with and without using coolant [1]. Turning is one of the most basic and common material removal processes, done via a rotating work piece, it along with machining has an overwhelming and increasing level of concern surrounding the use of metal working fluids (MWF's) during these metal removal operations as relayed by [2–6]. There have been numerous studies conducted on these concerns and the risks they impose on the operator and environment. Skin related problems have been

Z.A. Khan (✉) • M. Grover • M.H. Nazir
Faculty of Science and Technology, Sustainable Design and Research Centre (SDRC), Bournemouth University, Poole, UK
e-mail: zkhan@bournemouth.ac.uk; mattgrover@hotmail.com; hnazir@bournemouth.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_28

linked with direct exposure to the coolants as well as health risks associated with coolants becoming airborne [2, 7, 8], which was linked to bacteria and fungi colonizing within the cutting fluids and serving as a source of microbial toxins. With attention of dry machining successively brought to the field of environmentally friendly manufacturing by [9], it was soon made apparent of its potential advantages. Unfortunately, without the presents of a cooling agent, certain characteristics of the turned work piece can greatly suffer, mainly caused by the excessive generation of heat. This lack of coolant can affect some of the most important requirements for a turned work piece, making dry turning sometimes less effective, as discussed by [10]. For many years now the tooling company "Sandvik" has encouraged and developed dry turning for the industry and have made great success in terms of tool life and surface quality by producing more geometrically suitable and stable cutting tips. Flood cooling is the most widely used approach in industry for both milling and turning [10]. Although in some circumstance coolant can prolong the life of the cutting tip, higher costs may be present through material removal stages when coolant is applied, as cutting fluids *"impact both stationary and rotating elements within the machine tool system*" [6], as opposed to a possible reduction in the life of the cutting tip, which may come at a lower, overall expense to the company or metal worker. With mechanical energy being transferred into the cutting fluid, higher surface energy is obtained by the coolant which in turn can cause it to atomise through reduced stability. As suggested by [11, 12], reducing the fluid will in turn reduce the cutting force and improve surface finish.

Cost of coolants is not just a one off payment, but includes indirect costing that have been said to make up around 7–17% of the total manufacturing costs [4, 5]. The total cost of use for cutting fluids is comprised of several factors; first being the initial cost, top-up costs, life machine damage, health & safety issues, maintenance and most importantly and a growing concern, disposal costs [13]. To overcome this, minimal quantity lubrication (MQL) was developed and studied to reduce the amount of lubricants in metal removing operations due to issues of ecological, economical and most importantly occupational pressure [2]. Despite the reduction, MQL still has an undesired by-product of airborne particles, which increase the health risks of the operator. Long exposure to these airborne particles can result in "*health problems ranging in severity from mild respiratory illness to asthma and several types of cancer*" [7]. While reducing the cost of coolant, MQL involves additional costs to pressurise the air and technological support that is required with the process, so although it may be considered a more environmentally friendly option, cost saving or operator health is not a key advantage. Therefore, the objective of this report is to compare the tool wear, surface roughness and geometric accuracy under wet and dry conditions when turning, thus to gain a better insight and understanding on the decision whether to use MWF's and potentially reduce manufacturing costs and environmental hazards, as well as determining if the rate of tool wear outweighs the use of MFW's to the point that it makes it feasible. Despite the numerous studies on dry and hard turning and the attractiveness of it, "*. . . implementation in industry remains relatively low*" [14] as some outdated companies are still performing flood cooling where it is unnecessary, either because

they are unaware of the practice or do not know how to implement dry turning correctly in the right circumstances, which may be due to it being a relatively new processing technique with several questions remaining unanswered [14]. Certainly, coolant and lubricant is required in some aspects of machining and turning to remove excessive heat and chippings, but considerable costs and environment hazards can be reduced if dry turning is used appropriately when possible.

2 Background

Obtaining different levels of surface finish during metal turning should be dictated by its intended application, as high quality finishes and tight tolerances can induce increased machining times and costs, which may not be necessary or economical. In some circumstances a low surface roughness can be one of the most important requirements for many turned work pieces, where coolants and other cutting fluids are used to reduce heat dissipation to maintain surface quality and geometric accuracy throughout the material removal process. Such applications include interference fits and surfaces that are to be polished. A good surface roughness influences the materials mechanical properties whilst in service. Lower surface roughness can prevent premature fatigue failure, improve corrosion resistance, reduce friction, wear, noise and improve life of the product [15]. Accepting a slightly shorter tool life to potentially reduce the associated costs, as opposed to down time of machine tool to maintain or replace cutting fluids and its costs could be the less expensive choice [16]. In other instances, higher levels of surface roughness may also be desired, i.e. to allow coatings to adhere correctly to the part.

Surface finish is highly controlled by many different factors, including the cutting parameters [10, 17–19] tool type, rigidity of the lathe and geometry of the cutting tip, which includes rake and flank angles [15]. The surface finish also directly relates to and is an important measure of the overall quality of the part, as this influences the performance, mechanical properties and cost of production [20]. As [21] states "*Surface quality significantly improves fatigue strength, corrosion resistance, or creep life*", which stresses the importance of specifying surface requirements during design stages so that the tools maybe be set up appropriately to achieve it. One of the major causes of surface quality loss is through material build-up on the rake face of the cutting tip, also known as build-up-edge (BUE), which can be seen circled in Fig. 28.1. It is an unwanted, semi stable body of material on the cutting tool that is created by work piece material welding onto the tool during cutting. Layers of build-up weld to the tool face under the heavy pressure and heat generated at the tip of the tool face, also associated with lower cutting speeds and feed rates; it is therefore more common during dry turning. BUE has been linked in studies [21, 22] to causing low quality surface finish and cutting edge frittering when the built-up edge is torn away [23], as well as increasing the wear rate of the tool [24], although with an increase in surface cutting speed. The BUE phenomenon has a tendency to minimise wear by creating a protective layer on the tool [25], which is not always possibly with dry turning due to the limited controllability of excessive heat generation.

Fig. 28.1 Material build-up on the rake face of the cutting tip (BUE) (Amended from Balzers 2010)

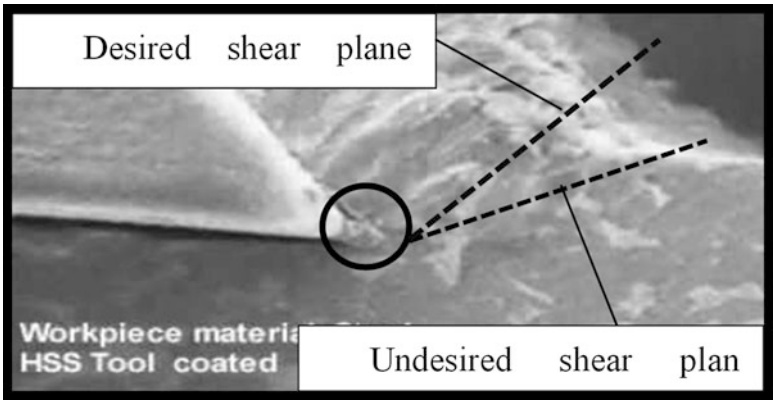


Fig. 28.2 Increased cutting speed and the shear plane angle effect on BUE (Amended from Balzers 2010)

Figure 28.1 (Right) shows the effect of the BUE being dragged down under the tool tip and becoming an imperfection on the turned face, while Fig. 28.1 (Left) shows a partial shearing of the surface also caused by the BUE. This can be overcome by increasing the cutting speed and the shear plane angle, which has been marked in Fig. 28.2. During testing procedures, to ensure that no unnecessary wear is taking place on the tool tip caused either by the BUE or by the cutting parameters, chip formation needs to be closely monitored. Ideally, chips should be of a helical shape and no more than 5 cm in length [25]. A blue appearance on the chips should be present, indicating the correct depth of cut and feed rate are being used and that sufficient heat is being taken from the work piece via the chip and in turn maintaining a high shear plane angle, thus causing the BUE to move off with the chips and not in the turned surface.

Due to the large influence on surface roughness, the phenomena of the build-up-edge has been investigated extensively, with the central focus being why and how it is bonded to the tool, but is still not completely understood as the mechanism of adhesion is very complex [21] as well as being a dynamic process and microscopic in nature [24]. However, three facts are determined undeniable: temperature of cutting zone and tool being lower than the work piece melting point, contaminant layers present between work piece and tool interface and if the welded work material is severely deformed. Several parameters have been studied that have been found to effect the formation of the build-up edge, as [21] discusses, tool geometry, tool material, machine tool, cutting fluid etc., most influential being tool geometry for it frictional effects in regards to rake angles.

3 Cutting Parameters

There have been numerous studies and investigations on the cutting parameters under wet and dry conditions to gain a better understanding on the numerous variables and their influences on the surface roughness, dimensional accuracy and other contributing factors that dictate the overall quality of the turned part. Experiments carried out by each study concluded with similar results, where cutting parameters, such as low cutting speed having no noticeable difference between dry and flood turning [15, 20, 26]. There were also other notable similarities within the results of the studies, where factors would be more beneficial in conjunction with other parameters. These results, as well as material suppliers' recommendation and professional guidance will be used and tailored to provide the best set up for the experiment, in terms of feed rate, depth of cut, cutting tip type, thus avoiding any unnecessary testing or factors that will corrupt the results, as they have already been proven to provide the best results.

4 Machine Setup

As specified on the material suppliers website [27], the through hardened En8 has a Vickers Hardness of 210-265. This can be used to determine the appropriate surface cutting speed, which when a carbide tip is being used, 91.44 m/min is recommended [17]. From this, the required RPM for the lathe can be calculated for a given diameter of work piece, which for a 30 mm diameter bar is 970RPM. The material could have been tested to provide a more accuracy reading, however this was considered unnecessary due its minimal influence on the RPM.

5 Experimental Procedure

EN8 Steel is to be used as the testing material that will be turned down. It is a very popular grade of through-hardened medium carbon steel, which in this case, is in the form of 30 mm diameter bar. EN8 is suitable for the manufacture of parts such as general-purpose axles and shafts, gears, bolts and studs. Little testing has been performed on low hardness steels in regards to tool wear, making EN8 an ideal candidate due to its chemical composition and popularity within the industry. Two unused cutting tips will be used to turn down two EN8 bars, one for each cutting condition. New tips are being used as company recommended cutting parameters are based around cutting tips in top condition that have no rake or flank wear, which will impede cutting performance. Cuts of 250 mm long and 0.7 mm deep will be made with each tip on the bars under the two conditions with a feed rate of 0.25 mm/rev at 970 RPM or the next closes speed on the lathe. This will ensure ample heat generation and exposure to the tip during the dry cut, making a notable and comparable influence on tool wear. A finishing cut in each condition will be taken at a reduced feed rate of 0.08 mm/rev. Although in many cases, a shallower cut is required to achieve a more desirable finish, the nature of EN8, with its lower HV when compared to materials that meet hard turning criteria and as studies suggested, will require the same depth of cut to allow the tool to penetrate the surface and maintain a high shear plane angle. For more of a comparison, a sample of the roughing and finishing cut in both conditions will be produced, cleaned and inspected under 3D surface analysis equipment on all surfaces of interest; this will also be the case for both cutting tips rack face and cutting edges. A micrometre will be used to measure the bar with a finishing cut of 25 mm diameter.

The coolant being used is specified for steel, stainless and titanium and is of an oil-based solution, which has been mixed between 7% and 10% concentration. It is considered a high performing coolant, which will provide adequate cooling during testing. As one factor of surface finish is vibration and due to the length of the work piece extending from the chuck, a live centre is used to stabilise the bar during turning so that any surface roughness caused by vibration is minimised allowing appropriate comparison to be made between tool wear and surface roughness caused by wet and dry cutting.

6 Intentions

This report hopes to determine the physical, environmental and practical elements of dry turning in regards to tool wear, the quality of surface finish and the dimensional accuracy of the cut, which will eventually lead to and aid in determining the point at which the rate of tool wear and error in dimensions out weights the cost of lubricants to consider using them while turning. It also intends to provide further information to make an informed decision into which method is most appropriate when faced with different applications and environments that the part is intended to be used in.

7 Results and Discussion

During the cutting procedure, chip formation for the dry condition was as expected and showed appropriate heat removal via the chips. However a jagged edge was present on the inside of the helical chip, which could potentially be caused by a too high feed rate. Despite this, alterations were not made to the setup, as other signs, such as excessive vibration where not present and suggested safe operating conditions and indicated no risk to breaking the cutting tip. When comparing the surface finishes of both wet and dry rough cuts, there was a notable difference, as illustrated in Figs. 28.3 and 28.4.

When sampled on a plane longitudinal to the work piece (perpendicular to the cut), both wet and dry had similar surface roughness's at 4.899 and 5.119 μm respectively. The similarity in roughness values is due to the peaks and troughs created by the cutting tip, which, as proven by the numerous studies is dictated by feed rate and tool radius. From an industrial point of view, these roughness values would be considered average during turning; when referring the quality of cut. This however, tends toward less frequent occurrence and perhaps unacceptable finishes [19]. Despite this, conditions of the wet cut are noticeably different on simple examination. When sampling on a plane parallel with the cut, and thus excluding peaks and troughs, a better understand on how the material has sheared is presented. Wet conditions greatly out performed dry, with 0.874 and 2.218 μm respectively.

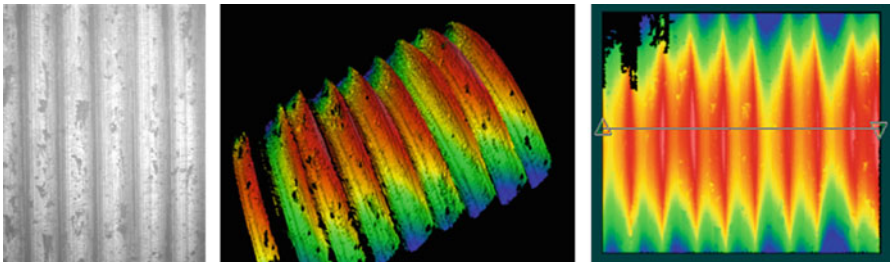


Fig. 28.3 3D surface analysis results for roughing wet cut

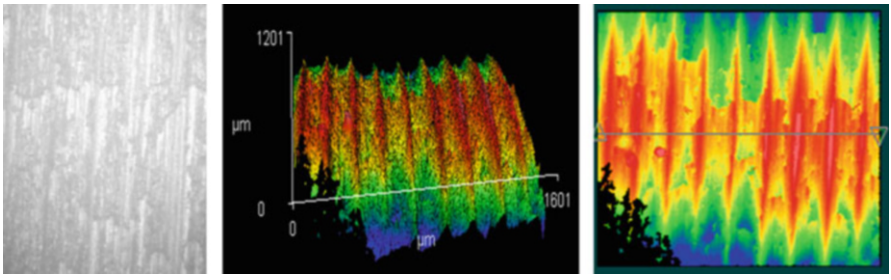


Fig. 28.4 3D surface analysis results for roughing dry cut

The level of disturbance and deterioration on the surface of the dry cut is a clear indication that excessive heat was present despite maximum heat removal via the chips, during the cut and consequently leading to a low shear plane angle, causing an unstable BUE that would have passed beneath the tool, causing the surface to tear. It must be noted that because of the cylindrical shape of the specimen, sampling was kept considerably short to avoid the curvature that would be perceived as a higher surface roughness. Unlike hard turning, these results conform to the findings of the several investigations on cutting parameters and BUE, as with the lower cutting speed for EN8, BUE is more likely to form [24].

As expected and in agreement with other studies, the reduced feed rate of 0.08 mm/rev for the finish cut improved surface roughness for both wet and dry conditions, with 0.559 and 1.139 μm respectively. With reduced feed rate, less material is being removed per revolution by the flank face and tool edge and therefore reduces heat and pressure. The now closer cutting lines form a more uniform surface, where peaks and troughs of the cuts are less profound. Dry finish turning does not produce an acceptable level of surface roughness comparable to polished surface or high tolerance interfacing parts, such as interference fits. Therefore, these finishes would not be recommended for applications where precision or aesthetic design considerations are needed.

On measuring the accuracy of the finishing cut, it is found that in dry conditions there was an overcut of −0.01 mm, while for wet cut conditions there was an undercut of +0.006 mm. Due to minor variations it would be incorrect to directly link them to the cutting conditions, although in some cases the heating of the metal can cause thermal expansion and thus causing a deeper cut to be taken by the tool. In this instance, the variations would be accepted in most machine shops for non-interfacing surfaces and can be put down to machine and human error.

With regard to tool wear, there were two anomalies present on the wet cutting tip (Fig. 28.6). Examination and investigation strongly suggests that these where present before cutting took place, as they are far too small to be considered notch wear and more possibly associated with the manufacturing of the tip.

On first inspections, it was clear that the dry cutting tip had encountered more direct and pin pointed wear on the rake face, (marked in Fig. 28.5). This would

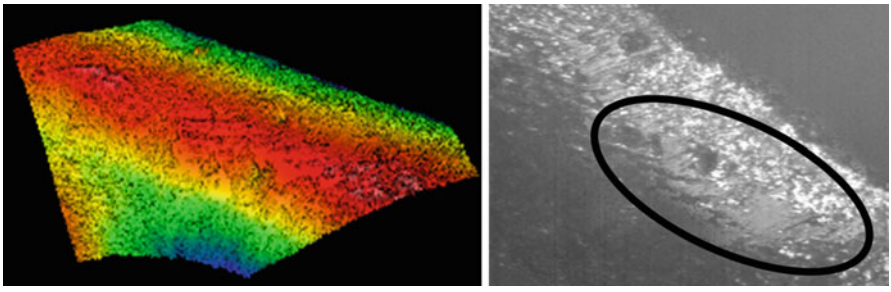


Fig. 28.5 3D surface analysis on cutting tip for wet cut conditions

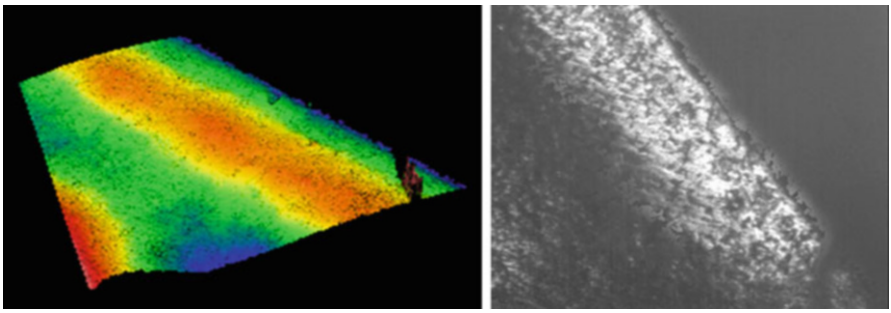


Fig. 28.6 3D surface analysis on cutting tip for dry cut conditions

have been caused by the hotter and harder chippings coming off the job with less deformation, while the cooled chips from the wet cut have deformed with less force against the rake face of the tool (Fig. 28.6), creating an overall smoother and distributed wear. Sampling was roughly taken 100 μm parallel to the cutting edge, over a distance of 400 μm on the rake face of the tool. Roughness values being 3.934 μm for wet and 5.121 μm for dry, with values before testing of 3.823 and 3.991 μm respectively. Results for edge wear and deformation amount were unobtainable; however values provided for crater wear on the rake side will provide a strong indication on the expected level of wear during the two conditions. Crater wear, seen in Figs. 28.5 and 28.6 is localised to the rake side of the insert and can lead to fracturing of the tip as the wear weakens the cutting edge. With a more distributed wear during wet condition, it is likely to prolong cutting tip failure when compared to dry conditions.

8 Conclusion

Caution should be taken with the results presented in this paper, as there are many factors that will contribute to the surface roughness. It can be concluded that although an increased wear rate is present during dry condition; the direct and indirect costs of coolant alone will outweigh the increased frequency for purchasing of new cutting tips. During rough cutting it was noted that a high surface roughness was present with dry conditions, but was still at an acceptable level for non-interfacing surfaces, which would suggest that for pure material removal operations, dry cutting is the most cost effective despite the slight increase in tool wear. The choice of wet and dry finish cutting highly depends on the intended application and the required amount of material needed to be removed. A compromise would be needed between the options either to use coolant or a slightly higher surface roughness with added manufacturing time. With coolant, it will be possible to further increase cutting feed rate without over heating the tool, however, for single

one off jobs, fluid is not justifiable and a slower feed rate should be employed with dry conditions. As a result of the slight increase in tool wear, it is recommended that for prolonged cutting jobs, dry conditions can be used for material removal with coolant applied for the finish turn as this will drastically reduce the volume of coolant used during the operation, while still providing an acceptable level of surface roughness. It should also be noted that heat generated during long cutting operations in dry conditions can possibly alter the material characteristic, particularly to smaller parts, which are less able to dissipate heat away, causing weakness in the material, through increased malleability, resulting in a change in cutting performance and strength and therefore affecting surface roughness.

The results discussed during this report have been derived from one off experiments that were performed on single sample test pieces for each condition, due to a limiting time frame. Given that, surface quality can be affected by numerous parameters. A stronger case could be presented, where additional results from multiple samples for each condition are performed and inspected to ensure results collected for this report are as accurate as possible. It would therefore be recommended that for further experimentation, at least three samples are produced in each condition, which can be combined to find the mean result and identify any errors that may have occurred, producing constant and reliable data.

References

1. Grover, M., Khan, Z.A.: The comparison on tool wear, surface finish and geometric accuracy when turning EN8 steel in wet and dry conditions. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 1093–1097. London, 2–4 July 2014
2. Khana, M.M.A., Mithua, M.A.H., Dharb, N.R.: Effects of minimum quantity lubrication on turning AISI 9310 alloy steel using vegetable oil based cutting fluid. J. Mater. Process. Technol. **209**(15–16), 5573–5583 (2009)
3. Silva, L.R.: Study on the behaviour of the minimum quantity lubricant – MQL technique under different lubricating and cooling conditions when grinding ABNT 4340 steel. J. Braz. Soc. Mech. Sci. **27**(2), 192–199 (2005)
4. Weinert, K.I., Sutherland, J.W., Wakabayashi, T.: Dry machining and minimum quantity lubrication. CIRP Ann. Manuf. Technol. **53**(2), 511–537 (2004)
5. Braz, J.: Application of cutting fluids in machining processes. J. Braz. Soc. Mech. Sci. **23**(2), 227–240 (2001)
6. The John W. Sutherland Research Page: Cutting fluid health hazards. Available: http://www.mfg.mtu.edu/cyberman/metal_fluids/health.html (2004). Accessed 16 Dec 2013
7. Tomac, N., Tonnessen, K., Rasch, F.O., Mikac, T.: A study of factors that affect the build-up material formation. In: Kuljanic, E. (ed.) AMST'05 Advanced Manufacturing Systems and Technology, vol. 486, pp. 183–192. Springer Vienna, New York (2005)
8. Greaves, I.A., Eisen, E.A., Smith, T.J., Pothier, L.J., Kriebel, D., Woskie, S.R.: Respiratory health of automobile workers exposed to metalworking fluid aerosols. Respir. Symptoms **32**(5), 450–459 (1997)
9. Klock, F., Eisenblaetter, G.: Dry cutting. CIRP Ann. Manuf. Technol. **46**(2), 519–526 (1997)
10. Davis, R., Madhukar, J.S., Rana, V.S.: Optimization of cutting parameters in dry turning operation of EN24 steel. Int. J. Emerg. Technol. Adv. Eng. **2**(10), 559–563 (2012)

11. Kouam, J., Songmene, V., Balazinski, M., Hendrick, P.: Dry, semi-dry and wet machining of 6061-T6 aluminium alloy. In: Ahmad, Z. (ed.) Dry, Aluminium Alloys – New Trends in Fabrication and Applications, pp. 199–221. InTech, Rijeka (2012)
12. Hardinge: Minimum Quantity Lubrication in Finish Hard Turning. Hardinge group, Georgia (2003)
13. Jemtech: Cutting fluid productivity – a question of attitude from supplier and customer. Jemtech. Available from: http://www.jemtech.co.uk/news/cutting-fluid-productivity.html (2011). Accessed 17 Dec 2013
14. Hardinge: Tool Life, Wear Rates, and Surface Quality in Hard Turning. Hardinge group, Georgia (2001)
15. Cassler, Z., Prato, Y., Muñoz-Escalona, P.: Built-up edge effect on tool wear when turning steels at low cutting speed. J. Mater. Eng. Perform. **13**(5), 542–547 (2004)
16. Graham, D., Huddle, D., Mcnamara, S.: Reducing cutting fluid use offers the chance for considerable cost savings. Tool life may even improve. In: Machining Dry Is Worth a Try. Modern Machine Shop, Cincinnati (2003). Available: http://www.mmsonline.com/articles/machining-dry-is-worth-a-try
17. Machine Shop 4: Cutting speeds, feeds and depth of cut. Fox Valley technical college, United States. Available from: http://its.foxvalleytech.com/MachShop4/Carbcut/Cutspeeds.htm (2000). Accessed 28 Jan 2014
18. Abouelatte, O.B., Mádl, J.: Surface roughness prediction based on cutting parameters and tool vibrations in turning operations. J. Mater. Process. Technol. **118**(1–3), 269–277 (2001)
19. Diniz, A.E., De Oliveira, A.J.: Optimizing the use of dry cutting in rough turning steel operations. Int. J. Mach. Tools Manuf. **44**(10), 1061–1067 (2004)
20. Rafai, N.H., Islam, M.N.: Comparison of Dry and Flood Turning in Terms of Quality of Turned Parts. In: Proceedings of the world congress on engineering, vol. 3, (2010)
21. Lou, M.S., Chen, J.C., Li, C.M.: Surface roughness prediction techniques for CNC end mill. J. Ind. Technol. **15**(1), 1–6 (1999)
22. Balzers: Balzers_CUT and Build up Edge formation. MPG. Available from: http://www.youtube.com/watch?v=uwh3ouvzSLk (2010). Accessed 28 Jan 2014
23. Sandvik: Built-up edge (B.U.E.). Available from: http://www.sandvik.coromant.com/en-gb/knowledge/general_turning/troubleshooting/tool_wear/pages/built-up-edge-(b.u.e.).aspx (2014). Accessed 29 Jan 2014
24. De Benedictis, K.: Read Your Chips. Modern Machine Shop, Cincinnati. Available from: http://www.mmsonline.com/articles/read-your-chips (1998). Accessed day month 2014
25. Subramanyam, K., Rangarao, P.V., Eswara Reddy, C.: Comparison of surface roughness and temperature between cubic boron nitride (CBN) and ceramic cutting tools when machining AISI52100 steel. Int. J. Adv. Eng. Res. Stud. **1**(2), 58–64 (2012)
26. L.J.Star incorporated: Surface finish Charts. L.J. Ohio: Star Incorporated. Available from: http://www.ljstar.com/design/surface_charts.aspx (2013). Accessed 9 Feb 2014
27. KV Steel Services LTD.: EN8 Steel. KV steel, Darlaston. Available from: http://www.kvsteel.co.uk/steel/EN8.html (2014). Accessed 28 Jan 2014

Chapter 29
Variable Selection Methods for Process Monitoring

Luan Jaupi

Abstract In the first stage of a manufacturing process a large number of variables might be available. Then, a smaller number of measurements should be selected for process monitoring. At this point in time, variable selection methods for process monitoring have focused mainly on explained variance performance criteria. However, explained variance efficiency is a minimal notion of optimality and it does not necessarily result in an economically desirable selected subset, as it makes no statement about the measurement cost or other engineering criteria. Without measuring cost many decisions will be impossible to make. In this article, we propose two new methods to select a reduced number of relevant variables for multivariate statistical process control that makes use of engineering, cost and variability evaluation criteria. In the first method we assume that a two-class system is used to classify the variables as primary and secondary based on different criteria. Then a double reduction of dimensionality is applied to select relevant primary variables that represent well the whole set of variables. In the second methodology a cost-utility analysis is used to compare different variable subsets that may be used for process monitoring. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. To do this, to any process monitoring procedure is assigned a score calculated as ratio of the cost at which it might be obtained to explained variance that it might provide. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables.

Keywords Cost-utility analysis • Covariance matrix • Dimension reduction • Eigenvalues • Influence function • Principal components • Process control • Variance efficiency

L. Jaupi (✉)
Conservatoire National des Arts et Métiers, Applied Statistics, 292 rue Saint Martin, Paris 75003, France
e-mail: jaupi@cnam.fr

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_29

1 Introduction

1.1 General Considerations

The aim of Statistical Process Control, SPC, is to bring a production process under control and keep it in stable condition to ensure that all process output is conforming. This under control state is achieved by monitoring the process through measurements of selected variables. When large number of variables are available, it is natural to enquire whether they could be replaced by a fewer number of measurements without loss of much information. Examples of situations in which variable selection is necessary can be found in [1, 2].

For process monitoring the variables should be selected according the needs of users and their ability to identify problems that may occur in the process. A two stage methodology to select a subset of variables that retains as much information on the full set of variables as possible, assuming that all variables are equally important according to engineering and economic criteria is given in [3]. However, in many cases measured variables are not equally important according to given criteria. For example, according to some engineering criteria some variables may be very important for the functionality of the part and others less important, or some variables may be easier and cheaper to carry out then others or some variables may be more efficient in waste reduction because their measurements are made at earlier points in the process. Neglecting this information in SPC would be counterproductive. At this point in time, there is a gap in the SPC literature devoted to statistical selection of variables in conjunction with given engineering or economic criteria.

1.2 Existing Methods

Multivariate process control charts have been increasingly popular to monitor many different industrial processes. One of the most popular multivariate control charts is based on Hotelling's T^2 statistic, [4–8]. Hotelling's T^2 statistic is easy to use and it can be shown to be the optimal test statistic for detecting a general shift in the process mean, [9]. However the control charts that are based on T^2 statistic are inadequate to monitor processes where many highly correlated variables are being collected. The main raison for this is that their shift-detection power would decline substantially if the number of monitored variables is greater than three. Therefore for processes where huge amounts of multidimensional data are available, the reduction of their dimensionality in terms of space and/or in terms of original variables has received much attention by a variety of industries.

The most common dimensionality reduction techniques in multivariate statistical process control, MSPC, in terms of space consists in projecting data into subspaces of low dimensions or in reducing the effective number of variables that need to

be measured. In projection methods the interest is to project the information that is contained in multidimensional data into subspaces of low dimensions in which process monitoring and interpretation of the results are much easier. Multivariate projection methods such as principal component analysis (PCA) and partial least squares (PLS) have been applied for processes where huge amounts of multivariate data are being collected, [1, 10–14]. Dimensionality reduction via PCA does not provide a real reduction of dimensionality in terms of the original variables because all the original variables are still involved in the retained principal components, PC, and they need to be measured in order to define the PCs.

In variable selection methodologies the interest is not only in reducing the dimension of the space but also in reducing the effective number of variables that need to be measured. Positively, any variable that is measured in a process provides information related to some aspects of the process and product quality. However, monitoring all variables of a complex process may be expensive and inefficient. Jolliffe in [15, 16] compares several variable selection methods based on PCA which were shown to be fast and efficient in simulated studies. However these selection procedures might be highly unreliable because they depend only on loadings. Different criteria to select a subset of principal variables that contains as much information of original variables as possible were discussed in [17, 18]. A two stage methodology to select a subset of variables that retains as much information on the full set of variables as possible, assuming that all variables are equally important according to engineering and economic criteria is given in [3]. But from a practitioner's point of view the measured variables generally are not equally important according to a given criteria. For example, according to some engineering criteria some variables may be very important for the functionality of the part and others less important, or some variables may be easier and cheaper to carry out then others or more efficient in waste redaction because their measurement are made at the earliest point in the process. There is a gap in the SPC literature devoted to statistical selection of variables in conjunction with given engineering or economic criteria.

1.3 Aim of Proposed Methods

In this article, we propose two new methods to select a reduced number of relevant variables for MSPC that makes use of engineering, cost and variability evaluation criteria. In the first method we assume that a two-class system is used to classify the variables as primary and secondary based on different criteria. Then a double reduction of dimensionality is applied to select relevant primary variables that represent well the whole set of variables. The selection methodology uses external information to influence the selection process. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables, thereby providing a SPC almost as efficient as we were monitoring all original variables. The proposed method is a

stepwise procedure. Various variable selection procedures might be used to select relevant primary variables. In this article we propose a backward elimination scheme, which at each step eliminates the less informative variable among the primary variables that have not yet been eliminated. The new variable is eliminated by its inability to supply complementary information for the whole set of variables. To achieve this we propose the use of Principal Components, PCs, which are computed using only the selected subset of primary variables, but represent well the whole set of variables. This strategy mitigates the risk that an assignable cause inducing a shift, that lies entirely in the discarded variables, will go undetected. To find such PCs we use Rao's approach on principal components of instrumental variables [18].

In the second methodology a cost-utility analysis is used to compare different variable subsets that might be used for process monitoring. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. To do this, to any process monitoring procedure is assigned a score calculated as ratio of the cost at which it might be obtained to explained variance that it might emanate. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables.

2 Method 1: Variable Selection with Pre-assigned Roles

2.1 Formulation

In the first stage of a manufacturing process a large number of variables might be available. Then, a smaller number of measurements should be selected for process monitoring. In what follows we suppose that $\mathbb{X} = (X_1, X_2, \ldots, X_m)$ is the vector of initial stage measured variables, with mean μ and covariance matrix Σ. We collect n observations and let X be the $n \times m$ matrix of in-control data. When a large number of measurements are available, it is natural to investigate whether they could be replaced by a fewer number of variables. In the proposed methodology we assume that a two-class system is used to classify the variables as *primary* and *secondary* based on different criteria. For example according to some measurement cost criteria some variables may be easier and cheaper to carry out then others or some variables may be more efficient in waste reduction because their measurement are made at earlier points in the process. Without loss of generality let $\mathbb{C}_1 = (X_1, X_2, \ldots, X_p)$ and $\mathbb{C}_2 = (X_{p+1}, \ldots, X_m)$ be the sets of primary and secondary variables respectively. We may write $\mathbb{X} = (\mathbb{C}_1, \mathbb{C}_2)$. Our goal is to find a subset $\mathbb{X}_1$ of c primary variables ($c \leq p$), which best in some sense represents the whole set of original variables $\mathbb{X}$. PCs that are based on the selected subset of primary variables are suggested for this purpose as an appropriate tool for deriving low-dimension subspaces which capture most of the information of the whole data set. For the case

$\mathbb{C}_1 = \mathbb{X}$, several selection methods have been suggested in different contexts (see for example [19, 20, 3, 17, 15, 21, 16, 22, 23]). Suppose that $\mathbb{X}_1$ is the selected subset of primary variables and similarly $\mathbb{X}_2$ the subset of remaining variables. We may write $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$. Let (μ_1, Σ_{11}) and (μ_2, Σ_{22}) denote the location scale parameters of $\mathbb{X}_1$, and $\mathbb{X}_2$ respectively. We have the following expressions for μ and Σ

$$\mu = (\mu_1, \mu_2)\ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{29.1}$$

Consider a transformation:

$$Y = \mathbb{X}_1 A \tag{29.2}$$

where A is a matrix of rank q. The residual dispersion matrix of X after subtracting its best linear predictor in terms of Y is

$$\Sigma_{res} = \Sigma - \Theta_1^t A\left(A^t \Sigma_{11} A\right)^{-1} A^t \Theta_1 \tag{29.3}$$

where $\Theta = (\Sigma_{11}, \Sigma_{12})$.

In this article we propose a variable selection procedure based on PCs, which are computed as linear combinations of selected subset, but are optimal with respect to a given criterion measuring how well each subset approximates all variables including those that are not selected. For a given q we wish to determine A such that the predictive efficiency of Y for X is maximum. Using as overall measure of predictive efficiency the trace operator we have the following solution: the columns of matrix A consist of q first eigenvectors of the following determinant equation:

$$\left|\left(\Sigma_{11}^2 + \Sigma_{12}\Sigma_{21}\right) - \lambda\Sigma_{11}\right| = 0 \tag{29.4}$$

Assuming that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_c$ are the ordered eigenvalues and denoting by $\alpha_1, \alpha_2, \ldots, \alpha_c$ the associated eigenvectors, the matrix A is given as following $A = \left(\alpha_1, \alpha_2, \ldots, \alpha_q\right)$, [18].

2.2 *Variability Evaluation Criteria*

There are several measures to summarize the overall multivariate variability of a set of variables. The choice of indices will depend on the nature and goals of specific aspect of data analysis but the most popular ones are based on trace operator, generalized variance and squared norm of the dispersion matrix. Al-Kandari and Jolliffe [24, 25], have investigated and compared the performance of several selection methods and their results showed that the efficiency of selection methods is dependent on the performance criterion. Furthermore they noted that it may be

not wise to rely on a single method for variable selection. In practice it is necessary to know how well Y approximates the whole data set X. A suitable criterion for this purpose is the proportion of variability explained by the best q space spanned by the selected subset $\mathbb{X}_1$ given by:

$$RX = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{trace\,(\Sigma)} \tag{29.5}$$

Classical Principal Components Analysis, PCA, results guarantee that the maximum value of the right hand of (29.5) is attained for $\mathbb{X}_1 = \mathbb{X}$. The index RX is useful to quantify how much information the selected variables have about the whole set of variables. However, it does not tell us how much information the selected variables have about the unselected ones. This information cannot be found in Σ_{res} but it can be found in conditional covariance matrix of subset $\mathbb{X}_2$ given Y, denoted as $\Sigma_{\mathbb{X}_2/\text{Y}}$ given by:

$$\Sigma_{\mathbb{X}_2/Y} = \Sigma_{22} - A^t \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} A \tag{29.6}$$

We then propose the use of a second variability evaluation criterion defined as:

$$RX_2 = 1 - \frac{\lambda'_1 + \lambda'_2 + \cdots + \lambda'_{m-c}}{trace\,(\Sigma_{22})} \tag{29.7}$$

where $\lambda'_1, \lambda'_2, \ldots, \lambda'_{m-c}$ are eigenvalues of $\Sigma_{\mathbb{X}_2/\text{Y}}$. The criterion RX_2 is similar to index R_{EX} defined in [6]. It grows both with the variance of the selected variables as well as with the variance of the unselected ones explained by the selected variables. If RX_2 is near zero it shows that the subspaces spanned by $\mathbb{X}_1$ and $\mathbb{X}_2$ are almost orthogonal and the sets of variables $\mathbb{X}_1$ and $\mathbb{X}_2$ describe different phenomena of the same process. Therefore a shift in the unselected variables could not be detected by the selected subset. Conversely, a high RX_2 value will guarantee that the selected variables may provide a SPC almost as efficient as if we were monitoring all m variables.

2.3 Variable Selection Algorithm

Various variable selection procedures might be applied to select relevant primary variables and then find PCs which are based on them but represent well the whole set of variables. Here we propose a backward elimination scheme:

1. Compute dispersion matrix of the whole data set.
2. Based on $\mathbb{C}_1$ calculate PCs that explain well the whole set of original variables $\mathbb{X}$, [18].
3. Looking carefully at eigenvalues and the cumulative proportions, determine the number of PCs to be used.

4. Remove each one among the p variables in $\mathbb{C}_1$ in turn, and solve p eigenvalue problems, (29.4), with (p-1) variables.
5. Find the best subset of size (p-1) according to selection criterion that is used and remove the corresponding variable.
6. Put p = (p-1) and continue backward elimination till stopping criteria are satisfied.

When selection procedure is stopped we have obtained the selected subset of primary variables $\mathbb{X}_1$.

3 Method 2: Variable Selection with Cost-Utility Analysis

3.1 Variance Recovery Cost Index

At this point in time, variable selection methods for process monitoring have focused mainly on the explained variance performance criteria. However, explained variance efficiency is a minimal notion of optimality and does not necessarily result in an economically desirable selected subset, as it makes no statement about the measurement cost or other engineering criteria. Without measuring cost many decisions will be impossible to make. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. To do this, to any process monitoring procedure is assigned a score calculated as ratio of the cost at which it might be obtained to explained variance that it might provide. Then, the ratio scores are compared to define the best economically desirable selected subset.

Let $\mathbb{X}_1$ be the selected subset under conditions examined and $F(\mathbb{X}_1)$ their associated cost, given by

$$F(\mathbb{X}_1) = \sum_{X_j \in \mathbb{X}_1} f_j \tag{29.8}$$

where f_j is the cost for measurement X_j.

To compare one use of resources with other possible uses, we propose variance recovery cost index, noted CR. The equation for CR is

$$CR = F(\mathbb{X}_1)/R \tag{29.9}$$

where $F(\mathbb{X}_1)$ is the cost and R is the variance recovery across conditions examined, for example given by (29.5) or (29.7). CR score attempts to define, how much, each unit of explained variance costs. Variable subsets for process monitoring can be ranked according to CR values. This allows easy comparison across different selected variable subsets, but still requires value judgments to be made about the quality of explained variance across the structure and information carried by the full set of original variables.

4 Control Charts Based on Influence Function

4.1 Influence Function

We assume that under a stable process the distribution of $\mathbb{X}$ is F, ideally multivariate normal. When special causes are present in the process $\mathbb{X}$ has an arbitrary distribution noted G. A distribution function which describes the two sources of variation in a process is the contaminated model, [7], given by:

$$\mathrm{F}_{\varepsilon \mathrm{H}} = (1 - \varepsilon)\,\mathrm{F} + \varepsilon \mathrm{G} \tag{29.10}$$

with $(0 \leq \varepsilon \leq 1)$.

If process is under control we have $\varepsilon = 0$. When process is not stable, roughly a proportion ε of output subgroups will be contaminants.

Let $\mathrm{T} = \mathrm{T(F)}$ be a statistical functional. The influence function $IF(x, T, F)$ of the statistical functional T at F is defined as the limit as $\varepsilon \to 0$ of

$$\{T\,[(1 - \varepsilon)\,F + \varepsilon \delta_x] - T(F)\}\,/\varepsilon \tag{29.11}$$

where δ_x denotes the distribution giving unit mass to the point $x \in R^p$. The perturbation of F by δ_x is denoted as

$$F_{\varepsilon x} = (1 - \varepsilon)\,F + \varepsilon \delta_x \;(0 \leq \varepsilon \leq 1) \tag{29.12}$$

As such the influence function measures the rate of change of T as F is shifted infinitesimally in the direction of δ_x, [26]. The importance about the influence function lies in its heuristic interpretation: it describes the effect of an infinitesimal contamination at point x on the estimate. Our idea is that output segments that have a large influence on monitored parameters show up the time when special causes are present in a manufacturing process. The influence functions may be calculated for almost all process parameters. Therefore, based on influential measures derived from them, multivariate control charts for different process parameters and with different sensitivities are be set up, [12, 27–30].

4.2 Control Charts

Assignable causes that affect the variability of the output do not increase significantly each component of total variance of $\mathbb{X}$. Instead, they may have a large influence in the variability of some components and small effect in the remaining directions. Therefore an approach to design control charts for variability consists to detect any significant departure from the stable level of the variability of each component. Based on $\mathbb{X}_1$ PCs that represent well the whole set of variables are

derived, [18]. To build up control charts one may use either the principal components or the influence functions of eigenvalues of dispersion matrix. The control limits of the proposed control charts are three sigma control limits as in any Shewhart control chart, (for details see [12, 27–30]).

5 Application

5.1 Case Study

The proposed methods will be illustrated by using data from a real production process, which manufactures bumper covers for vehicles. Bumper covers are molded pieces made of durable plastic designed to enhance the look and shape of the vehicle while hiding the real bumper. They are attached to the vehicle with fasteners. The current inspection procedure consists of measurements taken at 24 points. The variables that are measured are holes diameters. To fit well with the automobile's overall holes diameters have tight dimensional tolerances. But not all these variables are equally important according to engineering and economic criteria. Ten among them are very important because their deviations from target values lead to designs with less aesthetic fit of automobile's overall and they are very awkward to handle. Meanwhile for the remaining variables their deviations from target diameters can be handled easily by operators and lead to designs that fit well.

5.2 Variable Selection with Pre-assigned Roles

We applied our proposed variable selection methodology with pre-assigned roles to bumper cover manufacturing process. The number of elements in the sets of primary and secondary variables $\mathbb{C}_1$ and $\mathbb{C}_2$ are 10 and 14 respectively. In this article we used a backward elimination scheme, which at each step eliminates the less informative variable among the primary variables that have not yet been eliminated. The new variable was eliminated by its inability to supply complementary information for the whole set of variables. The subset of relevant variables was selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables, thereby providing a SPC almost as efficient as we were monitoring all original variables. The results showed that efficient monitoring of this process according to criterion RX in (29.5) could be attained by using only six primary variables. Shewhart control charts of influence functions of eigenvalues for the covariance matrix were used to monitor components of process variability. These influential control charts, accompanied with process logbook gave clear indications for all known assignable causes present in the process.

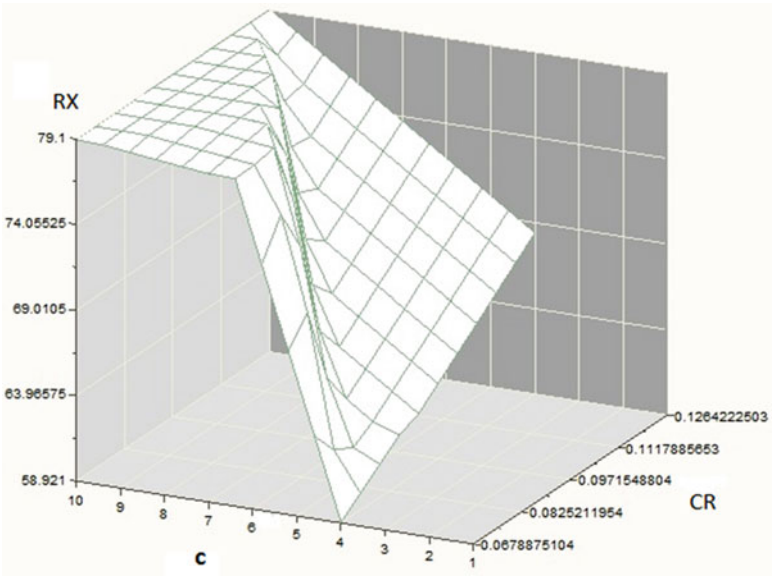


Fig. 29.1 Surface plot with Cartesian coordinates (c, CR, RX)

5.3 Variable Selection with Cost-Utility Analysis

To illustrate variable selection with cost-utility analysis we used only the measurements of ten primary variables. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. Variable subsets for process monitoring were ranked according to CR values in (29.9). Based on data from the cover bumper process, a surface plot with Cartesian coordinates (c, CR, RX) is displayed in Fig. 29.1. A subset of relevant variables that retains, to some extent, the structure and information carried by the full set of original variables should have high RX values and low CR and c values. This allows easy comparison across different selected variable subsets, but still requires value judgments to be made about the quality of explained variance across the structure and information carried by the full set of original variables. An inspection of surface plot in Fig. 29.1 shows that an effective use of resources is obtained for a selected subset with six primary variables.

6 Conclusion and Future Work

This article proposes two new methods to select a reduced number of relevant variables for multivariate statistical process control that makes use of engineering, cost and variability evaluation criteria. In the first method a double reduction of

dimensionality is applied to select relevant primary variables that represent well the whole set of variables. In the second methodology a cost-utility analysis is proposed to compare different variable subsets that may be used for process monitoring. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables. This strategy mitigates the risk that an assignable cause inducing a shift, that lies entirely in the discarded variables, will go undetected. Just like ordinary PCA the solution of the eigenvalue problem in (29.4) is not scale invariant, and therefore sometimes it is better to apply the above method to standardized data rather than raw data. In such cases the covariance matrices in their formulation are replaced by the corresponding correlation matrices.

References

1. Colosimo, B.M., Semeraro, Q., Pacella, M.: Statistical process control for geometric specifications: on the monitoring of roundness profiles. J. Qual. Technol. **40**, 1–18 (2008)
2. Woodall, W.H., Spitzner, D.J., Montgomery, D.C., Gupta, S.: Using control charts to monitor process and product quality profiles. J. Qual. Technol. **36**, 309–320 (2004)
3. Gonzalez, I., Sanchez, I.: Variable selection for multivariate statistical process control. J. Qual. Technol. **42**(3), 242–259 (2010)
4. Jackson, J.E.: A User's Guide in Principal Components. Wiley, New York (1991)
5. Sullivan, J.H., Woodall, W.H.: A comparison of multivariate control charts for individual observations. J. Qual. Technol. **28**(4), 398–408 (1996)
6. Woodall, W.H., Montgomery, D.C.: Research issues and ideas in statistical process control. J. Qual. Technol. **31**(4), 376–386 (1999)
7. Mason, R.L., Chou, Y.-M., Sullivan, J.H., Stroumbos, Z.G., Young, J.C.: Systematic patterns in $T2$ charts. J. Qual. Technol. **35**(1), 47–58 (2003)
8. Mason, R.L., Chou, Y.M., Young, J.C.: Detection and interpretation of a multivariate signal using combined charts. Commun. Stat. Theory Methods **40**, 942–957 (2011)
9. Mason, R.L., Tracy, N.D., Young, J.C.: Decomposition for multivariate control chart interpretation. J. Qual. Technol. **27**(2), 99–108 (1995)
10. Kourti, T., MacGregor, J.F.: Multivariate SPC methods for process and product monitoring. J. Qual. Technol. **28**, 409–428 (1996)
11. Kourti, T., Nomikos, P., MacGregor, J.F.: Analysis, monitoring and fault diagnosis of batch processes using multiblock and multi-way PLS. J. Process Control **5**, 277–284 (1995)
12. Jaupi, L.: Multivariate control charts for complex processes. In: Lauro, C., Antoch, J., Esposito, V., Saporta, G. (eds.) Multivariate Total Quality Control, pp. 125–136. Springer Physica Verlag Heidelberg (2001)
13. García-Muñoz, S., Kourti, T., MacGregor, J.F., Mateos, A.G., Murphy, G.: Troubleshooting of an industrial batch process using multivariate methods. Ind. Eng. Chem. Res. **42**, 3592–3601 (2003)
14. Ferrer, A.: Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process. Qual. Eng. **19**(4), 311–325 (2007)
15. Jolliffe, I.T.: Discarding variables in a principal component analysis I: artificial data. Appl. Stat. **21**, 160–173 (1972)
16. Jolliffe, I.T.: Principal Components Analysis, 2nd edn. Springer, New York (2002)

17. Krzanowski, W.: Selection of variables to preserve multivariate data structure, using principal components. Appl. Stat. **26**, 22–33 (1987)
18. RAO, C.R.: The use and interpretation of principal components in applied research. Sankhya, A **26**, 329–358 (1964)
19. Cadima, J.F.C.L., Jolliffe, I.T.: Variable selection and the interpretation of principal subspaces. J. Agric. Biol. Environ. Stat. **6**, 62–79 (2001)
20. Cumming, J.A., Wooff, D.A.: Dimension reduction via principal variables. Comput. Stat. Data Anal. **52**, 550–565 (2007)
21. Jolliffe, I.T.: Discarding variables in a principal component analysis II: real data. Appl. Stat. **22**, 21–31 (1973)
22. McCabe, G.P.: Principal variables. Technometrics **26**, 137–144 (1984)
23. Tanaka, Y., Mori, Y.: Principal component analysis based on a subset of variables: variable selection and sensitivity analysis. Am. J. Math. Manag. Sci. **17**(1 & 2), 61–89 (1997)
24. Al-Kandari, N.M., Jolliffe, I.T.: Variable selection and interpretation of covariance principal components. Commun. Stat. Simul. Comput. **30**, 339–354 (2001)
25. Al-Kandari, N.M., Jolliffe, I.T.: Variable selection and interpretation in correlation principal components. Environmetrics **16**, 659–672 (2005)
26. Hampel, F.R., Ronchetti, E.M., Rousseew, P.J., Stahel, W.A.: Robust Statistics – The Approach Based on Influence Functions. Wiley, New-York (1986)
27. Jaupi, L., Saporta, G.: Using the influence function in robust principal components analysis. In: Morgenthaler, S., Ronchetti, E., Stahel, W.A. (eds.) New Directions in Statistical Data Analysis and Robustness, pp. 147–156. Birkhäuser Verlag, Basel (1993)
28. Jaupi, L., Herwindiati, D., Durand, P., Ghorbanzadeh, D.: Short run multivariate control charts for process mean and variability. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2013, WCE 2013, pp. 670–674. London, 3–5 July 2013
29. Jaupi, L., Durand, P., Ghorbanzadeh, D., Herwindiati, D.E.: Multi-criteria variable selection for process monitoring. In: 59th world statistical congress, pp. 3550–3555. Hong Kong, Aug 2013
30. Jaupi, L.: Variable selection methods for multivariate process monitoring. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 572–576. London, 2–4 July 2014

Chapter 30
Grouping of Visible Terminals for Achieving High Throughput of Wireless Networks

Kengo Michishita and Yasushi Wakahara

Abstract Wireless networks have been widely and intensively used on a global scale. In accordance with such a use, the limitation of the total throughput of an access point (AP) for the network has become a problem. This problem is especially serious when the AP is accessed by a lot of wireless terminals because the total throughput of the AP becomes much smaller than expected due to the collision of accesses by the terminals. Although a lot of researches have been conducted to solve this problem, their results are far from satisfactory since the total throughputs achieved by the researches are rather limited. Thus, the purpose of this paper is to propose a novel approach to achieve high total throughput regardless of the number of terminals accessing an AP. This paper clarifies the main principles of the proposed approach with its detailed algorithm for the access control. Furthermore, this paper demonstrates that the throughput achieved by the proposed approach is much higher than that of the current IEEE standards and that the throughput by the proposed approach is close to its theoretical upper limit through some computer simulation results.

Keywords Access control • IEEE 802.11 • Terminal grouping • Throughput • Visible terminal • Wireless network

1 Introduction

Wireless networks based on IEEE 802.11 standards have been widely and intensively used on a global scale because of their usefulness and low cost in accordance with the progress of information society in the twenty-first century [1, 2]. Wireless networks can be operated in two modes. One is infrastructure mode and the other is ad hoc mode. In infrastructure mode, an access point (AP) is generally connected with a wired network and the AP is to be accessed by one or more wireless terminals. In ad hoc mode, the wireless network generally consists of wireless terminals and

K. Michishita (✉) • Y. Wakahara
Graduate School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku 113-8658, Tokyo, Japan
e-mail: k.michishita@cnl.t.u-tokyo.ac.jp; wakahara@nc.u-tokyo.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_30

the communication between any two of the terminals can be made either directly or through one or more other terminals. Some wireless networks in ad hoc mode are called MANET (mobile ad hoc network) where the terminals are assumed to make moves.

Wireless networks in both modes suffer from the deterioration of total throughput depending on various environment conditions, especially when there are a large number of wireless terminals [2–5]. And a lot of researches have been globally performed to solve or mitigate this throughput problem. One of the main differences between the two modes is whether the communication is of single or multi hop. In ad hoc mode, the communication between two terminals are often of multi-hop and the total throughput of such multi-hop communication has been widely and deeply studied [6] and as the result of the related researches a novel control method has recently been proposed to achieve high throughput that is very close to its theoretical upper limit [7, 8]. Thus, the research for achieving high throughput for wireless multi-hop communication is considered to have become mature enough.

Meanwhile, the research for achieving high throughput of a wireless network with an AP accessed by more than one terminal is not considered mature enough, since the total throughput achieved by the conventional access control methods is much lower than its theoretical upper limit. Considering that many of the currently widely used wireless networks are operated in infrastructure mode and that the level of the research for enhancing the throughput of wireless networks in infrastructure mode is not mature, this paper aims at achieving high throughput of a wireless network in infrastructure mode.

A lot of researches have been conducted in the literature to try to solve this throughput problem and they are presented and published in various conferences and journals [2–17]. Basic technologies for the access control of wireless terminals are distributed coordinated function (DCF) and point coordinated function (PCF), both of which are defined in IEEE 802.11 standards. The throughput by DCF becomes generally higher than that by PCF when the number of terminals is small because of the relatively larger polling overhead by PCF. Meanwhile, the throughput by PCF becomes higher than that by DCF when the number of terminals becomes large due to the increase in the collision probability of accesses by the terminals and also to the overheads of control frames such as RTS (request to send) and CTS (clear to send) in DCF. In the followings, some of the related researches, which are considered most representative, are described and discussed in concrete, and it is concluded that there is a strong need to invent a new access control method for a wireless network in infrastructure mode.

A simple approach to achieve high throughput regardless of the number of terminals is a hybrid combination of DCF and PCF [9]. This method defines a super-frame that consists of contention period for DCF and contention free period for PCF. The AP always monitors the accesses by terminals and calculates the throughputs that are achieved by DCF and PCF respectively in each super-frame and adjusts the durations of DCF and PCF in accordance with their throughputs in such a manner that the overall total throughput becomes higher in the following super-

frame. Basically this method takes DCF when the number of terminals is small and it takes PCF when the number is large. As easily understood, the throughput of this hybrid method is in principle limited by the throughputs of either DCF or PCF, and this method cannot achieve higher throughput than either DCF or PCF. As such, this method cannot overcome the basic problem of low total throughput of wireless networks in nature.

Another approach is to schedule the accesses of the terminals in a distributed manner. High performance DCF (HDCF) [10] is a method taking this approach and when a terminal accesses an AP and transmits a data frame, the frame indicates another terminal that is allowed to access the AP immediately after the end of the frame transmission followed by a predetermined guard time DIFS. HDCF can avoid overheads due to the exchange of RTS/CTS and also the backoff mechanism in DCF. Accordingly HDCF can achieve higher throughput than other conventional access control methods. In principle, HDCF can be made use of only if there are no hidden terminals in the coverage of the AP because each terminal needs to know and identify all the other terminals that may access the AP. In practice, however, there are many hidden terminals usually since the coverage areas of terminals are different from each other. In other words HDCF is not considered practical for a general wireless network.

There is an access control method based on the grouping of terminals and only the terminals of one group are allowed to access the AP at a time and the access opportunities for each group are given in turns [11]. This method is effective in mitigating the collision probability of terminals since the number of terminals that can access the AP at the same time becomes smaller in accordance with the number of groups. However, this method in practice cannot achieve much higher total throughput than expected because either the collision probability of terminals is not so small or the overhead due to the sequential control related to the turns of groups for the access to the AP is not so small depending on the number of groups and the numbers of the terminals in the individual group.

As described above, although there are lots of researches aiming at higher total throughput for a wireless network, the access control methods obtained as the results of these researches have not succeeded in solving a serious problem of low total throughput. The purpose of this paper is to propose a new method to achieve much higher throughput than DCF, PCF and other conventional access control methods and this proposed method aims at high total throughput that is close to its theoretical upper limit.

The rest of this paper is organized as follows. Section 2 presents a new access control method for an infrastructure mode wireless network to achieve high total throughput. Section 3 evaluates the throughput of the proposed method and demonstrates its superiority to the conventional methods and also the closeness of its throughput to the theoretical upper limit. Section 4 concludes this paper with some suggestions of its future extension.

2 Proposal of a New Access Control Method

As mentioned in Sect. 1, the deterioration of total throughput of a wireless network is due to time delay caused by the collision of frames and their retransmissions by the AP and the wireless terminals, and also by the control frame overheads of RTS and CTS exchanged by the AP and the terminals. These collisions are usually made by hidden terminals and it is important to reduce the probability of collision due to hidden terminals and to avoid the redundant or useless retransmission of frames in order to enhance the total throughput especially when the number of terminals becomes larger. It is also important to make RTS/CTS exchange unnecessary to reduce the related traffic overheads. This paper proposes a novel access control method to achieve such reduction and avoidance of collision and also to make RTS/CTS exchange unnecessary based on the following principles.

1. grouping of the terminals which are visible to each other so that there are no hidden terminals in each group, and selection of a group of terminals one by one for the AP access
2. application of DCF without exchange of RTS/CTS to the accesses by the terminals in each group
3. duration of the access by terminals in each group in proportion to the number of the terminals

Hereinafter, the following assumptions are made.

- The access control is performed by the AP in infrastructure mode based on IEEE 802.11 standards.
- The AP and all the terminals have non-directional antenna and their radio transmission ranges are all circle of the same size.
- The interference and the carrier sensing ranges are the same as the transmission range for the AP and every terminal.

2.1 Grouping of Terminals

An example of the terminal grouping where every group consists of only terminals visible to each other and there are no hidden terminals in every group is shown in Fig. 30.1. In Fig. 30.1, there are three groups of terminals and Groups 1, 2 and 3 have 3, 1 and 2 terminals, respectively, and the range of only the AP is depicted for simplicity. Since the AP and all the terminals have the same size transmission range, it is clear that there are no hidden terminals in each group. For example, Group 1 consists of terminals A, B and C, and each of the three terminals are within the transmission ranges of the other two terminals, and these three terminals are visible to each other. In other words, there are no hidden terminals in Group 1.

Theoretically, the number of groups can be made at most 6. Figure 30.2 shows the division of the transmission range of an AP into six areas of the same size and shape,

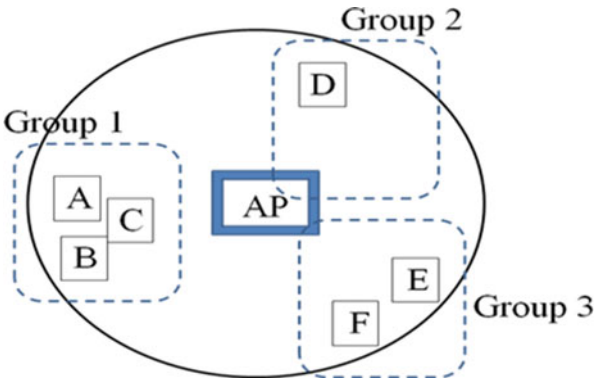


Fig. 30.1 An example of grouping of terminals

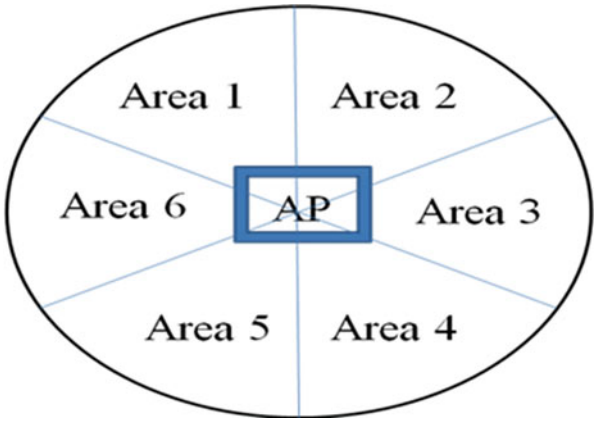


Fig. 30.2 Division of transmission range of an AP into six areas

where the central angle of each area is all 60°. The terminals in the transmission range of the AP are grouped so that every terminal in each of the six areas belongs to a same group. In this grouping, the longest distance between any two terminals in an area is equal to or less than the radius of the transmission range. Therefore all the terminals in each area are within the transmission range of each other and visible to each other, and thus there are no hidden terminals in each group.

In general, the larger the number of groups, the overhead for switching the turns of access among the groups becomes larger. The larger the number of terminals in each group, the collision probability by the terminals in individual group becomes larger. Thus, there is a tradeoff between the number of groups and the number of terminals in the individual group from the viewpoint of the access collision and the traffic overhead. However, the overhead for the switching is very small since the control frame used for polling the following group is of small size. Furthermore, the collision probability by the terminals in a group is also very small since there are no hidden terminals in each group and collision can be made only by two or more terminals which have the same backoff time that is the minimum among the terminals in the group because each terminal always performs carrier sensing

function. Thus, there is not a strong need to optimize the number of groups and the number of member terminals in individual group.

2.2 Visible Terminal Table (VTT)

In the proposed access control method, every terminal is equipped with a table named visible terminal table (VTT). VTT maintains a set of terminals that are within the transmission range of the owner terminal of the VTT and therefore all the terminals in the VTT are visible to its owner terminal. The VTT is constructed and maintained in the following manner.

In IEEE 802.11 standards, when a data frame is received by the AP or a terminal successfully, an ACK frame is always returned by the AP or the terminal. Thus, when a terminal overhears the exchange of a data frame and an ACK frame between the AP and another terminal, then the former terminal recognizes that the latter terminal is within the transmission range of the former terminal and is active and visible to the former terminal. In this manner, it is possible for a terminal to obtain the list of the terminals that are active within its transmission range and visible, and accordingly it is possible for a terminal to construct and maintain the VTT.

The time horizon by the proposed access control method is divided into super-frames and each super-frame in principle consists of access duration for each group terminals. Duration of a super-frame can be decided by considering the delay in the exchange of frames between the AP and the terminals. The delay can be designed depending on the number of terminals and also the trade-off between the processing burden for updating the grouping of terminals and the accuracy of the VTT information used for the update. The information in the VTT of a terminal is transmitted to the AP via a data frame from the terminal to the AP once per super-frame as far as the VTT is updated in the previous super-frame.

Figure 30.3 shows the access timings in more detail. Access can be made by not only terminals to the AP but also by the AP to terminals. Furthermore, if a new terminal comes into the transmission area of the AP and becomes active, then the terminal should be able to access the AP before the terminal into periods not only for each group and the AP but also for other terminals that do not belong to any groups. These terminals that do not belong to any groups will become members of some groups once they access the AP in the following super-frame based on the mechanism related to the VTT mentioned above.

2.3 Algorithm for Grouping of Terminals

The AP collects and maintains all the latest information in the VTTs of all the terminals in every super-frame. Making use of the information, the AP groups the terminals according to the algorithm whose flow chart is shown in Fig. 30.4 and updates the groups whenever the information is updated.

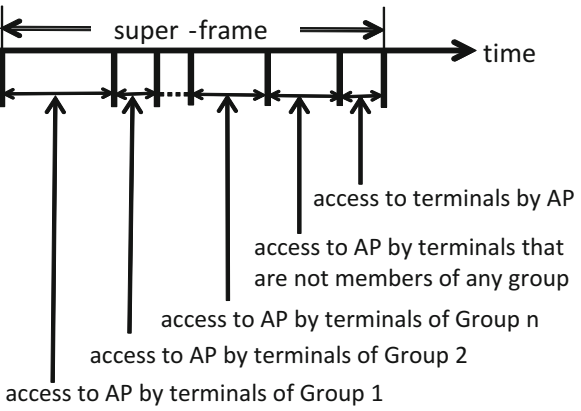


Fig. 30.3 Access by the terminals and the AP

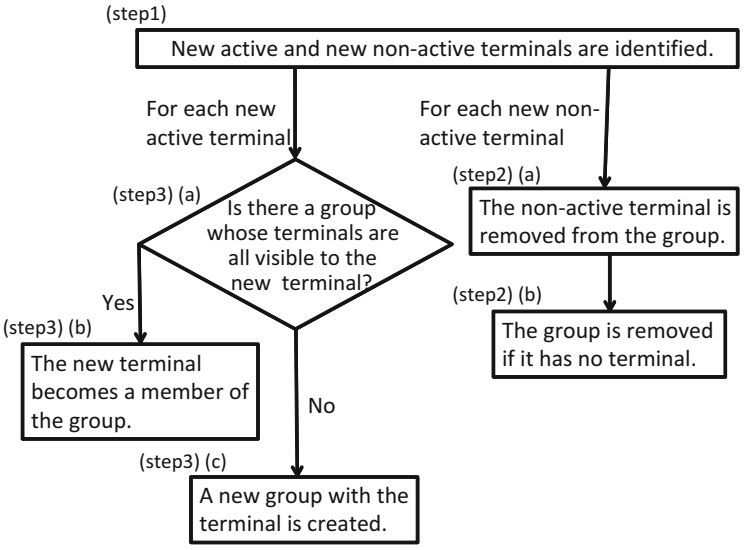


Fig. 30.4 Flow chart of the grouping algorithm

It is possible to modify (step1) so that the judgment on the state change of a terminal from active to non-active is made only after the terminal has not exchanged any data frames during the last and more super-frames in order to avoid misjudgment: e.g. the disability of frame transmission by a terminal due to congestion of frame traffic instead of the real non-activeness of the terminal.

2.4 Access Control by the Proposed Method

The AP of a wireless network performs grouping of terminals according to the algorithm described above by the proposed access control method. As its result, the AP determines the time duration of access by the terminals of each group. The time duration should be determined in such a manner to achieve the fairness among the terminals. The fairness can be defined in different ways depending on the policy of the management of the wireless network. Since the management policy is not a subject of this paper, the fairness is simply defined in this paper in accordance with the number of active terminals in the groups as formulated below, where the active terminals are defined as the terminals that have data to transmit.

$$T_i = T \times A_i / A,$$

where

- T_i and T denote the access time duration assigned to group i and the time duration of a super-frame which is equal to a period of grouping cycle, respectively and
- A_i and A denote the number of active terminals in group i and their total sum, respectively.

At the start of access duration of each group, the AP broadcasts a control message, similar to a poll in PCF, indicating the group and its member terminals. On receiving the control message, all the active terminals in the designated group will start contending for the transmission of frames based on DCF. Since there are no hidden terminals in the group, the possibility of collision by the terminals is very low because of carrier sensing function of every terminal, and thus RTS/CTS exchange becomes unnecessary in the proposed method, which leads to the efficient use of the time and accordingly to high total throughput.

After the access duration ends of a group, the AP broadcasts another control message designating another group. By sequentially designating different groups by the control messages of the AP, it becomes possible that every group will be given an opportunity to transmit data frames alternatively and sequentially.

It should be noted that the number of active terminals of each group is defined as the number in the previous period. By such definition, there can be some error in the number of active terminals of a group in practice and in order to avoid wasting of time due to the error, the AP changes the turns of group access by broadcasting a new control message automatically if there is no access by the terminals in the group for a predefined interval time.

3 Evaluation of Proposed Access Control Method

3.1 Evaluation Method and Simulation Scenario

The total throughput of the proposed access control method for a wireless network in infrastructure mode is simulated by making use of a network simulator NS-2 [18]. For the comparison, DCF with RTS/CTS, DCF without RTS/CTS, PCF and Hybrid access control method [9] are also simulated. The simulation is repeated for 20 times with randomly selected different locations of terminals to show the average value of total throughputs with their 95% confident intervals. It should be noted that the terminals do not make any moves in the simulation for simplicity reason. The total throughput is defined as the total number of user data bits in the data frames exchanged per second between every one of the terminals and the AP.

The main parameters and their values in the simulation scenario are summarized in Table 30.1.

In the simulation, data frames are generated by every active terminal at the rate sufficiently high so that there is no waiting time for the generation before the start of transmitting the frames by the terminal. Data frame transmissions are initiated by only the terminals and the simulation of the case where the AP initiates the

Table 30.1 Simulation scenario

Parameters	Values
IEEE standard	802.11 g
Antenna	Non-directional
Wave propagation model	Two ray ground
Transport layer protocol	UDP (user datagram protocol)
UDP application	CBR (continuous bit rate)
Packet size	1,500 byte
Contention window size	31–1023
Data generation rate	20 Mb/s
PLCP data rate	1 Mb/s
Basic rate	1 Mb/s
Data frame rate	54 Mb/s
Duration of a super-frame	0.8 s
Transmission/interference/carrier sensing range	200 m
Size of preamble and PLCP_header	192 byte
Size of MAC frame	1,578 byte
Size of RTS	44 byte
Size of CTS	38 byte
Size of ACK	38 byte
Size of control frame	44 byte
Duration of SIFS	10 μs
Duration of slot	20 s

transmission of data frames has not been performed in the following evaluation since the simulation results of the total throughput are considered basically the same regardless of whether the data frame transmission is initiated by the terminals or the AP.

3.2 Evaluation Results and Their Discussions

Figures 30.5 and 30.6 show the total throughputs versus the number of active terminals when there are 40 and 4 terminals in the transmission range of an AP, respectively. The solid lines and the dotted lines denote the total throughputs obtained as the results of the simulation and their theoretical upper limits, respectively.

The theoretical upper limits of the total throughputs are derived assuming no collision of accesses by the terminals. As an example, the theoretical upper limit by the proposed method Th_{p} is calculated by the following formula:

$$Th_{\mathrm{p}} = p\_size/\left(T_{\mathrm{DIFS}} + T_{\mathrm{backoff}} + T_{\mathrm{data}} + T_{\mathrm{SIFS}} + T_{\mathrm{ACK}}\right),$$

where

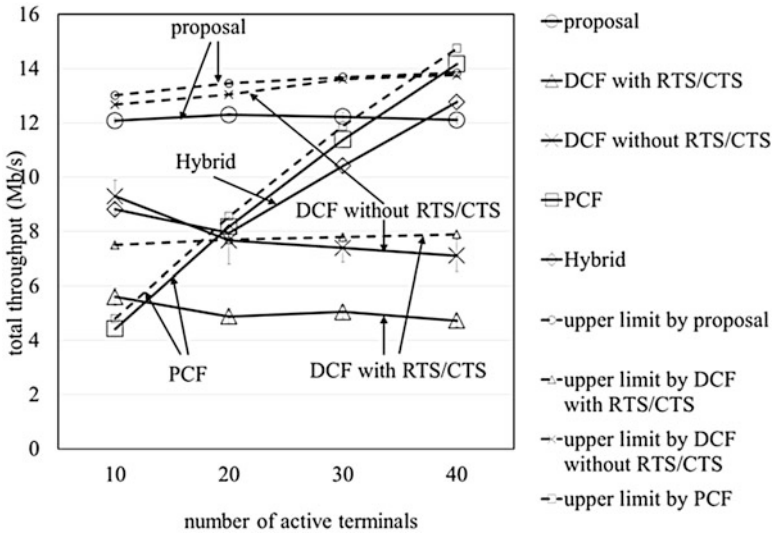


Fig. 30.5 Throughput of various access control methods versus the number of active terminals out of 40 terminals

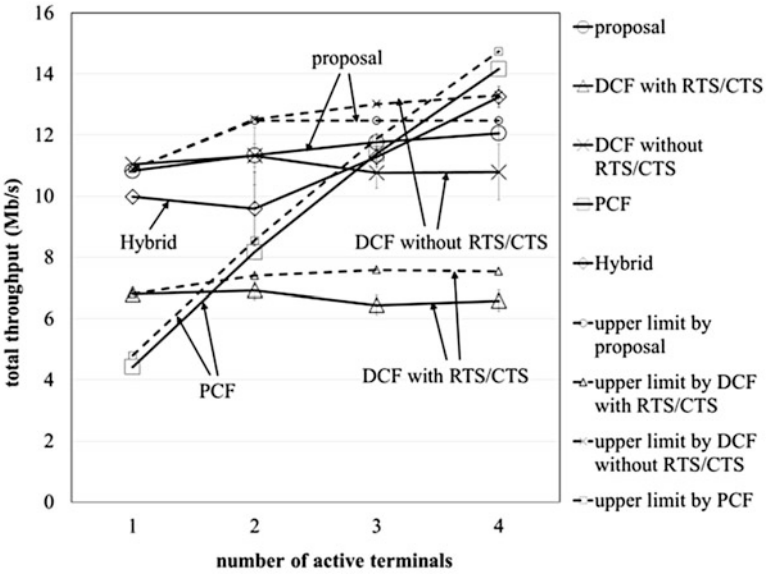


Fig. 30.6 Total throughput of various access control methods versus the number of active terminals out of four terminals

p_size = size of data in bits,
T_{DIFS} = time duration of DIFS
$T_{backoff}$ = time duration of backoff
T_{data} = time duration of data frame,
T_{SIFS} = time duration of SIFS, and
T_{ACK} = time duration of ACK frame.

According to Fig. 30.5, the total throughput of PCF is highest only when the number of active terminals is very close to the total number of terminals and the total throughput of the proposed method is highest in all the other cases. Especially when the number of active terminals becomes smaller, the total throughput of the proposed method is much higher than all the other methods. In general, the number of active terminals is far smaller than the total number of terminals, since each terminal will not continue to be active to transmit or receive data frames for a long time and it usually transmits or receives data frames intermittently in practice. Thus, the proposed method is considered by far the best among the simulated access control methods in terms of the total throughput. It should be noted that the proposed method is characterized by its stable high total throughput regardless of the number of active terminals.

The difference between the total throughput of PCF and its theoretical upper limit is very small because there is no collision of frames in PCF and the difference comes from only the overheads of polling messages. The difference between the

total throughput of the proposed method and its theoretical upper limit is small because the collision probability is very small due to the principle that each group consists of only visible terminals. On the other hand, the total throughput of DCF without RTS/CTS is much higher than that of DCF with RTS/CTS because of no traffic overheads due to RTS/CTS. The difference between the total throughput of DCF without RTS/CTS and its theoretical upper limit is very large because of large collision possibility and the resultant retransmission of frames with longer backoff. Furthermore, the difference between the total throughput of DCF with RTS/CTS and its theoretical upper limit is rather large because of some possibility of collision and the resultant retransmission of frames with longer backoff. The total throughput of Hybrid method is limited by either that of PCF or DCF as expected.

According to Fig. 30.6 with 4 terminals in total, when the number of active terminals is small, the collision probability becomes small even for DCF and therefore DCF without RTS/CTS and the proposed method outperform other methods if the number of active terminals becomes small. However, the feature of stable highest total throughput is maintained by the proposed method with the exception of the case where the number of active terminals is very close to the total number of terminals. Since the probability of such exception is very low in practice as mentioned above, the proposed access control method is concluded as the best in terms of the total throughput of a wireless network in infrastructure mode based on IEEE 802.11 standards.

In the simulation of the proposed method, the number of groups was usually 4 or 5 and the maximum number of groups was 7, which means there is some room to improve the grouping algorithm by making the number of groups at most 6 to minimize the overhead for switching the groups accessing the AP. The throughputs of different terminals were almost the same for all the cases by the proposed method. In other words, the fairness among the terminals can be maintained regardless of the groups by the proposed grouping algorithm.

In the above simulation, it is assumed that only the terminals have data to transmit. In practice, however, the AP is very likely to have data transmit. If the AP is taken as a terminal in the above simulation, the evaluation results become applicable for such a practical case.

4 Conclusion

This paper has addressed a problem of low total throughput for a wireless network with a lot of terminals accessing an AP in infrastructure mode based on IEEE 802.11 standards. The main contribution of this paper is a proposal of a novel access control method by terminals based on groups of only visible terminals and with no hidden terminals, which makes the collision probability by the terminals in each group very small and as such the dominant reason for the low total throughput problem is mostly solved. Some computer simulations were conducted to demonstrate the effectiveness of the proposed method and the total throughput achieved by the

proposed method is close to its theoretical upper limit. It is thus concluded that the proposed access control method for a wireless network is basically a best solution to the throughput deterioration problem. The proposed access control method is characterized by its stable high total throughput regardless of the number of active terminals. This characteristic is practically very useful, since it is not easy to control the number of active terminals and the number can change randomly and arbitrary in actual wireless networks.

Some further study on or extensions to the proposed access control method are as follows to make its applications wider and to achieve even higher total throughput.

- Improvement of the grouping algorithm to optimize the numbers of groups and the terminals of the individual groups.
- Evaluation and extension to the grouping algorithm to cope with moves of the terminals.
- Extension to the grouping algorithm to cover latest IEEE 802.11 standards such as 802.11 ac and 802.11ad [19].
- Evaluation and extension to the access control to cope with the case with more than one AP, where the assignment of different channels should be taken into account.
- Evaluation and extension to the access control to cope with the case where the interference/carrier sensing rage is much larger than the transmission range.
- Evaluation and extension to the access control to cover QoS (quality of service) other than the total throughput such as time delay with its variation and loss probability of frames.

References

1. Michishita, K., Wakahara, Y.: High throughput access control based on groups of visible terminals for wireless network. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 739–744. London, 2–4 July 2014
2. IEEE Standard for Information technology: Telecommunications and information exchange between systems Local and metropolitan area networks–specific requirements Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications. IEEE Std 802.11™-2012
3. Zhai, H., Kwon, Y., Fang, Y.: Performance analysis of IEEE 802.11 MAC protocols in wireless LANs. Wiley Commun. Mob. Comput. **4**, 917–931 (2004)
4. Shah, S.H., Chen, K., Nahrstedt, K.: Available bandwidth estimation in IEEE 802.11-based wireless networks. IEEE Trans. Mobile Comput. **7**(10), 1228–1241 (2008)
5. Eldin, S.S., Nasr, M., Khamees, S., Sourour, E., Elbanna, M.: Performance enhancement of IEEE 802.11n wireless LAN using irregular LDPCC. In: Proceedings of the sixth international conference on wireless and optical communications networks, WOCN'09, pp. 312–316. Cairo, Egypt, 28–30 April 2009
6. Vaze, R., Heath, R.W.: Transmission capacity of ad-hoc networks with multiple antennas using transmit stream adaptation and interference cancellation. IEEE Trans. Inf. Theory **58**(2), 780–792 (2012)

7. Yao, X., Wakahara, Y.: Synchronized multi-hop protocol with high throughput for an IEEE 802.11 multi-hop wireless network. In: Proceedings of IEEE smart communications in network technologies, pp. 136–141. Paris, 17–19 June 2013
8. Yao, X., Wakahara, Y.: Application of synchronized multi-hop protocol to time-variable multi-rate and multi-hop wireless network. In: Proceedings of APNOMS conference, pp. 1–6. Hiroshima, 25–27 Sept 2013
9. James, D.X., Ergen, M., Varaiya, P., Puri, A.: Improving the aggregate throughput of access points in IEEE 802.11 wireless LANs. In: Proceedings of 28th annual IEEE international conference on local computer networks. LCN'03, pp. 682–691. Bonn, Germany, 20–24 Oct 2003
10. Luis, A., Ferrus, R., Agusti, R.: WLAN throughput improvement via distributed queuing MAC. Commun. Lett. IEEE **9**(4), 310–312 (2005)
11. Yasuda, K., Tagashira, S., Fujita, S.: Adaptive MAC scheme for wireless LAN based on dynamic group constructions. In: Seventh international conference on parallel and distributed computing, applications and technologies, PDCAT'06, pp. 300–305. Taipei, Taiwan, 4–7 Dec 2006
12. Al-Mefleh, H., Chang, J.M.: High Performance Distributed Coordination Function for Wireless LANs. Networking. In: Proceedings of 7th international IFIP-TC6 networking conference, pp. 812–823. Singapore, 5–9 May 2008
13. Mare, S., Kotz, D., Kumar, A.: Experimental validation of analytical performance models for IEEE 802.11 networks. In: Second international conference on communication systems and networks (COMSNETS), pp. 1–8. Bangalore, India, 5–9 Jan 2010
14. Yang, B.: Comparison of IEEE802.11 standards. Directed Research Project Report, University Of Alberta (2012)
15. Jansons, J., Barancevs, A., Petersons, E., Bogdanovs, N.: IEEE802.11a standard performance in mobile environment. Int. J. New Comput. Archit. Appl. **2**(3), 496–499 (2012)
16. Wang, Q., Runser, K.J., Scharbarg, J.-L., Fraboul, C., Sun, Y., Li, J.: Delay distribution of IEEE802.11 DCF: a comparative study under saturated conditions. In: Proceedings of the 10th ACM symposium on performance evaluation of wireless ad hoc, sensor, & ubiquitous networks, PE-WASUN, pp. 41–48. Barcelona, Spain, 3–7 Nov 2013
17. Juliet, F.-J.O., Stuart, W.: Empirical performance evaluation of enhanced throughput schemes of IEEE802.11 technology in wireless area networks. Int. J. Wirel. Mob. Netw. **5**(4), 171–185 (2013)
18. The Network Simulator – ns-2. http://nsnam.isi.edu/nsnam/index.php/User_Information
19. Perahia, E., Gong, M.X.: Gigabit wireless LANs: an overview of IEEE 802.11ac and 802.11ad. ACM SIGMOBILE Mob. Comput. Commun. Rev. **15**, 23–33 (2011)

Chapter 31
Experimental Study on RSS Based Indoor Positioning Algorithms

Hélder David Malheiro da Silva, José Augusto Afonso, and Luís Alexandre Rocha

Abstract This work compares the performance of indoor positioning systems suitable for low power wireless sensor networks. The research goal is to study positioning techniques that are compatible with real-time positioning in wireless sensor networks, having low-power and low complexity as requirements. Map matching, approximate positioning (weighted centroid) and exact positioning algorithms (least squares) were tested and compared in a small predefined indoor environment. We found that, for our test scenario, weighted centroid algorithms provide better results than map matching. Least squares proved to be completely unreliable when using distances obtained by the one-slope propagation model. Major improvements in the positioning error were found when body influence was removed from the test scenario. The results show that the positioning error can be improved if the body effect in received signal strength is accounted for in the algorithms.

Keywords Fingerprinting • Linear least squares • Localization • Map matching • Received signal strength • Weighted centroid • Wireless sensor networks

1 Introduction

Localization capability in wireless sensor networks (WSN) brings spatial information to sensor data and enables numerous added value applications. Localization can be used in the most various contexts, from geodesic routing to antenna beam forming, or to detect soil temperature and pinpoint the origin of a wildfire.

In outdoors environment, the global positioning system (GPS) is capable of offering an adequate service to the majority of applications. Device size is no longer an issue in WSN due to the miniaturization of GPS hardware. Remaining disadvantages of this approach relate to energy consumption and node price when using this technology in WSN.

Regarding indoors environment, GPS is not reliable due to the signal attenuation. Ultra-wideband is a technology with potential to solve the problem of indoor

H.D.M. da Silva (✉) • J.A. Afonso • L.A. Rocha
Centro Algoritmi, University of Minho, Guimarães 4800-058, Portugal
e-mail: helderdavidms@gmail.com; jose.afonso@dei.uminho.pt; larocha@dei.uminho.pt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_31

location due to its high accuracy when inferring distances between devices [1]. However, and despite large standardization efforts (e.g., the IEEE 802.15.4a standard), a fully compliant commercial device for sale is unavailable. Since no mass market is currently in place, prices for available proprietary hardware are very high.

Received signal strength (RSS) based positioning is a popular approach in WSNs since RSS is readily available with the radio module. Due to typical WSN energy and computational profiles, low complexity positioning solutions are desired. As such, researchers seek to find balance between accuracy and computational complexity.

This work, which is a revised and extended version of our previous work [2], concerns the implementation of positioning systems (PS) in WSN that best fit the indoor scenario. We experimentally compare positioning calculation using map matching, approximate positioning and exact positioning algorithms in an indoor test scenario. We also study the effect of the body in the performance indicators.

Map matching solutions are mainly used in large areas, such as office settings and warehouses with several divisions. Our work differs from the usual approach, since the fingerprinting solution is implemented in a smaller predefined space of a room, without walls in between access points.

2 Background

2.1 RSS Based Indoor Positioning Systems

An overview of technologies used in positioning systems is available in [3]. Ultrasound, ultra-wideband, radio-frequency identification (RFID) and RSS based systems are among the most used technologies for indoor positioning. Accuracies span from 5 m (RSS) to a few centimeters (ultrasound).

RSS systems are known for the low reliability when inferring distances from measurements. Filtering techniques are a solution for dealing with RSS reliability under noisy environment conditions. These techniques also stand as the common solution for integration of heterogeneous positioning systems, in order to provide more accurate location estimation. Kalman filters [4] and particle filters [5] are the usual approaches; however, since these solutions need high computational capacity, they are usually not compatible with WSNs. Instead, filtering is typically accomplished by averaging multiple measurements, thus positioning accuracy is sacrificed in the tradeoff for lower computational demands, longer lifespan of sensor nodes and faster positioning update rates when desired.

Propagation models are an important topic in RSS based systems, for which [6] presents a general overview. Several efforts have been made to characterize radio signal propagation [7], developing many propagation models. For indoor settings, the one-slope [8] and the multiwall [9] models are frequently used in state-of-the-art. Several types of fading affect these signals [10], where attenuations as high as

15 dBm are reported [11] due to the human body. We refer the reader to our previous work [2] for a more detailed description of this topic concerning the context of this work.

2.2 Map Matching

Two phases compose the system originally implemented by Bahl et al. [12]. In the offline phase, data relating position and RSS from access points (AP) is gathered from the site on to a database, in order to create a radio map. In the online phase, mobile nodes report to a server the RSS from APs in range. The server compares signatures so a match (or the closest to) can be found, thus pinpointing the mobile node's position.

In Ref. [13], a comprehensive study on fingerprinting is presented. Authors conclude that map density translates to higher accuracy with a nonlinear behavior in increasing the number of calibration points. The direction faced when collecting samples, also studied [12], is crucial and greatly improves system accuracy.

Approaches to facilitate creation of radio map in the offline phase have been conducted. Authors in [14] use propagation models to ease the process of creating the radio map. Ray-tracing modeling is another solution to obtain the attenuation values of signal propagation [15].

2.3 Approximate Positioning

The approximate positioning method uses parameters or metrics that can be used to infer proximity to a known location. The weighted centroid localization (WCL) is a well-known approximate positioning method, which presents low complexity and good robustness to noise. Bulusu et al. implemented this method in [16], where node connectivity was the metric used to infer distance. Given a set of beacon nodes in the network possessing knowledge of their location, the position of sensor nodes can be estimated by calculating the centroid of all beacon node coordinates for which the sensor is in range of.

LANDMARC [17] uses RSS readings in their approximate positioning method. Tag readers report RSS from moving RFID tags, along with RSS from reference tags. Reference tags are fixed and their RSS is used as means of comparison between that of the movable tags to infer proximity. In a more recent work [18] authors further improve LANDMARC's positioning error to a 1-m accuracy with a signal reporting cycle of 2 s.

Hop count positioning algorithms such as DV-Hop [19] can use RSS as a metric to infer distance for each hop. In [20], authors achieve less than 10 % radio coverage error.

2.4 Exact Positioning

The exact positioning method involves the determination of angles or distances between a sensor node and multiple known reference points. Triangulation and trilateration (or multilateration) are the typical methods employed to determine the sensor position. Distance estimates are usually obtained by measuring the time of arrival (TOA), time difference of arrival (TDOA) or the round trip time of flight (RTOF) [21].

The linear least squares method (LLS) [22] is the most used exact positioning algorithm in WSNs, due to the simple closed form solution.

Measuring the propagation time of a transmission is a more robust method when compared to measuring the signal strength. However, in WSN this generally involves adding extra hardware, increasing energy consumption. The main goal in this work is the implementation of low power and real-time sensor node positioning in indoor environment. As such, addition of extra hardware is avoided and RSS measurements are performed for distance estimates.

3 Materials and Methods

3.1 Hardware

Texas Instruments CC2530DK development kit was used in this work. We refer the readers to our previous work [2] for a more detailed description.

The test scenario is composed by four anchor nodes and one sensor node. Each anchor node is composed by a CC2530 evaluation module and a battery board powered by two AA batteries. The sensor node is composed by a development board and an evaluation module.

3.2 Experimental Setup

The anchor's role is to broadcast beacon messages periodically, so sensor nodes can receive these messages and locate themselves. Our main test bed is a room with 10×4.7 m free space area, as shown in Fig. 31.1.

Anchors are placed in the corners of the mentioned area on top of a stand, 1.2 m above ground. The stands used are made of plastic, so no extra interferences affect the radio messages.

Numbered from 0 to 3, each anchor broadcasts one beacon message periodically. Using the sequence number in the beacon messages, the sensor node detects lost beacons during data collection and inserts a value of -127, indicating an invalid RSS sample. Calculations are performed in an offline phase.

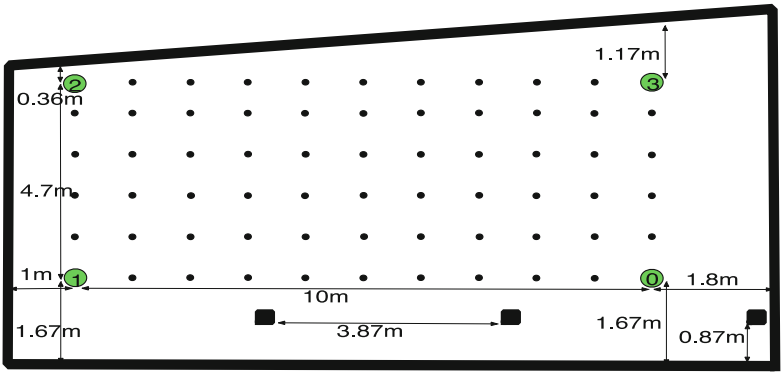


Fig. 31.1 Experimental setup. Anchor locations are depicted as *green circles*, along with distances to walls. *Black dots* indicate calibration points. A calibration point was also taken at each anchor location. The three supporting beams on the *bottom right* side of the figure are depicted as *squares*

The one-slope propagation model used in this scenario was calibrated on site. Details on the calibration method used can be found in [2].

3.3 Map Matching

The radio map was created with a grid resolution of one squared meter. Since our positioning area is 4.7 m wide, the last column of the grid has a smaller resolution of 0.7 m^2. A total of 66 grid points covered our test field. A calibration point was collected at each grid point and for each body orientation (e.g., north, west, south and east), amounting to a total of 264 calibration points. Each point is composed by true position (x and y with origin on anchor 0), body orientation and average RSS obtained from 100 RSS samples from all four anchor nodes.

During the online phase, the sensor node obtains and stores RSS samples. At the end of a test run (e.g.: after collecting 100 samples), data is uploaded to a PC running MATLAB and the position is computed. The weighted k-nearest neighbor (WKNN) algorithm [12] uses (31.1) to find the distance in signal space between a RSS sample and each calibration point.

$$D_{SS} = \left[\sum_{i=1}^{N} \left| R_{map}(i) - R_s(i) \right|^p \right]^{\frac{1}{p}} \tag{31.1}$$

N is the number of anchor nodes in range and p is the norm used. The $R_{map}(i)$ is the RSS stored for anchor i in a calibration point of the radio map and $R_s(i)$ is the RSS sampled in the online phase for anchor i. After computing the distances for all calibration points, the K smallest distances are used to estimate the node's position using (31.2), where p_i is the coordinates of each calibration point.

$$\widehat{x} = \frac{\sum_{i=1}^{K} w_i \times \overrightarrow{p}_i}{\sum_{i=1}^{K} w_i};\ w_i = \frac{1}{D_1} \tag{31.2}$$

The weight applied to each neighbor found in the search process is simply the inverse of the signal space distance.

3.4 Approximate Positioning

In this type of positioning, the only information needed by a node to calculate its position is the coordinates of each anchor node in range. The position estimate is calculated using (31.3):

$$\widehat{x} = \frac{\sum_{i=1}^{B} w_i \times \overrightarrow{L}_i}{\sum_{i=1}^{B} w_i};\ w_i = \left| \frac{1}{(R_p)^e} \right| \tag{31.3}$$

Where L_i is the coordinates of each anchor node and R_p is the radio parameter used to calculate the weight. In this work, both the RSS and the distance using a propagation model were used to calculate the weights, in two different approaches. The exponent e allows an adjustment of the importance of the weight applied to each anchor node's RSS.

3.5 Exact Positioning

The Linear Least Squares method is an exact positioning technique, which computes the position of a node using a set of three or more non-collinear distance measurements (in the two dimensional case). Each measurement produces an equation of the form illustrated in (31.4):

$$(x - x_n)^2 + (y - y_n)^2 = {d_n}^2 \tag{31.4}$$

Several measurements produce a system of equations, which has no solution when circles don't intersect. To find a solution to this system, first a linearization of the system of equations is obtained by subtracting the location of the first anchor node from other locations. This cancels the unknown squared terms, and a linear system of the form $\mathbf{Av} = \mathbf{b}$ is obtained, as shown in (31.5), (31.6) and (31.7):

$$\mathbf{A} = 2 \times \begin{pmatrix} x_1 - x_2 & y_1 - y_2 \\ x_1 - x_3 & y_1 - y_3 \\ \cdots & \cdots \\ x_1 - x_n & y_1 - y_n \end{pmatrix} \tag{31.5}$$

$$\mathbf{b} = \begin{pmatrix} {d_2}^2 - {d_1}^2 + {x_1}^2 - {x_2}^2 + {y_1}^2 - {y_2}^2 \\ {d_3}^2 - {d_1}^2 + {x_1}^2 - {x_3}^2 + {y_1}^2 - {y_3}^2 \\ \cdots \\ {d_n}^2 - {d_1}^2 + {x_1}^2 - {x_n}^2 + {y_1}^2 - {y_n}^2 \end{pmatrix} \tag{31.6}$$

$$\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{31.7}$$

Since the vector **b** may be located outside the plane defined by matrix **A**, the solution is to find the projection of **b** onto **A**, thus minimizing the Euclidean distance (or squared error), using (31.8).

$$\mathbf{v} = \left(\mathbf{A}^\mathbf{T} \times \mathbf{A}\right)^{-1} \times \left(\mathbf{A}^\mathbf{T} \times \mathbf{b}\right) \tag{31.8}$$

4 Results

Two sets of samples were collected, with one set being obtained with the user's body near the receiving antenna (BP set), the other set without the body influence (BNP set). A set is composed by several test runs; each test run contains 100 RSS samples. Position estimation is computed for each sample in a test run, thus no averaging was used in the results presented.

All sample sets were taken in positions where a calibration point exists. The BP set is composed by 79 test runs, from which 66 were taken facing the north direction. The remaining 13 test runs were randomly chosen across the positioning area, with different orientations. The BNP set is composed by 12 test runs randomly chosen and do not have an orientation associated since the body is not present.

The height of the sensor nodes is the same as the anchor nodes (1.2 m above ground). The mean error (ME) and standard deviation (STD) of the absolute error (Euclidean distance between the calculated position and the true position) were the metrics chosen as primary performance indicators.

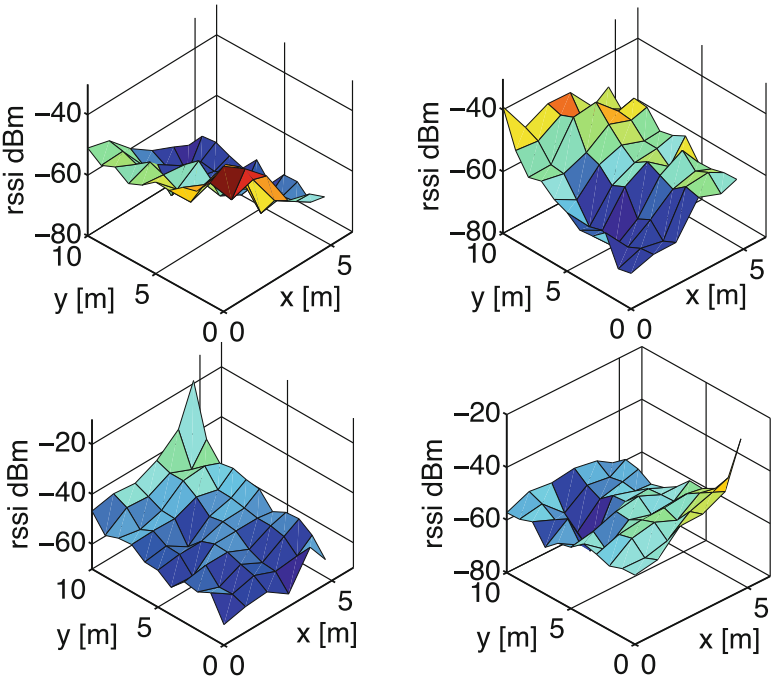


Fig. 31.2 Average RSS fingerprint map. *Top left*: anchor 0; *top right*: anchor 1, *bottom left*: anchor 2 and *bottom right*: anchor 3

4.1 Map Matching

The radio map is a representation of the propagation conditions that the algorithms were subject to. Figure 31.2 illustrates the average RSS for each of the anchor nodes obtained from all calibration points.

The RSS values from Fig. 31.2 were obtained by averaging all calibration points in a given x and y position for all four directions. The values depicted clearly correlate with the position of the anchor nodes, where the strongest RSS values appear in the area where the anchor is located.

Two parameters were tested in the map matching solution: the number of neighbors K and the norm used p. The ME and STD are presented in Fig. 31.3.

The body influence is presented for each of the p-norms tested. In the BP case, the ME variation between $K = 1$, equivalent to nearest neighbor (NN) algorithm, and the other values of K is not significant. This can be explained due to the positioning system area and calibration point density. Since the area is small and the density of calibration points is high, the NN algorithm tends to perform as good as WKNN. Other works, such as [13], also pointed out this outcome, yet under a different environment. Note that a map matching solution with NN as the positioning

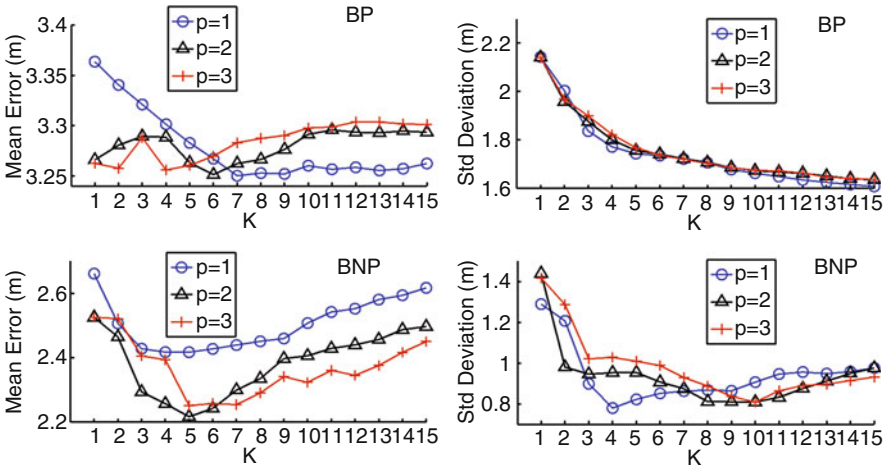


Fig. 31.3 ME (*left side*) and STD (*right side*) for different values of K and p. At the *top* is displayed the BP case, at the *bottom* is displayed the BNP case

algorithm only needs to find one nearest neighbor, which is computationally faster than the WKNN case.

In the BNP case, the value K has a more important influence than in the BP case, where for $p = 2$ and $K = 5$, ME reaches a minimum of approximately 2.2 m. This scenario where body influence is not present is, of course, a best-case scenario, which does not happen when the system is to be used by a person. Yet, it shows a boundary of positioning error that deterministic frameworks can provide in this environment, if accounting the body influence in the position calculation.

The STD values exhibit a monotonic decrease, with the increase of K in the BP case. Differences between norms are negligible. In the BNP case, STD values reach a minimum of 0.8 m for $p = 1$ and $K = 4$.

4.2 Approximate Positioning

RSS (RWCL) and distance using the one-slope path loss model (DWCL) are tested as weights in the WCL algorithm. In the RWCL, the exponent e was varied. Results are presented in Fig. 31.4.

In contrast with other works [23, 24], we found the optimum e parameter between 2 (BP) and 6 (BNP), where a tradeoff between the ME and STD exists. As the parameter e increases beyond 4 in the BP case, and beyond 6 in the BNP case, ME and STD also increase. With a high e value, the position is strongly influenced by the anchor node with the greater RSS reading. In limit conditions, the calculated position would be the same as that of the anchor node with higher RSS in the field. Again, body influence plays a very important role. As an example, for an exponent

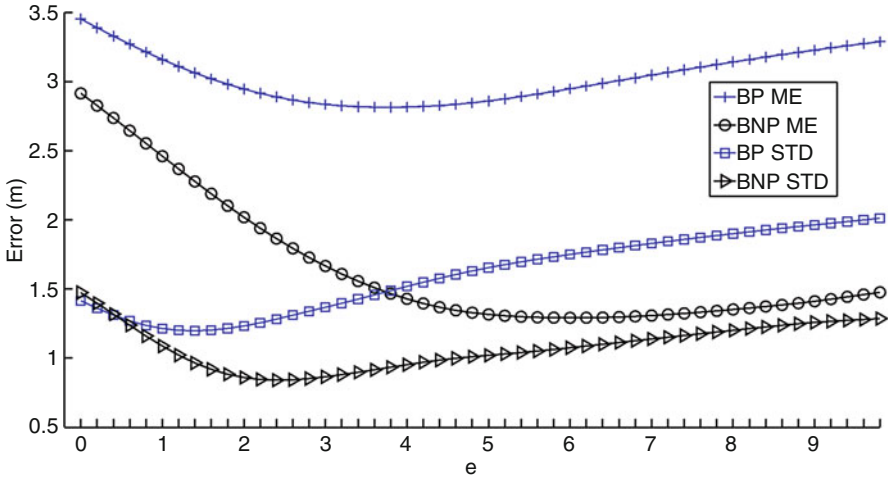


Fig. 31.4 ME and STD for different values of exponent *e*, for both BP and BNP case

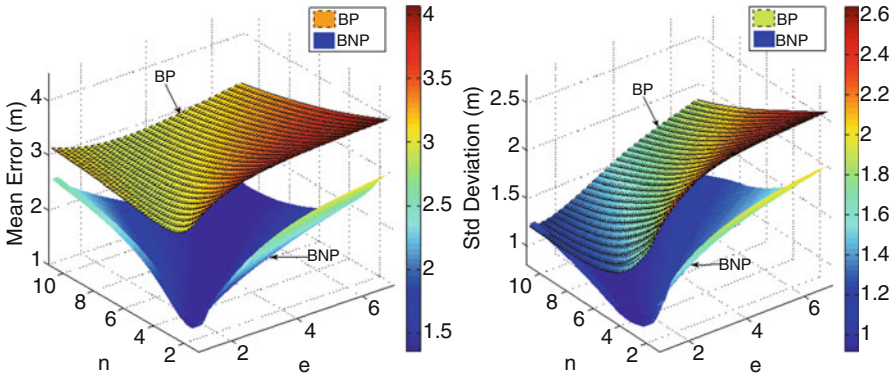


Fig. 31.5 DWCL ME (*left side*) and STD (*right side*) for different values of path loss exponent n and parameter e

of $e = 4$, the ME in the BNP case is approximately half of the ME in the BP case. In the case of STD, an improvement of more than 50 % in the BNP case is also achieved.

In the DWCL algorithm, two parameters can be varied: exponent e and the path loss exponent n. Results are presented in Fig. 31.5.

The minimum ME of 1.36 m is achieved ($n = 2.2$, $e = 1.4$) in the BNP case, while in the BP case, minimum ME was 2.92 m ($n = 3.4$, $e = 1$). Body influence increases the error by a factor slightly higher than 2.

There is a balance between parameters, due to n and e balancing each other, which can be seen as the "saddle" effect in Fig. 31.5.

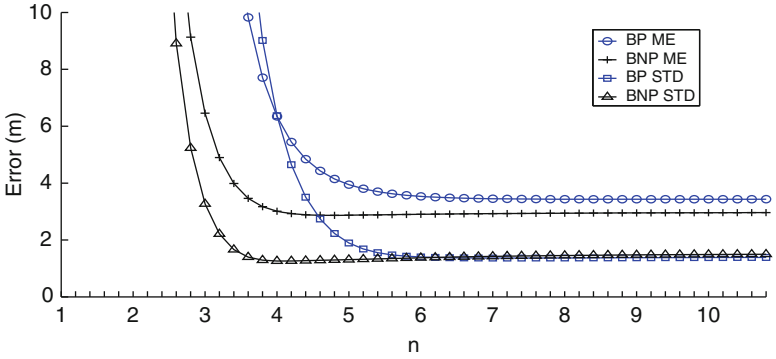


Fig. 31.6 LLS ME and STD for different values of path loss exponent n

The value of $n = 2.2$ obtained in the BNP case is also very similar to the value obtained in [2] of $n = 2.19$, which validates the use of linear regression as an appropriate method of determining path loss exponent when in LOS conditions.

4.3 Exact Positioning

The influence of the parameter n of the one-slope model, used in the RSS to distance conversion, was tested. Results for the LLS algorithm are depicted in Fig. 31.6.

Increasing the value of n produces a dampening effect on the error, since the estimated circumferences around each anchor node become smaller. Even though the ME and the STD decrease as n increases, the algorithm exhibits a saturated behavior, has can be seen for values of n higher than 6.

Positioning error increases rapidly for values of n smaller than 4. For a value of $n = 2.19$, as obtained for the one-slope model used in this work, the ME rises to around 1,000 m, many orders higher than the positioning area itself, which renders the algorithm useless.

4.4 Algorithm Comparison

For the algorithm comparison, the best parameter values for each of the algorithms were considered. To have a frame of reference, a fictitious positioning algorithm, called static center position (SCP) was added to each CDF plot. This algorithm simply returns the center position of the PS area, for any input. The CDF plots for WKNN ($k = 5$ $p = 2$ for BNP case, $k = 1$ $p = 2$ for BP case) and LLS ($n = 6$ for BNP case, $n = 9$ for BP case) algorithms; RWCL ($e = 6$ for BNP case, $e = 3.4$ for BP case) and DWCL ($n = 2.2$ $e = 1.4$ for BNP case, $n = 3.4$ $e = 1$ for BP case) are presented in Fig. 31.7.

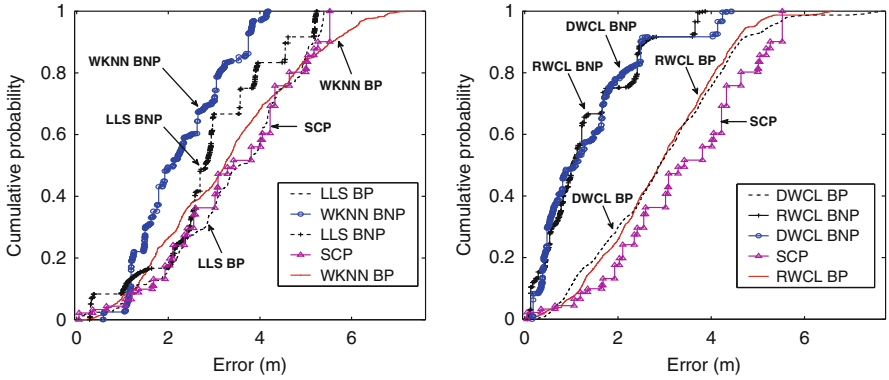


Fig. 31.7 Cumulative distribution functions for WKNN and LLS algorithms (*left side*), RWCL and DWCL algorithms (*right side*)

Regarding the WKNN algorithm, the body influence is evident, with a 30 % improvement for an error of 3 m. The body has a bigger impact on WCL than in the map matching solution, yet the WCL algorithms present slightly better results than WKNN when under body influence. When body is not present, WCL produces the best position estimates of all algorithms tested. Considering a probability of around 70 %, WCL improves from an accuracy of 4 m (BP case) to approximately 1.8 m (BNP case).

RWCL and DWCL obtained equivalent performances, which implies that RSS is the best weighting solution in WCL for our setup, since it is simpler than using a propagation model.

LLS had the worst performance, where the BNP case performed at the same level of the BP case for the other algorithms. When compared with SCP, LLS can even sometimes perform worse.

4.5 Body Influence on RSS

We collected two sets of measurements inside an anechoic chamber. These measurements were obtained by placing an anchor node and a sensor node two meters apart. In the first set of measurements, which we will call static mode, the node is placed on top of a plastic stand and there is no body influence; in the second set of measurements, which we will call dynamic mode, the node is attached to the user's body.

In the static mode, several readings are obtained in different sensor node orientations (approximately 15° between readings). In the dynamic mode, the user performs a 360° turn for approximately 60 s. The module of the RSS values for both modes is presented in Fig. 31.8.

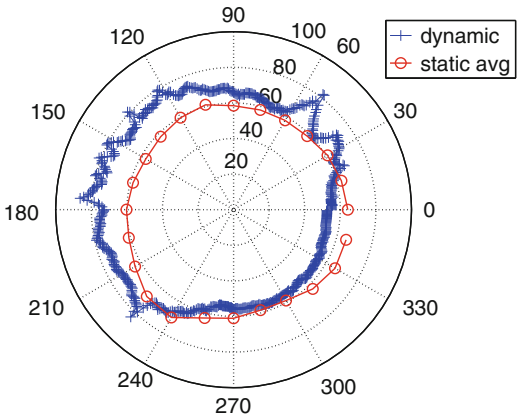


Fig. 31.8 Module of RSS values obtained inside the anechoic chamber

The user is facing the anchor node in the 0° direction. From Fig. 31.8, despite the attenuation peaks that occurred between 30° and 60°, we can see there is a trend in the mean value of the dynamic mode. Mean attenuation value rises when the user rotates past the 90° direction and until 180°. Mean attenuation decreases between 180° and 270°, point from which the mean value starts to exhibit the same trend as in the 0° direction.

5 Discussion

Propagation models typically model large-scale fading LOS propagation. Body influence, NLOS between nodes and multipath dominated environments induce large variations in RSS, which are not accounted for in the propagation model. The comparison between the results obtained for the BP and BNP case demonstrate how strong the body influence is. Also, measurements obtained in the anechoic chamber also suggest that body influence is important when estimating the position using RSS.

The use of propagation models proved to be unreliable in the case of the LLS algorithm. Distances estimated by the propagation model used as weights in approximate positioning algorithms produced acceptable results, yet they did not surpass results using RSS alone as weight.

Although more information from the propagation environment is embedded in the map matching solution, which includes body orientations, the results obtained did not compensate such effort when compared to WCL algorithm. Approximately 2 h were needed to collect all calibration points in our small test environment. If a bigger area were involved, the offline phase map creation would be harder to accomplish without resorting to other mapping techniques.

In the BP case, performance obtained from the WCL solutions is equivalent to the map matching solution. WCL solutions provided the best position estimates in

the BNP case. This, associated with the fact that RWCL solution does not require prior calibration and setup, makes this type of positioning the best possible under our test conditions.

The LLS based on propagation model solution provided the weakest results. Clearly, LLS algorithm cannot be used with RSS measurements in such an environment. LLS algorithm needs more accurate methods to detect distance between nodes.

6 Conclusions and Future Work

From the results obtained we can conclude that the RWCL solution provides overall better results than map matching, with the advantage of having lower complexity and easier setup. The LLS is an inappropriate solution when using RSS to estimate position in indoor environment. Distances estimated from propagation models are severely affected by biases that heavily depend on factors such as body orientation, LOS/NLOS condition, multipath between nodes and proximity to other objects, walls or obstructions. Approximate positioning algorithms tend to perform better in this kind of environment due to its error resilience.

All algorithms showed poor positioning capabilities when body influence is present. When body influence is removed, positioning accuracy improves drastically, with the exception of LLS. Between all three types of positioning algorithms, body influence impact was small in the LLS case, medium in the map matching solution, and highest in the WCL algorithms.

Anchor node placement is a very important issue in RSS positioning systems that has not been addressed in this work. A minimum number of anchor nodes were employed, assuring always a total of four non-collinear points. Increasing the number of anchor nodes in the test area is another possibility to further reduce positioning error. This measure needs to be taken with caution in the case of WKNN, since increasing number of anchor nodes also increases algorithm complexity.

As future work, we intend to integrate the RSS indoor positioning capability in our wireless posture monitoring system (WPMS) [25]. The WPMS is a motion capture system that uses information from multiple inertial and magnetic sensors placed in the user's body. The objective is to provide location information, which, together with the body posture, will characterize not only how the user is moving but also his location.

Acknowledgment Helder D. Silva is supported by the Portuguese Foundation for Science and Technology under the grant SFRBD/78018/2011.

References

1. Karapistoli, E., Pavlidou, F.-N., Gragopoulos, I., Tsetsinas, I.: An overview of the IEEE 802.15.4a standard. IEEE Commun. Mag. **48**(1), 47–53 (2010)
2. Silva, H.D., Afonso, J.A., Rocha, L.A.: Experimental performance comparison of indoor positioning algorithms based on received signal strength. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2014, WCE 2014, pp. 677–683. London, 2–4 July 2014
3. Kohoutek, T. K., Mautz, R., Donaubauer, A.: Real-time indoor positioning using range imaging sensors. In: Proceedings of SPIE 7724: Real-Time Image and Video Processing, Brussels, Belgium, vol. 7724, pp. 77240K–77240K–8 (2010)
4. Glanzer, G., Bernoulli, T., Wiessflecker, T., Walder, U.: Semi-autonomous indoor positioning using MEMS-based inertial measurement units and building information. Position. Navig. Commun. **2009**, 135–139 (2009). 2009 6th work
5. Wang, H., Lenz, H., Szabo, A., Bamberger, J., Hanebeck, U.D.: WLAN-based pedestrian tracking using particle filters and low-cost MEMS sensors, 2007 4th Work. Position. Navig. Commun. **2007**, 1–7 (2007)
6. Forkel, I., Salzmann, M.: Radio propagation modelling and its application for 3G mobile network simulation. In: 10th Aachen symposium signal theory, pp. 363–375. Aachen (2001)
7. Damosso, E., Correia, L.M.: Cost Action 231: Digital Mobile Radio Towards Future Generation Systems, Final Report. Brussels, Belgium (1999)
8. Zhang, Z., Wan, G., Jiang, M., Yang, G.: Research of an adjacent correction positioning algorithm based on RSSI-distance measurement. In: 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Shanghai, pp. 2319–2323 (2011)
9. Yeong, S.-Y., Al-Salihy, W., Wan, T.-C.: Indoor WLAN monitoring and planning using empirical and theoretical propagation models. In: 2010 Second International Conference on Network Applications Protocols and Services, Kedah, pp. 165–169 (2010)
10. Rappaport, T.S.: Wireless Communications: Principles and Practice. Prentice-Hall, Englewood Cliffs (1996)
11. Miluzzo, E., Zheng, X., Fodor, K., Campbell, A. T.: Radio characterization of 802.15.4 and its impact on the design of mobile sensor networks. In: 5th European Conference on Wireless Sensor Networks (EWSN '08), Bologna, Italy, pp. 171–188 (2008)
12. Bahl, P., Padmanabhan, V. N.: RADAR: an in-building RF-based user location and tracking system. In: INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Tel Aviv, Israel, vol. 2, pp. 775–784 (2000)
13. Honkavirta, V., Perala, T., Ali-Loytty, S., Piche, R.: A comparative survey of WLAN location fingerprinting methods. In: 2009 6th workshop on positioning, navigation and communication, vol. 2009, pp. 243–251 (2009)
14. Mestre, P., Coutinho, L., Reigoto, L., Matias, J., Serodio, C.: Hybrid technique for fingerprinting using IEEE802.11 wireless networks. In: International Conference on Indoor Positioning and Indoor Navigation IPIN, Guimarães, Portugal, pp. 1–7 (2011)
15. Bosisio, A. V.: Performances of an RSSI-based positioning and tracking algorithm. In: 2011 International Conference on Indoor Positioning and Indoor Navigation IPIN, Guimarães, Portugal, pp. 1–8 (2011)
16. Bulusu, N., Heidemann, J., Estrin, D.: GPS-less low-cost outdoor localization for very small devices. Pers. Commun. IEEE **7**(5), 28–34 (2000)
17. Ni, L.M., Liu, Y., Lau, Y.C., Patil, A.P.: LANDMARC: indoor location sensing using active RFID. Wirel. Netw. **10**(6), 701–710 (2004)
18. Yinggang, X., Jiaoli, K., Zhiliang, W., Shanshan, Z.: Indoor location technology and its applications base on improved LANDMARC algorithm. In: Proceedings of the 2011 Chinese Control and Decision Conference, CCDC, Mianyang, no. 2, pp. 2453–2458 (2011)

19. Tian, S., Zhang, X., Liu, P., Sun, P., Wang, X.: A RSSI-based DV-hop algorithm for wireless sensor networks. In: 2007 International conference on wireless communications, networking and mobile computing, pp. 2555–2558, Sept 2007
20. Nagpal, R., Shrobe, H., Bachrach, J.: Organizing a global coordinate system from local information on an Ad Hoc sensor network. In: Proceedings of the 2nd international conference on information processing in sensor networks, pp. 333–348 (2003)
21. Hightower, J., Borriello. G.: A survey and taxonomy of location systems for ubiquitous computing. IEEE Computer, Seattle, Tech. Rep. UW-CSE 01-08-03 (2001)
22. Savvides, A., Han, C.-C., Strivastava, M.B.: Dynamic fine-grained localization in Ad-Hoc networks of sensors. In: Proceedings of the 7th annual international conference of mobile computing network – MobiCom '01, pp. 166–179 (2001)
23. Blumenthal, J., Grossmann, R., Golatowski, F., Timmermann, D.: Weighted centroid localization in zigbee-based sensor networks. In: 2007 IEEE international symposium on intelligent signal processing, pp. 1–6 (2007)
24. Reichenbach, F., Timmermann, D.: Indoor localization with low complexity in wireless sensor networks. In: 2006 IEEE international conference on industrial informatics, pp. 1018–1023 (2006)
25. Silva, H.D., Macedo, P., Afonso, J.A., Rocha, L.A.: Design and implementation of a wireless sensor network applied to motion capture. In: 1st portuguese conference on wireless sensor networks (CNRS 2011). Coimbra (2011)

Chapter 32
Study and Simulation of Protocols of WSN Using NS2

Ouafaa Ibrihich, Salah-ddine Krit, Jalal Laassiri, and Said El Hajji

Abstract Wireless sensor networks are becoming significantly vital to many applications, and they were initially used by the military for surveillance purposes. One of the biggest concerns of WSNs is that they are defenseless to security threats. Due to the fact that these networks are susceptible to hackers; it is possible for one to enter and render a network. However, WSN presents many challenges. These networks are prone to malicious users attack, because any device within the frequency range can get access to the WSN. There is a need for security mechanisms aware of the sensor challenges (low energy, computational resources, memory, etc.). Thus, this work aims to simulate a secure routing protocol for WSN by using trusted frame works called SAODV. The Trust Scheme evaluates the behavior of all nodes by establishing a trust value for each node in the network that represents the trustworthiness of each one thereby identifies and eliminates the malicious nodes. It also observes node's mobility, number of neighbors each node has, number of packets generated and forwarded by the neighboring nodes, and the past activity of the node.

Keywords Ad Hoc • AODV • Attacks • NS2 • Routing protocols • Security • Wireless Sensor Networks (WSN)

O. Ibrihich (✉) • S. El Hajji
Laboratory of Mathematics, Computing and Applications, Department of Mathematical and Computer Sciences, Faculty of Sciences, University of Mohammed-V, Rabat, Morocco
e-mail: wafaa.ibrihich@gmail.com; elhajji@fsr.ac.ma; http://www.fsr.ac.ma/mia/elhajji.htm

S.-d. Krit
Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Agadir, Morocco
e-mail: krit_salah@yahoo.fr

J. Laassiri
Faculty of Sciences of Kenitra, Department of Computer Sciences, Ibn Tofail University, Morocco
e-mail: laassiri.jalal@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_32

1 Introduction

A wireless sensor network is a collection of nodes organized into a cooperative network [1, 2]. Each node consists of processing capability (one or more microcontrollers, CPUs or DSP chips), may contain multiple types of memory (program, data and flash memories), have a RF transceiver (usually with a single Omni-directional antenna), have a power source (e.g., batteries and solar cells), and accommodate various sensors and actuators. The nodes communicate wirelessly and often self-organize after being deployed in WSN fashion. Wireless networks are gaining popularity to its peak today, as the users want connectivity in terms of wireless medium irrespective of their geographic position. There is an increasing threat and various attacks on the Wireless Network.

However, each node in the network has the ability to discover its neighbors and to construct routes to reach other nodes in the collection. Like other networks, sensor networks are vulnerable to malicious attack; however, the hardware simplicity of these devices makes defense mechanisms designed for traditional networks infeasible. This work explores the Denial-of-Service (DoS) attack, in which a sensor node is targeted [3].

2 Wireless Sensor Network: An Overview

The characteristics of WSNs are discussed from two perspectives: from the nodes that make up the network, and from the network itself.

The very idea of a wireless network introduces multiple venues for attack and penetration that are either much more difficult or completely impossible to execute with a standard, wired network. This inherent limitation makes WSNs especially sensitive to several key types of attacks. In contrast to resource-rich networks such as the Internet, a WSN is less stable, more resource-limited, subject to open wireless communication, and prone to the physical risks of in-situ deployment. These factors increase the susceptibility of WSNs to distinct types of attacks.

Attacks can be performed in a variety of ways, most notably as denial of service attacks, but also through traffic analysis, privacy violation, node takeover, attacks on the routing protocols, and attacks on a node's physical security, which all are out of the scope of this research. In this paper, we first address some common denial of service attacks and then describe the most famous defensive strategies against them.

Security in WSNs can be defined as the method of protecting a prospective application against all known types of attack. Attacks including denial-of-service (DOS), traffic analysis, multiple identity/node replication, confidentiality and physical tampering are all areas for concern within WSN security architecture design, it is extremely important to ensure that all known attacks are defended against when designing a security system for a WSN. The success of the application will depend largely upon its reliability and robustness against attack [4].

3 Attacks in Wireless Networks

Wireless networks are more susceptible to attacks because of their shared physical medium, open transmission of radio frequencies [5].

3.1 Vulnerabilities of Wireless Sensor Networks

A typical wireless sensor network is expected to give a certain data that the user is actively enquiring about after some amount of time. Many attack schemes tend to stop the proper performance of sensor networks to delay or even prevent the delivery of data requested by user. Despite the fact that the term attack usually refers to an adversary's attempt to disrupt, undermine, or destroy a network, a Denial-of-Service (DoS) attack refers to any event that diminishes or eliminates a network's ability to perform its expected function. Such a technique may be helpful in specific applications such as utilizing the best of these attacks to find the weak tips of presented protocols at different layers. These attacks consequently would expose weaknesses that lead to effective countermeasures.

3.2 Denial of Service Attack in Wireless Sensor Networks

The aim of DoS attack is to make services unavailable to legitimate users, and current network architectures allow easy-to-launch and hard-to-stop DoS attacks. Particularly challenging are the service-level DoS attacks, whereby the victim links are destroyed and flooded with legitimate-like requests attack, in which wireless communication is blocked by malicious radio interference. These attacks are overwhelming even for massively resourced services, and effective and efficient defenses are highly needed.

Denial of Service (DoS) is a common type of cyber-attack over the Internet. The purpose of DoS is to make a computer's resources unavailable to its intended users. One way to launch a DoS attack is by sending malformed traffic to the target or by sending a huge amount of normal traffic which will overload the target's buffer. To be more effective, attackers often use many compromised machines, rather than just one, as a source for the attack. There are many security attacks which are considered under in Dos [6].

3.3 Problematic

Denial-of-Service attack in wireless sensor network occurs due to intentional intrusion attack or unexpected node failure [7, 8]. Various software bugs, unex-

pected sensor node failure, exhausted power supply system, environmental disaster, complication in data transmission and communication or even intentional intruder attack may execute DoS attack. Often the outsiders try to weaken or destroy a network or cause an interruption in secure data communication by sending loads of unnecessary data packets to the victim nodes and therefore exhibit DoS attack [9].

3.4 Attacker's Distribution

Clearly, if only one node on the border of the network is attacked, the impact on performance metrics that determine the "health" of the network will be minimal. On the other hand, if the attacked node is a one through which many routes must pass, the impact of the attack will be more noticeable; assuming that attackers are poorly informed, though it is fair to expect that they wouldn't be able to distinguish a border node from an internal node. For this reason, we assume that every node in the network is equally likely to be attacked. In our model, we divide the whole network into k certain attack zones where k represents the previously estimated number of attackers. Each such zone shows the zone of attack or the territory of the attack node. Zone size is -controlled by the number of nodes in the network which defines a minimum bound on the number of serving attackers to cause the desired effect in degrading network performance characterized by decreasing the throughput at the sink and increasing the corresponding delay of the delivered data [3].

Figure 32.1: below demonstrates the division of the network in to k attack zones where k equals the number of attackers represented by red circles, are legitimate sensors, represented by circles in black.

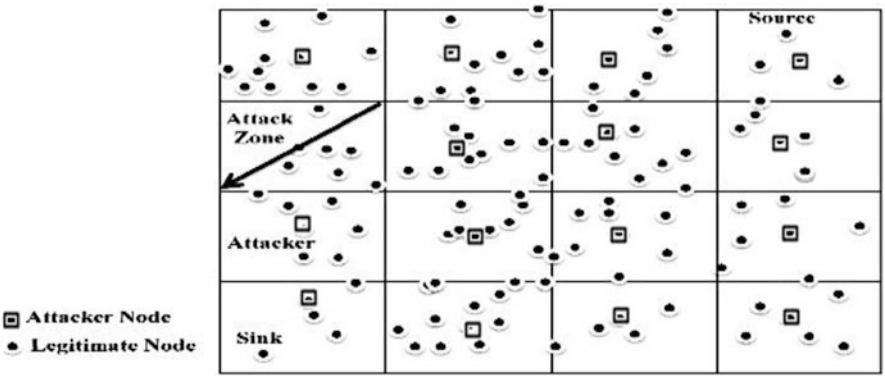


Fig. 32.1 Attacker distribution into attack cells throughout the network

4 Simulation Setup

4.1 Simulation Tools

The NS-2 simulator is a discrete event-driven network simulator, which is popular with the networking research community [10]. It was developed at the University of California at Berkeley and extended at Carnegie Mellon University, CMU, to simulate wireless networks [11]. These extensions provide a detailed model of the physical, link layer behavior of a wireless network, and allow arbitrary movement of nodes within the network. It includes numerous models of common Internet protocols including several newer protocols, such as reliable multicast and TCP selective acknowledgement. Additionally, different levels of configurations are present in NS-2 due to its open source nature, including the capability of creating custom applications and protocols as well as modifying several parameters at different layers.

The simulator is written in C++, accompanying an OTCL script language based on Tcl/Tk. The researcher defines the network components such as nodes, links, protocols and traffic using the OTCL script. NS-2 uses OTCL as the interface to the user (Fig. 32.2). This script is then used with NS, the simulator, to conduct the desired simulation, and as a result outputs traces at different selective layers. The

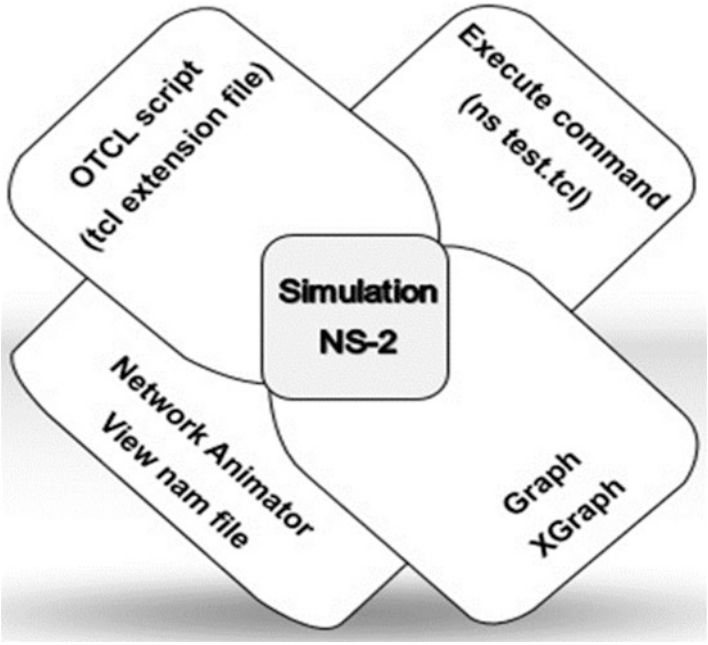


Fig. 32.2 Simulation overview

Table 32.1 Simulation parameters for scenario1

Parameters	Value
Simulator	NS-2 (version 2.35)
Channel type	Channel/Wireless channel
Radio-propagation model	Propagation/TwoRayGround
Network interface type	Phy/WirelessPhy
MAC type	Mac/802.11
Interface queue type	Queue/DropTail\PriQueue
Link layer type	LL
Antenna model	Antenna/OmniAntenna
Maximum packet in ifq	20
Number of mobile node	50
Traffic type	TCP
Simulation time	500 s
Routing protocols	AODV
Nominal bit rate	2 Mb/s
Node speed	1–15 m/s
Transmission rate	4 packets/s
Area of simulation	1,000 m * 300 m

output data within the trace output files is then filtered and extracted using statistical analysis software like excel/access program. The extracted relevant data is then used to evaluate performance by manipulating various metrics such as delays, throughput, overheads etc.

4.2 Simulation Environment

We simulated DoS attack in NS-2.35. We also use NAM visualization tool to show the Network animator (Table 32.1).

The following are the configurations set as per the assumed simulation context:

For describing the working of DoS attack three steps are discussed as Route Request (RREQ), Route Reply and Propagation of route reply (PREP):

In order, the malicious node M0 begins by broadcasting a RREQ (Route REQuest) message (malicious packets). Route request will be broadcasted in the manner of multi node hops. In Fig. 32.3 during path discovery process, sender broadcasts RREQ to its neighboring nodes i.e. 1, 5, 10, 15, and 20. The neighboring nodes will forward RREQ further to their neighbors.

After getting the route request to destination from the sender, destination will unicast a route reply (RREP) packet to source node 0 (Fig. 32.4).

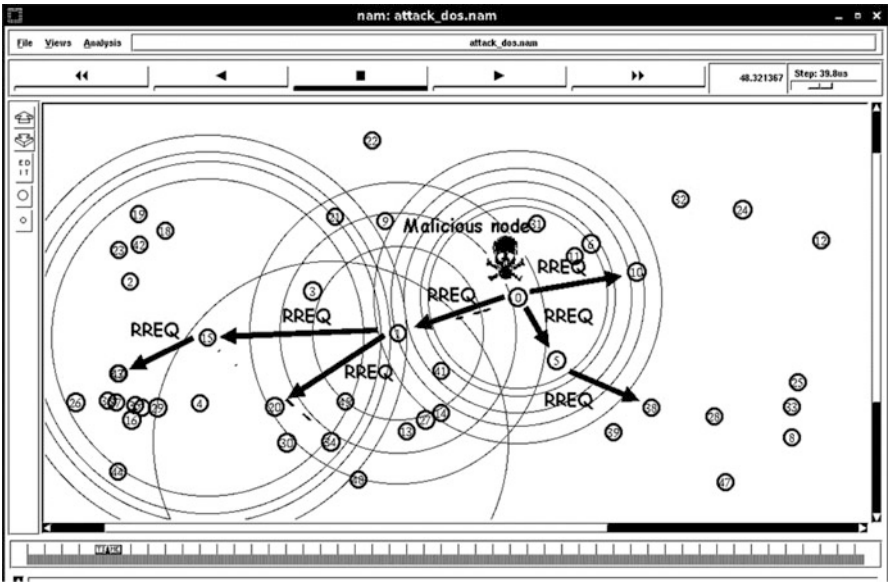


Fig. 32.3 Malicious node sending Route REQuests to get access to different nodes

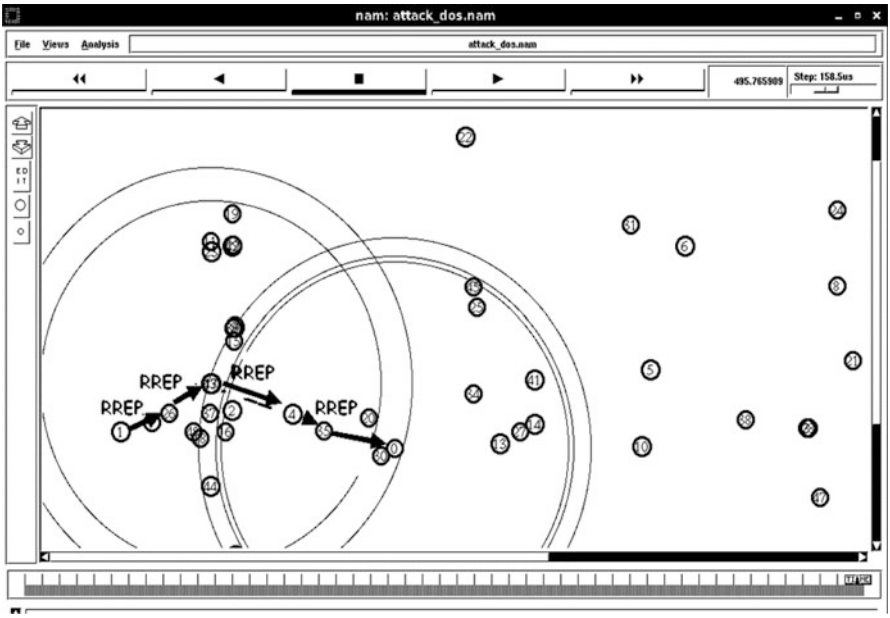


Fig. 32.4 RREP message from destination to malicious node

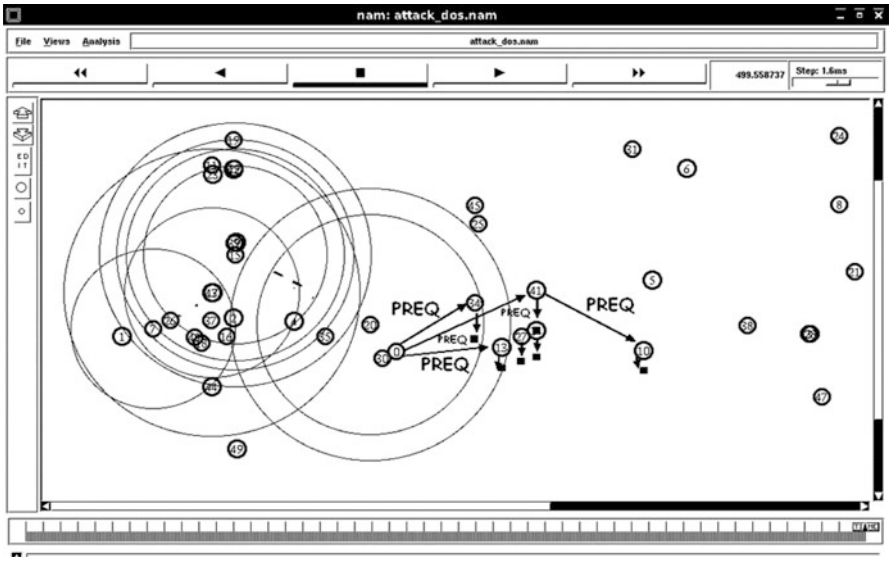


Fig. 32.5 PREQ message from malicious node to destination

Whenever node 0 detects a link break from link layer, the source and end nodes are notified by propagating an RERR packet similar (malicious packets) to different nodes (Fig. 32.5).

The attacker can allow his replica nodes to randomly move or he could move his replica nodes in different patterns in an attempt to frustrate proposed scheme. We also assume that the base station is a trusted entity. This is a reasonable assumption in mobile sensor networks, because the network operator collects all sensor data and can typically control the nodes operation through the base station. Thus, the basic mission of the sensor network is already completely undermined if the base station is compromised.

5 AODV and SAODV Routing Protocols

5.1 Ad-hoc on Demand Distance Vector Routing Protocol

AODV is a reactive routing protocol which creates a path to destination when required. Routes are not built until certain nodes send route discovery message as an intention to communicate or transmit data with each other [12]. This routing protocol uses two phases. In phase one route discovery is done. In phase two route maintenance is done. It uses three control messages namely (Figs. 32.6, 32.7, and 32.8):

Route Requests (RREQs)

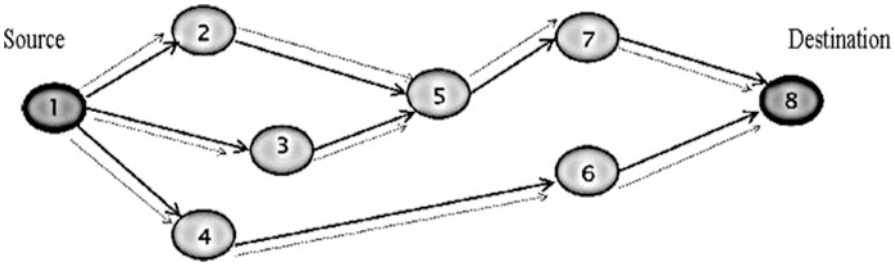


Fig. 32.6 Propagation of route request (PREQ) packet

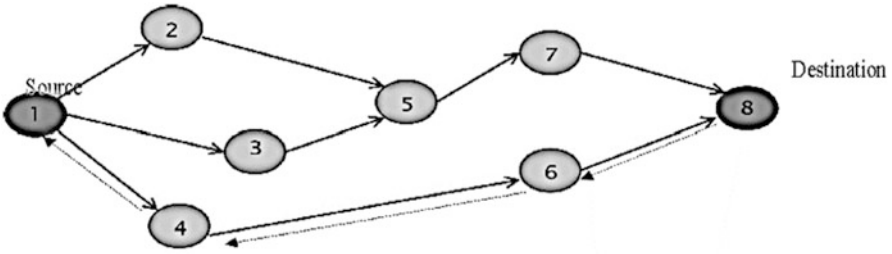


Fig. 32.7 Propagation of route reply (PREP) packet

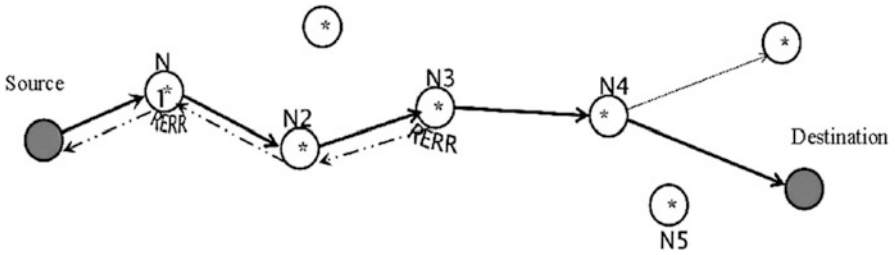


Fig. 32.8 Route errors (RERRs)

Route Replies (RREPs)
Route Erros (RERRs)

The RREQ and RREP messages are used in phase one whereas RRER control message is used in phase two. The steps to be followed in AODV protocols are as follows [13]:

i. Source node broadcasts RREQ message. It contains source and destination address, sequence number and broadcast id.
ii. If the next node is the destination then it replies with RREP message or else message is forwarded to next node.

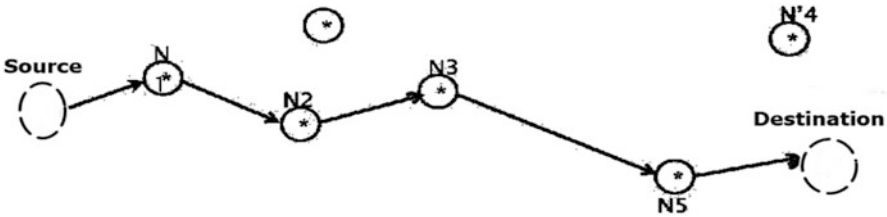


Fig. 32.9 Route maintenance

iii. When forwarding the RREQ message node maintains broadcast id, source address and maintains a reverse route.
iv. Sequence number helps in route updation and helps in getting fresh enough route to the destination.
v. Destination node on receiving RREQ then sends a unicast RREP message to the source node on the same path that was created during RREQ.

5.2 Secure AODV Routing Protocol

SAODV is an extension to AODV. It uses asymmetric cryptography to secure AODV's routing messages. SAODV uses Digital Signatures to protect the non-mutable data in the RREQ and RREP messages. The four basic operations performed for the Route Establishment are 1. Route Discovery 2. Route Request 3. Route Reply and 4. Route Maintenance (Fig. 32.9).

Before entering the network, each node obtains a public key certificate from a trusted certificate server. There are End-to-end authentication between source and destination and Hop-to-hop authentication between intermediate nodes. Hash chains are used in SAODV to authenticate the hop count of the AODV routing. Source broadcasts signed RDM (Route Discovery Message) along with its own certificate. RDM contains the source IP address, along with a source-specific nonce (to detect duplicates) [14].

6 Related Works

In this paper, first I want to show the comparison between the packet loss of AODV and SAODV. We have different number of nodes and simulation parameter by which we can do analysis. The graph shown the packet loss and packet received for AODV and SAODV. Further, we can change the simulation parameter and time and see the changes in graph.

Table 32.2 Simulation parameters for scenario2

Parameters	Value
Simulator	NS-2 (version 2.35)
Channel type	Channel/Wireless channel
Radio-propagation model	Propagation/TwoRayGround
Network interface type	Phy/WirelessPhy
MAC type	Mac/802.11
Interface queue type	Queue/DropTail\PriQueue
Link layer type	LL
Antenna model	Antenna/OmniAntenna
Maximum packet in ifq	30
Area for simulation	200 * 200
Number of mobile node	10
Traffic type	TCP
Simulation time	50 s
Routing protocols	AODV, SAODV

We can have different parameter for better results. The Table 32.2 shows the simulation parameters for our scenario:

After simulation, NS2 outputs a trace file, which can be interpreted by many tools, such as NAM and Xgraph. We create a simulation scenario using NS-2 Scenario Generator [15]. Table 32.2 shows the network parameter definition in the TCL file. The first parameter tells the simulator that nodes transmits and receives packets through wireless channels. We have used the IEEE 802.11 standard, which specifies the media access control and the physical layer [16].

The Fig. 32.10 shows the X graph of comparison between AODV and SAODV. By the Figure, we see that as the simulation start the packet received and packet loss is initially zero, because initially there is no CBR connection and nodes taking their right place. As the CBR connections establish between the nodes the number of packet received increases but no packet loss is there, it means all generated packets are being received by the nodes. But the packet loss increases substantially on the simulation time increases (in AODV). Finally, the packet received is more than the packet loss and nodes taking their right place (in SAODV). As the CBR connections establish the number of packet lost increases very much as compare to packet received. It shows that the nodes are dropping mostly generated packets.

From the graph (Fig. 32.11), it is clear that the throughput of dropping packets (coming from the malicious node) at the destination node is low using the AODV protocol. The throughput of dropping packets at the receiving node becomes high using the secure AODV. This means that using this secure routing protocol allow rejecting malicious packets. Then malicious node will not be able to send malicious packets on the WSN.

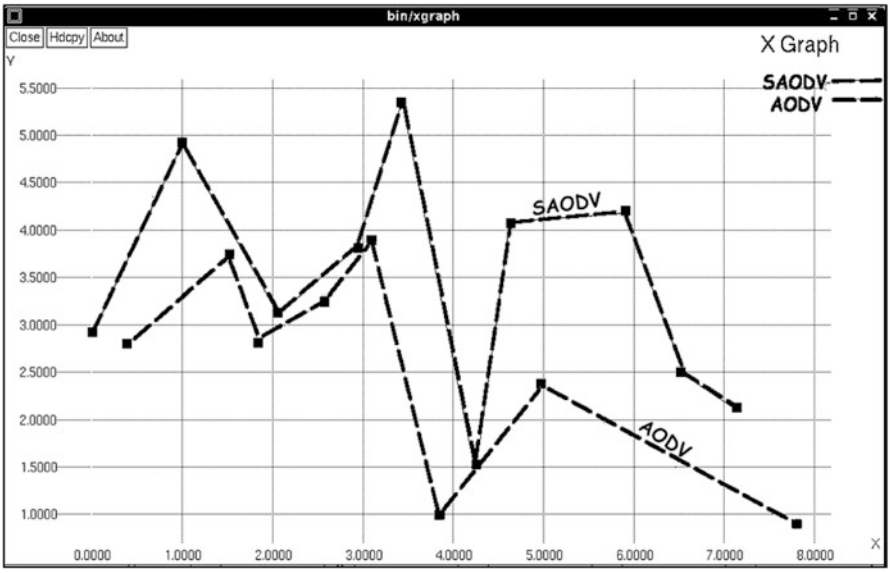


Fig. 32.10 Throughput vs. no of nodes

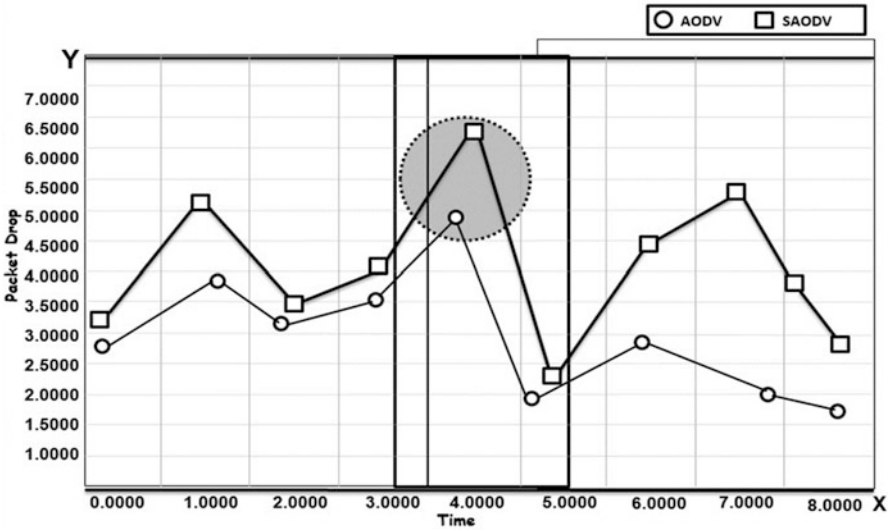


Fig. 32.11 Throughput of dropping packets

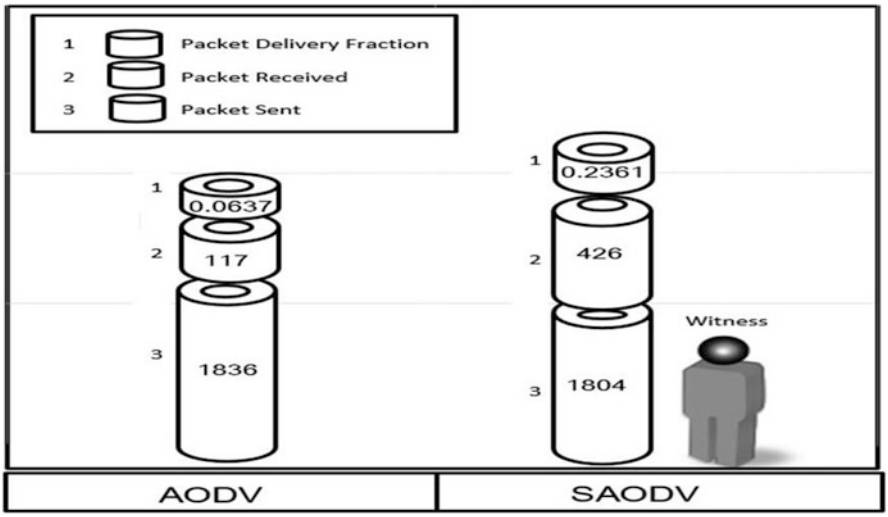


Fig. 32.12 Packet delivery ratio vs. no of nodes (sent and received)

6.1 Packet Delivery Fraction

The packet delivery ratio in this simulation is defined as the ratio between the number of packets sent by constant bit rate sources and the number of received packets by the CBR sink at destination.

From the graph (Fig. 32.12), SAODV performs better than AODV in case of packet delivery fraction and goodput because the number of nodes is less and no periodic update is maintained in SAODV.

7 Conclusion and Future Works

Secure routing is vital to the acceptance and use of sensor networks for many applications, but we have simulated that currently routing protocols for these networks are insecure. In this paper, we have used an approach based on changing the packet transmission frequency in the AODV protocol, which is a reactive protocol. The goal is to avoid the DoS attack. We have shown an attack scenario, which we have simulated using NS-2.

For future work, we intend to simulate security protocols for aggregation and localization algorithms in WSN's, then add a switching technique from one protocol to another, based on sensor states (energy, mobility, connectivity, vicinity, etc.). At the end, we will have a secure and context aware protocol.

References

1. Ibrihich, Ouafaa, Krit, Salah-ddine, Laassiri, Jalal, Hajji, Said El: In: Comparative Analysis of secure routing in WSN, Proceedings of the World Congress on Engineering 2014, vol. I, WCE 2014, pp. 734–738. London 2–4 July 2014
2. Ilyas, M.: The Handbook of Ad-Hoc Wireless Networks. CRC Press, Florida (2008)
3. Huda Bader Hubboub.: Denial of Service Attack in Wireless Sensors Networks (2010)
4. Boyle, D. Newe. T.: Securing Wireless Sensor Networks: Security Architectures, vol. 3, no. 1, January 2008
5. Jain, S., Dr. Jain, S.: Detection and prevention of wormhole attack in mobile Ad hoc networks. Int. J. Comput. Theory Eng. **2**(1), 78–86 (2010). February
6. Kessler, G.C.: Defences Against Distributed Denial of Service Attacks. November 2000. http://www.garykessler.net/library/ddos.html. Retrieved 30 Oct 2008
7. Wood, A.D., Stankovic, J.A.: Denial of service in sensor networks. Computer **35**(10), 54–62 (2002)
8. Blackert, W.J., Gregg, D.M., Castner, A.K., Kyle, E.M., Hom, R.L., Jokerst, R.M.: Analyzing interaction between distributed denial of service attacks and mitigation technologies. In: Proceedings of DARPA Information Survivability Conference and Exposition, vol. 1, pp. 26–36. April 2003
9. Halim, T., Islam, Md. R.: A study on the security issues in WSN. Int. J. Comput. Appl. (0975–8887). **53**(1), (2012). Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.5658&rep=rep1&type=pdf
10. Fall, K., Varadhan, K.: Editors ns notes and documentation. The VINT Project, UC Berkeley, LBL, USC/ISI, and Xerox PARC, 4 Nov 2011. Available: http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf
11. Panc Pancardo, P., Dueñas, J.C.: A proposal for System Architecture to Integrate Scarce resources Wireless Sensors Networks into Ubiquitous Environments [Online]. Available http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-208/paper23.pdf
12. Mistry, M., Jinwala, D.C.: Improving AODV protocol against black hole attacks. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. II (2010)
13. Sreedhar Dr, C., Madhusudhana Verma, S., Kasiviswanath Dr, N.: Potential security attacks on wireless networks and their countermeasure. Int. J. Comput. Sci. Inf. Technol. **2**(5), 76–89 (2010)
14. Vigneswari, P., Jayashree, R.: Comparative analysis of AODV and trusted AODV (TAODV) in MANET. Int. J. Adv. Inf. Sci. Technol. **10**(10), 49–56 (2013)
15. Huang, J., Shahnasser, H.: A preprocessor Tcl script generator for NS-2 communication network simulation. International Conference on Communications and Information Technology (ICCIT 2011), San Francisco State University, San Francisco, pp. 184–187, 2011
16. Haboub, R., Ouzzif, M.: Secure routing in WSN. Int. J. Distrib. Parallel Syst. **2**(6), (2011)

Chapter 33
Building Successfull Brand on the Internet

Tina Vukasović

Abstract Innovation in social networking media has revolutionized the world in twenty-first Century. Social networking media presents potentially opportunities for new forms of communication and commerce between marketers and consumers. Objective of the study is to analyze the effective communication strategy through social networking media. Survey was conducted randomly among Facebook user community, by sending questionnaire through online to collect the individual opinion from the respondents. The total population is social networking user community, but to collect the effective data the sampling is constrained to the target population like young adults, graduates within the age of 18 years to 55 years. The sampling size is 400. The paper presents research results and internet marketing activities that have contributed to building a relationship with the brand. It is necessary to study the effectiveness of brand communication strategy followed in social networking media which are mainly accessed by Slovenian users. This study would help the advertisers to understand the effective communication strategy to communicate their brand among the users. In recent trend of marketing in social networking sites, various brand communications are widely used to attract targeted leads. So, this study would help to know the effectiveness of communication and strategy done through social networking media which make the target audience to participate in this kind of advertising.

Keywords Brand • Fast moving consumer goods • Internet • Marketing • Market research • Social media • Social networking sites

T. Vukasović (✉)
International School for Social and Business Studies, Mariborska cesta 7, Celje SI – 3000, Slovenia

DOBA Faculty, Prešernova 1, Maribor SI – 2000, Slovenia
e-mail: tina.vukasovic@mfdps.si

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_33

1 Introduction

The Internet is transforming the business environment, creating new challenges and opportunities. This chapter provides an overview of the Internet and its defining characteristics, highlighting the key developments that have contributed to its explosive growth and its impact on the business environment. In its current form, internet is primarily a source of communication, information and entertainment, but increasingly, it also acts as a vehicle for commercial transactions. Since the explosion of the web as a business medium, one of its primary uses has been for marketing. Soon, the web could become a critical distribution channel for the majority of successful enterprises. One among them is marketing and spreading brand communication through Social networking sites [18]. The Internet provides the opportunity for companies to reach a wider audience and create compelling value propositions never before possible (e.g. Amazon.com's range of 4.5 million book titles), while providing new tools for promotion, interaction and relationship building. It is empowering customers with more options and more information to make informed decisions. The Internet also represents a fundamental shift in how buyers and sellers interact, as they face each other through an electronic connection, and its interactivity provides the opportunity for brands to establish a dialogue with customers in a one-to-one setting. As such, the Internet is changing fundamentals about customers, relationships, service and brands, and is triggering the need for new brand-building strategies and tools [3].

1.1 Building Successfull Brand on the Internet

The Internet is changing the brand environment or "brandscape". This chapter explores new strategies and tools for building brands on the Internet, including the interactive approach to attracting customers and building loyalty. Building a strong brand is a complex task. The brand building process starts with the development of a strong value proposition. Once this has been established, the next step is to get customers to try the brand. If the offering is developed properly, it should provide a satisfactory experience and lead to a willingness to buy again. To entice trial and repeat purchase requires triggering mechanisms, which are created through advertising, promotion, selling, public relations, and direct marketing. The company needs to communicate the values of the brand and then reinforce brand associations to start the wheel of usage and experience, and keep it turning. Through the combination of the stimulus of consistent communications and satisfactory usage and experience, brand awareness, confidence and brand equity are built. Traditionally, in addition to providing added value, brands were a substitute for information – a way for customers to simplify the time-consuming process of search and comparison before deciding what to buy. However, the Internet makes search and comparison much easier. This threatens to undermine the value of brands.

On the other hand, the logic of the Internet cuts another way. Transactions on the Internet require customers to provide detailed personal information – names, addresses, credit card numbers, etc. Generally, people have concerns about sharing personal information. In addition, the intangible nature of the Internet, and the fact that customers are buying goods that, in most cases, they have never handled or seen (except on-screen), has placed greater importance on trust and security. People only tend to transact with sites they know and trust – sites that provide a wealth of information and make comparison shopping easy, where the user feels a part of, and sites that understand the user's needs and preferences. This highlights the surfacing of information and relationships as key sources of added value in the Internet economy. Customers derive added value through the provision of information on the products or services they buy, as well as on topics of interest related to the brand and product characteristics. Traditionally, brands have been developed in an environment whereby a company creates a brand, and projects it onto a third party intermediary (the media). In response, many unnamed customers develop a "relationship" with the brand. The Internet, on the other hand, offers interactivity, whereby the company can establish a dialogue and interact with individual customers on a one-to-one basis. In doing so, a company can listen, learn, understand and relate to customers, rather than simply speaking at customers. This creates the opportunity for companies to build stronger relationships than previously attainable. However, this also poses a challenge as these relationships may take on a life and character of their own [3]. The differences between the traditional approach and the one-to-one approach are outlined in Table 33.1.

The Internet gives companies control over all their interactions with customers and therefore, brand-building must focus on the end-to-end customer experience – from the promises made in the value proposition, to its delivery to the customer. In maximising the customer experience, companies have to find innovative ways of leveraging the information and relationship building characteristics of the Internet.

Table 33.1 The emerging brand building environment

Traditional approach	One – to – one approach
Monologue	Dialogue
Public	Private
Mass	Individual
Anonymous	Named
Adversarial	Collaborative
Focused primarily on one –off transactions	Focused on relationship over time
Remote research	Intimate learning
Manipulative, "stimulus-response" approach	Genuine need driven, service approach
Standardised	Customised

Source: Cleland [3], 46

1.2 The Social Media

Over the last decade, social networks have changed communications, shifting the way we consume, produce and interact with information, based on explosive migration to the web. Social media websites like Facebook (FB), Twitter etc. have created huge impetus on the communication platform with the end customers of different products and services that was lacking in the traditional medium. Social networking sites are used as marketing tool by marketers in creating brand relationship. Huge growth of customer base enables marketers to explore social media networking sites as new medium of promoting products and services and resulting in reduction in clutter of traditional medium advertising of reaching the mass customers and not realizing the actual ROI. Social networking sites are more collaborative and interactive in comparison to traditional media followed by marketers. Social media advertising is a paid form of brand, service or business promotion and requires a proper and planned communicative message and budget. Advertising is customer centric in nature. Customers play an important role in any major or minor communication because they are the one who are going to decide the fate of the advertising communication. Some benefits of social network advertising include [11]:

1. Popularizing your brand, idea or service to the target group.
2. Informing target audience about your brand or service's presence in the market.
3. Encouraging healthy competition in the market.
4. Providing social benefits for the brand.
5. Making the audience to interact and keep them intact with the brand.

Advertising on internet provides a major contribution to brand competition in the market. Advertising here not only provides information about a product or service but also promotes innovation. Besides it also facilitates customer satisfaction. Big and small companies, individuals of all walks of life, major and minor events, concepts, etc., nowadays lay their base on social network advertising to get recognized in the market [24]. Social media foster communities where people tend to gather around a common goal or shared interest and interact regularly. Join the conversation, but remember that as a member of the community, you need to do a fair share of listening. Engage in conversations with community members, share ideas and actively participate. Recruiters should be transparent about their connection to the employment brand because creating an authentic brand is one of the most crucial pieces to online success [13]. Facebook, Twitter etc. have become a personal, product and corporate branding hub in the world. Every brand that exists on social networking sites has the same core features and benefits, such as the ability to create a page, share resources, add multimedia and much more [6]. The effective brand communication strategies are analyzed to find the impact among the users.

1.3 Social Networking Sites

Social networking is the classifying and expanding of individuals into specific groups or communities creating connections through individuals, groups or organizations [4, 19]. The word "social" refers to a people-to-people interaction and may consist of a set of customers, employees or organizations which have a relationship that holds them together [8]. Examples of famous global Social networking sites are Facebook with over 500 million active users and LinkedIn with over 90 million users worldwide in 2011. Social networking sites facilitate meeting people, finding like minds, sharing content [4, 19].

A social networking site creates network communication among the user community. Though social networking site serves for communication purposes among special interest groups, the marketing strategy has also entered this medium for its reach. People get exposed to various kinds of brand communication through this media. They tend to interact with the brand and also get awareness about the brand and its service in an interesting way [15]. In recent trend of marketing in social networking sites, various brand communications are widely used to attract targeted leads [11]. Social networking sites are more collaborative and interactive in comparison to traditional media followed by marketers.

1.4 Brand Pages on Social Networking Sites

Social networking sites, like Facebook, Twitter or Netlog, provide the opportunity for brands to create their own online profile. They are called "brand pages", "fan pages", "groups" or "profile pages" depending on the network. According to Lee [14], "brands become members of the social network like other users". Network users have then the opportunity to associate with a brand. The communication on these pages can go in four different directions: brand to member, member to brand, member to member, member to outsider [7]. Unfortunately, most advertisers still use these social network tools to push product information rather than to invite people to interact. Brand pages, are, when used properly, "a priceless medium to gauge what your marketplace is saying about you and/or your company" [9]. Practitioners should understand that they share control of the brand with consumers who want to engage with them in a rich online dialogue. Brand pages on Social networking sites qualify on all the characteristics of online communities: they are online, not limited to business transactions and allow information exchanges and influence games between members [12]. Trusov et al. [19] even freely use Social networking sites and online communities interchangeably. The appearance of Social networking sites features has introduced a new organizational framework for online communities and a vibrant new research context [1]. Facebook, Twitter etc. have become a personal, product and corporate branding hub in the world. Every brand that exists on social networking sites has the same core features and benefits, such

as the ability to create a page, share resources, add multimedia and much more [6]. The effective brand communication strategies are analyzed to find the impact among the users.

2 Building Successfull Brand by Using Social Media

Food industry in Slovenia is one of the most competitive international branches and is in its mature stage of development. In order to maintain their competitive position on the market, companies have to constantly prove themselves by providing supplies of innovative products and processes. New products and strong brand play a very important role in this industry. It is large and competitive industry with many active companies [22, 23]. The producers of high-volume products are aware that a company's leading role on the market is ensured through strong and established brands and loyal consumers. Today, the labelling with brands is such a strong factor that there are hardly any products without a brand. Various literatures provide tons of advices and rules on how to create or shape a successful brand. On the other hand, decisions regarding the policy of brands are far from being straight-forward and require numerous researches and considerations [21, 23]. Objectives of the study in this paper is to analyze the effective communication strategy through social networking media. It is important to study the effective way of communication in branding the product in social networking media and analyze its reach among the consumers. The paper explores the consumer engagement practices adopted by social networking media for building the brand. Dialogue between consumers and the brand is presented in the paper on the case of a leading brand in the category of fast moving consumer goods on Slovenian market. So, this study would help to know the effectiveness of communication and strategy done through social networking media which make the target audience to participate in this kind of advertising. This study would help the advertisers to understand the effective communication strategy to communicate their brand among the users.

Today Internet is widely spreading as a communication media in Slovenia. Emergence of the information super highway has revolutionized the way media is created and consumed. Previously media used to be created by media firms who are the content generators as well as the content owners. And the content used to broadcast to the consumers by the media. This concept has undergone an elementary change, now anyone can create content best known as User Generated content and share it with others using platforms like Blogging, Social Networks, YouTube etc. Now the consumers of the media have converted into media creators and the content also distributed on the internet through social networking and people connect with this content through comments. Social networking has become more popular among everyone. Consumers are getting more connected and communicative with their networks and technology allows them to voice their opinions rapidly. Today Slovenian consumers can make use of facilities like E-banking-retailing-shopping etc. with more confidence and trust as never before. Internet has proved to be more than just

emails and Google search. With the advent of networking media, broadband and Web 2.0 now many people are joining the social networks like Facebook, Twitter and thus it is easier for the marketers to spread the word about them over the network. To bridge the gap between the consumers, organization, marketing and media planning people there is the need of uniformity and trust in the social media. Social media, community networking, blogging, twitting, etc. were beyond anticipation for large mass of people. E- Shopping and E- bookings, E-learning and online dating have achieved good heights and became popular among Slovenian consumers. Another area of opportunity for building brand is Viral marketing. Viral marketing, also known as word-of-mouth (WOM) or "buzz marketing", is the tactic of creating a process where interested people can market to each other. Implementing an effective viral marketing strategy is a good way to get people talking about your business. A successful viral marketing campaign can be developed through social networking media like Twitter, Facebook etc. With the information available on online network the marketers have the knowledge of the needs and wants of different level of customers. Word-of-mouth is a particularly powerful medium, as it carries the implied endorsement from a friend. The Internet, with its e-mail lists, web sites, chat rooms and bulletin boards, makes communication tighter, and word-of-mouth even more effective. As a result, viral marketing is an effective tool in getting a message out fast, with a minimal budget and maximum effect. If a company can provide a strong enough incentive for customers to share their lists of personal contacts, whether for communications or community, they will have a powerful viral opportunity at their disposal. A good virus will look for prolific hosts (such as students) and tie into their high frequency social interactions (such as e-mail and messaging) [20, 23].

Next chapter provides some information about leading Slovenian brand in the category of fast moving consumer goods. Chapter also provides a mix of activities that have been carried out as part of marketing campaign by using social networking media.

2.1 Leading Slovenian Brand in the Category of Fast Moving Consumer Goods

Due to data protection and providing discreet company for analyzed brand we used in this paper instead of brand name label X. So Brand X is leading Slovenian brand in the category of fast moving consumer goods, precisely in the category of chicken sausage products. In Slovenia and in the region South-East Europe it has been an "icon" for 38 years: it is present on the market since 1974. It became first chicken pariser sausage on the European market, best selling sausage in its category. The brand X is consistent in its quality, its recipe remained unchanged from the very beginning. Brand X is a synonym for pleasure, emotions and quality. Brand X became the first customer's choice: leading brand in the region in terms of sales

and market share. Brand X personality is matching the characteristics of brand X target group – generation Y (witty, funny, dynamic, full of energy), which always needs something new and challenging. Brand name X doesn't bring any negative associations neither regionally, neither by any consumer group. The Slovenian brand is considered as a trustworthy market – leader in the chicken sausage category with a lot of positive associations: packaging, reasonable price, wide product range with long tradition. Slovenian brand X stands for values like trust, safety, quality and loyalty.

Advertising campaign were designed on irresistible desire for brand X. Basic guideline of the market communication campaign was the orientation towards consumers and their benefits. The next guideline was the creativity as the most powerful marketing tool to create a brand. After defining the marketing and communicative goals and target group to which the message is intended, what followed was the stage of defining the implementation of the creative strategy. The key element of the creative strategy was humor with the core communications like "Brand X is not mad, we are mad about brand X".). Advertising campaign took place in the media in the time from 1.4. to 31.7.2012. Campaign was spread across web site for brand X, internet media with web advertising on Windows Live Messenger, 24ur.com, izklop.com, Cosmopolitan, Igre123.com, Ringaraja.net, Facebook.com, YouTube.com.

The marketing goals of the campaign were:

1. To retain 46.4 % volume and 61.6 % value share for brand X on an annual basis in category of poultry pariser sausages.
2. To achieve high (90 %) campaign recognition by the target group.

The communicative goals of the campaign were: to increase the involvement of consumers and their engagement with the brand X. Measured through key performance indicators for the period from 1.4. to 31.7.2012.

1. Web site for brand X:
   - To increase the attendance of web site for brand X with basic 3.367 unique visitors and 4.581 visits in the previous 3 month period to 100.000 different visitors and 200.000 page views.
2. Facebook:
   - Brand X activate in the Facebook environment – to increase the number of Facebook fans: from 5.882 fans before campaign to 20.000 fans after campaign.
   - To increase the level of interaction with the posts (min. 500 comments, min. 1,000 likes).
   - To increase the number of female Facebook fan (25+ years) of 100 % (key decision makers on daily and weekly purchases).
3. Fun Club for brand X:
   - To acquire at least 10.000 new email contacts for future direct communication.

4. YouTube:
   – To achieve at least 50.000 views on video for brand X.

The second phase of the campaign was included different viral videos "Mad about brand X". The videos were posted on YouTube.

2.2 Research Design

To analyze and find the effectiveness of communication strategy to building a relationship with the analyzed brand, communication through social networking media was done with the survey method in the research.

2.2.1 Methods of Data Collection and Sample

Survey was conducted randomly among Facebook user community, by sending questionnaire through online to collect the individual opinion from the respondents. Non probability sampling technique is used to collect the opinion from the online respondents. The total population is social networking user community, but to collect the effective data the sampling is constrained to the target population like young adults, within the age of 18 years to 55 years. The sampling size is 400. The sample consisted of 250 women and 150 men. The majority of respondent were between 25 and 45 years old. More than half of them had finished high school (56 %), 24 % had a higher, high or more education, 8 % had a primary school and 12 % of respondents completed secondary school. 85 % of respondents currently living in cities, visit rural areas regularly, at least once a month. The respondents are interested in using internet and social networking sites often or very often, suggest that the survey respondents provide an interesting study group for this issue.

2.2.2 Data Analyses

Analysis of quantitative data provided by mentioned questionnaire used the Statistical Package for Social Sciences (SPSS 17.0). The data obtained from the survey were analysed with univariate analysis in order to check distributions of frequencies and to detect possible errors occurring during the research and/or data entry. The level of comparison was set at, 0.05. Prior to hypotheses testing, factor analysis was conducted to determine the independent variables of the study. Ambiguous items were eliminated from the survey by varimax rotation. The data collected from the surveys was put through a validity assessment (KMO = 0.898; Bartlett's Test sign = 0.000) thus revealing that the sample of the study was appropriate for factor analysis and that there is a strong relationship between the variables. The reliability

analysis of the research instrument yielded a Cronbach's Alpha value of 0.89 and a significance level of 0.000.

2.2.3 Research Hypotheses

The survey aims to examine a number of research hypotheses formulated based on the literature review. More specifically, the research hypotheses are:

H1: More than 80 % of respondents are aware of social networking sites
H2: More than 90 % of respondents are aware of social networking sites for brand X.
H3: More than 70 % of young population (between 18 and 35 years old) have high level of awareness about internet marketing campaign for brand X.
H4: More than 80 % of user agreed that the communication strategy used in brand X communication creates impact on brand effectively.
H5: The target group for brand X is young population, between 18 and 35 years old.
H6: An effectiveness of communication and strategy done through social networking media could increase brand relationship with young people.

3 Results and Discussion

3.1 Awareness of Social Networking Sites and Internet Marketing Campaign for Brand X

From this result, it is observed that nearly 99 % of the internet users are aware of social networking sites and only 1 % of them are clueless. Although the concept of computer-based communities dates back to the early days of computer networks, only some years after the advent of the internet online social networks have met public and commercial in a successive manner. At the most basic level, an online social network is an internet community where individuals interact, often through profiles that represent their selves to others [5]. Social networks have grown rapidly, and some like Facebook, Youtube, have achieved the mass market and penetrated in a few months since their inception, such applications have infringed their users in different strategy to interact with more people [11]. Also from this results it is observed that nearly 95 % of the internet users are aware of internet marketing campaign for brand X. More than 72 % of young population (between 18 and 24 years old) had high level of awareness about internet marketing campaign for brand X. Based on the presented results the hypothesis 1 and 2 were confirmed. More than 82 % of young population (between 18 and 35 years old) had high level of awareness about internet marketing campaign for brand X. Based on the presented results the hypothesis 3 was confirmed.

3.2 Accessibility of Ads in Internet Marketing Campaign for Brand X

Only 3 % of the total samples say that they have never accessed or shown interest to the ads displayed in social networking sites for brand X. 87 % of the respondents use to access often and were interested to listen to the advertisements for brand X, 10 % of the respondents use to access sometimes and were interested to listen to the advertisements for brand X (Table 33.2). Today's customers want to be engaged differently than in years past and many traditional marketing tactics simply do not work anymore. Social media marketing is a revolutionary way to build solid relationships with customers long before first contact with fun, attractive messages and interactions [2].

3.3 Brand Communication That Attracts the Users

Interactive fan page for brand X attracts 45 % of users. 35 % of users are pulled their interests towards game for brand X and 15 % of users listened to viral video ads, 5 % of users are interested to traditional banner ads. Most of the social networking websites are enabling brands to engage the right people in the right conversation at right time [17]. Nowadays communication on branding in social networking sites is more personal, contentious, fascinating and influencing among the user community (Table 33.3).

Table 33.2 Accessibility of ads in internet marketing campaign for brand X

Accessibility of ads in internet marketing campaign for brand X	%
Often	87
Sometimes	10
Never	3

Table 33.3 Brand communication that attracts the users

Brand communication that attracts the users	%
Interactive fan FB page for brand X	45
Banner ads for brand X	5
Web game for brand X	35
VIDEO ads for brand X	15

3.4 Impact on Building Brand Relationship with Target Group

According to the respondents, 95 % of user agreed that the communication strategy used in brand X communication creates impact on brand effectively and also which could help it to recall the same often and interactively. Based on the presented results the hypothesis 4 was confirmed. The remaining respondents almost 5 %, says that it does not create much impact on brand relationship but still effective for other communication purpose like sharing and chatting information. Based on research results the target group for brand X is young population, between 18 and 24 years old. The hypothesis 5 was confirmed.

3.5 Key Performance Indicators Campaign for Analyzed Brand

To analyze and find the effectiveness of communication strategy to building a relationship with the analyzed brand some key performance indicators campaign were analyzed. The finding of the study states that the internet marketing campaign for brand X was effective and resulted in the process of building a relationship with the brand. Key performance indicators campaign for brand X are:

- After campaign volume market share has been raised to 48.3 % and value market share for brand X has been raised to 63.7 % in the category of poultry pariser sausages on Slovenian market [16].
- The awareness of internet marketing campaign for brand X was 92 %, so the campaign for brand X has been recognized.
- We activated 178.682 unique visitors and reached 377.832 page views [10].
- With the mentioned marketing campaign brand X gained more than 20.000 Facebook fans; at the end of the campaign, brand X had 20.510 Facebook fans [10].
- We increased the level of interaction with the posts: we had got 938 comments and 2.841 likes [10].
- The number of female Facebook fans, 25+ years has increased by 431.63 % (from 784 to 3.384 Facebook fans) [10].
- We achieved 12.212 new email contacts for future direct communication
- We achieved 61.047 views on video “Mad about brand X” [10].

Based on the presented results the hypothesis 6 was confirmed. We can conclude that the marketing and communicative goals of the campaign for brand X were achieved.

4 Conclusions

With the advent of Internet Technology and social media revolution it's a must for the Slovenian company to take help of social networking websites for the sales generation as well as brand promotion to achieve global competiveness. Social networking sites users of Facebook and Twitter have become a personal, product and corporate branding hub in Slovenia nowadays in digital era. Every brand that exists on social networking sites has the same core features and benefits, such as the ability to create a page, share resources, add multimedia and much more. Social networking sites are filled with potential users who are mainly young adults. They spend more time in these networking sites due to heavy commercial contents, entertainment and social gathering. So, product or service communicators throng their ads in these areas with more and more interactive and with fascinating factors so that their brand identity is developed among the right choice of focused audience. Advertisers and brands uses social networking sites as the major resource for their promotion and developing brand identity among the focused market. This study would help the advertisers to understand the effective communication strategy to communicate their brand among the users. This paper highlights the need for marketers' presence in social media, the value that can be yielded from a well-designed internet marketing campaign and strategies for penetrating into the new media segment.

References

1. Brandt, C., Dessart, L., Pahud de Mortanges, C.: The impact of brand pages on brand page commitment and brand commitment in the context of Social Networking Sites (2010). Retrieved from http://www.brandmanagement.usi.ch/pag/private/Papers/saturday/session5/socialmediaandbrand/Brandt_brand%20pages_FINAL_20110130.pdf
2. Borges, B.: Marketing 2.0: Bridging the Gap between Seller and Buyer Through Social Media Marketing, pp. 45–63. Wheat Mark, Tucson (2009)
3. Cleland, R.S.: Building successful brands on the internet. A Dissertation Submitted in Partial Fulfilment of the Requirements of a Masters in Business Administration (MBA), University of Cambridge (2000)
4. Constantinides, E., Fountain, S.: Web 2.0: Conceptual foundations and marketing issues. J. Direct Data Digit. Market. Pract. **9**(3), 231–244 (2008)
5. Donath, J., Boyd, D.: Public displays of connection. BT Technol. J. **22**(4), 71–82 (2004)
6. Eric, E.: Growth Puts Facebook in Better Position to Make Money (2008). http://venturebeat.com/2008/12/18/2008-growth-puts-facebook-in-better-position-to-makemoney
7. Godin, S.T.: We Need You to Lead Us. Penguin Group, New York (2008)
8. Haythornthwaite, C.: Social networks and Internet connectivity effects. Inform. Commun. Soc. **8**(2), 125–147 (2005)
9. Holzner, S.: Facebook Marketing: Leverage Social Media to Grow Your Business. Que Publishing, United States of America (2009)
10. Internet: Google Analytics, Facebook Insights, YouTube.com Analytics
11. Jothi, P.S., Neelamalar, M., Shakthi Prasad, R.: Analysis of social networking sites: A study on effective communication strategy in developing brand communication. J. Media Commun. Stud. **3**(7), 234–242 (2011)

12. Kim, J.W., Choi, J.Q.W., Han, K.: It takes a marketplace community to raise brand commitment: the role of online communities. J. Market. Manage. **24**(3–4), 409–31 (2008)
13. Leary, R.: Developing a Social media Strategy (2009). Retrieved from www.kenexa.com
14. Lee, B.: Web 2.0 attracts brands looking for customer interaction. Admap **2007**, 30–33 (2007)
15. Nicole, K.: Building a Brand Through Social Networks (2007). Retrieved from http://mashable.com/2007/05/08/brand-social-networks/
16. Nielsen, Panel trgovin. Research report (2012)
17. Shih, C.: The Facebook Era. Tapping online social networks to build better products, reach new audiences and selling more stuff. Prentice Hall Publication, pp. 85–128 (2009)
18. Thompson, S.H.T.: Attitudes toward online shopping and the internet. Behav. Inform. Technol. **21**(4), 259–271 (2002)
19. Trusov, M., Buckling, R.E., Pauwels, K.: Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. J. Market. **73**, 90–102 (2009)
20. Vukasović, T.: Building successful brand by using social networking media. J. Media Commun. Stud. **5**(6), 56–63 (2013)
21. Vukasović, T.: Buying decision-making process for poultry meat. Br. Food J. **112**(2), 125–139 (2010)
22. Vukasović, T.: Searching for competitive advantage with the brand extension process. J. Prod. Brand Manage. **21**(7), 492–498 (2012)
23. Vukasović, T., Strašek, R.: A study on effective communication strategy in developing brand communication: analysis of social networking site. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, WCE 2014, pp. 690–693. London 2–4 July 2014
24. Zarrella, D.: The Social media marketing. O'Reilly Media (2010)

Chapter 34
An Active Integrated Zigbee RFID System with GPS Functionalities for Location Monitoring Utilizing Wireless Sensor Network and GSM Communication Platform

Farhana Ahmad Poad and Widad Ismail

Abstract An active integrated Radio Frequency Identification (RFID) system that operates in 2.45 GHz ISM band frequency is developed to support indoor and outdoor real-time location monitoring by utilizing Global Positioning System (GPS) in Wireless Sensor Network (WSN) and Global System for Mobile (GSM) communication platform. The proposed active RFID system is based on an automated switching mechanism between indoor and outdoor location and the capabilities of the system is extended by providing a contactless communication between the tagged items or persons and the monitoring station. There are two types of communication protocol; Reader Talk First (RTF) and Tag Talk First (TTF) involved in the proposed RFID system. The effectiveness of the proposed RFID system is evaluated based on the communication protocols implemented and the capability of the proposed RFID reader to read multiple tags is tested by analyzing the tag collection process in the RFID system. From the results, it is shown that the system with TTF protocol is better than the RTF protocol in terms of data collision and average time delay while performing transmission and reception processes.

Keywords Global positioning system • Global system for mobile • Radio frequency identification • Reader talk first • Tag talk first • Wireless sensor network

F.A. Poad (✉)
Department of Communication Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor 86400, Malaysia
e-mail: farhana@uthm.edu.my

W. Ismail
Auto_ID Laboratory, School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Seberang Perai Selatan, Pulau, Pinang 14300, Malaysia
e-mail: eewidad@usm.my

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_34

1 Introduction

Hajj (pilgrimage) is the largest gathering of Muslim worldwide and has a unique characteristics with regards to the people who attend it (pilgrims). According to figures released by the Central Department of Statistics and Information, more than three million pilgrims visited Saudi Arabia during the 2012 Hajj (www.saudiembassy.net, [14]). A total of 3,161,573 took part in the annual pilgrimage, with 1.4 million from Saudi Arabia and the majority around 1.7 million visiting from overseas and around 26 thousand pilgrims are from Malaysia and the total was up by eight percent in 2013. Such a setup poses a real challenge to the authorities in managing the crowd, tracking missing pilgrims and identifying lost, dead and injured pilgrims. In such a scenario, there is a need for a robust system for pilgrims to identify and track their locations especially during medical emergencies. However, the system is not limited to Hajj pilgrims application, but also suitable for other applications such as supply chain management [7], animal tracking [12], asset tracking [6], solid waste monitoring [2], crowd control application [15] and the most popular application demand have recently been is to support information and communication technologies in collaboration during emergency response [11] and disaster management [3].

Passive and active RFID systems have been tested in the past by [8] with limited success and other approaches such as image based tracking system are not suitable for a large crowd [5]. Thus the idea using WSN is introduced [9] to provide the location tracking for Hajj pilgrims. However, the work done by Mohandes et al. (2011) only focused on providing location data for outdoor environment. Thus, this research work combined the active RFID system and WSN platform in order to come out with a new RFID system consists of a modified active RFID reader and active RFID tag embedded with GPS and GSM in WSN platform. The proposed RFID system is developed by applying an automated switching mechanism in order to provide identification and sequence location detection for the indoor and outdoor locations, respectively. The RSS value is chosen to calculate the distance and perform the location tracking since it is an inexpensive Radio Frequency (RF) based approach with low configuration requirements and can be retrieved from the active RFID tag itself without using any external hardware devices [4]. This work, however, focused on the system design, implementation and testing of the proposed RFID system in order to evaluate the effectiveness of the proposed RFID system.

2 System Design

The work presented here is an extended version of previous research work done by Poad et al. [10] which focused on the automated switching mechanism for indoor and outdoor location tracking with embedded RFID and GPS in WSN platform. The GPS receiver covered outdoor location tracking that is fulfilled by satellite system,

while the active RFID tag provides an identification for each tag holder and covered indoor location tracking especially near or inside buildings, which cannot be tracked by GPS technology. In WSN, one of the methods uses to localize tag location is by measuring RSS, and translates the RSS value into the distance between reader, routers and tags. Since the RSS does not need any hardware modification in order to extract the value, it has been utilized in the proposed RFID system to estimate the location of the active RFID tag for indoor location tracking. The work presented by Poad et al. [10] is later extended to include the function of GSM communication in order to provide an alternative way when the active RFID tag is out of wireless network coverage. Thus make the active RFID system is a contactless system that can support two different types of communication technology on a single platform. Figure 34.1 shows the block diagram of the previous work done by Poad et al. [10], while Figs. 34.2 and 34.3 show the extended version of the embedded RFID tag and RFID reader presented in this work.

The proposed embedded RFID tag will be given to each person that consists of 2.45 GHz active integrated ZigBee RFID embedded with GPS and GSM technology to provide tracking for the indoor and outdoor location which utilized WSN platform. However, the GSM technology will be activated only when the embedded RFID tag is out of wireless network coverage. All the information gathered from RFID tags will be networked to RFID reader connected to the host computer at the monitoring station via WSN platform, otherwise the location data from the embedded RFID tag will be sent through short messaging system (SMS) to an authorized person via GSM platform. The RFID tag periodically sends out location data obtain from GPS receiver to RFID reader at the monitoring station to track and trace the movement and sequence location of the tagged items or persons only when the GPS having valid signals from satellites. However, if there are no valid signals from satellites, the location tracking will be done with the RFID tag using

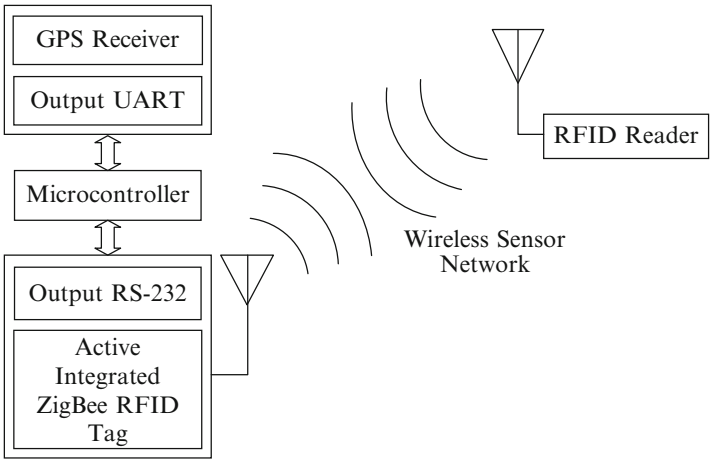


Fig. 34.1 Block diagram of existing embedded hardware (Poad et al. [10])

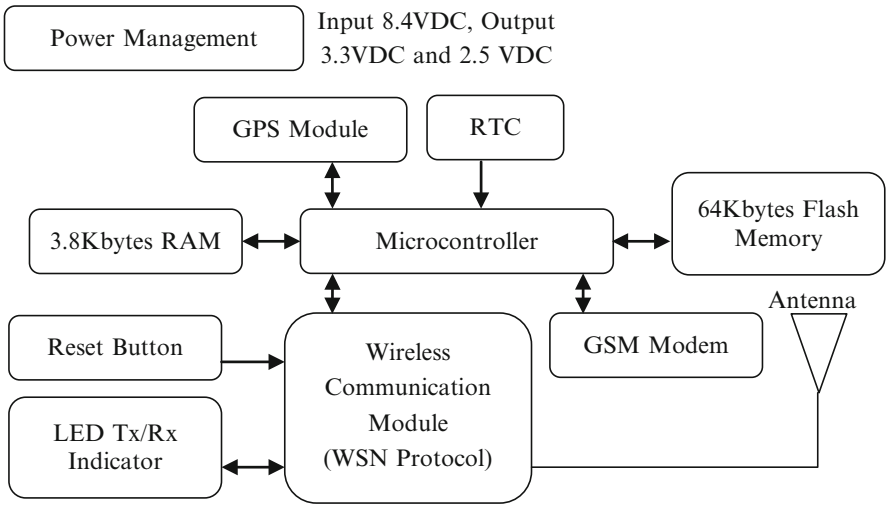


Fig. 34.2 Block diagram of the proposed embedded active RFID tag

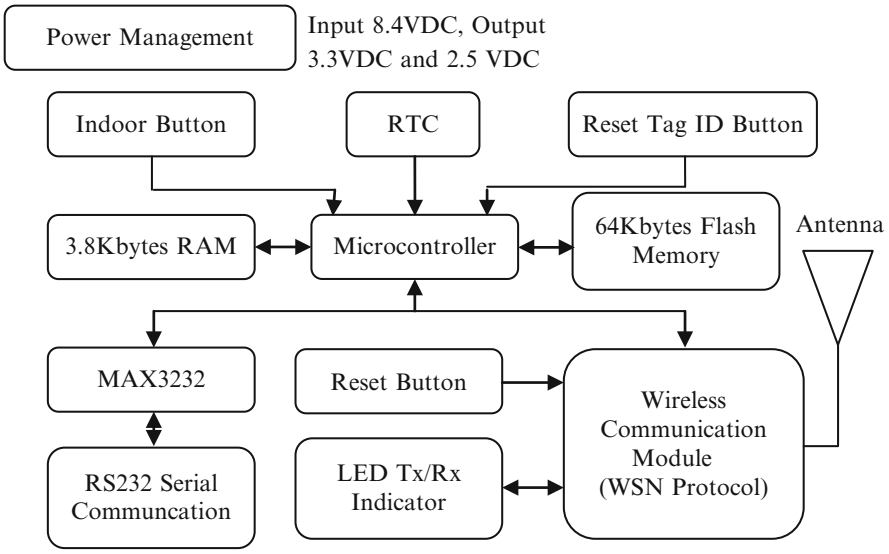


Fig. 34.3 Block diagram of the modified active RFID reader

the RSS values retrieved from the wireless module. The RSS value is periodically or manually request by the RFID reader from embedded RFID tag, which can be used later to calculate distance between the RFID reader and the RFID tag. Instead of tracking indoor and outdoor, the RFID tag node identification (ID) can be reset from the RFID reader based on user requirement and this application contributes to machine to machine (M2M) communication without human intervention.

Previously, the experimental studies are done based on the propagation in an indoor and outdoor environment by extracting the RSS value of the embedded RFID tag to analyze the differences in propagation between indoor and outdoor [10]. Moreover, a comparison also has been made between standalone RFID tag and embedded RFID tag to study the performance before and after of embedment of GPS functionalities. In this work, the experimental study related to data collision is performed in order to study the effectiveness of the proposed RFID system while performing transmission and reception processes and to ensure that the proposed RFID reader has the capability of multiple reading at one time by analyzing the tag collection process.

3 Collision Data Performance Analysis

The goal of this test is to ensure that the modified active RFID reader has the capability of multiple reading at one time by analyzing the tag collection process in the RFID system. The RFID system used the CSMA-CA algorithm for collision avoidance implementation [4]. Therefore, non anti-collision protocol cannot be performed. As the distance between reader and tag increase, the reader should identify a number of tags at one time. When there is more than one tag situated within the read range of the reader, all the tags could send data at the same time, which could possibly lead to mutual interference. This event will cause data loss and it is defined as a collision [13]. Therefore the collision data performance has to be evaluated in order to study the effectiveness of the proposed RFID system.

3.1 Experimental Setup

There are two protocols involved in developing the communication between embedded RFID tag and active RFID reader which are RTF and TTF. The active RFID reader is programmed with API mode while the embedded RFID tag is programmed in AT mode. The anti-collision test has been conducted for both communication protocols. The test is conducted in the laboratory for RTF protocol, where all the equipment's are placed on the table such as computers, power supplies and measurement devices, while for TTF, the test is conducted at outdoor environment (field) where the GPS can provide the location data. The tag collection process has been repeated ten times in order to get the average data for analyzing purposes. The arrangement of the test bed has been shown in Fig. 34.4, where tag 1 is situated at 90° to Tag 3, while Tag 2 is 45° to Tag 1 and Tag 3.

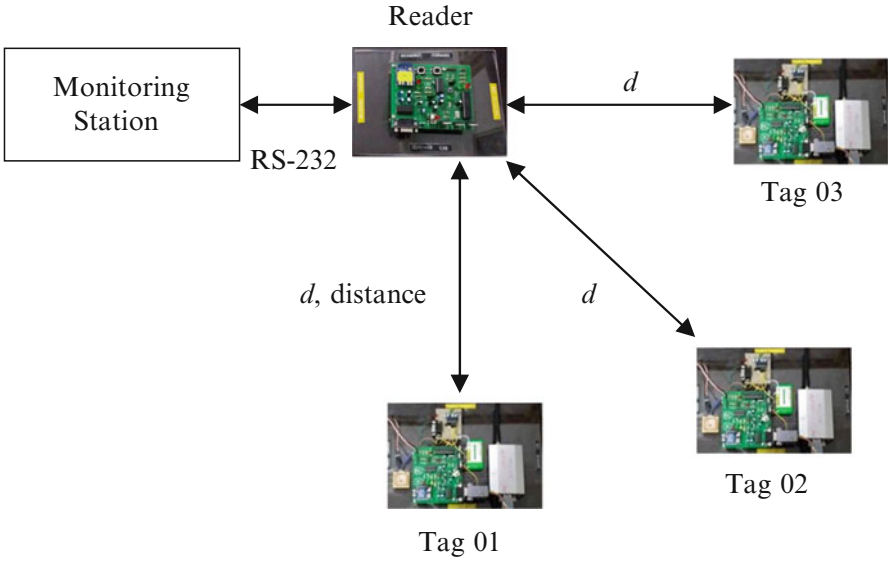


Fig. 34.4 Anti-collision performance test experimental setup

3.2 Anti Collision for RTF Protocol

The anti collision performance test for RTF is done by sending the collection command from the reader to the tags available in the network. The tags received the collection command, thus send an acknowledgment status and the requested data to the reader. This protocol is known as RTF. In this experimental study, three tags were used as a sample and the RFID reader broadcast a collection command for twenty times (once a minute). The tags received the collection command, send an acknowledgment to the reader and perform the reader request to measure the RSS in AT command mode. The anti-collision test for the tags with RTF protocol is conducted at 5 and 10 m distances due to the limited space of indoor locations. The reader will receive all the data send by the tags if no collision occurs. Thus, the percentage of data received will be 100 %. The percentage of data received can be calculated using Eq. 34.1

$$\text{Percentage of Data Received} = \frac{\text{Number of Data Received}}{\text{Number of Data Sent}} \times 100\,\% \qquad (34.1)$$

Data received from the tags are randomly received and displayed on the serial terminal as shown in Fig. 34.5. The variables of analyzing the data; 5 and 10 m distances have been utilized in order to see whether there are any differences in data losses. The tag collection process is performed for ten rounds to get an appropriate data variation. The data received are summarized in term of percentage as described

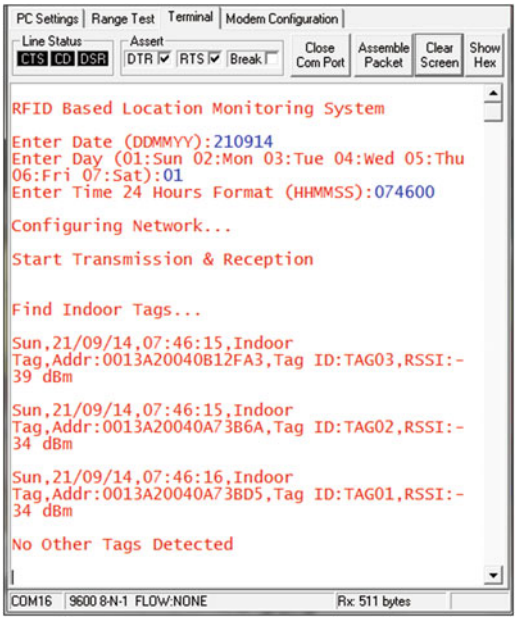


Fig. 34.5 Data randomly received from three embedded RFID tags based on RTF protocol

Table 34.1 The Percentage of data received at 5 m read range (RTF)

	Number of data received									
	1	2	3	4	5	6	7	8	9	10
Tag 01	16	16	19	17	19	17	17	17	17	15
Tag 02	17	18	18	17	17	19	19	18	18	18
Tag 03	18	17	20	14	15	17	17	17	17	18
Total data received (/60)	51	51	57	48	51	53	50	52	52	51
% Data received	85	85	95	80	85	88.3	83.3	86.7	86.7	85
Average (%)	86.2									

in Table 34.1. Each tag has sent 20 data to the RFID reader, but due to the collision, the reader only received 47 out of 60 total data in the first round of tag collection process. The total data received in the second and third round are 50 and 46 out of 60 total data. The reader has received 49 total data in the fourth round. The total data received in fifth round increased by three compared to the total data received in the fourth round. In the sixth round, the reader received 50 out of 60 total data. The reader received 47 total data in the seventh round and in the eighth round, the number of total data received is remained same with previous. The total data received in ninth and tenth round are 51 and 53, respectively. The average percentage of the data received for 5 m read range is about 81.99 %. The anti-collision test has been extended from 5 to 10 m distance between the tag and reader. The data received percentage of the RFID system for 10 m read range is summarized in Table 34.2. The Percentage of data received at 5 m read range (RTF) collection process has been

Table 34.2 The percentage of data received at 10 m read range (RTF)

	Number of data received									
	1	2	3	4	5	6	7	8	9	10
Tag 01	15	16	16	17	18	19	16	15	17	18
Tag 02	17	16	13	16	16	16	17	17	18	18
Tag 03	15	18	17	16	18	15	14	15	16	17
Total data received (/60)	47	50	46	49	52	50	47	47	51	53
% Data received	78.3	83.3	76.7	81.7	86.7	83.3	78.3	78.3	85	88.3
Average (%)	82									

repeated for ten times and each tag sends 20 data (once a minute) to the reader. In the first and second round, the reader received 51 out of 60 total data from the three tags. The number of total data received by the reader in the third round is increased by six compared to the first round. The total data received from reader is 48 and 51 in fourth and fifth round respectively. The reader received 53 total data in the sixth and seventh round. In the eighth round, total data received is 50 out of 240. The total data received in ninth round increased by two compared to eighth round, while the total data received in the last round are 51 out of 60. The average percentage of data received for 10 m distance is 86.16 %.

From the results, it can be concluded that the efficiency of anti-collision for the tags with AT command mode system is 81.99 % and 86.16 % for 5 and 10 m distance, respectively. The results of data collection for 5 and 10 m distances are compared and it is show that the number of data losses increases as the distance between tag and reader increases.

3.3 Anti Collision for TTF Protocol

The anti collision performance test for TTF is done by sending the location data provided by the GPS receiver from embedded RFID tag to the RFID reader. Later, the RFID reader sends an acknowledgement packet to all tags that are successfully sent their packet, thus extracted and displayed the data from RFID reader to a host computer. This protocol is known as TTF. Three tags are used as a sample in the experimental study. The amount of data sent is set to be 70 for each tag. The tag collection process is performed for ten rounds to get an appropriate data variation. The anti-collision test for the tags with TTF protocol is conducted at 5 and 25 m distances. The efficiency of anti-collision for the tags with TTF protocol is 100 % at 5 m and 100 % at 25 m distance, respectively. Data transmitted from the embedded RFID tags are randomly received and displayed on the serial terminal as shown in Fig. 34.6. The percentages of data received are summarized in Tables 34.3 and 34.4 for 5 and 25 m, respectively. Since the tag developed with TTF mechanism is a non AT command mode, thus the system provided better anti-collision performance

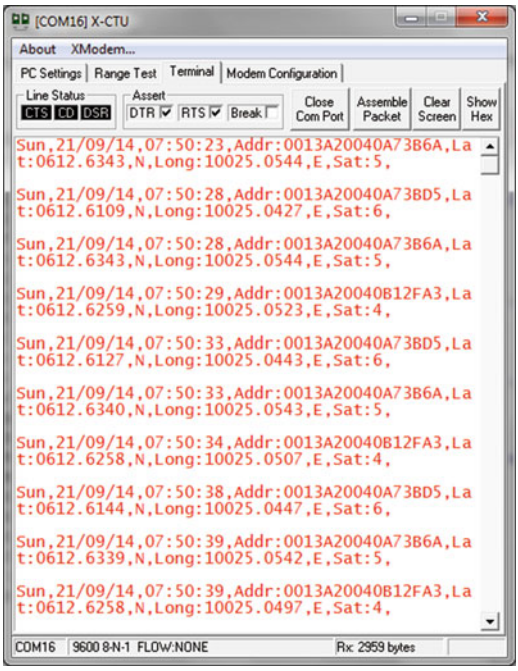


Fig. 34.6 Data randomly received from three embedded RFID tags based on TTF protocol

Table 34.3 The percentage of data received at 5 m read range (TTF)

	Number of data received									
	1	2	3	4	5	6	7	8	9	10
Tag 01	70	70	70	70	70	70	70	70	70	70
Tag 02	70	70	70	70	70	70	70	70	70	70
Tag 03	70	70	70	70	70	70	70	70	70	70
Total data received (/60)	210	210	210	210	210	210	210	210	210	210
% Data received	100	100	100	100	100	100	100	100	100	100
Average (%)	100									

Table 34.4 The percentage of data received at 25 m read range (TTF)

	Number of data received									
	1	2	3	4	5	6	7	8	9	10
Tag 01	70	70	70	70	70	70	70	70	70	70
Tag 02	70	70	70	70	70	70	70	70	70	70
Tag 03	70	70	70	70	70	70	70	70	70	70
Total data received (/60)	70	70	70	70	70	70	70	70	70	70
% Data received	100	100	100	100	100	100	100	100	100	100
Average (%)	100									

compared to tag developed with RTF protocol. The variables of analyzing the data; 5 and 25 m distances have been utilized in order to see whether there are any differences in data losses due to the increased in range.

4 Conclusion

The extended version of an active RFID system is successfully implemented and the performance of the system is evaluated through the anti collision performance in order to study the effectiveness of the proposed RFID system while performing transmission and reception processes and to ensure that the proposed RFID reader has the capability of multiple reading at one time. The tag collection process has been repeated ten times for data analyzing purposes and the percentage of the data losses are calculated. Since there are two communication protocols; RTF and TTF, thus the anti-collision test is conducted for both communication protocols. In RTF protocol, three tags are used in the analysis, where the RFID reader starts the transmission by broadcasting the collection command for twenty times (once a minute). The tags received the collection command and perform the reader request to measure the RSS in AT command mode. The anti-collision test for the tags with RTF protocol is conducted at 5 m and 10 m distances due to the limited space of indoor locations. The results show that the efficiency of anti-collision for the tags with the AT Command mode system is 81.99 % and 86.16 % for 5 m and 10 m distance, respectively. The amount of data losses increased as the distance between tag and reader are increased. Based on these findings, it can be concluded that if a network having more than two devices operates in AT command mode approach, the data transmission and reception will face problems with data collision and time switching between in and out of the command mode.

In the TTF protocol, three tags are used in the analysis and the tags are triggered by a GPS receiver and send the location data to the proposed RFID reader. The anti-collision test for the tags with TTF protocol is conducted at 5 and 25 m distances, respectively. The efficiency of anti-collision for the tags with TTF protocol is 100 % at 5 m and 100 % at 10 m distance, respectively. By comparing the result between RTF and TTF protocols, the TTF protocol shows that the anti-collision is better with 100 % data received than the RTF protocol with 81.99 % efficiency. In comparison with Alejandro et al [1] , the anti-collision efficiency is nearly 100 % by using the proposed CSMA-MS algorithm using TTF protocol. In terms of time delay between transmission and reception for both protocols, the RTF protocol takes about 87.7 ms to complete the data transmission and reception, while the TTF protocol takes about 61.2 ms average delay to complete the transmission and reception between RFID reader and tag. Table 34.5 shows the average time delay in millisecond (ms) for RTF and TTF protocols.

From the results, it can be seen that the tag developed with TTF protocol provided better anti-collision performance and average time delayed compared to tag developed with RTF protocol.

Table 34.5 The average time delay for RTF and TTF protocol

Protocol		Average time (ms)	Total average time (ms)
RTF	Transmit and receive acknowledgement	43	87.7
	Reply to reader request and receive acknowledgement	44.7	
TTF	Transmit and receive acknowledgement	61.2	61.2

Acknowledgments The authors would like to thank the Malaysia Ministry of Higher Education (LRGS fund) for sponsoring the research and development of this project. Also, special appreciation to Prof. Dr. Mohamad Kamal Abdul Rahim from Universiti Teknologi Malaysia, head of the LRGS project and Associate Professor Dr. Alyani Ismail, Universiti Putra Malaysia for the support.

References

1. Alejandro Palomo López, M., Victoria Bueno Delgado, Esteban Egea López, Juan J. Alcaraz Espín, Javier Vales Alonso, CSMA Multi-stage Anti-collision Protocol for Active RFID Systems, RFID Technology-Concepts, Applications, Challenges, Proceedings of the 4th International Workshop, IWRT 2010, pp. 23–35, Funchal, June 2010
2. Arebey, M., Hannan, M.A., Basri, H., Begum, R.A., Abdullah, H.: RFID and Integrated Technologies for Solid Waste Bin Monitoring System, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2010, WCE 2010, pp. 29–33, London, 30 June–2 July 2010
3. Chatfield, A., Wamba, S.F. Tatano, H.: E-Government Challenge in Disaster Evacuation Responses: The Role of RFID Technology in Building Safe and Secure Local Communities, 43rd Hawaii International Conference on System Sciences (HICSS 2010), pp. 1–10, Hawaii, May 2010
4. Digi International, XBee/XBee PRO ZB RF Modules Product Manual (2014), pp. 1–158, Updated on 24 October 2014
5. Germa, T., Lerasle, F., Ouadah, N., Cadenat, V., Devy, M.: Vision and RFID-Based Person Tracking in Crowds from A Mobile Robot, 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 5591–5596
6. Hannan, M.A., Mustapha, A.M., Hussain, A., Basri, H.: Intelligent Bus Monitoring and Management System. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2010, WCE 2010, pp. 37–43, London, 30 June–2 July 2010
7. Shi, J., Li, Y., He, W., Sim, D.: SecTTS: A secure track & trace system for RFID-enabled supply chain. Comput. Indust. **63**(6), 574–585 (2012)
8. Mohandes, M.: RFID based system for pilgrims identification & tracking. J. Appl. Comput. Electromag. Soc. (ACES) **25**(3), 273–282 (2010)
9. Mohandes, M., Haleem, M. A., Abul-Hussain, A., Balackrishnan, K.: Pilgrim Tracking using Wireless Sensor Network, Workshop International Conference Advance Information Network Application (WAINA 2011), pp. 325–328, Singapore, 22–25 Mar 2011
10. Poad, F.A., Ismail, W.: Automated Switching Mechanism for Indoor and Outdoor Propagation with Embedded RFID and GPS in Wireless Sensor Network Platform. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 710–714, London, 2–4 July 2014

11. Ran Xu, Lili Yang, Shuang-Hua Yang.: Architecture Design of Internet of Things in Logistics Management for Emergency Response, Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things, IEEE International Conference on Cyber Physical and Social Computing, (iThings/CPSCom), pp. 395–402, Beijing, 20–23 August 2013
12. So–Hyeon Kim, Do Hyeun Kim, Hee Dong Park: Animal Situation Tracking Service Using RFID, GPS and Sensors, 2nd International Conference on Computer and Network Technology (ICCNT), pp. 153–156, Bangkok, 23–25 April 2010
13. Tao, C., Li, J.: Analysis and Simulation of RFID Anti-collision Algorithms. In: The 9th International Conference on Advanced Communication Technology, pp. 697–701, 12–14 Feb 2007
14. www.saudiembassy.net, The Royal Embassy of Saudi Arabia, Washington, DC. Accessed 31 Oct 2014
15. Yamin, M., Mohammadian, M., Xu Huang, Sharma, D.: RFID Technology and Crowded Event Management, 2008 International Conference on Computational Intelligence for Modelling Control & Automation, pp. 1293–1297, Vienna (2008)

Chapter 35
Leveraging MMWAVE Technology for Mobile Broadband/Internet of Things

Oluwadamilola Oshin, Oluyinka Oni, Aderemi Atayero, and Babasanjo Oshin

Abstract Today, almost every individual possesses at least one internet-connected device. According to Cisco, there were over 12.5 billion devices in 2010 alone. It has been predicted that 25 billion devices will be connected by 2015, and 50 billion devices by 2020; all contributing towards the Internet of Things (IoT). This rapid increase exposes the obvious need for enhancements in various underlying technologies. IPv6 for example, has been developed to provide 340 undecillion IP addresses, and 3GPP LTE and its further enhancements provides impressive high bitrates cost-efficiently. That been said, there is still a limit on the amount of data that can go through a frequency channel. Therefore, the surge in demand for data by the billions of devices emphasizes the need to re-visit spectrum planning. Beginning with a review on the success of unlicensed spectrum operations, this work looks into the potentials of complementing the licensed frequency bands with unlicensed by tapping into the advantages of millimeter wave access technology.

Keywords IEEE802.11 • Internet of Things (IoT) • mmwave • LTE • Spectrum • Unlicensed • WLAN

1 Introduction

Spectrum is regarded as the life-wire of the telecommunications industry [1] but is also a very scarce and expensive asset. Therefore increasing the capacity of a wireless cellular network does not automatically imply increasing the bandwidth. LTE-Advanced (LTE-A), Release 10, supports bandwidth increase through Carrier Aggregation of its Release 8/9 carriers, yielding a maximum of 100 MHz bandwidth. However, there is still a limit on the amount of data that can go

O. Oshin (✉) • O. Oni • A. Atayero
EIE Department, Covenant University, Km 10, Idiroko Road, Ota, Ogun State, Nigeria
e-mail: damilola.adu@covenantuniversity.edu.ng; oluyinka.oni@covenantuniversity.edu.ng; atayero@covenantuniversity.edu.ng

B. Oshin
Sigma Plc, 29, Durban Street, Wuse 2, Nigeria, Abuja, Nigeria
e-mail: oshincit@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_35

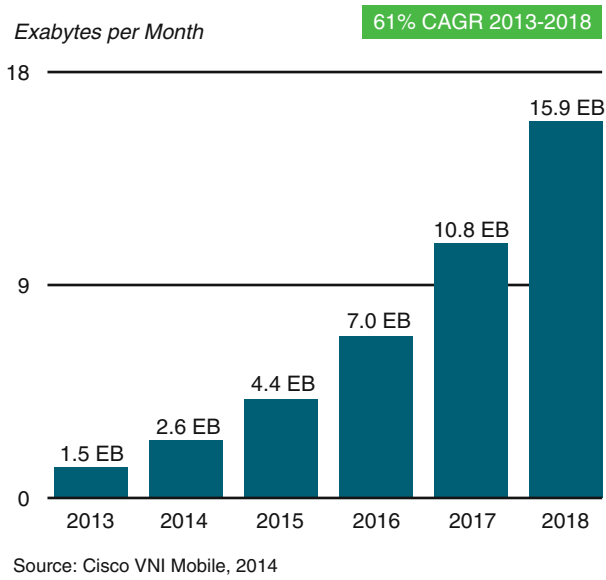


Fig. 35.1 Mobile data traffic growth [5]

through a frequency channel. All cellular network providers, cutting across evolving generations of access technologies, are restricted to utilizing carrier frequencies within the 700 MHz and 2.6 GHz bands, referred to as the most valued spectrum [2]. Within this limited spectrum range, they all attempt to provide wireless high-speed data rates and low delay services to customers. Hence, the adoption of LTE has largely been dependent on refarming of spectrum previously provisioned for GSM [3, 4] so as to take advantage of the potentials of these frequency bands.

With the unending increase in demand for mobile data (as seen in Fig. 35.1) by increasing billions of devices, a looming shortage of bandwidth is envisaged in the near future. If this issue is not addressed early, the surge in demand for data will likely overtake the capability of the current wireless networks to meet the demand.

At the ITU World Radio Conference (WRC) in 1992, 230 MHz of new radio spectrum was identified for IMT-2000 towards initial implementation and commercialization of 3G in the year 2000. Also, at the WRC in 2007, radio spectrums below 1 GHz and above 2 GHz were identified for IMT-A (4G) [6]. 4G only gained significant grounds in 2013/2014 with 331 LTE networks commercially launched, expected to get to 350 by the end of 2014 [7]. This gives an idea on the amount of time it takes for spectrum procurement and licensing after identifying potential spectrum, by the regulatory bodies. With the exponentially increasing demand rate for mobile broadband by the exponentially increasing devices, this time-line for spectrum licensing is unacceptable to meet the demand.

Therefore, unlicensed spectrum is the quickest and potentially sustainable route of supporting this rapid growth/ network expansion. It may be integrated into the network to complement services offered through licensed bands.

2 Unlicensed Spectrum – Brief History

The electromagnetic spectrum represents the range of all frequencies possible for electromagnetic radiation. It is within this range that several applications are derived. Figure 35.2 illustrates a simplified electromagnetic spectrum chart.

Wireless spectrum refers to the frequency range from 3 kHz to 300 GHz, where the United States (US) controls through the Federal Communications Commission (FCC), ownership and purpose of use of given sets of frequencies. When a carrier is said to be providing her services within a given frequency range, it means this carrier has obtained (through payment) a license to operate within the allotted frequency range in a given area.

In 1985, the US FCC decided to make available several bands of this spectrum for use without the need for a license. Spread Spectrum (SS) technology was used as the modulation technology – to spread the radio waves over a wide range of frequencies making the signals less prone to interferences [9]. This unlicensed spectrum found applications in extending the capabilities of existing hardware/services offered by the licensed spectrum.

2.1 IEEE802.11 Standards

In 1990, an IEEE committee called IEEE802.11 was set up to work on a standard to provide Wireless LAN services – by defining the over-the-air (OTA) interface between an access point and a wireless client, or between two or more wireless clients. By 1997, this standard was ready and published. It made use of the 2.4 GHz unlicensed spectrum, supporting a maximum data rate of 2 Mbps and channel

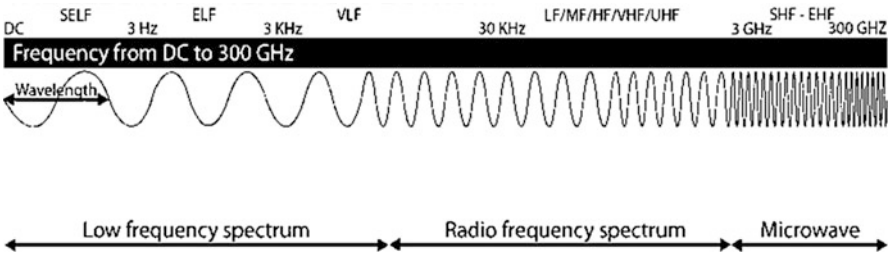


Fig. 35.2 The electromagnetic spectrum [8]

bandwidth of 20 MHz. In a very short time, this data rate became too slow for many applications; the committee continued to work to improve the standard. The next standards released were the IEEE802.11a and IEEE802.11b. The former supported 54 Mbps data rate (due to OFDM as its modulation scheme) within a channel bandwidth of 20 MHz but an operating frequency of 5 GHz while IEEE802.11b supported 11 Mbps and operating on the 2.4 GHz spectrum. The IEEE802.11b standard was cheaper, covered more distance and could penetrate obstructions; hence, gained more popularity than the IEEE802.11a. By June 2003, another variant was released, IEEE802.11g; this was an improvement which combined the advantages of IEEE802.11a and IEEE802.11b. It supported maximum data rate of 54 Mbps but had an operating spectrum of 2.4 GHz. As the demand for higher data rates, number of devices connecting to access points and the need to access sophisticated services (such as interactive gaming and high-definition video streaming) increased, the IEEE802.11 WG was setup in 2007 to proffer solutions to the foreseen challenges. The target of this group was to increase the data rate of operations under the 5 GHz bands to 1 Gbps shared by devices connected to an access point while able to support 500 Mbps for a single link. The second target was to achieve single-link data rate of up to 1 Gbps in 60GHz spectrum. However, another variant, IEEE802.11n was released by October 2009. IEEE802.11n was developed to operate at both 2.4 and 5 GHz spectrum bands. It supports wider channel bandwidth up to 40 MHz, uses OFDM and multiple antenna techniques (MIMO) – offering up to 600 Mbps data rate. In response to the target for very high throughput in the Gigabit class, in January 2013, a new amendment to WLAN standards was approved – IEEE802.11ad (Wi-Gig). This standard was designed to operate at 60 GHz; however, it had a feature which enabled the connected devices transition between the 60 band to the 2.4 and 5GHz spectrum bands. This ensured constant connectivity using the best conditions. The IEEE802.11ad supports up to 7 Gbps but is easily obstructed by water, walls and other factors; hence is most applicable for room use. Furthermore, another WLAN standard, IEEE802.11ac (Gigabit Wi-Fi), was released in January 2014 – supporting up to 7 Gbps within the 5 GHz spectrum band. This standard uses OFDM for its modulation and supports flexible channel assignments adding bandwidths of 80 and 160 MHz to the previously existing bandwidths. It also supports smart antenna techniques, such as MU-MIMO (Multi-user MIMO) and also Transmit Beamforming technology. Table 35.1 below gives a summary of the IEEE802.11 Wireless LAN standards till date.

2.2 Bluetooth

In 1998, the Bluetooth Special Interest Group was formed with the aim of developing a standard to allow pairing of devices for the purpose of data transfer and sharing. The first standard was released in 1999 and several improvements have followed in subsequent years. Like some Wi-Fi standards, Bluetooth operates in the

Table 35.1 Summary of IEEE802.11 standards

	802.11	802.11a	802.11b	802.11g	802.11n	802.11ad	802.11ac
Year	1997	1999	1999	2003	2009	2013	2014
Frequency	2.4 GHz	5 GHz	2.4 GHz	2.4 GHz	2.4, 5 GHz	60 GHz	5 GHz
Data rate	2 Mb/s	54 Mb/s	11 Mb/s	54 Mb/s	600 Mb/s	7 Gb/s	7 Gb/s
Modulation	DSSS, FHSS	OFDM	DSSS	DSSS, OFDM	OFDM	SC, Low power SC, OFDM	OFDM
Channel BW	20 MHz	20 MHz	20 MHz	20 MHz	20, 40 MHz	2.16 GHz	20, 40, 80, 160 MHz
Adv antenna technology	N/A	N/A	N/A	N/A	MIMO	Adaptive Beamforming	MU-MIMO, Beamforming

2.4 GHz frequency band but transfers data over shorter distances (100 m). It engages FHSS as its modulation technology at 1,600 hops/s. It was intended for personal area networks but is starting to fit into more sophisticated applications such as sensors for health and in-vehicle systems [10]. In 2011, enhancements called Bluetooth Smart and Bluetooth Smart Ready were introduced.

2.3 Radio Frequency Identification (RFID)

An RFID is a wireless device with two components: a tag and a reader. The reader emits and receives data and identity information through radio waves from the tag. RFID tags can either be passive (powered by readers) or active (powered by batteries). Being subject to local regulations, RFID also operates on four major unlicensed frequency bands: 125–148 kHz, 13.56 MHz, 915 MHz and 2.45 MHz.

The use of unlicensed spectrum by these reviewed technologies and others has led to several applications in industrial and medical equipment, inventory systems, remotely-controlled car door openers or garage door openers, wireless keyboards and very many others. Therefore, devices using unlicensed spectrum keeps increasing as innovators come up with new use cases. Also, CEA research found that of all installed number of tablets/smartphones with cellular capabilities, only about half of the owners subscribe (pay) for cellular internet connectivity [11]. Today, as the IoT ecosystem expands to Internet of Everything (IoE), through increasing number of communicating devices, processes, people and data, the unlicensed spectrum cannot but play the vital role as the backbone through which bits of information are transmitted from one device to another.

However, it is obvious that majority of these technologies rely on spectrum bands within the Low Frequency (LF) and a bit of Super High Frequency (SHF) bands.

Due to this growing vast majority, existing bandwidths have become insufficient. Therefore, researchers are beginning to consider the Millimeter Wave (mmwave) bands i.e. the 30–300 GHz spectrum, not only for wider bandwidth availability but gigabit speed for mobile broadband connectivity. The IEEE802.11ad standard operating at 60 GHz frequency, already presents some of the potentials of this spectrum with a 2.16 GHz bandwidth and up to 7 GHz data rate.

3 Millimeter Wave (MMWAVE) Technology

The mmwave frequencies are actually a part of the frequencies referred to as microwave frequencies; they represent frequencies further into microwave frequencies. As mentioned earlier, the mmwave frequencies are within the 30–300 GHz frequency bands, although, the industry considers mmwave frequencies to be from 10 GHz [12]. Due to the high frequencies, wider bandwidths are available resulting in multi-gigabit speeds. However, as frequency increases, received power drops unless offset by an increase in a combination of transmit power, transmit antenna gain and receive antenna gain. In other words, the high frequencies consequentially result in shorter wavelengths, therefore, shorter range capabilities. Mmwave propagation also faces challenges such as shadowing, in NLOS applications. Taking 60 GHz as an example, Fig. 35.3 is a plot of the free space pathloss in dB, showing increasing pathloss as the range increases.

It is a common myth that propagation at mmwave frequencies suffers severe attenuation due to rainfall and the atmosphere (oxygen) but experiments in real-life

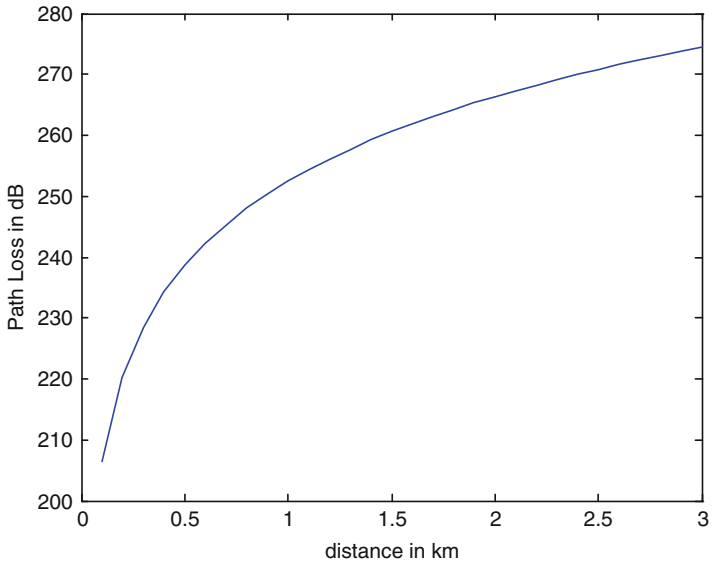


Fig. 35.3 Free space pathloss at 60 GHz

scenarios have shown that neither of these factors significantly affects propagation at certain mmwave frequencies. Also, with advancements in IC technology – enabling smaller antenna elements for operations in the mmwave frequency bands coupled with high gain and steerable antennas at the mobile and access points/base station [13], mmwave technology is able to support kilometres of range. The shorter wavelengths also enable advanced antenna techniques such as Massive MIMO and adaptive Beamforming. With these technologies, mmwave has great potentials in radio astronomy, mobile communications and wireless backhaul, satellite-satellite link, and other applications [14].

There is also ongoing research in implementing a fully integrated mm-wave overlay system over the existing cellular system [15].

3.1 Performance at MMWAVE Frequencies

The following frequency bands have been identified and analyzed extensively to be suitable for adoption to deliver multi-gigabit data rates offering up to multi- gigabits in bandwidth too [12, 14, 16, 17]: 23, 28, 38, 40, 46, 47, 49, and 60 GHz and E-band frequencies (70, 80 and 90 GHz). Some of these bands have been studied to obtain some propagation characteristics.

In [14], the authors identified potential mmwave frequency bands and also presented their respective available bandwidths. The 60 GHz band is said to be available worldwide, with 5 GHz bandwidth common to several countries. However, it suffers about 20 dB/km pathloss due to the atmosphere. At 70 GHz (E-band frequency), light rain causes 1 dB/km attenuation while heavy rain yields 10 dB/km attenuation. In [12, 17], studies were carried out on 28 and 73 GHz bands, for a cell size of 200 m in a densely populated urban environment, and transmitters were situated at rooftops of up to five stories high buildings. It was observed that at 73 GHz, heavy rainfall yielded 2 dB attenuation while 28 GHz bands yielded 1.4 dB attenuation. It was also observed in [16] that atmospheric absorption has no significant effect on 28 and 38 GHz frequency signals within the 200 m cell size. Penetration losses were also observed: they were seen to be way higher for outdoor propagation than indoors, hence the need for access points for effectual handoffs into buildings. At 73 GHz, a symbol rate of 1.536 Gsym/s was obtained with peak data rate of 15.7 Gbps using MIMO.

4 Conclusion

Mmwave technology has been proven by researchers to be a viable and potential solution to enhance capacity and data rate for full support of the IoT ecosystem and future mobile broadband needs. The size of bandwidth future communications require is domiciled at these high frequency bands. With appropriate support-

ing technologies, the flaws of mmwave frequencies can be circumvented. The strength of mmwave frequencies has been authenticated by the introduction of IEEE802.11ad.

References

1. Merz, P.: Optimizing Spectrum Usage for 2020 and Beyond. Nokia, Johannesberg Summit (2014)
2. Idaho National Laboratory: Idaho scientist creates solution for looming broadband shortage (2013) www.inlportal.inl.gov/portal/server.pt/community/newsroom/257/feature_story_details/1269?featurestory=DA_612083
3. Adu, O.I., Idachaba, F.E., Alatishe, A.: Refarming 1800 MHz GSM spectrum to LTE: the effects on coverage based on pathloss estimation. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 673–676. London, 2–4 July 2014
4. Gillet, J.: Spectrum refarming at 1800 MHz key to LTE device adoption. GSMA Intelligence (2012)
5. CISCO: Cisco Visual Networking Index: Mobile Data Traffic Forecast Update, 2013–2018 (2014)
6. ITU: Spectrum for IMT. www.itu.int/ITU-D/.../Spectrum-IMT.pdf
7. GSA: GSA updates on the LTE ecosystem, confirms smartphones driving LTE devices growth (2014). www.gsacom.com
8. Electromagnetic Field (EMF) Limits. Accessed November 2014. www.wirelesspowerconsortium
9. Intel: Helping Define IEEE 802.11 and other Wireless LAN Standards. Accessed 2014. www.intel.com
10. Bluetooth: Welcome to bluetooth technology 101: a brief tutorial on Bluetooth wireless technology (2014). www.bluetooth.com
11. Valuing Unlicensed Spectrum, April 2014. www.ce.org
12. Rangan, S., Rappaport, T.S., Erkip, E.: Millimeter-wave cellular wireless networks: potentials and challenges. Proc. IEEE **102**(3), 66–385 (2014)
13. Rappaport, T.S., Ben-Dor, E., Murdock, J.N., Qiao, Y.: 38 GHz and 60 GHz Angle-dependent propagation for cellular and peer-to-peer wireless communications. In: Proceedings of IEEE International Conference Communications, pp. 4568–4573. June 2012
14. Pi, Z., Khan, F.: An introduction to millimeter-wave mobile broadband systems. IEEE Commun. Mag. **49**(6), 101–107 (2011)
15. European Commision: Millimetre-wave evolution for backhaul and access. MiWEBA (2013–2014)
16. Rappaport, T.S., et al.: Millimeter wave mobile communications for 5G cellular: it will work! IEEE Access **1**, 335–349 (2013)
17. Rappaport, T.S.: Millimeter wave wireless communications: the renaissance of computing and communications. IEEE International Conference on Communications, Sydney. June 2014

Chapter 36
Fast Power-Efficient Techniques for Collision Detection in Wireless Sensor Networks

Fawaz Alassery, Walid Ahmed, Mohsen Sarraf, and Victor Lawrence

Abstract Recently a lot of research effort has been focused on Wireless Sensor Networks (WSNs) due to its various applications. Over the last few years, several techniques have been proposed for investigating the power consumption which represents one of the most challenges and main concerns in designing WSNs. Power consumption of nodes in WSNs has a great effect on the lifetime of network nodes which are difficult to replace or recharge their batteries. In this context, this paper represents a receiver approach for alleviating power consumption of WSNs. Unlike other power consumption techniques, instead of decoding every received signal at the receiver which consume too much power our approach studies the histograms of sensors' transmitted signals to detect collisions, so the receiver can determine when the transmitted signals can be decoded without wasting precious power decoding transmitted signals suffering from collisions. We also present a complexity and power-saving comparison between our novel approach and a conventional full-decoding algorithm in order to demonstrate the significant power and complexity saving advantage of our approach.

Keywords Efficient techniques in WSNs • Low computational complexity algorithms • Packets collision • Power consumption techniques • Signal statistics in WSNs • WSNs protocols

1 Introduction

Wireless Sensor Networks (WSN) consist of many sensor nodes distributed in various environments in order to perform specific tasks such as passive localization, target tracking, systems control, healthcare monitoring, air pollution and temperature monitoring, irrigation management and water monitoring, etc. [1]. In many cases each node in a WSN has a limited power source which is a small battery. After the initial deployment of nodes in an environment, the nodes must be active for a

F. Alassery (✉) • W. Ahmed • M. Sarraf • V. Lawrence
Electrical and Computer Engineering Department, Stevens Institute of Technology, Hoboken, NJ 07037, USA
e-mail: falasser@stevens.edu

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_36

long period of time. Therefore, power saving techniques may play a very important role in order to extend the lifetime of WSN nodes [2].

In term of power consumption in WSN transmission, it is obvious that the transmissions in a WSN follow different stages either in a transmitter or a receiver. Each stage has its own electronic circuit which consumes some power. Power consumption starts from baseband processing of data, amplification, filtering, modulation and RF front-end stages of the transmitted signal in a transmitter to RF front-end and demodulation processes in the receiver. In general, the sensor node has hardware blocks which cause energy loss due to the specific function that is performed by that block [3].

One of the main sources of overhead power consumption in wireless sensors is collision detection, since until the access node has expended the required power and processing-time to detect/decode the received packet, it wouldn't know that the packet has suffered a collision. Current collision detection mechanisms have largely been revolving around direct demodulation and decoding of the received packets and deciding on a collision based on some form of a frame error detection mechanism, such as a CRC check [4]. The obvious drawback of full-detection of a received packet is the need to expend a significant amount of energy and processing computation in order to fully decode a packet, only to discover that the packet has suffered a collision. In this paper, we propose a suite of novel, yet simple and power-efficient techniques to detect a collision without the need for full decoding of the received packet. Our novel approach aims at detecting collision through fast examination of the signal statistics of a short snippet of the received packet via a relatively small number of computations over a small number of received IQ samples. Hence, operating directly at the output of the receiver's analog-to-digital-converter (ADC) and eliminating the need to pass the signal through the entire demodulator and decoder line-up. Accordingly, our novel approach not only reduces processing complexity and hence power consumption, but it also reduces the latency (or delay) incurred to detect a collision since it operates on only a small number of samples in the beginning of a received packet instead of having to buffer and process the entire packet as is the case with a full-decoding approach. Furthermore, our approach does not require any special pilot or training patterns. It operates directly on random data, i.e., the received packet as is. We also present a complexity and power-saving comparison between our novel approach and conventional full-decoding (for select coding schemes) to demonstrate the significant power and complexity saving advantage of our approach. In addition, we show that with a relatively short measurement period, our scheme can achieve low False-Alarm and Miss probabilities, resulting in a reliable collision-detection mechanism. We also demonstrate how to tune various design parameters in order to allow a system designer multiple degrees of freedom for design trade-off and optimization.[1]

[1]For the remainder of this paper, we shall refer to our proposed approach as the "Statistical Discriminator, or SD" method. We shall also refer to the traditional full-decoding methods as "FD" methods.

The remainder of this paper is organized as follows. Section 2 investigates related works. Section 3 describes our proposed system. In Sect. 4 we define the algorithm and metrics, and show how to select the system thresholds. In Sect. 5, we compare the computational complexity of our metrics against commonly used decoding technique (i.e. Soft Output Viterbi Algorithm-SOVA). In Sect. 6 we show and discuss the results of the simulations used to model our system for different scenarios. Finally, we offer our conclusions in Sect. 7.

2 Related Works

Many techniques have been introduced in various studies aimed at maximizing WSN node lifetime by reducing power consumption. Variety of definitions for WSNs lifetime is introduced based on network connectivity, coverage, application requirements, and number of active nodes [5]. Power efficient techniques in WSNs have been categorized into five classes briefly introduced in the following:

First class is the power efficient techniques that focus on reducing the data processed and transmitted from the source sensor. In [6], authors use clusters in order to aggregate the information being transferred. They proposed LEACH (Low-Energy Adaptive Clustering Hierarchy) which is a clustering based protocol aimed to distribute the energy load among the WSNs nodes. Another data reduction strategy is proposed in [7] where it is based on avoiding transfer of the information to undesirable sensors. This can be done via defining a smaller dominating set of sensors when two hops are considered. Moreover, data reduction can be resulted from compression algorithms explained thoroughly in [8] where authors investigate compression algorithms applicable in WSNs such as coding by ordering, pipelined in-network compression, low-complexity video compression and distributed compression.

Second class of power efficient techniques in WSNs deals with controlling the topology via tuning the transmission power while maintaining the connectivity of the network. In this context, authors in [9] present a Local Minimum Spanning Tree (LMST) algorithm to control the wireless multi-hop topologies. In this algorithm each node builds its own LMST independently using locally collected data. This algorithm leads to further increase in network capacity and a significant reduction in power consumption. Furthermore, an Adaptive Transmission Power Control (ATPC) algorithm for WSNs is proposed in [10], where each node in the network builds a model which describes the correlation between the transmission power and link quality with its neighboring nodes.

Reducing unsuccessful end-to-end transmissions and avoiding nodes that consume too much power in routing packets of WSNs is the third class of power efficient techniques. Some protocols in this class use the advantages of mobility and broadcast communication to reduce the power consumption when sending packets to a sink node. Others protocols use the geographicalcoordination of source nodes

to determine their position when building the route that connects them to destination nodes [1]. A survey on energy aware routing protocols in WSN is provided in [11–13].

3 System Description

We assume a WSN where a number of intermediate sensors are deployed arbitrarily to perform certain functionalities including sensing and/or collecting data and then communicating such information to a central (access) sensor node. The central node may process and relay the aggregate information to a backbone network.

There are N wireless sensors that communicate to the central sensor node, where at any point in time, multiple sensors may accidentally transmit simultaneously and cause a collision. Without loss of generality, we shall assume for the sake of argument that one sensor is denoted a "desirable" sensor, while the rest of the colliding sensors become "interferers".

A commonly accepted model for packet arrivals, i.e., a packet is available at a sensor and ready to be transmitted, is the well-known Bernoulli-trial-based arrival model, where at any point in time, the probability that a sensor has a packet ready to transmit is $P_{Tr.}$[2]

Upon the receipt of a packet, the central node processes and evaluates the received packet and makes a decision on whether the packet is a collision-free (good) or has suffered a collision (bad). In this paper, we propose a suite of fast collision detection techniques where the central node evaluates the statistics of the received signal's IQ samples at the output of the receiver's analog-to-digital converter (ADC) directly using simple discrimination metrics, as will be explained in more detail in the following section, saving the need to expend power and time on the complex modem line-up processing (e.g., demodulation and decoding). If the packet passes the discrimination metric test, it is deemed collision-free and undergoes all the necessary modem processing to demodulate and decode the data. Otherwise, the packet is deemed to have suffered a collision, which in turn triggers the central node to issue a NACK message per the mechanism and rules mandated by the specific multiple-access scheme employed in the network.

4 Algorithm Description

As mentioned earlier, our proposed algorithm is based upon evaluating the statistics of the received signal at the receiver ADC output via the use of a simple statistical discrimination metric calculation that is performed on a relatively small portion of

[2]The actual design details and choice of the multiple access mechanism, e.g., slotted or un-slotted Aloha, are beyond the scope of this paper and irrelevant to the specifics of the techniques proposed herein.

the received IQ packet samples. The resulting metric value is then compared with a pre-specified threshold to determine if the statistics of the received samples reflect an acceptable signal-to-interference-plus-noise ratio (SINR) from the decoding mechanism perspective. If so, the packed is deemed collision-free and qualifies for further decoding. Otherwise, the packet is deemed to have suffered a collision with other interferer(s) and is rejected without expending any further processing/decoding energy. A repeat request may then be issued so the transmitting sensors to re-try depending on the MAC scheme. In other words, the idea is to use a fast and simple calculation to determine if the received signal strength (RSS) is indeed due to a single transmitting sensor that is strong enough to achieve an acceptable SINR at the central node's receiver, or the RSS is rather due to the superposition of the powers of multiple colliding packets, hence the associated SINR is less than acceptable to the decoding mechanism.

Let's define the k^{th} received signal (complex-valued) IQ sample at the access node as:

$$y_k = x_{0,k} + \sum_{m=1}^{N-1} x_{m,k} + n_k$$

where

$$y_k = y_{k,I} + jy_{k,Q}, \quad j = \sqrt{-1},$$
$$x_{0,k} = x_{0,k,I} + jx_{0,k,Q}$$

is a complex-valued quantity that represents the k^{th} IQ sample component contributed by the desired sensor, while

$$x_{m,k} = x_{m,k,I} + jx_{m,k,Q};\ m = 1, \ldots, N-1$$

is the k^{th} IQ sample component contributed by the m^{th} interfering (colliding) sensor. Finally, $n_k = n_{k,I} + jn_{k,Q}$ is a complex-valued Additive-White-Gaussian Noise (AWGN) quantity (e.g., thermal noise).

We propose three time-averaging statistical discrimination (SD) metrics that are applied to the envelope value, $|y_k| = \sqrt{y^2{}_{k,I} + y^2{}_{k,Q}}$, of the received IQ samples at the central node as follows:

1. Entropy (Logarithmic) based metric:

$$\begin{aligned} Log_{metric} &= \langle \log_{10} |y_k| \rangle \\ &= \tfrac{1}{K} \sum_{k=0}^{K-1} \log_{10} |y_k| \end{aligned} \tag{36.1}$$

2. Moment based metric:

$$\begin{aligned} Moment_{metric} &= \langle |y^t{}_k| \rangle \; ; \; t = 3, 4, 5, \ldots \\ &= \tfrac{1}{K} \sum_{k=0}^{K-1} |y^t{}_k| \end{aligned} \tag{36.2}$$

3. Signal Dynamic-Range Maximum-Minimum based metric:

$$MaxMin_{metric} = \frac{\max\limits_{k=0,\ldots,K} \{|y_k|\}}{\min\limits_{k=0,\ldots,K} \{|y_k|\}} \tag{36.3}$$

The computed statistical discrimination metric is then compared with a pre-specified threshold value that is set based on a desired signal-to-interference-plus-noise ratio (*SINR*) cut-off assumption, $SINR_{cut\_off}$. That is (and as will be described in more detail later in this paper), a system designer pre-evaluates the appropriate threshold value that corresponds to the desired $SINR_{cut\_off}$. If the SD metric value is higher than the threshold value, then the SD metric value reflects a SINR that is less than $SINR_{cut\_off}$ and the packet is deemed not usable, and vice-versa. Accordingly, a "False-Alarm" event occurs if the received SINR is higher than $SINR_{cut\_off}$ but the SD algorithm erroneously deems the received SINR to be less than $SINR_{cut\_off}$. On the other hand, if the SD algorithm deems the SINR to be higher than $SINR_{cut\_off}$ while it is actually less than $SINR_{cut\_off}$, a "Miss" event is encountered. Miss and False-Alarm probabilities directly impact the overall system performance as will be discussed in the following sections. Therefore, it is desired to minimize such probabilities as much as possible.

4.1 Threshold Selection

The decision threshold is chosen based on evaluating the False-Alarm and Miss probabilities and choosing the threshold values which satisfy the designer's requirements of such quantities. For example, we generate a 100,000 Monte-Carlo simulated snapshots of interfering sensors (e.g., 1 30 sensors with random received powers to simulate various path loss amounts) where for each snapshot we compute the discrimination metric value for the received total signal plus interference plus noise. In addition the simulator sweeps a range of threshold values for the snapshot at hand and determines if, for each threshold value, there would be an event of a False-Alarm or a Miss in order to count the probabilities of such events. At the end of the simulations the False-Alarm and Miss probabilities are plotted versus the range of evaluated threshold values which enables the designer to determine a satisfactory set point for the threshold.

5 Power Saving and System Throughput Analysis

To analyze the power saving of our proposed SD system we introduce the following computational complexity metrics:

$$F_B = S + P_{miss} F \tag{36.4}$$

$$F_G = S + (1 - P_{FA}) F \tag{36.5}$$

In above formulas, S is the number of computational operations incurred in our proposed approach, while F is the number of computational operations incurred in a full-decoding approach, P_{miss} and P_{FA} are the probabilities of Miss and False-Alarm events respectively. Hence, F_B represents the computational complexity for the case where the central node makes a wrong decision to fully-decode the received packet (i.e., declared as a collision-free packets) while the packet should has been rejected (i.e., due to collision). On the other hand, F_G is the computational complexity for the case where the central node makes a correct decision to fully decode received packet.[3]

In addition, and for the comparison purposes, we introduce the following formulae in order to compare the computational complexity saving achieved by our proposed SD approach (i.e. T_{SD}) over the FD approach (i.e. T_{FD}):

$$T_{SD} = F_B P_{collision} + F_G P_{no\_collision} \tag{36.6}$$

$$T_{FD} = F \tag{36.7}$$

In above formulae, $P_{collision}$ and $P_{no\_collision}$ are the probabilities of collision and no-collision events respectively. $P_{no\_collision}$ and $P_{no\_collision}$ have been obtained via Monte-Carlo simulation as follows: A random number of interfering sensors (maximum of 30 sensors) is generated per a simulation snapshot, where each sensor is assumed to have a randomly received power level at the access node (to reflect a random path loss/location effect). The generation of the interfering sensors is based on a Bernoulli trial model where it is assume that the probability of a packet available for transmission at a sensor (hence the existence/generation of the sensor for the snapshot at hand) is equal to α. If the total SINR is found to be worse than the cut-off limit, a collision is assumed and vice-versa. For our numerical example in this section we used $\alpha = 0.3$ and $SINR_{cut_off} = 5$dB. Also, we typically generate more than 100,000 snapshots in order to achieve a reliable estimate of the collision

[3] Our system throughput is defined as $Throughput = (1 - P_{FA})_{SD}$; Where P_{FA} denotes the False-Alarm probability.

probabilities. For the aforementioned choices of α and $SINR_{cut\_off}$, we found the collision probabilities to be $P_{collision} = 0.3649$ and $P_{no\_collision} = 0.6351$.

5.1 Comparing with Full Decoding

In order to assess the computational complexity of our SD scheme, we first quantize our metrics calculation in order to define fixed-point and bit-manipulation requirement of such calculations. We also assume a look-up table (LUT) approach for the logarithm calculation. Note that the number of times the algorithm needs to access the LUT equals the number of IQ samples involved in the metric calculation. Thus, our algorithm only needs to perform addition operations as many times as the number of samples. Hence, if the number of bits per LUT word/entry is equal to M at the output of the LUT, our algorithm needs as many M-bit addition operations as the number of IQ samples involved in the metric calculation.

As a case-study, we compare the complexity of our SD scheme with the complexity of a FD algorithm assuming a Soft Output Viterbi Algorithm (SOVA). SOVA has been an attractive choice for WSNs [14]. Authors in [15] measure the computational complexity of SOVA (per information bit of the decoded codeword) based on the size of the encoder memory. It has been shown in [15] that for a memory length of λ, the total computational complexity per information bit can be estimated as:

$$F_{\text{SOVA}} = 3 \times 2^{\lambda} + 9\,(\lambda + 1) + 16 \tag{36.8}$$

In contrast, our SD system does not incur such complexity related to the size of the encoder memory. In addition, our SD system avoids other complexities required by a full decoding such as time and frequency synchronization, Doppler shift correction, fading and channel estimation, etc., since our SD scheme operates directly at the IQ samples at the output of the ADC "as is". Finally, the FD approach requires buffering and processing of the entire packet/codeword while our SD scheme needs only to operate on a short portion of the received packet.

Now let's compute the computational complexity for our SD approach using the logarithmic (entropy) metric. Let's assume that the IQ ADCs each is D bits. Also, let's assume a $(\cdot)^2$ operation is done through a LUT approach to save multiplication operations. In addition, let's also assume that the square-root, $\sqrt{}$, is also done through a LUT approach. Hence, each of the I^2 and Q^2 operations consume of the order of D bit-comparison operations to address the $(\cdot)^2$ LUT. Then, if the output of the LUT is G bits, it follows that we need about G bit additions for an $I^2 + Q^2$ operation. Let's assume that the $\sqrt{}$ LUT has G bits for input addressing and K output bits. Then, we need about $G + 1$ bit-comparison operations to address the $\sqrt{}$ LUT. Let's assume a $log\,(\cdot)$ is also done througha K-bit-input/L-bit-output LUT. Hence,

a $\log\left(\sqrt{I^2+Q^2}\right)$ costs about K bit-comparison operations to address the $log\,(\cdot)$ LUT. Finally, for simplicity, let's assume that a bit comparison operation costs as much as a bit addition operation. Accordingly, the total number of operations needed to compute the $log\,(\cdot)$ for one IQ sample is:

$$2D + G + (G+1) + K = 2D + 2G + 1 + K \tag{36.9}$$

If we assume the IQ over-sampling rate (OSR) to be Z (i.e., we have Z samples per information symbol), then we need about $Z \times L$ bit additions to add the Z $\log(\cdot)$ values for every information symbol. Hence, for one information symbol, we need a total of:

$$(2D + 2G + 1 + K) \times Z + Z \times L = (2D + 2G + 1 + K + L)\, Z \tag{36.10}$$

Now if we assume an M-ary modulation (i.e., $\log_2(M)$ information bits are mapped to one symbol), then the computational complexity per information bit can be computed as:

$$S/\text{InfoBit} = \frac{(2D + 2G + 1 + K + L)\, Z}{\log_2(M)} \tag{36.11}$$

For example, in order to show the complexity saving of our SD algorithm, let's assume a QPSK modulation scheme ($M = 4$). Also, let's assume $Z = 2$ (two samples per symbol), and $D = G = K = L = 10$ bits, which represents a good bit resolution. Also, let's assume a memory size of $\lambda = 6$ for the SOVA decoder. Using the formulae (36.8), it follows the SOVA FD approach costs 271 operations per an information bit while our entropy (logarithmic) SD approach based on formula (36.11) costs only 61 operations per an information bit, which represents a 77 % saving on the computational complexity.

In addition, in a no-collision event, the SD algorithm check would represent a processing overhead. Nonetheless, our SD approach still provides a significant complexity saving over the FD approach as demonstrated by the following example. Table 36.1 shows the probability of Miss and False-Alarm to be 0.0762 and 0.0684, respectively for QPSK and a 50 bits measurement period.[4] Now, based on formulae (36.4) and (36.5), F_B and F_G (per information bit) for our SD algorithm will equal:

$$F_B = S + P_{miss}\, F = 61 + 0.0762 \times 271 = 82 \text{ Operations/Info Bit}$$
$$F_G = S + (1 - P_{FA})\, F = 61 + (1 - 0.0684) \times 271 = 314 \text{ Operations/Info Bit}$$

[4]The measurement period is 50 bits and the modulation scheme is QPSK, so the number of symbols is 25.

Table 36.1 QPSK – logarithm metric

Logarithm metric							
QPSK							
No # samples	Threshold point	Miss prob (%)	FA prob (%)	Tolerance SNR (dB)	Sample rate	Quantize level	Period length
25	14.7	32.88	33.33	1	2	4	25
100	15	26.02	24.98	1	8	10	25
25	14.6	25.79	26.47	1.5	2	4	25
100	15	17.11	16.27	1.5	8	10	25
50	14.6	25.60	24.72	1	2	4	50
200	14.7	16.23	15.97	1	8	10	50
50	14.4	15.76	16.24	1.5	2	4	50
200	14.6	7.62	6.84	1.5	8	10	50
200	14.2	8.44	9.10	1	2	4	200
800	14.3	2.06	2.58	1	8	10	200
200	14.1	2.02	1.93	1.5	2	4	200
800	14.2	0.14	0.16	1.5	8	10	200
500	14.1	1.99	1.96	1	2	4	500
2,000	14.2	0.15	0.20	1	8	10	500
500	14.1	0.09	0.11	1.5	2	4	500
2,000	14.2	0.00	0.00	1.5	8	10	500
1,000	14.1	0.52	0.25	1	2	4	1,000
4,000	14.1	0.00	0.00	1	8	10	1,000
1,000	14.1	0.00	0.00	1.5	2	4	1,000
4,000	14.1	0.00	0.00	1.5	8	10	1,000

For the comparison purposes between our SD algorithm and SOVA FD algorithm, formulae (36.6) and (36.7) are used to find the computational complexity when no-collision is detected:

$$\begin{aligned} T_{SD} &= F_B P_{collosion} + F_G P_{no\_collosion} \\ &= 82 \times 36.49\% + 314 \times 63.51\% \\ &= 230 \ \text{Operations/Info Bit} \\ T_{FD} &= F \\ &= 271 \ \text{Operations/Info Bit} \end{aligned}$$

Hence, the complexity savings (in number of operations per information bit) becomes:

$$\Delta_{SD}\% = (T_{FD} - T_{SD}) / T_{FD} = (271 - 230) / 271 = 15.12\%$$

Note that the above complexity saving calculation, in fact, represents a lower bound on the saving since the above calculation did not take into account the modem

Table 36.2 QPSK – 3rd moment metric

3rd moment metric							
QPSK							
No # samples	Threshold point	Miss prob (%)	FA prob (%)	Tolerance SNR (dB)	Sample rate	Quantize level	Period length
25	116.4	34.10	33.55	1	2	4	25
100	117	26.86	27.31	1	8	10	25
25	116.2	26.99	26.88	1.5	2	4	25
100	117.1	18.86	18.80	1.5	8	10	25
50	116.7	27.77	27.87	1	2	4	50
200	117	20.97	21.20	1	8	10	50
50	116.5	19.16	19.05	1.5	2	4	50
200	117.1	10.97	11.35	1.5	8	10	50
200	116	10.91	10.54	1	2	4	200
800	116.2	5.29	5.15	1	8	10	200
200	115.8	3.57	3.57	1.5	2	4	200
800	116	1.36	1.44	1.5	8	10	200
500	115.7	3.33	2.95	1	2	4	500
2,000	115.5	1.22	1.12	1	8	10	500
500	115.2	0.43	0.48	1.5	2	4	500
2,000	115.1	0.16	0.14	1.5	8	10	500
1,000	115.4	0.84	0.78	1	2	4	1,000
4,000	115.4	0.00	0.00	1	8	10	1,000
1,000	115.2	0.00	0.00	1.5	2	4	1,000
4,000	115.2	0.00	0.00	1.5	8	10	1,000

line-up operational complexity in order to demodulate and receive the bits in their final binary format properly (i.e., synchronization, channels estimation, etc.).

The performance of our technique can be tuned as desired by a system designer. Tables 36.1, 36.2 and 36.3 provide performance comparisons for various examples where the system designer may choose to reduce the measurement period (e.g., to 25 or 50 bits) at the expense of increasing the Miss and False-Alarm probabilities, or may increase the throughput by using a longer estimation period in order to improve the accuracy of the statistical discriminator performance and reduce the Miss and False-Alarm probabilities.

6 Results and Discussion

We have generated 100,000 simulation snapshots where each snapshot generates a random number of sensors up to 30 sensors with random power assignments (or equivalently path loss, i.e., assignments). All proposed metrics exhibit robust performance. In our study, we have evaluated QPSK modulation scheme versus

Table 36.3 QPSK – maximum to minimum based metric

Maximum to minimum based metric							
QPSK							
No # samples	Threshold point	Miss prob (%)	FA prob (%)	Tolerance SNR (dB)	Sample rate	Quantize level	Period length
25	550	35.23	36.33	1	2	4	25
150	1,100	26.02	24.98	1	8	10	25
25	550	30.79	31.47	1.5	2	4	25
150	1,050	22.11	21.37	1.5	8	10	25
50	800	34.60	33.52	1	2	4	50
200	1,550	27.23	26.97	1	8	10	50
50	1,350	28.76	27.24	1.5	2	4	50
200	1,050	21.01	20.14	1.5	8	10	50
200	1,450	27.44	28.10	1	2	4	200
800	1,550	18.93	18.34	1	8	10	200
200	1,450	19.84	17.34	1.5	2	4	200
800	1,500	12.34	11.54	1.5	8	10	200
500	1,550	23.43	22.21	1	2	4	500
2,000	1,600	12.15	12.20	1	8	10	500
500	1,550	15.09	14.65	1.5	2	4	500
2,000	1,650	8.90	8.34	1.5	8	10	500
1,000	1,700	19.32	20.65	1	2	4	1,000
4,000	1,750	11.34	11.78	1	8	10	1,000
1,000	1,700	13.34	12.32	1.5	2	4	1,000
4,000	1,850	5.55	5.93	1.5	8	10	1,000

various measurement durations, sampling rates and metric numerical (fixed-point) quantization levels to reflect the effects of practical implementation constraints.

Our proposed algorithm has a low sensitivity to deviations of the received SINR from the assumed set-point which is 5 dB (i.e. $SINR_{cut\_off}$). The algorithm works reliably and able to determine if the packet is in collision or not. That is if the SINR is well below or above the set-point, the received signal statistics are expected to also be less confusing to the discriminator anyway and the algorithm shall perform reliably.

Figures 36.1 and 36.2 show the Miss and False-Alarm probabilities versus the choice of the metric comparison threshold level (i.e., above which we decide the packet is in collision or not) for the logarithmic and 3rd moment metrics when signals belong to QPSK modulation scheme respectively. As shown in the figures, the intersection point of the False Alarm and Miss curves, can be a reasonable point to choose the threshold level in order to have a reasonable (or balanced) consideration of the Miss and False-Alarm probabilities, but certainly a designer can refer to the Appendix in order to choose an arbitrarily different point for a different criterion of choice.

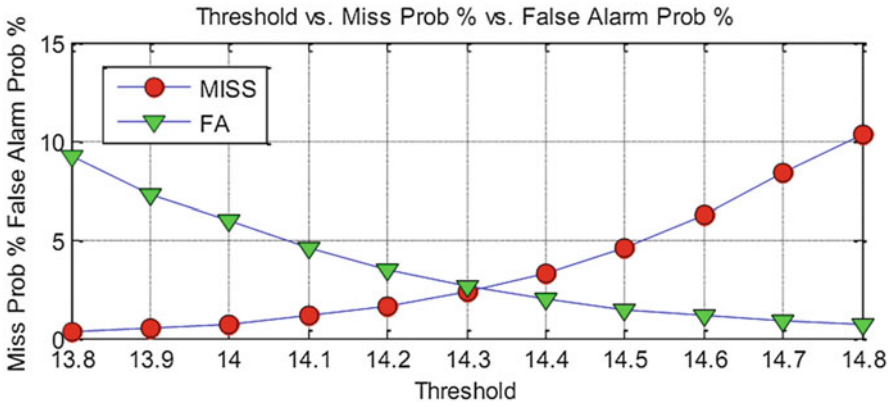


Fig. 36.1 A miss probability =2.06 % vs. false-alarm probability = 2.58 % vs. threshold = 14.3, tolerance SINR = 1 dB up/below cutoff SINR = 5 dB, logarithm metric, QPSK, quantization level = 10, sampling rate = 8, measurement period length = 200 bits

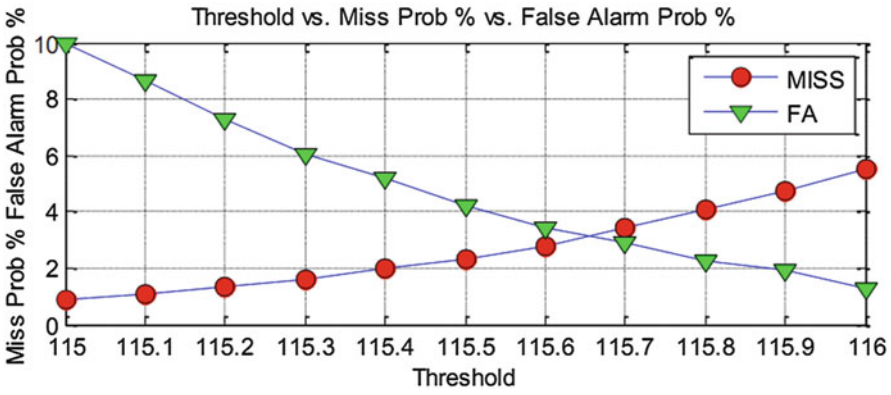


Fig. 36.2 A miss probability =3.33 % vs. false-alarm probability = 2.95 % vs. threshold = 115.7, tolerance SINR = 1 dB up/below cutoff SINR = 5 dB, 3rd moment metric, QPSK, quantization level = 4, sampling rate = 2, measurement period length = 500 bits

7 Conclusion

In this paper we analyze the performance of a novel power saving algorithm for WSNs. Our proposed SD algorithm is based on studying the statistics of received signals and hence the receiver can make a fast decision to decode or reject a packet. In addition, our SD algorithm is based on three simple discrimination metrics which have low computational complexities as well as short measurement period requirements. Also, our SD algorithm minimizes the delay when decoding the received packet, while most full decoding algorithms need to expend a significant amount of energy and processing complexity in order to fully-decode a packet, only

to discover the packet is illegible due to a collision. The analysis and associated figures/tables presented in this paper can be regarded as a designer's guide for achieving significant power saving with low-complexity and low-throughput loss.

Acknowledgment This work has been partially sponsored by Stevens Institute of Technology, Hoboken, NJ, USA and Taif University of the Kingdom of Saudi Arabia. Some of this work has been presented in part at The World Congress on Engineering, 2–4 July, 2014, London, U.K. [16].

References

1. Soua, R., Minet, P.: A survey on energy efficient techniques in wireless sensor networks. Wireless and Mobile Networking Conference (WMNC), 2011 4th Joint IFIP, pp. 1–9, Toulouse, 26–28 Oct 2011
2. Damaso, A., Freitas, D., Rosa, N., Silva, B., Maciel, P.: Evaluating the power consumption of wireless sensor network application using models. Sensors **13**, 3473–355 (2013)
3. Sandra, S., Jaime, L., Miguel, G., Jose, F.T.: Power saving and energy optimization techniques for wireless sensor networks (invited paper). J. Commun. **6**(6), 439–459 (2011)
4. Ustunel, E., Hokelek, I., Ileri, O.: A cross-layer goodput enhancement considering CRC coding and ARQ dynamics. 2012 IEEE Symposium on Computers and Communications (ISCC), pp. 000023–000028, Cappadocia, 1–4 July 2012
5. Mahfoudh, S., Minet, P.: Survey of energy efficient strategies in wireless ad hoc and sensor networks. Seventh International Conference on Networking, 2008. ICN 2008, pp. 1–7, 13–18 355, Cancun, Apr 2008
6. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 2000, vol. 2, p. 10, Hawaii, 4–7 Jan 2000
7. Ingelrest, F., Simplot-Ryl, D., Stojmenovic, I.: Smaller connected dominating sets in ad hoc and sensor networks based on coverage by two-hop neighbors. 2nd International Conference on Communication Systems Software and Middleware, 2007. COMSWARE 2007, pp. 1–8, Bangalore, 7–12 Jan 2007
8. Kimura, N., Latifi, S.: A survey on data compression in wireless sensor networks. International Conference on Information Technology: Coding and Computing, 2005. ITCC 2005, vol. 2, pp. 365 8–13, vol. 2, Las Vegas, 4–6 Apr 2005
9. Li, N., Hou, J.C., Sha, L.: Design and analysis of an MST-based topology control algorithm. IEEE Trans. Wirel. Commun. **4**(3), 1195–1206 (2005)
10. Lin, S., Zhang, J., Zhou, G., Gu, L., Stankovic, J.A., He, T.: Adaptive transmission power control for wireless sensor networks. Fourth International Conference on Embedded Networked Sensor Systems. ACM, 2006, SenSys'06, pp. 223–236, Boulder, Oct 2006
11. Kwon, S., Shroff, N.B.: Energy-efficient interference-based routing for multi-hop wireless networks. Proceedings of INFOCOM 2006. 25th IEEE International Conference on Computer Communications, pp. 1–12, Barcelona, Apr 2006
12. da S Araújo, H., Holanda Filho, R., Filho, R.H.: WSN routing: an geocast approach for reducing consumption energy. Wireless Communications and Networking Conference (WCNC), 2010 IEEE, pp. 1–6, Sydney, 18–21 Apr 2010
13. El-Aaasser, M., Ashour, M.: Energy aware classification for wireless sensor networks routing. 2013 15th International Conference on Advanced Communication Technology (ICACT), pp. 66–71, PyeongChang, 27–30 Jan 2013
14. Woo Tae Kim, Sang Jae Bae, Seog Geun Kang, Eon Kyeong Joo: Reduction of computational complexity in two-step SOVA decoder for turbo code. Global Telecommunications Conference, 2000. GLOBECOM '00. IEEE, vol. 3, pp. 1887–1891, San Francisco, Nov 2000

15. Robertson, P., Villebrun, E., Hoeher, P.: A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain. IEEE International Conference on Communications, 1995. ICC '95 Seattle, 'Gateway to Globalization', vol. 2, pp. 1009–1013, vol. 2, Seattle, Jun 1995
16. Alassery, F., Ahmed, W.K.M., Sarraf, M., Lawrence, V.: Efficient power saving algorithm to detect collision based on statistics of received packets in wireless sensor networks. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 659–666, London, 2–4 July 2014

Chapter 37
Power Aware Virtual Path Routing Protocol for Cognitive Radio Ad Hoc Networks

Farhan Mahmud, Qurratulain Minhas, Hasan Mahmood, Zia Muhammad, and Hafiz Malik

Abstract One of the difficult challenges in Cognitive Radio Ad Hoc Networks (CRAHNs) is the throughput maximization in an environment with uncertain availability of spectrum resources and spatial diversity. Opportunistic spectrum utilization can alleviate the degradation in throughput by employing dynamic routing algorithm and accomplishing Virtual Path Routing (VPR). The VPR aims at finding the most reliable path for multi-hop communication between Secondary Users (SUs) in the presence of Primary Users (PUs) and other interfering SUs. By joint routing and dynamic spectrum access with interference avoidance, VPR selects the path, which ensures optimal link throughput with minimum interference. In this chapter, two utility functions are proposed for cognitive networks routing. The first utility function incorporates probabilistic Signal-to-Interference Ratio (SIR), PU influence and channel switching time delay. The second proposed utility is based on minimizing power consumption to improve battery lifetime. The VPR performance is compared with other well known works in terms of throughput, Bit Error Rate (BER) and Packet Arrival Delay (PAD). The results suggest VPR provides better end-to-end throughput, BER and PAD by avoiding zones of PU presence and mitigating interference effects of neighboring SUs. It is observed that applying a power control scheme such as water-filling, the VPR shows improved throughput. The BER for VPR with water-filling is also reduced compared to the previously suggested scheme such as Gymkhana alongwith the efficient power consumption.

Keywords Cognitive radio ad hoc network • CR routing • Packet arrival delay (PAD) • Probabilistic signal-to-interference ratio (SIR) • Spectrum switching • Virtual path routing

F. Mahmud (✉) • Q. Minhas • H. Mahmood • Z. Muhammad
Department of Electronics, Quaid-i-Azam University, Islamabad, Pakistan
e-mail: farhanmbutt@yahoo.com; qminhas@qau.edu.pk; hasan@qau.edu.pk; mzia@qau.edu.pk

H. Malik
Department of Electrical and Computer Engineering, University of Michigan-Dearborn, Dearborn, MI, USA
e-mail: hafiz@umich.edu

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_37

1 Introduction

Cognitive Radio Ad Hoc Network (CRAHN) [3] is a distributed multi-hop mechanism consisting of transceivers, known as the Primary Users (PUs) and Secondary Users (SUs). The PUs are owners of spectrum with preference for its usage, while SUs are unlicensed users who can utilize the unused bands for communication between the nodes [13]. The SUs opportunistically use available Spectrum Opportunities (SOPs) and establish communication. SOPs can be defined as the set of channels unoccupied by PUs and are therefore, open to SUs. CRAHN is a highly dynamic network paradigm where a SU can use a particular SOP until PU is absent and it has to immediately relinquish the used spectrum band as soon as PU becomes active.

CRAHNs are different from 'classical' ad hoc networks, because the communication among SUs heavily depends on the PU presence. In ad hoc networks, the nodes consider only the interference effects of other network nodes [9] and adopt techniques to avoid packet loss [11] by introducing cooperation among nodes. On the contrary, in CRAHNs, the spectrum usage and route maintenance entirely rely on the activity nature of the PUs. It can vary from occasional mobility and channel usage to highly mobile and sporadic in terms of spectrum utilization [6]. The local spectrum resource fluctuations due to infrequent nature of PUs and hop-by-hop interference caused by SUs have a deep impact on the overall link quality. These factors pose a major challenge for throughput optimization in CRAHNs [4].

In this work, we introduce utility function that senses the PU and SU activity probabilistically. The routing scheme introduced in this chapter, namely **Virtual Path Routing (VPR)** [10], works for throughput maximization based on link stability by considering the PU's presence and neighboring SUs' interference. The consideration of interference effects of other SUs is important because high level of interference on a particular channel may render it useless. The interference factors of all the network entities is considered statistically by computing the probabilistic SIR on the possible route. The proposed protocol tends to minimize the power utilization by penalizing paths with frequent channel switching.

The simulations show that VPR can effectively avoid zones of PU presence along with selection of channel providing optimal SIR, which leads to better throughput, less BER and reduction in PAD. The results are further improved when we apply water-filling for power allocation in SUs.

The rest of this chapter is organized as follows. In Sect. 2 we study the previously proposed routing algorithms based on PU and SU interference avoidance and throughput maximization. In Sect. 3 we discuss system model. In Sect. 4 we present the VPR protocol and its algorithm in detail. Section 5 presents the simulation results and Sect. 6 concludes the chapter.

2 Literature Review

Based on spectrum knowledge, CRAHN routing protocols are broadly classified into two major classes: *Network wide spectrum information* based solutions and *local spectrum information* based algorithms. In the former case, spectrum occupancy information is available to the nodes locally through network graphs or to the central controlling unit. While in the later case, spectrum availability data is constructed locally at each node in the network [19].

Interference constraints form the basis of work presented in [18]. It is a decentralized algorithm, in which minimum interference level avoidance is analyzed from PU's perspective. The paper lays down the basic principles for multi-hop route selection in CRAHNs, however SU interference impact is not discussed.

Minhas et al. in [12] present a game theoretic approach to introduce cooperation among SUs in a cognitive sensor network. The SUs cooperate to carefully analyze the spectrum to be used, as it has impact over the entire network. On the contrary in this work, all the SUs are considered as independent in their choice of channels. Their channel selection is observed probabilistically and the best channel is selected based on maximum SIR.

The protocol in [15] considers both per hop spectrum availability and source to destination shortest distance during route selection. The proposed algorithm creates a run time forwarding mesh consisting of shortest paths between source and destination and selects the best path to maximize throughput. However with increase in network size and highly mobile SU nodes, the performance of the said protocol degrades. Another spectrum aware technique is proposed in [16], which integrates the dynamic route functionality with per hop channel utilization to optimize throughput. In CRAHNs, channel availability changes hop-by-hop and SPEAR [16] addresses this heterogeneity by combining spectrum sensing with channel reservation for collision-free cooperative routing. In contrast, the routing protocol proposed in this chapter does not rely on cooperative scenario among SUs. It probabilistically analyzes the spectrum availability and forms the route to avoid interference.

Authors in [5] have introduced a routing protocol, which aims at maximizing the throughput by using better bandwidth utilization and avoiding paths crisscrossing with PU receivers. The proposed algorithm defines two classes based on the preference given to PU. In the first class, decreasing end-to-end latency is given precedence over PU interference avoidance. In the second class, safeguarding PU communication is given more importance. It considers several metrics during spectrum selection stage but ignores the interference effect of neighboring SUs, which is discussed in this work.

Spectrum and Energy Aware Routing (SER) protocol proposed in [7] aims to efficiently use energy resources using TDMA style channel-time slot assignment to establish route. The utility function introduced in [7] selects the node, which satisfies the minimum threshold residual energy level. SER protocol generates Route Recovery (RREC) and Route Error (RRER) messages for route maintenance. The

paper does not discuss the significant effect of overheads associated with these messages. Under a highly dynamic CRAHN scenario, increase in overheads may reduce the efficient use of energy resources of the network nodes.

A clustered based technique; United Nodes [17] opts for paths offering the least PU interference for maximizing throughput. Nodes run clustering algorithm and adjust themselves in clusters. Clustering algorithm considers node position, communication efficiency and interference for clustering. The metric in [17] considers the interference of PUs and SUs, but link restructuring depends on PU interference only. It does not account for the change in spectrum availability due to neighboring SUs' presence after PU disruption period is over.

Link stability based routing protocol named Gymkhana is proposed in [1, 2]. The Gymkhana method calculates path connectivity by considering the second smallest eigenvalue of Laplacian of a graph. During route formation, Gymkhana avoids zones of PU presence only, it considers neighboring SUs as inactive. Unlike Gymkhana we consider both the impact of PUs and SUs during the route formation.

3 System Model

In this work, we assume cognitive radio network model, consisting of N_s SUs and N_p PUs, in a uniformly distributed environment. Each PU p uses a particular spectrum band ch, where $ch = 1, \ldots, N_p$ and has transmission range r_p. We assume that each PU p uses the channel ch probabilistically defined by activity factor a_p. The activity factor a_p is classified by average *on* and *off* transmission. Mathematically it is defined as:

$$a_p = \frac{\bar{t}^p_{on}}{\bar{t}^p_{on} + \bar{t}^p_{off}} \tag{37.1}$$

$\bar{t}^p_{on}$ is the average time duration during which PU p is using ch for transmission and $\bar{t}^p_{off}$ is the average silent duration. The probability that p will not use its licensed spectrum is $1 - a_p$. Thus the probability a_p, also called as the PU activity factor, is responsible for determining the suitable route for the SUs. Each SU s ($s = 1, \ldots, N_s$) has a transmission range r_s and can use a channel for communication if that channel is not in use by PU p or it is not under its influence area. A simple path loss model is considered, as defined in [8]. The transmission power of SU i is given by P_i. The power at the receiver is given as,

$$P_{ri} = \frac{P_i \lambda^2}{d^\alpha} \tag{37.2}$$

The received power is represented by P_{ri}, the transmitted power is P_i, d is the distance between transmitter and receiver, λ is the wavelength, and α is the path loss coefficient. As it is a simplified path loss model, therefore propagation exponent α is assumed to be 2.

Link gain g_{ij} for transmitting SU s_i and relay/receiver s_j is defined as [8]:

$$g_{ij} = \frac{\lambda^2}{d_{ij}^2} \tag{37.3}$$

modifying (37.2) in terms of link gain, we can write,

$$P_{ri} = P_i g_{ij} \tag{37.4}$$

The SIR between transmitting SU s_i and receiving SU s_j is measured, probabilistically and is given as,

$$f[SIR]_{ch}^{ij} = \frac{(1-a_p)g_{ij}P_i}{\sum_{n=1,n\neq i,j}^{N_s}(1-a_p)g_{jn}P_n + a_p g_{jp}P_p} \tag{37.5}$$

where, $1-a_p$ is the channel usage probability of SU s_i and nth SU. The second term $a_p g_{jp} P_p$ in the denominator of (37.5) shows the probabilistic interference of PU p, transmitting with P_p power. The selection of power levels play an important role in route selection by affecting the values of SIR function. Higher transmission power increases the interference level and results in poor performance in terms of battery life and SIR. Hence, the transmission power of relaying users must be adjusted to improve the performance of selected route.

The probabilistic SIR of (37.5) is channel based so each SU computes $f[SIR]$ for all the channels. It depends on the random activity of neighboring SUs and their interference. It also considers the influence and interference of licensed PU for that channel. The impact of each SU and the PU is inversely proportional to the square of the distance between interfering and transmitting nodes. We analyze the system for different power options, initially for constant transmission power, then randomly varying power levels and finally for power levels assigned according to water-filling power control. According to the assumption, the distance and the channel usage probability play a major role in affecting the SIR between s_i and s_j. Moreover, the transmission power of network users also affects the SIR significantly.

The transmission power of network users play an important role, as increase in transmission power results in increasing interference. However, too low transmission power results in weak links with higher packet drop rates. Hence, the route selection process must involve relays with sufficient power for successful packet transmission with reduced interference. One simple and effective method is to employ water-filling algorithm which assigns power to the users on a channel according to the link gain and threshold SNR. When applying water-filling to the SUs involved in a route, the power vector for channel ch can be written as [14]:

$$P_{ch} = (I - H_{ch})^{-1}\gamma_{ch} \tag{37.6}$$

where, I is the interference matrix, defined as:

$$I = \begin{cases} 1, & \text{if } s_i \text{ and } s_j \text{ choose } ch, \\ 0, & \text{otherwise.} \end{cases} \tag{37.7}$$

γ_{ch} is the threshold SNR and $\gamma_i = \frac{\gamma_{th} N_o}{g_{ii}}$. The channel vector H_{ch} is given by $H_{ch} = (g_{ij})_{ch}$, where s_i and s_j choose the same channel ch. These power levels are assigned to the cognitive users based on their respective channel choices.

We analyze the network for power aware routing and employ two different utility functions for implementing the VPR. The previous formulation assumed constant power for all network users and no mechanism for power control was proposed [10]. This work reduces the interference level for the selected route by updating the power levels of the network user for improved throughput.

4 Virtual Path Routing

The objective of the VPR path establishment is to deliver information across the route that tries to avoid "hurdles," i.e. at the same time abstain from unstable areas, where the risk of disconnection is high. These "hurdles" are in the form of zones (or channels) occupied by active PUs and high activity SUs. For illustration, a cognitive network is schematically presented in Fig. 37.1 with three PU influence areas shown by three different semi circles. PUs can only use only their licensed spectrum, therefore PU_1, PU_2 and PU_3 can use only channel 1, 2 and 3 respectively. A path between source Sx and destination Dx is drawn with per hop channel use information by straight lines labeled as 3. As it can be seen from the figure, VPR avoids PU interference by selecting a different channel in its zone.

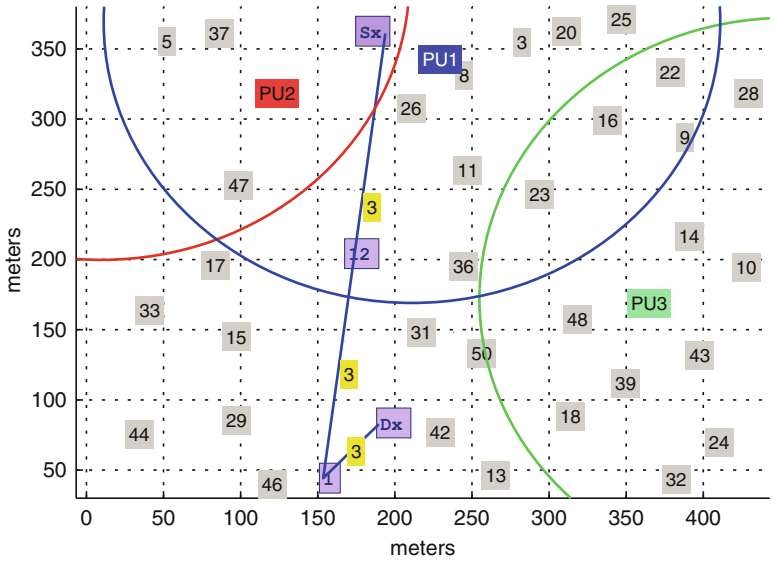


Fig. 37.1 Network diagram showing path formation of VPR protocol

VPR comprises of two phases. In the first phase, source initiates route discovery using *VPR Protocol* to gather information of all possible paths between source and destination. In the second stage of route selection, the destination chooses a path using *VPR Algorithm*.

4.1 VPR Protocol

The VPR requires information that is available via AODV protocol [16]. It is assumed that SUs are able to sense the influence of each PU and measure its activity factor a_p, where $p = 1, \ldots, N_p$. Moreover it is also assumed that each SU s_i is able to compute SIR with other SU s_j, where $i, j \subseteq N_s$. In this work, an assumption is made that each SU can probabilistically access a channel *ch* with probability $1 - a_p$. Therefore all the available channels can be utilized by the SUs with some probability based on PU activity a_p.

The SUs utilize the important information of $f[SIR]$ and a_p, for calculating the link establishment possibility. These two important parameters are used to compute utility and store in a *local utility* vector $\mathcal{LU}$. Each SU maintains and periodically updates its own *local utility* vector. The length of $\mathcal{LU}$ vector is equal to the number of SUs, with each generic element $\mathcal{LU}(i)$, $(i = 1, \ldots, N_s)$ indicating the utility computed between two SUs. The formulation of $\mathcal{LU}$ is discussed in detail in the *VPR Algorithm*.

A route request is initiated by a source node $\mathcal{S}$ in order to update the routing table and renew information about the destination node $\mathcal{D}$. The criteria adopted for candidate path selection is maximum SIR on that path with minimal PU presence. Following this criteria, all possible paths leading to the destination are computed. Let K be the number of paths found between source and destination. Our quest is to find a path that provides highest SIR and improved throughput among these possibilities.

The RREQ, which reaches $\mathcal{D}$ for the kth path contains three important information vectors.

- Node IDentity vector, NID_k, which contains the IDs of the relay nodes k. The nodes are assigned a unique tag. NID_k initializes with source ID and last element is the ID of destination node.
- Path Local Utility vector, $\mathcal{PLU}_k$, every SU on path k appends its local utility vector $\mathcal{LU}$ value in Path Local Utility Vector. The values stored in $\mathcal{PLU}_k$ are used by destination to compute the utility of the path.
- Path Channels vector, $PathCh_k$, contains information about the possible hop-by-hop channel selection for a path k. This hop-by-hop channel adoption is based on the optimal value of local utility $\mathcal{LU}$ between two communicating SUs s_i and s_j on a path k.

Each SU receiving the RREQ packet, checks if its ID is already added in *NID*. If the ID exists, the packet is dropped to avoid looping. In other case, the node enters

its ID in the *NID* and forwards the packet to the next candidate node. By using this methodology, destination $\mathcal{D}$ receives the number of packets that are within the threshold. The destination node uses the information received through RREQs to run VPR algorithm and selects the best path and channel that ensures maximum SIR and minimum PU activity. After selecting the best route, specific packets used for replies are sent back via each node participating in the forward route.

4.2 VPR Algorithm

The destination $\mathcal{D}$ runs the algorithm for route selection on the basis of information stored in RREQ packet. Out of the three basic information RREQ packet holds, $\mathcal{LU}$ has the fundamental importance in the formulation of VPR Algorithm. Here we discuss in more detail the mathematical composition of local utility vector and the factors affecting it.

- ***Local Utility Vector*** ($\mathcal{LU}$): Each SU s_i calculates the utility of choosing a particular channel *ch*, where $ch = 1, \ldots, N_p$, for communication with SU s_j, where $s_i, s_j \subseteq N_s$. This utility is based on the SIR, which exists between s_i and s_j and the PU influence on the said nodes. Mathematically, this utility is defined as:

$$\mathcal{LU}_{ch}^{ij} = \frac{f[SIR]_{ch}^{ij}}{I_p^i + I_p^j + \eta} \tag{37.8}$$

  where, I_p^i and I_p^j are the PU p influence factors on SU s_i and s_j, respectively.

$$I_p^i = \begin{cases} a_p, & \text{if } 0 \leq r_{s_i} \leq r_p, \\ 0, & \text{otherwise.} \end{cases} \tag{37.9}$$

  $f[SIR]_{ch}^{ij}$ is channel based SIR, probabilistically computed between two SUs, s_i and s_j for channel *ch*. η is a channel noise constant.

  Each SU s_i calculates local utility with all other SUs s_j $(i \neq j = 1, \ldots, N_s)$ for all the channels available and selects the maximum local utility value. This maximum value is stored in s_i's local utility vector $\mathcal{LU}$. The information of corresponding channel *ch* is stored in *local channel* list. This local channel list provides channel usage information to $PathCh$ vector during route formation.

During RREQ initialization, source node $\mathcal{S}$ selects m best utility values from its local utility vector $\mathcal{LU}$ and forwards RREQ packets towards corresponding SUs. Thus m number of routes are selected between $\mathcal{S}$ and relaying nodes. The value of m can vary depending on network congestion and nodes' activity. During network congestion, value of m is increased so that more RREQ packets are forwarded to

increase chances of successful path formation. When network dynamics are low, value of m is reduced to reach out only a selected group of nodes as chances of route formation are high with less usage of energy resources. Each RREQ packet contains utility information between source and mth relaying node along with corresponding channel and node IDs information. The name ***Virtual Path*** comes from the fact that the algorithm opportunistically selects the path and channel (or virtual path) while relaying a RREQ packet. The relaying nodes receiving the RREQ packet selects the next best hop according to their local utility vector and broadcast RREQ to next candidate hops after inserting their local utility, channel and node ID information. Multiple route request packets are received by the destination pertaining to best possible routes.

The destination node $\mathcal{D}$ chooses the path that offers least channel switching delay with optimal end-to-end local utility. The path utility function is defined as:

$$\mathcal{U}1_k = \frac{\sum_{h=1}^{H_k} \mathcal{PLU}_k^h}{\tau_{sw}} \tag{37.10}$$

where $\sum_{h=1}^{H_K} \mathcal{PLU}_k^h$ is the sum of path local utility values of all the hops h on path k, with total number of hops H_K, $(h = 1, \ldots, H_K)$. $\mathcal{PLU}_k^h$ is the maximum local utility among all the available channels in a hop on path k. It can be defined as:

$$\mathcal{PLU}_k^h = \max_{ch}(\mathcal{LU}_{ch}^{NID_k(l,l+1)}) \tag{37.11}$$

τ_{sw}, shows the switching time delay for path k. An important parameter considered is the amount of time required to vacate the band that is under use by other transmitters. The number of available channels determine the delay in switching between channels, if required.

The above utility function computes the routes based on the SIR and channel switching. However, in order to reduce the power consumption, we propose another utility function to improve the battery lifetime of the network users. The VPR algorithm is improved by incorporating power efficient utilities for routing. The new utility function proposed includes the cost of routing in terms of power consumption, which assists in improving the battery life of users.

$$\mathcal{U}2_k = \frac{\sum_{h=1}^{H_k} \mathcal{PLU}_k^h}{\tau_{sw}} - \sum_{h=1}^{H_k} P_h \tag{37.12}$$

where, P_h is the power utilized during the transmission of each hop.

The best value of $\mathcal{U}_k$ is considered from K possible paths. The utility in (37.10) is given by:

1. Total hops in a path k.
2. Number of channel switching occurring on path k.
3. PU activity and neighboring SU interference along the path k.

The goal of VPR is to optimize throughput by selecting the path that minimizes above three points and maximize SIR. After selecting the best route, RREP message is sent back along the selected route to initiate communication between source and destination. The utilities $\mathcal{U}1_k$ and $\mathcal{U}2_k$ assist in improving the SIR and reduce the transmission power for the selected route.

5 Simulation Results

In this section, we experimentally evaluate the performance of VPR. We consider the cognitive ad hoc network topology, in which the nodes are uniformly distributed and include parameters associated with it in the analysis. We compare the working of VPR with *Gymkhana* routing protocol [1]. We analyze the efficiency of the two protocols based on PU avoidance and maintaining optimum throughput along with BER and packet delay performance.

The VPR utility function that is defined in (37.10) selects the path with maximum utility value and is compared with Gymkhana routing protocol. The Gymkhana utility function, defined as U^c [2], selects the path with maximum utility based on path connectivity. Gymkhana considers the affect of PUs to select the path.

The routing protocols are evaluated by generating different cognitive networks. The protocols under consideration derive all the possible paths between a given source and destination based on their respective criteria. A traffic session is simulated on the secondary network in which not only PUs can be active based on their activity factor a_p, but neighboring SUs are also active and can use any available spectrum based on a_p.

We assume that N_s SUs and N_p PUs are uniformly distributed in a square deployment region of area 500 m^2. During simulation, 100 data packets are sent across the 'best' routes selected by the two routing protocols. PU activity factor values during the experiment are $a_1 = 0.7, a_2 = 0.3$ and $a_3 = 0.4$. Other simulation parameters are reported in Table 37.1.

Figures 37.2 and 37.4 report the performance of VPR and Gymkhana routing schemes in terms of throughput and BER. The resulting plots are for 100 networks, averaging them in case of 50 SUs and 3 PUs. The activity factors are kept constant

Table 37.1 Simulation parameters

Parameters	Symbols	Values
Number of secondary users	N_s	50
Number of primary users	N_p	3
No of channels	ch	3
Primary user activity factors	a_p	0.2–0.7
Range of secondary users	r_s	75 m
Range of primary users	r_p	150 m
Channel switching time delay	τ_{sw}	25 ms

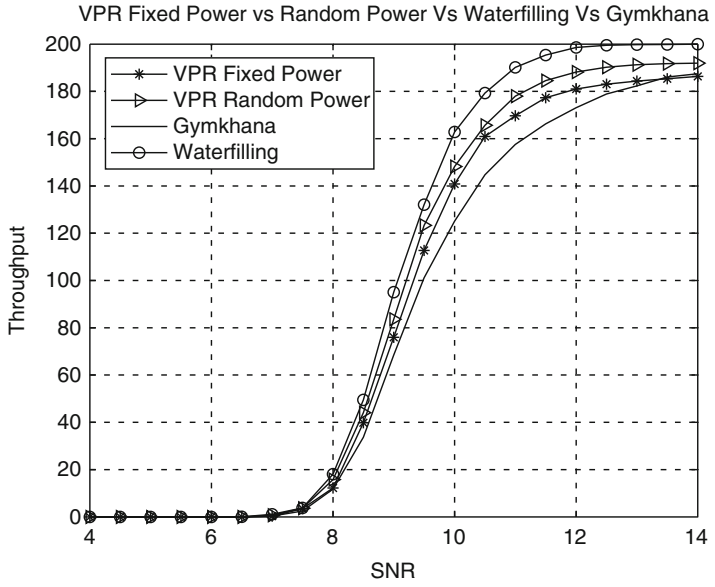


Fig. 37.2 Throughput comparison between VPR and Gymkhana

with increasing Signal-to-Noise Ratio (SNR). VPR has better performance if its average value is considered.

Figure 37.2 shows the throughput for the end-to-end paths. This is dependent on the SNR for the selected routes for the two methods. As it can be seen that when noise levels are high the success ratio of correct packet reception at destination is low and with increase in signal power, the throughput starts improving. But as seen from this figure, the improvement in case of VPR is more as compared to the other routing protocol. In case of VPR the data is affected by noise only as it successfully avoids the PUs and neighboring SUs. Gymkhana avoids the PU interference but the interference from SUs corrupts the data.

Performance comparison of throughput perspective is shown in Fig. 37.2, which shows that VPR obtains better percentage throughput than Gymkhana because it successfully avoids the interference affects of SUs on the received data. The percentage throughput for the two routing schemes is computed for 10 dB SNR and averaged for 50 different networks as shown in Fig. 37.3.

Figure 37.4 shows the BER for the routing algorithms, plotted against SNR. In order to perform the experiment, 50 random networks are generated with simulation parameters reported in Table 37.1. The SNR values are incremented from 4 to 16 dB. From Fig. 37.4, it is evident that at low SNR values the performance of both protocols is almost same. As the SNR increases, the noise affects decrease, but the interference effects from other network entities remains the same. Therefore, with increasing SNR, BER of VPR reduces as interference effects on VPR are less as compared to Gymkhana.

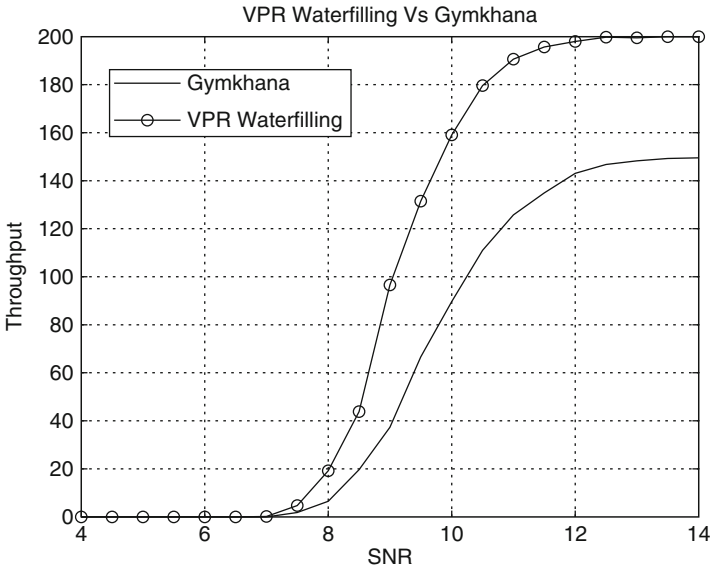


Fig. 37.3 Throughput comparison for utility $\mathcal{U}2_k$

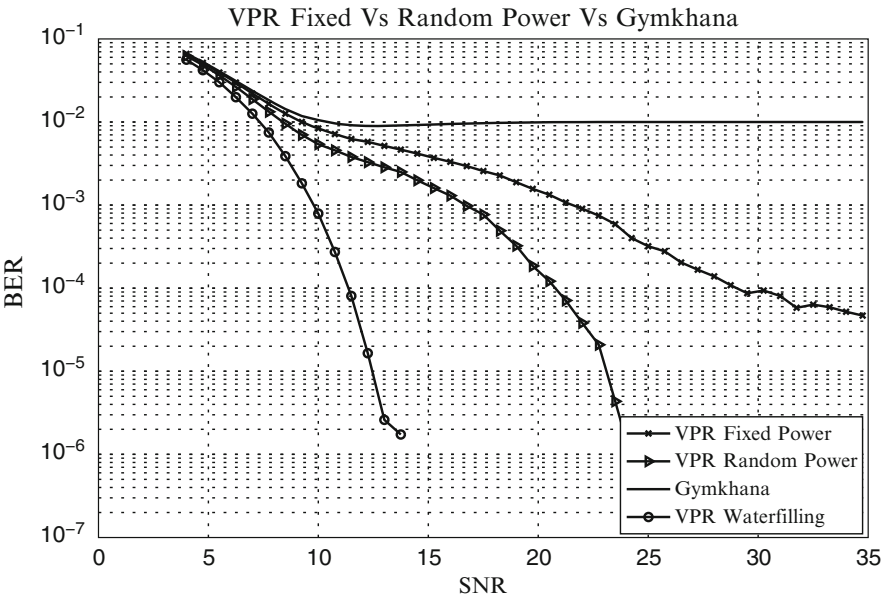


Fig. 37.4 Bit error rate (BER)

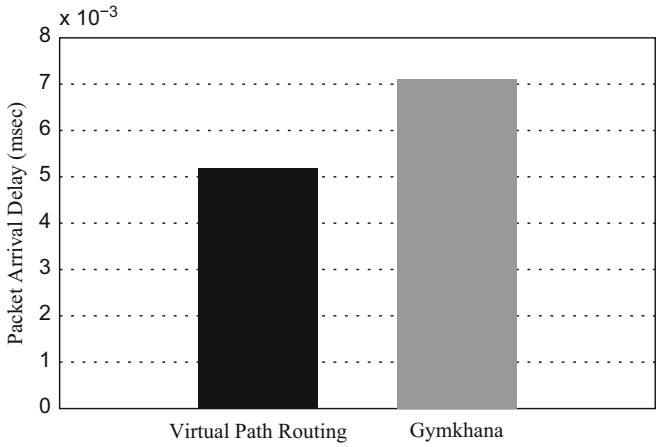


Fig. 37.5 Packet arrival delay (milliseconds)

Figure 37.5 reports the performance in terms of PAD. PAD is defined as the delay in the reception of packet at destination due to number of hops, PU presence and channel switching time delay. From the analysis shown in Fig. 37.5, it is evident that due to efficient route selection, VPR is able to avoid zones of PU presence. The number of hops on the selected path are optimal and channel switching is not profound. Gymkhana is able to avoid the PU affected zones but the slight increase in the PAD is due to higher number of hops on Gymkhana selected route.

From the simulation results, we observe that the VPR performance is considerably improved by using water-filling for assigning power to the secondary users in the cognitive network. The proposed utility function $\mathcal{U}2_k$ involves power to accommodate the reduction in power for improved battery lifetime.

6 Conclusions

In this chapter, we present interference and spectrum aware routing algorithm named as ***Virtual Path Routing (VPR)***. The VPR is a dynamic spectrum access routing solution which opportunistically selects paths offering least interference. The interference from both PUs and neighboring SUs is considered during route selection. We introduce a mathematical model, which combines the effects of PU activity and the neighboring SUs' interference impact for route selection. Secondary nodes in VPR take channel level decisions during path formation by selecting the channel that offers best SIR and is not under PU's use.

BER, throughput and latency are important parameters to measure the efficiency of any protocol. In CRAHNs, throughput, BER and PAD are directly affected

by the presence of PUs. PUs affect the channel availability for SUs. Therefore communicating SUs not only face interference and contention from PUs but also neighboring SUs.

VPR performance is evaluated through simulations in a CRAHN scenario and comparison is made with Gymkhana routing protocol. In case of Gymkhana, it is observed that this protocol routes data based on minimum hops and considers the interference effects of PUs only. On the contrary, VPR considers the interference effects of all network entities as it is evident from throughput, BER and PAD results. Moreover, extending the VPR scheme for power aware routing allows network users to save power resources and also improve the BER by lowering interference over the network. The incorporation of water-filling power scheme provides efficient power consumption and improved throughput.

References

1. Abbagnale, A., Cuomo, F.: Gymkhana: a connectivity based routing scheme for cognitive radio ad-hoc networks. In: IEEE International Conference on Computer Communications, INFOCOM, San Diego (2010)
2. Abbagnale, A., Cuomo, F.: Leveraging the algebraic connectivity of a cognitive network for routing design. IEEE Trans. Mob. Comput. **7**, 1163–1178 (2012)
3. Akyildiz, I., Lee, W., Chowdhury, K.: CRAHNs: cognitive radio ad hoc networks. Ad Hoc Netw. **7**(5), 810–836 (2009)
4. Brik, V., Rozner, E., Banerjee, S., Bahl, P.: DSAP: a protocol for coordinated spectrum access. In: First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), Baltimore, pp. 611–614 (2005)
5. Chowdhury, K., Akyildiz, I.: CRP: a routing protocol for cognitive radio ad hoc networks. IEEE J. Sel. Areas Commun. **29**(4), 794–804 (2011)
6. Federal Communications Commission: Spectrum policy task force technical report (2002)
7. Kamruzzaman, S., Kim, E., Jeong, D.: Spectrum and energy aware routing protocol for cognitive radio ad hoc networks. In: IEEE International Conference on Communications, Kyoto, pp. 1–5 (2008)
8. Mahmood, H., Comaniciu, C.: Adaptive spreading/coding gains for energy efficient routing in wireless ad hoc networks. In: Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05), Stockholm, vol. 4, pp. 2454–2458 (2005)
9. Mahmood, H., Comaniciu, C.: Interference aware cooperative routing for wireless ad hoc networks. Ad Hoc Netw. **7**(1), 248–263 (2009)
10. Mahmud, F., Minhas Q.-A., Mahmood, H., Muhammad, Z., Malik, H.: Connectivity driven virtual path routing protocol for cognitive radio ad hoc networks. In: Proceedings of the World Congress on Engineering 2014 (WCE 2014), London, 2–4 July 2014. Lecture Notes in Engineering and Computer Science, pp. 667–672 (2014)
11. Minhas, Q.A., Mahmood, H., Malik, H.: Incentive driven cooperation to avoid packet loss in multihop ad hoc networks. In: International Conference on Emerging Technologies (ICET), Islamabad, pp. 1–6 (2012)
12. Minhas, Q., Tawhid, M., Mahmood, H.: Efficient power and channel allocation strategies in cooperative potential games for cognitive radio sensor networks. Int. J. Distrib. Sens. Netw. **2014**, 9 (2014). Article ID 130130. http://dx.doi.org/10.1155/2014/130130
13. Mitola, J.: Cognitive radio, an integrated architecture agent for software defined radio. Ph.D. dissertation thesis, KTH (2000)

14. Nie, N., Comaniciu, C.: Adaptive channel allocation spectrum etiquette for cognitive radio networks. Mobile Netw. Appl. **11**(6), 779–797 (2006). Springer, New York
15. Pefkianakis, I., Wong, S. H., Lu, S.: SAMER: spectrum-aware mesh routing in cognitive radio networks. In: Proceedings of the 3rd IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Chicago, pp. 1–5 (2008)
16. Sampath, A., Yang, L., Cao, L., Zheng, H., Zhao, B.Y.: High throughput spectrum-aware routing for cognitive radio based ad hoc networks. In: Proceedings of the International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), Singapore (2008)
17. Talay, A., Altilar, D.: United nodes: cluster-based routing protocol for mobile cognitive radio networks. IET Commun. **5**(15), 2097–2105 (2011)
18. Xie, M., Zhang, W., Wong, K.: A geometric approach to improve spectrum efficiency for cognitive relay networks. IEEE Transactions on Wireless Communications **9**(1), 268–281 (2010)
19. Xin, C., Ma, L., Shen, C.C.: A path-centric channel assignment framework for cognitive radio wireless networks. Mob. Netw. Appl. **13**(5), 463–476 (2008)

Chapter 38
Performance Evaluation of VoIP QoS Parameters Using WiFi-UMTS Networks

Mahdi Hassan Miraz, Muzafar Aziz Ganie, Maaruf Ali, Suhail Ahmed Molvi, and AbdelRahman Hamza Hussein

Abstract Simulation of VoIP (Voice over Internet Protocol) traffic through UMTS (Universal Mobile Telecommunication System) and WiFi [Wireless Fidelity] (IEEE 802.11x) in isolation and combined are analysed for the overall Quality of Service (QoS) performance. The average jitter of the VoIP traffic transiting the WiFi-UMTS network has been found to be lower than that of either solely through the WiFi and the UMTS networks. It is normally expected to be higher than traversing through the WiFi network only. Both the subjective MOS (Mean Opinion Score) and the objective packet end-to-end delay were also found to be much lower than expected through the heterogeneous WiFi-UMTS network.

Keywords Mean Opinion Score (MOS) • Quality of Service (QoS) • SIP (Session Initiation Protocol) • UMTS • Voice over Internet Protocol (VoIP) • WiFi

1 Introduction

Due to the preponderance of mobile internet usage, it is difficult to completely ignore the importance of using VoIP technology to make both voice and telephone network calls. With the continued proliferation and increasing widespread global

M.H. Miraz (✉)
Glyndŵr University, Wrexham, UK

University of Ha'il, Ha'il, Kingdom of Saudi Arabia
e-mail: m.miraz@glyndwr.ac.uk; m.miraz@uoh.edu.sa

M.A. Ganie • S.A. Molvi
University of Ha'il, Ha'il, Kingdom of Saudi Arabia

Pacific University (PAHER), UDAIPUR, Rajasthan, India
e-mail: m.ganie@uoh.edu.sa; s.molvi@uoh.edu.sa

M. Ali
University of East London, London, UK
e-mail: M.Ali3@uel.ac.uk

A.H. Hussein
University of Ha'il, Ha'il, Kingdom of Saudi Arabia, India
e-mail: ar.hussein@uoh.edu.sa

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_38

market penetration of mobile and fixed telecommunication networks, such as BGAN (Broadband Global Area Network), VSAT(Very-small-aperture terminal), UMTS, LTE (Long Term Evolution) and WiMAX (Worldwide Interoperability for Microwave Access), the digitized data has to traverse multiple networks. The user now has to consequently experience various types of signal degradations, such as: latency and jitter whose values differ from that experienced over the PSTN (Public Switched Telephone Network). The primary goal of this research is to ascertain to what extent the QoS of the VoIP traffic varies while traveling through the latest heterogeneous generation of networks. To achieve this aim, the objectives include: (1) Designing different network scenarios using the OPNET (Optimized Network Engineering Tool) modeler; (2) Simulating them to assess their performance; (3) Analysing the results obtained through the simulation. The first two scenarios consisted of several VoIP clients exchanging data through UMTS-to-UMTS and WiFi-to-WiFi networks. The QoS components of VoIP traffic such as the: MOS, Availability, Throughput, Distortion, Cross-talk, Attenuation, Link Utilization, Loss, Jitter, Packet end-to-end Delay, Packet Delay Variation and Echo – were examined and analysed. In the third scenario, the VoIP traffic travelled through heterogeneous networks i.e. UMTS-to-WiFi. The results previously obtained from the homogeneous scenarios were then analysed and compared with it.

This paper is an extension of a paper that was presented at WCE 2014 [1].

2 Background

2.1 VoIP (Voice over Internet Protocol)

VoIP [2] is the transmission of packetized voice traffic over a digital data network. This mode of transmission being set-up as a shared virtual circuit differs from the analogue non-shared reserved switched circuit connection. The packets now have the freedom of network travel to take any route to reach its final destination and also to arrive in any order with consequent differing delays. This efficient use of channels also means having to contend with several disadvantages such as that the packet can experience often considerable delays or never arrive at all (packets being dropped) due to severe traffic congestion. The flexibility of multiple routing does give the advantage of cheaper or even free routing of VoIP traffic. Sending digital data as packets mean that digitized multimedia information can now be transmitted over a common shared broadband network.

Thus the mainly software based VoIP technology has the major advantage of inexpensive scalability compared to other mainly hardware implemented telephony systems. However, since being a mainly software based system – it is vulnerable to increasingly targeted attacks from hackers currently in the form of worms and viruses. Some security countermeasures to address these problems are discussed in [3].

The rollout of the 3G (third Generation) network has helped considerably in the lead to convergence [4] especially between the internet, fixed and mobile services – terrestrial including maritime and satellite. Global mobile internet access anywhere, along with broadband multimedia usage, is becoming more prevalent, especially with the spread of WiFi, WIMAX and femtocells. This capability has been facilitated by UMTS Rel. 5 using the IP Multimedia Subsystem (IMS). The user data and signaling is transported using the IMS packet based control overlay network. The IETF (Internet Engineering Task Force) SIP (Session Initiation Protocol) was adopted by the 3GPP (Third Generation Partnership Project) for the call setup session of VoIP and other IP-based multimedia communication. The current UMTS and IEEE 802.11 WiFi networks fully support real-time services such as VoIP [5].

2.2 SIP (Session Initiation Protocol)

The entire call setup procedure from the beginning to the end requires the exchange of signaling and control information between the calling and called parties involved. This already complicated process becomes even more so where roaming and hence mobility is involved through a heterogeneous network. Different devices have differing capabilities and the seamless flow of information between them needs *a priori* information about their capabilities before the efficient full-flow of information can take place. This is handled by the SIP [6] application layer control protocol working alongside the existing other protocols. The destination and source "user agents" discover each other and establish the seamless connection between them based on shared properties using SIP. SIP creates any necessary proxy servers where needed by intermediary nodes dealing with such events as: invitation, registration and other such call connection, maintenance and termination requests. Mobility features are also catered for, including: name mapping; redirection and the maintenance of an external visible location invariant identifier [7]. SIP is both session and device independent, capable of handling a wide range of multimedia data exchange. SIP also has other very useful functionalities, such as: the ability to setup a multicast call; removal of participants; user availability; session setup; user capabilities; user location and session management.

2.3 Quality of Service (QoS) Parameters

The flexibility of the data networks' data handling capability over that of the traditional fixed infrastructure telephone services puts it at a great advantage [5]. The QoS for VoIP can be measured using several different types of subjective and objective measures, such as: the Mean Opinion Score (MOS), as shown in Table 38.1; the jitter and the end-to-end delay.

Table 38.1 The subjective Mean Opinion Score (MOS)

Quality scale	Score	Listening effort scale
Excellent	5	No effort required
Good	4	No appreciable effort required
Fair	3	Moderate effort required
Poor	2	Considerable effort required
Bad	1	No meaning understood with effort

The MOS is calculated using a non-linear mapped R factor [9] as shown in Eq. (38.1), below:

$$\text{MOS} = 1 + 0.035R + 7 \times 10^{-6} \left[R\,(R - 60)\ \ (100 - R)\right] \tag{38.1}$$

Where:

$R = 100 - I_s - I_e - I_d + \text{A}$
I_s : voice signal impairment effects;
I_e : impairment losses suffered due to the network and codecs
I_d : impairment delays particularly mouth – to – ear delay.
A : Advantage factor (attempts to account for caller expectations)

The '*jitter*' "is the variation in arrival time of consecutive packets" [10], evaluated over a period of time [8]. It is the signed maximum time difference in the one way delay over a set time period. Wherever buffers are used they can both over-fill and under-fill, resulting in packets being discarded. Let $t(i)$ be the time transmitted at the transmitter and $t'(i)$, the time received at the receiver, the jitter is then defined as:

$$\text{Jitter} = \max_{1 \le i \le n} \left\{\left[t'(n) - t'\,(n-1)\right] - \left[t(n) - t\,(n-1)\right]\right\} \tag{38.2}$$

The '*Packet end-to-end delay*' "is measured by calculating the delay from the speaker to the receiver [including the] compression and decompression delays" [10]. D_{e2e}, the total voice packet delay, is calculated thus:

$$D_{e2e} = D_n + D_e + D_d + D_c + D_{de} \tag{38.3}$$

where D_n, D_e, D_d, D_c and D_{de} represent the network, encoding, decoding, compression and decompression delay, respectively.

The International Telecommunication Union – Telecommunication (ITU-T) offers guidelines for these as shown in Table 38.2 [10].

The '*Packet Delay Variation*' (PDV) is an important factor to consider in network performance degradation assessment as it affects the overall perceptual voice quality. Higher packet delay variation leads to congestion of the packets causing

Table 38.2 Guideline for the voice quality [10]

Network parameter	Good	Acceptable	Poor
Delay (ms)	0–150	150–300	>300
Jitter (ms)	0–20	20–50	>50

more network overheads causing further degradation in voice quality. The PDV is the variance of the packet delay, which is given by:

$$\text{PDV} = \left\{ \sum_{i-1}^{n} \left(\left[t'(n) - t(n) \right] - \mu \right)^2 \right\} / n \tag{38.4}$$

Where: μ is the average delay of n selected packets.

2.4 WiFi (Wireless Fidelity)

The contention based IEEE 802.11x Wireless Local Area Network or WLAN is derived from the non-wireless Ethernet 802.3 Local Area Network (LAN) access technology. The Layer 1 (physical) and Layer 2 (data link) operate over two frequency bands of 2 GHz and 5 GHz. 802.11b (11 Mbit/s) and 802.11 g (54 Mbit/s) are two common standards with a range of between 80 and 100 m. Contention will reduce the practical bitrates and affect the QoS for real-time traffic, especially VoIP. The large packet headers, for both the WiFi and VoIP, constrain the payload capacity and affect the QoS further especially in times of network congestion. WiFi has now become cheap and comes as an installed hardware standard feature on most network devices. WiFi can be found in: public transportation, public spaces such as airports and railway stations, domestic and industrial applications [11].

2.5 UMTS (Universal Mobile Telecommunication System)

The 3GPP (3rd Generation Partnership Project) UMTS takes its foundational architecture from the GSM (Global System for Mobile Communications, originally Groupe Spécial Mobile) network [12], viz.:

- The USIM or UMTS SIM card is backwards compatible in a GSM handset;
- The CCITT (Consultative Committee International Telegraph Telecommunications) Signalling System Number 7 (SS7) MAP (Mobile Application Part) protocol of UMTS is an evolution of the GSM MAP;
- Similar but enhanced versions of the circuit and packet transmission protocols are utilised;
- Seamless mobility during hard handovers is facilitated by special procedures during circuit and packet switching.

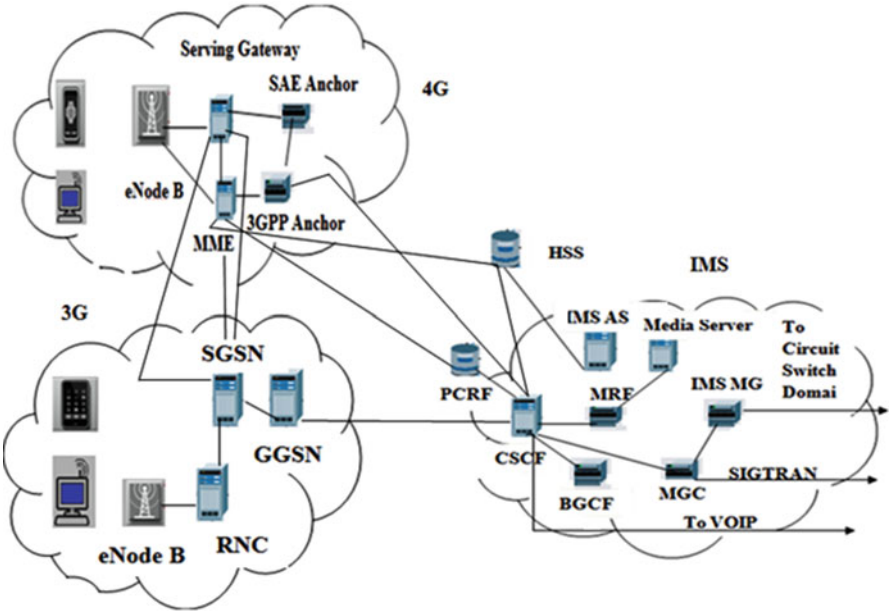


Fig. 38.1 Sample 3G and 4G (3G LTE/SAE [system architecture evolution] network)

3G (as shown in Fig. 38.1) offers many of these features as defined by IMT (International Mobile Telecommunication):

- Flexibility: In terms of services, deployment, technology and air interfaces, the 3G IMT-2000 (International Mobile Telecommunication 2000) standard, also known as UMTS in Europe and J-FPLMTS (Japanese Future Public Land Mobile Telecommunications Services) in Japan, allows for five radio interfaces based on three different multiplex access technologies: FDMA (Frequency Division Multiplex Access), TDMA (Time Division Multiplex Access) and CDMA (Code Division Multiplex Access);
- Affordability: To ensure global uptake by consumers, it must be affordable by the consumers;
- Compatibility with existing systems: 3G must be able to work with previous mobile telephone generations to allow for global roaming as 2G systems still continue to grow and thrive globally;
- Modularity of Design: Scalability allows for coverage of growing population centres including implementing new services with the possible minimal cost.

2.5.1 HSPA (High-Speed Packet Access)

UMTS offers HSPA [13], which is a combination of HSDPA (High-Speed Downlink Packet Access) and HSUPA (High-Speed Uplink Packet Access). The 3GPP also

defines, Enhanced HSPA or HSPA+ (3GPP Rel-7) and Rel-8 Long Term Evolution (LTE) – also known as Evolved UMTS Terrestrial Radio Access (EUTRA). 4G (Fourth Generation) is currently being deployed [14].

3 Literature Review

A previous study has shown that VoIP does adversely affect the throughput of both WiFi and WiMAX [11, 15, 16]. However, packet loss and jitter is only experienced in a WiFi network. Common OPNET network simulation software parameters used for performance studies include: "MOS, end-to-end delay, jitter, and packet delay variation" [8]. A study over High Speed Packet Access (HSPA) [15] has shown that not all VoIP implementations are the same as they have differing quality effects on the voice.

A strategy used to overcome the quality problems of congestion in WiFi when using VoIP is known as SQoSMA (Session-based Quality-of-Service Management Architecture), Quality Assurance of Voice over WLANs [17]. SQoSMA tries to overcome the bandwidth limitation problems by combining the control and data planes in order to mitigate congestion events by choosing the lowest bitrate adaptive audio codec. It then implements a call stopping procedure when necessary to reduce congestion even further.

A similar strategy [18] used an edge VoIP gateway between the Internet Cloud and the WLAN to select the most appropriate variable speech coding rate (16-64, Kbit/s) whilst maintaining overall QoS of the speech traffic.

Transmission delay reduction of VoIP traffic was implemented through a Transmission Control Protocol (TCP) Friendly Rate Control (TFRC) algorithm based 802.11e network using EDCF (Enhanced Distributed Coordination Function)/HCF (Hybrid Coordination Function) scheme [18].

The use of routing and labelling may also be employed to help the rapid passage of the VoIP real-time traffic through the WLAN. The approach was to implement a distributive packet aggregation technique to reduce the time and number of hops through the network [19].

A study in subjective voice quality measure [20] looked at the E-Model and the PESQ (Perceptual Evaluation of Speech Quality) which combined the advantages of both to form the AdmPESQ (Advanced Model for Perceptual Evaluation of Speech Quality) measure. AdmPESQ is especially suitable for heterogeneous networks with differing packet losses and delay parameters.

VoIP continues to grow in popularity as it offers the services of the traditional Public Switched Telephone Network (PSTN) whilst incorporating added value features at competitive rates. Many successfully leading competing companies now offer this service over often heterogeneous networks, though this does affect the QoS. Furthermore, VoIP is now the target of many computer hacker attacks, Materna [21] has categorized these into four types:

1. Service Availability;
2. Service Integrity;
3. SPIT (Spam over Internet Telephony) and
4. Eavesdropping.

Network outage must be kept ideally at zero level as this may critically affect an organization, such as the vital operation of a hospital or nuclear power station. Thus "Service Availability Attacks" must be completely stopped. Downtime means financial loss and unplanned maintenance costs too. Thus the IP Telephony network must be robustly protected against all known forms of attacks such primarily the "Denial of Service" (DOS) types of attacks as well as from viruses and worms. Attacks will affect the service ranging from quality of service deterioration to at worst, the total loss of service. Customers demand the highest level of service so voice quality has to be maintained to an acceptable intelligibility.

VoIP services are more sensitive and hence vulnerable than computers (which are protected more securely) to attacks due to their lower thresholds and immunity. Thus any attacks, for example, a worm, will have a far greater impact on a VoIP network then on a traditional computer network. A computer may be slowed down but the VoIP network may totally crash and suffer from a complete network outage.

The paper presents the results of VoIP traffic through a WiFi and a UMTS network and also in between them, especially in regards to the effect on the voice QoS.

4 Research Methods

An academic environment often constrains research to simulations because of limited resources. However, much can be learned from this for future deployment in the real world for initial testing. Jack Burbank describes "Modeling and Simulation (M&S)" as an acute component in the "design, development and test and evaluation (T&E)" process. According to him, "It is almost always preferable to have insight into how a particular system, individual device, or algorithm will behave and perform in the real world prior to its actual development and deployment." [22] Simulation offers many advantages such as the opportunity to scale a network in a virtual environment thus saving considerable costs [23]. Comparison of technologies is easily achieved in simulations. Our project make use of the OPNET Modeler as it integrates a wide range of technologies and protocols [24], as well as comprising a "development environment" to facilitate M&S for various types of networks for our studies.

In the first scenario, two WiFi subnets, Hawaii and Florida, were deployed. These were configured with a SIP server credential connected through an IP cloud, shown in Fig. 38.2. In the second scenario, two UMTS subnets, New York and California, were used instead of the WiFi subnets. In the third scenario, one of the UMTS subnet (New York) was replaced by a WiFi subnet (Hawaii). The subnets implemented in the project are shown in Table 38.3.

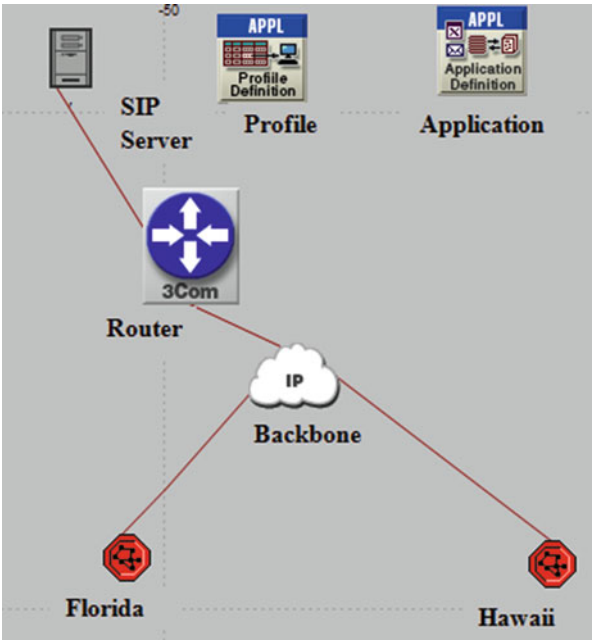


Fig. 38.2 WiFi network scenario

Table 38.3 Devices of the subnets deployed

Subnet name	Scenario	Base St. type	Work station type	No. of workstations
Hawaii	WiFi	WiFi	Mobile	4
Florida	WiFi	WiFi	Mobile	4
New York	UMTS	UMTS	UMTS workstation	4
California	UMTS	UMTS	UMTS workstation	4
Hawaii	WiFi	WiFi	Mobile	4
California	UMTS	UMTS	UMTS workstation	4

In the second scenario, two UMTS subnets, New York and California, were used instead of the WiFi subnets. In the third scenario, one of the UMTS subnet (New York) was replaced by a WiFi subnet (Hawaii). The subnets implemented in the project are as given in Table 38.3.

Workstation in both the UMTS and Wi-Fi network models are configured to run the VoIP application. This VoIP Application is defined to generate one voice frame per packet and to run as an 'Interactive Voice' service. This application is defined in the application profile to run in serial mode. Calls to workstations are based on random generation and are exponentially distributed with average duration of three (3) minutes. The call inter-arrival time are exponentially distributed. UMTS has two major divisions, namely the UMTS Terrestrial Radio Access Network (UTRAN) and the Core Network (CN), as shown in Fig. 38.3. The UTRAN is a combination of two parts: the Radio Network Controller (RNC) and the Node-B. The UTRAN

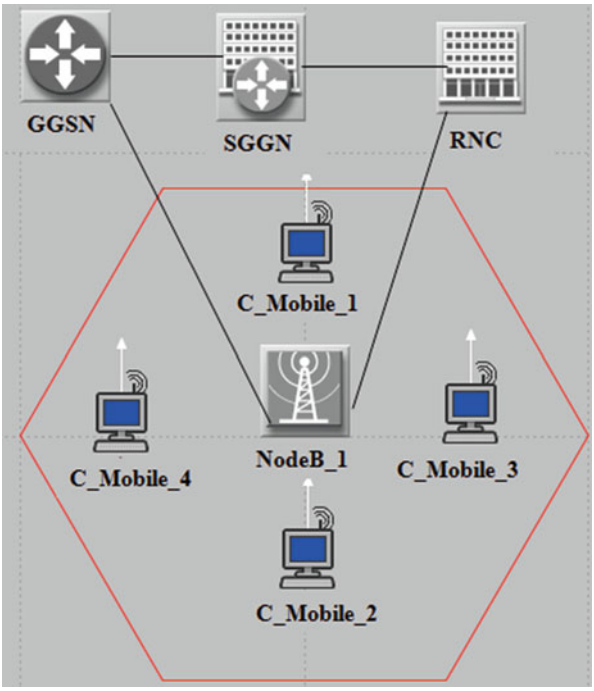


Fig. 38.3 Inside view of a UMTS subnet

handles all radio related functionalities. The CN is responsible for maintaining the subscriber data and for switching the voice and data connections.

In Fig. 38.3, GGSN is the Gateway GPRS (General Packet Radio Service) Support Node; SGGN is the Serving GPRS Support Node and the RNC is the Radio Network Controller.

5 Results

The results were collected simulating all three scenarios for one (1) hour. Figure 38.4a shows the average jitter of all three scenarios as overlaid graphs. Whereas Fig. 38.4b presents the average jitter graphs as stacked. The UMTS jitter has been found to be much higher (around 0.10 s) than that of the other two scenarios (around 0.0 s). In fact, the jitter graph for UMTS had no results for about the first 5 min. This is most probably due to the network convergence period. It is also to be noted that the mixed scenario presents negative jitter during this time due to the same reason. Negative jitter indicates that the packets arrived before the interval time during that period.

It is quite interesting that the average jitter is not only very steady in both the WiFi and WiFi-UMTS scenarios but also remains nearly zero (0). As the simulation

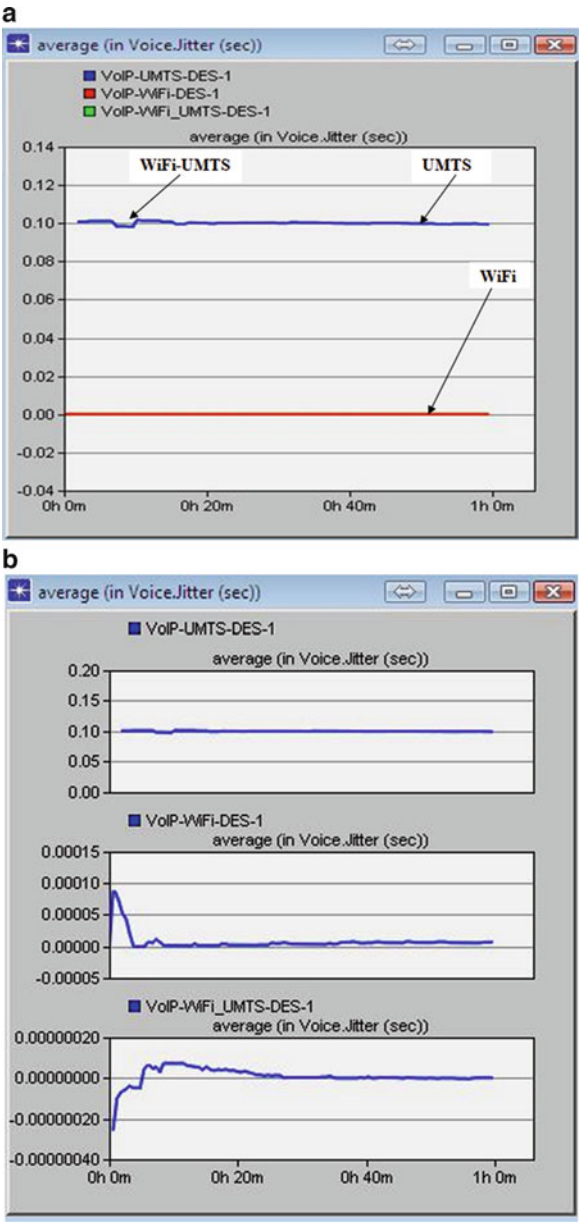


Fig. 38.4 (**a**) Average VoIP jitter (overlaid). (**b**) Average VoIP jitter (stacked)

was configured to run on the basis of random call generation and there is no direct handover involved, the average jitter of the WiFi-UMTS scenario should preferably always stay somewhere in-between the WiFi and UMTS jitter times. This result, hence, is very thought provoking and demands further study.

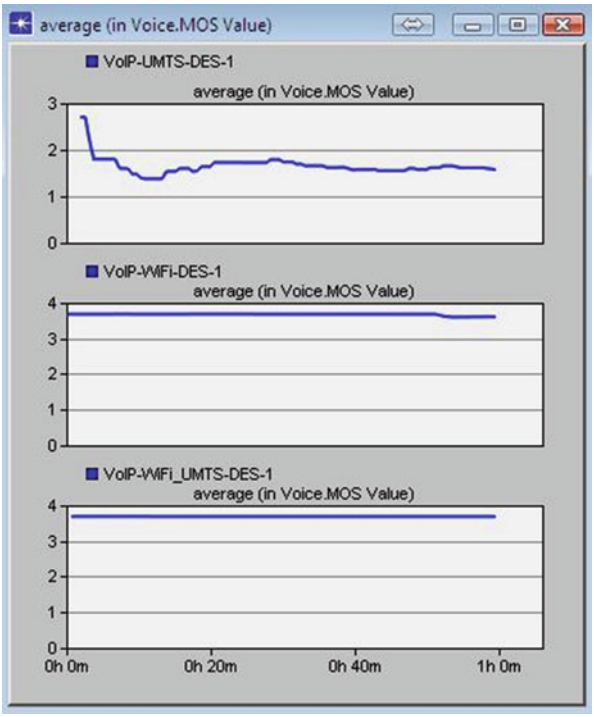


Fig. 38.5 Average MOS (overlaid)

This similar tendency has also been observed whilst comparing the MOS performance, as shown in Fig. 38.5. In terms of the MOS, although the call generation was exponentially distributed, both the WiFi and WiFi-UMTS networks observe a similar level of performance. This remains very close to the subjective value of four (4), over the complete simulation period. UMTS, on the other hand, suffers from not only a lower level of MOS, but also an unsteady level.

The MOS of these scenarios varied between: 1.5 and 2.8. So, considering the MOS score alone, it can be concluded that: (1) both the WiFi and the WiFi-UMTS networks surpass the UMTS network and (2) despite the MOS of the WiFi-UMTS network should ideally remain somewhere near the mid-point of the WiFi and UMTS MOS graphs, it exhibits a greater performance than that.

In terms of the packet end-to-end delay, unexpectedly, the WiFi-UMTS network provides better services than either the WiFi or the UMTS networks, as shown in Fig. 38.6. The similar behaviour has also been observed while comparing the Packet Delay Variation, as shown in Fig. 38.7. These results are very significant as it is expected that in the mixed (WiFi-UMTS) network case, the variables should remain in the middle under the ideal situation and might sometimes provide an inferior performance in the worst case scenario. Furthermore, in depth research is required to identify the reasons behind this phenomenon.

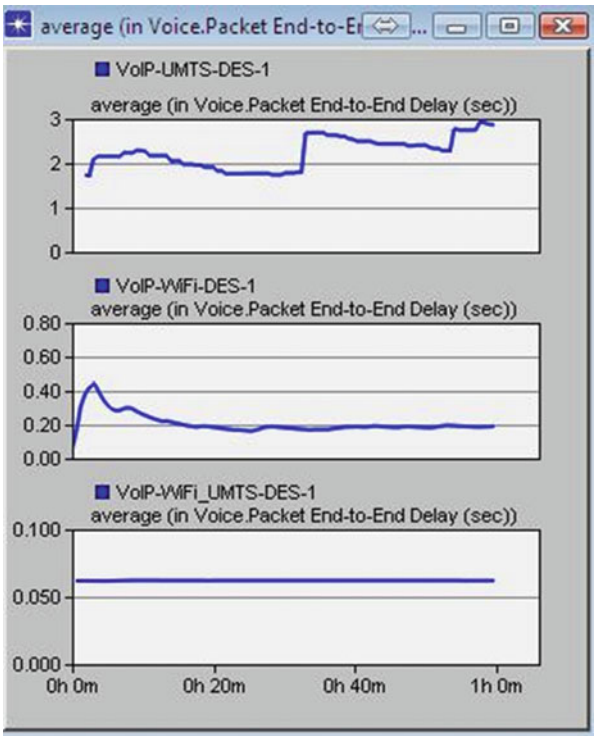


Fig. 38.6 Packet end-to-end delay

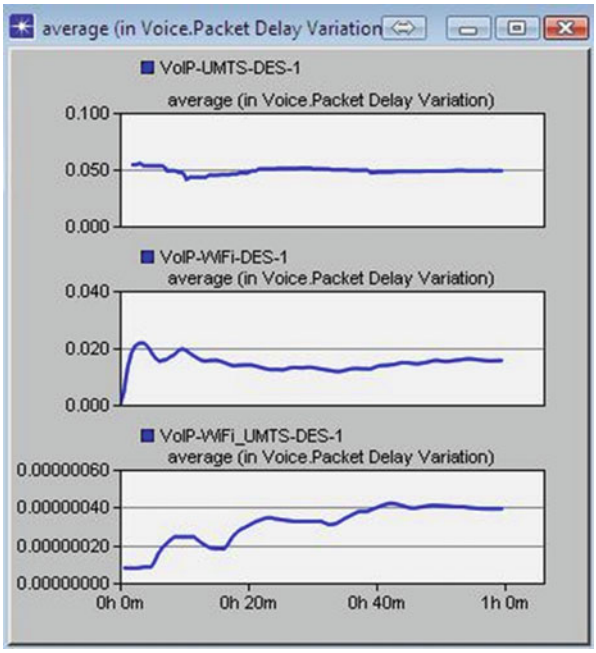


Fig. 38.7 Packet delay variation

6 Conclusion and Future Work

The paper reported the preliminary findings of VoIP traffic through the WiFi, UMTS and WiFi-UMTS networks. To begin with, two different scenarios were deployed (WiFi and UMTS) and tested, where VoIP call generation and termination took place in the respective homogenous networks. Later, an additional scenario was added to the project: VoIP calls being generated using the WiFi network terminated using the UMTS network and vice-versa. While considering the Packet Delay Variation and the packet end-to-end delay, it has been found that the WiFi-UMTS heterogeneous network provides better services than the WiFi and UMTS networks separately. These results are unexpected as described in the discussion section of this paper. Further detailed research is definitely necessary to identify the reasons behind these phenomena.

It is quite interesting that the average jitter is not only very steady in both the WiFi and WiFi-UMTS scenarios but also remains nearly zero (0). As the simulation was configured to run on the basis of random call generation with no direct handovers, the average jitter of the WiFi-UMTS scenario should preferably always stay somewhere in-between the WiFi and UMTS networks. This result, hence, is very perplexing and requires further investigation.

Considering the MOS, the UMTS network suffers not only from poor performance, but also remains unsteady. In addition to that, the MOS performance of both the WiFi and WiFi-UMTS networks surpass that of the UMTS network and also the MOS of the WiFi-UMTS network exhibits a greater performance than expected.

The future of the project aims to include the study of other network technologies and techniques covering: CDMA, EDGE, GSM, GPRS, LTE and 4G. The effect of VoIP negotiating both a non-heterogeneous and heterogeneous network will be of one particular major focus of this ongoing research initiative. The study will be further extended to investigate the effect on other parameters such as: throughput, queuing delay and the packet drop rate. The network parameters will be carefully scrutinized for their optimization to improve the overall network efficiency.

References

1. Miraz, M.H., Molvi, S.A., Ali, M., Ganie, M.A., Hussein, A.R.H.: Analysis of QoS of VoIP traffic through WiFi-UMTS networks. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2014, pp. 684–689. WCE 2014, London, 2–4 July 2014
2. Ebbinahaus, R.: VoIP lessons: look who's calling. IET Commun. Eng. **1**(5), 28–31 (2013). October–November 2013 [Online]. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1244763
3. Gerea, F.: Security in VoIP. J. Knowl. Manag., Econ. Inf. Technol. **2**(2), 1–10 (2012). [Online]. http://www.scientificpapers.org/information-technology/security-in-voip/
4. Gerea, F.: Security in VoIP. J. Knowl. Manag., Econ. Inf. Technol. **2**(2), 1–10 (2012) [Online]. http://www.scientificpapers.org/information-technology/security-in-voip/

5. Islam, S., Rashid, M., Tarique, M.: Performance analysis of WiMax/WiFi system under different codecs. Int. J. Comput. Appl. **18**(6), 13–19 (2011) [Online]. http://www.ijcaonline.org/archives/volume18/number6/2290-2973
6. Rosenberg, J., et al.: SIP: session initiation protocol. IEFT RFC Editor, pp. 1–194 (2002) [Online]. http://ietfreport.isoc.org/all-ids/draft-ietf-sip-rfc2543bis-08.pdf
7. Pandya, R.: Emerging mobile and personal communication systems. IEEE Commun. Mag. **33**(6), 44–52 (1995) [Online]. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00387549
8. Jadhav, S., Zhang, H., Huang, Z.: Performance evaluation of quality of VoIP in WiMAX and UMTS. In: Proceedings of the 12th IEEE International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2011, pp. 375–380. Gwangju, 20–22 October 2011 [Online]. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6118553
9. The International Telecommunication Union – Telecommunication (ITU-T): G.107: The E-model: a computational model for use in transmission planning, Rec., vol. G.107. February 2014 [Online]. http://www.itu.int/rec/T-REC-G.107
10. Yu, J., Al-Ajarmeh, I.: Call admission control and traffic engineering of VoIP. In: Proceedings of the Second IEEE International Conference on Digital Telecommunications (ICDT 2007), pp. 11–16. San Jose, 1–5 July 2007 [Online]. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4270577
11. Qureshi, M.A., Younus, A., Saeed, M., Sidiqui, F.A., Touheed, N., Qureshi, M.S.: Comparative study of VoIP over WiMAX and Wi-Fi. Int. J. Comput. Sci. Issues **8**(3), 433–437 (2011) [Online]. http://ijcsi.org/papers/IJCSI-8-3-1-433-437.pdf
12. Lescuyer, P., Lucidarme, T.: Evolved Packet System (EPS): The LTE and SAE Evolution of 3G UMTS. John Wiley & Sons Ltd., Chichester (2008)
13. Bhebhe, L., Parkkali, R.: VoIP performance over HSPA with different VoIP clients, Springer, New York (2010)
14. The mobile broadband evolution: 3GPP release 8 and beyond HSPA+, SAE/LTE and LTE-ADVANCED, 3G Americas, pp. 8–9, February (2009)
15. Ohrtman, F.: WiMAX Handbook: Building 802.16 Wireless Networks. McGraw-Hill, New York (2005)
16. Hrudey, W., Trajković, L.: Mobile WiMAX MAC and PHY layer optimization for IPTV. Math. Comput. Model. **53**(11–12), 2119–2135 (2011) [Online]. http://www.sciencedirect.com/science/article/pii/S0895717710003821
17. Tebbani, B., Haddadou, K., Pujolle, G.: Quality assurance of voice over WLANs (VoWLANs) with differentiated services, Institut Télécom and Springer-Verlag, New York (2010)
18. Trad, A., Ni, Q., Afifi, H.: Adaptive VoIP transmission over heterogeneous wired/wireless networks, INRIA, Planete Project 2004 Route des Lucioles, BP-93, 06902 Sophia-Antipolis
19. Niculescu, D., Ganguly, S., Kim, K., Izmailov, R.: Performance of VoIP in a 802.11 wireless mesh network, NEC Laboratories America, Princeton (2006)
20. Rodrigues, D., Cerqueira, E., Monteiro, E.: QoE assessment of VoIP in next generation networks, University of Coimbra, Department of Information Engineering, 3030–290 Coimbra, Federal University of Para, Rua Augusto Corra, 01, 66075–110, Belém (2009)
21. Materna, B.: VoIP insecurity. IET Commun. Eng. **4**(5), 39–42 (2006) [Online]. http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?reload=true&arnumber=4049981
22. Burbank, J.: Modeling and simulation: a practical guide for network designers and developers. IEEE Commun. Mag. **47**(3), 118 (2009)
23. Maria, A.: Introduction to modeling and simulation. In: Winter Simulation Conference, IEEE Computer Society, Washington, DC (1997)
24. OPNET Modeler [Online]. http://www.opnet.com/solutions/network_rd/modeler.html

Chapter 39
Extended Performance Studies of Wi-Fi IEEE 802.11 A, B, G Laboratory Wep Point-to-Multipoint and Point-to-Point Links

J.A.R. Pacheco de Carvalho, C.F. Ribeiro Pacheco, A.D. Reis, and H. Veiga

Abstract Wireless communications using microwaves are increasingly important, such as Wi-Fi. Performance is a most fundamental issue, leading to more reliable and efficient communications. Security is equally very important. Laboratory measurements were performed on several performance aspects of Wi-Fi (IEEE 802.11a, b, g) WEP point-to-multipoint links. Our study contributes to performance evaluation of this technology, using available equipments (DAP-1522 access points from D-Link and WPC600N adapters from Linksys). New detailed results are presented and discussed, namely at OSI levels 4 and 7, from TCP, UDP and FTP experiments: TCP throughput, jitter, percentage datagram loss and FTP transfer rate. Comparisons are made to corresponding results obtained for WEP point-to-point and Open point-to-multipoint links. Conclusions are drawn about the comparative performance of the links.

Keywords IEEE 802.11a • IEEE 802.11b • IEEE 802.11g • Wi-Fi • Wireless network laboratory performance measurements • WLAN • WEP point-to-multipoint links

1 Introduction

Contactless communication techniques have been developed using mainly electromagnetic waves in several frequency ranges, propagating in the air. Examples of microwave based and laser based wireless communications technologies are Wi-Fi and FSO, respectively, whose importance and utilization have been growing.

J.A.R.P. de Carvalho (✉) • C.F.R. Pacheco • A.D. Reis
Unidade de Detecção Remota, Universidade da Beira Interior, Covilhã 6201-001, Portugal
e-mail: pacheco@ubi.pt; a17597@ubi.pt; adreis@ubi.pt

H. Veiga
Centro de Informática, Universidade da Beira Interior, Covilhã 6201-001, Portugal
e-mail: hveiga@ubi.pt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_39

Wi-Fi provides for versatility, mobility and favourable prices. The importance and utilization of Wi-Fi have been growing for complementing traditional wired networks. Both ad hoc and infrastructure modes are used. In this case an access point, AP, permits communications of Wi-Fi devices (such as a personal computer, a wireless sensor, a PDA, a smartphone, a video game console, a digital audio player) with a wired based LAN through a switch/router. In this way a WLAN, based on the AP, is formed. Wi-Fi has reached the personal home, where a WPAN permits personal devices to communicate. Point-to-point (PTP) and point-to-multipoint (PTMP) configurations are used both indoors and outdoors, requiring specific directional and omnidirectional antennas. Wi-Fi uses microwaves in the 2.4 and 5 GHz frequency bands and IEEE 802.11a, 802.11b, 802.11g and 802.11n standards [1]. As the 2.4 GHz band becomes increasingly used and interferences increase, the 5 GHz band has received considerable attention, although absorption increases and ranges are shorter.

Nominal transfer rates up to 11 (802.11b), 54 Mbps (802.11a, g) and 600 Mbps (802.11n) are specified. CSMA/CA is the medium access control. There are studies on wireless communications, wave propagation [2, 3], practical implementations of WLANs [4], performance analysis of the effective transfer rate for 802.11b point-to-point links [5], 802.11b performance in crowded indoor environments [6].

Performance evaluation is a crucially important criterion to assess the reliability and efficiency of communication. In comparison to traditional applications, new telematic applications are specially sensitive to performances. E.g. requirements have been pointed out, such as: 1–10 ms jitter and 1–10 Mbps throughput for video on demand/moving images; jitter less than 1 ms and 0.1–1 Mbps throughputs for Hi Fi stereo audio [7].

Wi-Fi security is very important. Microwave radio signals travel through the air and can be easily captured by virtually everybody. Therefore, several security methods have been developed to provide authentication such as, by increasing order of security, WEP, WPA and WPA2. WEP was initially intended to provide confidentiality comparable to that of a traditional wired network. A shared key for data encryption is involved. In WEP, the communicating devices use the same key to encrypt and decrypt radio signals. The CRC32 checksum used in WEP does not provide a great protection. However, in spite of its weaknesses, WEP is still widely used in Wi-Fi communications for security reasons, mainly in PTP links. WPA implements the majority of the IEEE 802.11i standard [1]. It includes a message integrity check, MIC, replacing the CRC used in WEP. WPA2 is compliant with the full IEEE 802.11i standard. It includes CCMP, a new AES-based encryption mode with enhanced security. WPA and WPA2 can be used in either personal or enterprise modes. In this latter case an 802.1× server is required. Both TKIP and AES cipher types are usable and a group key update time interval is specified.

Several performance measurements have been made for 2.4 and 5 GHz Wi-Fi open [8, 9], WEP [10, 11], WPA [12] and WPA2 [13] links, as well as very high speed FSO [14]. In the present work new Wi-Fi (IEEE 802.11a, b, g) results arise, using WEP, through OSI levels 4 and 7. Performance is evaluated in laboratory measurements of WEP PTMP links using new available equipments. Comparisons

are made to corresponding results obtained for WEP PTP and Open PTMP links. The present work is an extension of [11].

In prior and actual state of the art, several Wi-Fi links have been investigated. Performance evaluation has been considered as a crucially important criterion to assess communications quality. The motivation of this work is to evaluate performance in laboratory measurements of WEP PTMP links using available equipments. Comparisons are made to corresponding results obtained for WEP PTP and Open PTMP links. This contribution permits to increase the knowledge about performance of Wi-Fi (IEEE 802.11a, b, g) links [4–6]. The problem statement is that performance needs to be evaluated under security encryption and several topologies. The solution proposed uses an experimental setup and method, permitting to monitor, mainly, signal to noise ratios (SNR) and noise levels (N) and measure TCP throughput (from TCP connections) and UDP jitter and percentage datagram loss (from UDP communications).

The rest of the paper is structured as follows: Sect. 2 presents the experimental details i.e. the measurement setup and procedure. Results and discussion are presented in Sect. 3. Conclusions are drawn in Sect. 4.

2 Experimental Details

The measurements used a D-Link DAP-1522 bridge/access point [15], with internal PIFA *2 antenna, IEEE 802.11a/b/g/n, firmware version 1.31 and a 100-Base-TX/10-Base-T Allied Telesis AT-8000S/16 level 2 switch [16]. The wireless mode was set to access point mode. Two PCs were used having a PCMCIA IEEE.802.11a/b/g/n Linksys WPC600N wireless adapter with three internal antennas [17], to enable PTMP links to the access point. In every type of experiment, interference free communication channels were used (ch 36 for 802.11a; ch 8 for 802.11b, g). This was checked through a portable computer, equipped with a Wi-Fi 802.11a/b/g/n adapter, running NetStumbler software [18]. 128 bits WEP encryption was activated in the AP and the wireless adapters of the PCs, using a shared key composed of 26 ASCII characters. The experiments were made under far-field conditions. No power levels above 30 mW (15 dBm) were required, as the wireless equipments were close.

A versatile laboratory setup has been planned and implemented for the PTMP measurements, as shown in Fig. 39.1. At OSI level 4, measurements were made for TCP connections and UDP communications using Iperf software [19]. For a TCP connection (TCP New Reno, RFC 6582, was used), TCP throughput was obtained. For a UDP communication with a given bandwidth parameter, UDP jitter and percentage loss of datagrams were determined. Parameterizations of TCP packets, UDP datagrams and window size were as in [13]. One PC, 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 is the smooth mean of differences between consecutive transit times, was continuously computed by the server, as specified by the real time protocol RTP, in

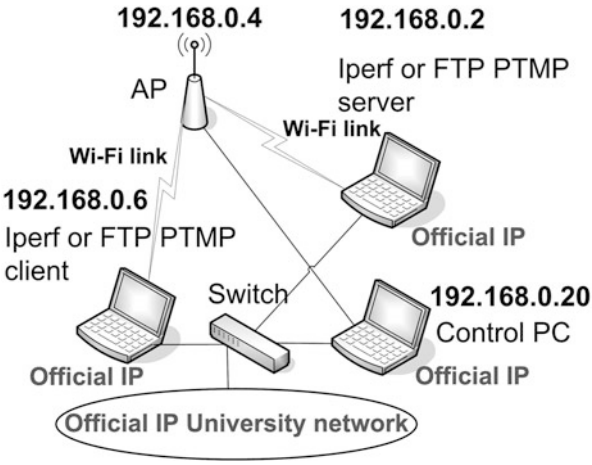


Fig. 39.1 Experimental laboratory setup scheme [11]

RFC 1889 [20]. Another PC, with IP 192.168.0.20, was used to control the settings in the AP. The scheme of Fig. 39.1 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. The server and client PCs were HP nx9030 and nx9010 portable computers, respectively, running Windows XP. They were configured to optimize the resources allocated to the present work. Batch command files have been written to enable the TCP, UDP and FTP tests.

The results were obtained in batch mode and written as data files to the client PC disk. Each PC had a second network adapter, to permit remote control from the official IP University network, via switch.

3 Results and Discussion

The access point and the wireless network adapters of the PCs were manually configured for each standard IEEE 802.11a, b, g with typical nominal transfer rates (1, 2, 5.5, 11 Mbps for 11b; 6, 9, 12, 18, 24, 36, 48, 54 Mbps for 11a, g). For every fixed transfer rate, data were obtained for comparison of the laboratory performance of the WPA PTMP and PTP links at OSI levels 1 (physical layer), 4 (transport layer) and 7 (application layer) using the setup of Fig. 39.1. For each standard and every nominal fixed transfer rate, an average TCP throughput was determined from several experiments. This value was used as the bandwidth parameter for every corresponding UDP test, giving average jitter and average percentage datagram loss.

At OSI level 1, noise levels (N, in dBm) and signal to noise ratios (SNR, in dB) were monitored and typical values are shown in Fig. 39.2.

The main average TCP and UDP results are summarized in Table 39.1, both for WEP PTMP and PTP links. The statistical analysis, including calculations of

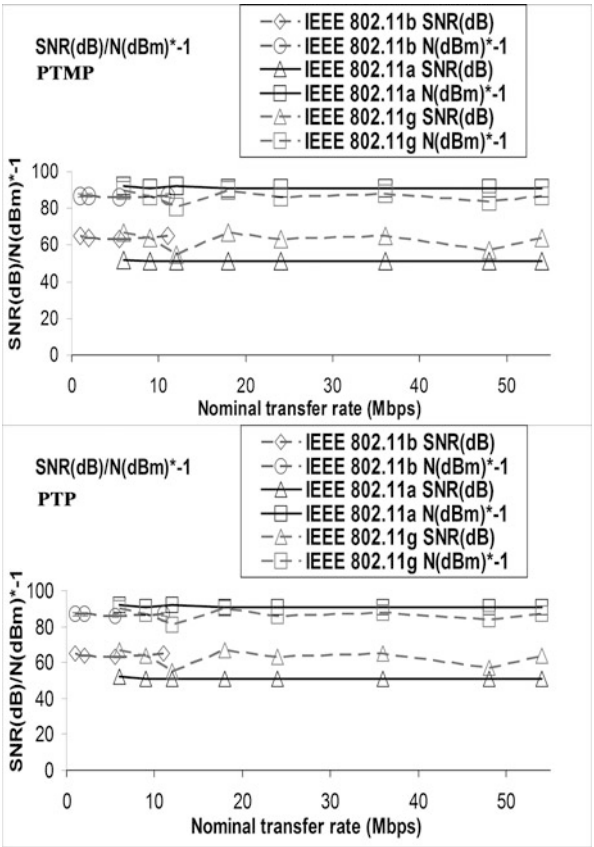


Fig. 39.2 Typical SNR (dB) and N (dBm); WEP PTMP and PTP links

Table 39.1 Average Wi-Fi (IEEE 802.11a, b, g) WEP results; PTMP and PTP links

Link type	PTMP			PTP		
Parameter/IEEE standard	802.11b	802.11a	802.11g	802.11b	802.11a	802.11g
TCP throughput (Mbps)	$1.1 + -0.0$	$7.4 + -0.2$	$6.5 + 0.2$	$3.0 + 0.1$	$15.3 + 0.5$	$14.7 + -0.4$
UDP-jitter (ms)	$7.7 + -2.4$	$2.8 + -0.2$	$4.3 + -0.6$	$5.3 + 0.3$	$2.5 + 0.3$	$2.6 + 0.2$
UDP-% datagram loss	$1.2 + -0.2$	$1.2 + -0.1$	$1.7 + -0.2$	$1.2 + 0.2$	$0.7 + 0.1$	$1.6 + 0.4$

confidence intervals, was carried out as in [21]. In Fig. 39.3 polynomial fits were made (shown as y versus x), using the Excel worksheet, to the 802.11a, b, g TCP throughput data for PTMP and PTP links, respectively, where R^2 is the coefficient of determination. It gives information about the goodness of fit. If it is 1.0 it means a perfect fit to data. It was found that, on average, the best TCP throughputs are for 802.11a and PTP links ($15.3 + -0.5$ Mbps, versus $7.4 + -0.2$ Mbps for PTMP).

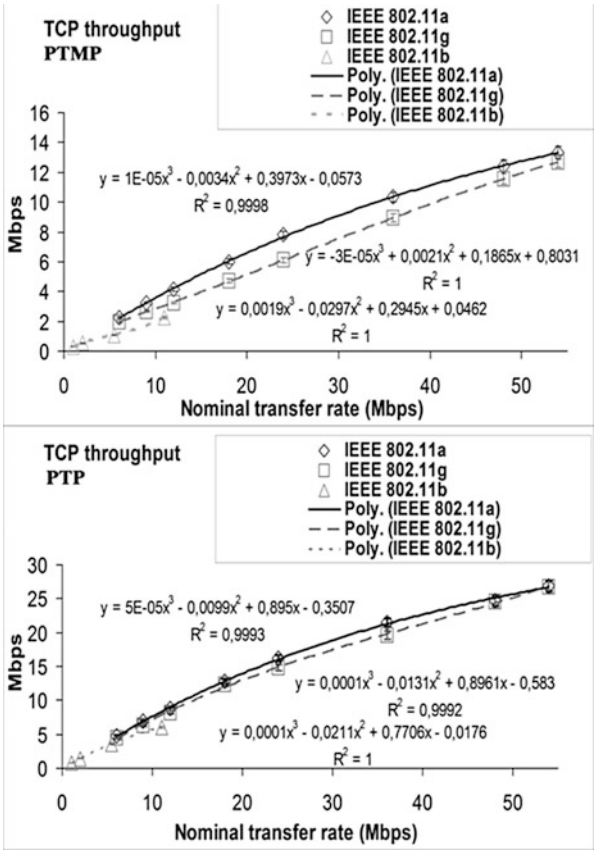


Fig. 39.3 TCP throughput results (y) versus technology and nominal transfer rate (x); WEP PTMP and PTP links

On average TCP throughput for Open PTMP links ($7.6 + -0.2$ Mbps) was found slightly better than for WEP PTMP links ($7.4 + -0.2$ Mbps). In Figs. 39.4 and 39.5, the data points representing jitter and percentage datagram loss were joined by smoothed lines. It was found that, on average, the best jitter performances are for 802.11a and PTP links ($2.5 + -0.3$ ms). On average, jitter performance was found similar for Open PMTP ($3.1 + -0.3$ ms) and WEP PTMP links ($2.8 + -0.2$ ms). Concerning percentage datagram loss, the best performance was for 802.11a and PTP links ($0.7 + -0.1$ %, versus $1.2 + -0.1$ % for PTMP). On average, percentage datagram loss performance was found better for Open PTMP links ($1.0 + -0.1$ %) than for WEP PTMP links ($1.2 + -0.1$ %). Generally, in comparison to PTP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for PTMP links. Generally, in comparison to Open PTMP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for WEP PTMP links, where data length is increased.

At OSI level 7 we measured FTP transfer rates versus nominal transfer rates, configured in the access point and the wireless network adapters of the PCs, for the

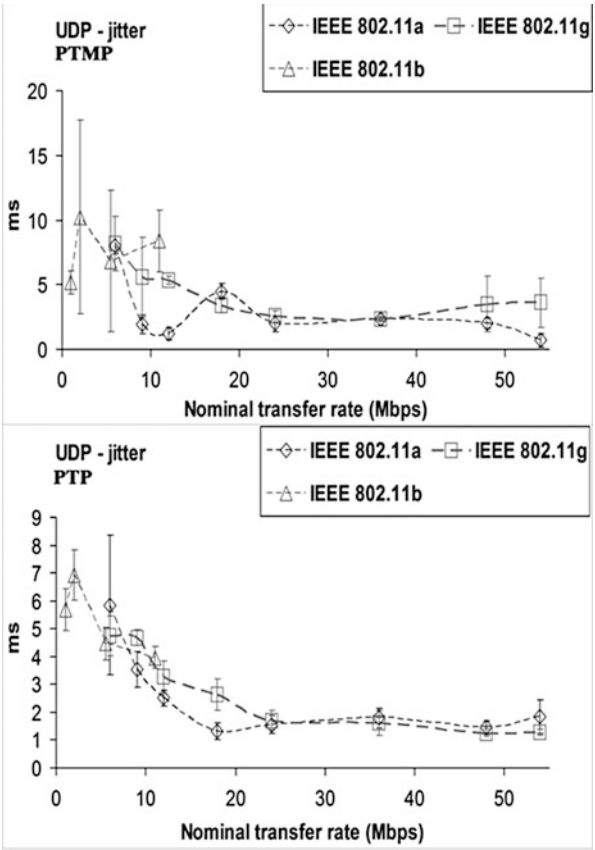


Fig. 39.4 UDP – jitter results versus technology and nominal transfer rate; WEP PTMP and PTP links

IEEE 802.11a, b, g standards. The result for every measurement was an average of several experiments involving a single FTP transfer of a binary file with a size of 100 Mbytes. The FTP results show the same trends found for TCP throughput.

4 Conclusions

In the present work a versatile laboratory setup arrangement was planned and implemented, that permitted systematic performance measurements of new available wireless equipments (DAP-1522 access points from D-Link and WPC600N adapters from Linksys) for Wi-Fi (IEEE 802.11a, b, g) in WEP point-to-multipoint links.

Through OSI layer 4, TCP throughput, jitter and percentage datagram loss were measured and compared for every standard and WEP PTMP and PTP links.

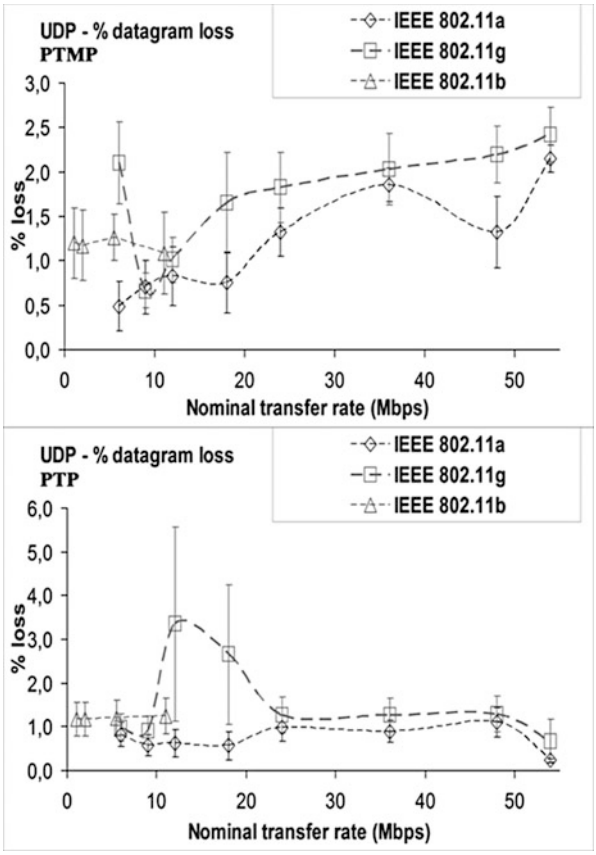


Fig. 39.5 UDP – percentage datagram loss results versus technology and nominal transfer rate; WEP PTMP and PTP links

It was found that, on average, the best TCP throughputs are for 802.11a and PTP links. On average TCP throughput for Open PTMP links was found slightly better than for WEP PTMP links. It was found that, on average, the best jitter performances are for 802.11a and PTP links. On average, jitter performance was found similar for Open PMTP and WEP PTMP links. Concerning percentage datagram loss, the best performance was for 802.11a and PTP links. On average, percentage datagram loss performance was found better for Open PTMP links than for WEP PTMP links. Generally, in comparison to PTP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for PTMP links, where the access point has to maintain links between PCs. Generally, in comparison to Open PTMP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for WEP PTMP links, where data length is increased.

At OSI layer 7, FTP performance results have shown the same trends found for TCP throughput.

Further work involving additional performance studies is planned using several equipments, topologies, security settings and noise conditions, 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)/PEst-OE/FIS/UI0524/2014(ProjectoEstratégico-UI524-2014) are acknowledged.

References

1. IEEE 802.11a, 802.11b, 802.11g, 802.11n, 802.11i standards; http://standards.ieee.org/getieee802. Accessed 10 Jan 2011
2. Mark, J.W., Zhuang, W.: Wireless Communications and Networking. Prentice-Hall, Upper Saddle River (2003)
3. Rappaport, T.S.: Wireless Communications Principles and Practice, 2nd edn. Prentice-Hall, Upper Saddle River (2002)
4. Bruce III, W.R., Gilster, R.: Wireless LANs End to End. Hungry Minds, New York (2002)
5. Schwartz, M.: Mobile Wireless Communications. Cambridge University Press, Cambridge (2005)
6. 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, 27–30 Nov 2006
7. Monteiro, E., Boavida, F.: Engineering of Informatics Networks, 4th edn. FCA-Editor of Informatics Ltd, Lisbon (2002)
8. 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 (2008)
9. Pacheco de Carvalho, J.A.R., Veiga, H., Gomes, P.A.J., Ribeiro Pacheco, C.F., Marques, N., Reis, A.D.: Wi-Fi point-to-point links- performance aspects of IEEE 802.11a, b, g laboratory links. In: Ao, S.-I., Gelman, L. (eds.) Electronic Engineering and Computing Technology. Lecture Notes in Electrical Engineering, vol. 60, pp. 507–514. Springer, Dordrecht (2010)
10. Pacheco de Carvalho, J.A.R., Veiga, H., Marques, N., Ribeiro Pacheco, C.F., Reis, A.D.: Wi-Fi WEP point-to-point links- performance studies of IEEE 802.11a, b, g laboratory links. In: Ao, S.-I., Gelman, L. (eds.) Electronic Engineering and Computing Technology. Lecture Notes in Electrical Engineering, vol. 90, pp. 105–114. Springer, Dordrecht (2011)
11. Pacheco de Carvalho, J.A.R., Ribeiro Pacheco, C.F., Reis, A.D., Veiga, H.: Laboratory performance measurements of IEEE 802.11b, g WEP PTP links. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 724–727. WCE, London, 2–4 July 2014
12. Pacheco de Carvalho, J.A.R., Veiga, H., Ribeiro Pacheco, C.F., Reis, A.D.: Performance studies of Wi-Fi IEEE 802.11a, g WPA point-to-multipoint links. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, pp. 1415–1419. WCE, London, 3–5 July 2013
13. Pacheco de Carvalho, J.A.R., Veiga, H., Marques, N., Ribeiro Pacheco, C.F., Reis, A.D. Laboratory performance of Wi-Fi IEEE 802.11b, g WPA2 point-to-point links: a case study. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2011, pp. 1770–1774. WCE, London, 6–8 July 2011
14. Pacheco de Carvalho, J.A.R., Marques, N., Veiga, H., Ribeiro Pacheco, C.F., Reis, A.D.: Experimental performance evaluation of a Gbps FSO link: a case study. Proceedings of WINSYS 2010- International Conference on Wireless Information Networks and Systems, pp. 123–128. Athens, 26–28 July 2010

15. DAP-1522 wireless bridge/access point technical manual. http://www.dlink.com (2010). Accessed 15 Jan 2012
16. AT-8000S/16 level 2 switch technical data. http://www.alliedtelesis.com (2009). Accessed 10 Dec 2010
17. WPC600N notebook adapter user guide. http://www.linksys.com (2008). Accessed 10 Jan 2012
18. NetStumbler software. http://www.netstumbler.com (2005). Accessed 21 Mar 2011
19. Iperf software. http://dast.nlanr.net (2003). Accessed 10 Jan 2008
20. Network Working Group: RFC 1889-RTP: a transport protocol for real time applications. http://www.rfc-archive.org (1996). Accessed 10 Feb 2008
21. Bevington, P.R.: Data Reduction and Error Analysis for the Physical Sciences. Mc Graw-Hill, New York (1969)

Chapter 40
Deterministic Ethernet Using a Network Traffic Oscillator

Yuen Kwan Mo, Mark S. Leeson, and Roger J. Green

Abstract This chapter discusses the structure of *deterministic* Ethernet which helps to improve the security and accessibility of services in a time critical communication system. Frame switching is the method of connecting frames to their destination. A disorganised frame switching network increases the fluctuation of application response times and creates weakness in a sensitive network. Dynamic frame traffic oscillation allows the frame workload to be more organised and manageable for networking by isolating external influences that disrupt network service. This concept, which we term *Critical Networking,* organises networks and reduces network resource wastage, such as overheads, thereby reducing application network frame delay by carefully planning the type of frame transmission and the available network resource. The frame traffic is handled by a network traffic oscillator (NTO), which creates deterministic time response in frame transmission. Both concepts when combined remove the need to assign arbitrary priority numbers in frames, and overheads for consecutive application transmissions and hides sensitive overhead information within every frame. The concept is illustrated using a simulation of real-time traffic in an airfield scenario.

Keywords Cloud computing • Critical networking • Deterministic Ethernet • Interlayer networking • Network traffic oscillator • SESAR

1 Introduction

The Ethernet standard has a strict frame structure [1] but requires upper OSI layers such as the network and transport layers for organising traffic shape and size. Frames appear to be roaming the network randomly when observed from perspective of recorded traffic. Frame transmissions are not however random, rather this randomness is caused by other external factors such as unusual network switching schemes in the upper OSI layers. This traffic disparity is not recorded in the frame overhead field [2]. A unified traffic shaping platform transmitting

Y.K. Mo (✉) • M.S. Leeson • R.J. Green
School of Engineering, University of Warwick, Coventry CV4 7AL, UK
e-mail: Y-K.T.Mo@Warwick.ac.uk; mark.leeson@Warwick.ac.uk; roger.green@Warwick.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_40

in a coherent manner across all OSI fields, spanning switching and application transmission, promotes punctuality in frame transmission and ultimately reduces frames lost from timeouts. This method also reduces the use of overheads by keeping track of the frame, and removes unnecessary additional frame control overhead. This was originally difficult to achieve using the current OSI model, where each OSI layer works independently [3]. This problem is a common issue amongst all real-time communication networks. One contemporary example is a car based controller area network (CAN bus) [4], where there is frame congestion due to poor network routing and planning, thus frames do not arrive in sequence. Priority frame switching has been introduced to allow important frames to be transmitted first, but this only temporarily reduces some of the important frame delays on a selection of applications. Time Triggered Ethernet (TTE) is an architecture protocol designed for critical time window transmissions [5]. This protocol focuses primarily on maintaining a target time transmission window by forcing frames to be delivered within a slotted window. This is useful in ad-hoc point to point communications, where the frame transmission rate is low and there are many network resources to support every frame. A busy real-time star topology network with unknown frame arrival rate and payload size creates conditional switching probabilities that remove the deterministic nature of Ethernet communication and ultimately distort the uniform time response between communications. The safety critical nature presented in this work is for airfield communications, where the European Union Single European Sky Air Traffic Management Research (SESAR) project is progressing toward a Europe-wide unified air traffic control infrastructure [6]. SESAR proposes a management information model known as the System Wide Information Architecture (SWIM) [6], which combines many traffic streams (concerning flights, weather and so on) into one centralised data pool. This pool is then available for the appropriate subscriber access, such as airlines and air traffic control. The proliferation of connections inside this network opens up many issues of operation, maintenance and security. The major network security issue is middle-man cyber-attacks [7]. Critical networking offers enhanced transparency and control of network operation. Better security measures such as advanced firewall detection and blocking will benefit from this concept and higher efficiency of frame payload encryption can be used (while frame overhead is reduced). Busy real-time airfield communication systems such as the SWIM architecture have many delay variations in a star network. This is the key factor for designing a traffic transmission and management system that removes this delay uncertainty and promotes security. Priority Ethernet frame queuing [8] addresses some of the issues of traffic control using an arbitrary priority number within an Ethernet frame to correct this stochastic delay variance with data throughput. This method works by shifting priority frames to be transmitted first, but only partially fixes the problem as there are conflicting priority orders within each frame. Here, we propose an alternative method which we term a network traffic oscillator (NTO) [9] that always ensures the same frame transmission sequence for every application.

We have found that creating an oscillation in frame transmissions with varying frame size management (traffic load) produces a deterministic arrival rate that

increases the precision of transmission, thus removing unnecessary delays. Current network switches and nodes have buffers to handle unknown frame arrival rates. The NTO uses a master buffer and flow traffic controller from the transmitter for as long as this arrival rate is maintained; there is no need for any additional queue which increase frame transmission delay.

Furthermore, an NTO corrects these modern communication delay problems by dynamically rearranging network resources to suit different network traffic conditions, this rearrangement is universally known across all transmission nodes. This chapter investigates the NTO in four parts. Section 2 contains an explanation of the NTO design, Sect. 3 describes the Ethernet simulation by employing an NTO per server application in a star network created from a SWIM environment, Sect. 4 contains the deterministic response time server application results and the discussion of creating predictable traffic loads in the system and findings; Finally conclusions are presented in Sect. 5 where we summarise the benefits of the NTO approach for real-time critical networking, and the achievement of predictable packet arrival rate without dedicated circuits in the presence of random background traffic. This is a significant step forward in the delivery of a deterministic service using modern switching technology and the Ethernet protocol.

2 Network Traffic Oscillator

The operation of the NTO is to alternate frame transmission in two forms, the transformation between the number of frames or *packets per second* P(t) and the frame size or *payload per packet* R(t). Packets are traditionally referred to as a network layer unit and frames for the data-link layer, but here this arbitrary division is overcome to connect the two layers and create a critical network suitable for continuous real-time communication. The second form, R(t), measures the frame size as a larger frame size requires greater resultant network loading, which translates to larger buffers and longer queues per node. Similarly, when many frames occur in one time instance, P(t) increases network workload. Managing the frames by oscillating between the number of frames and the frame size, can dramatically improve the network performance by actively matching and swapping network resources to the appropriate application transmission. A network that is overburdened with large frames will continue to exceed each frame service time unless these are broken down into smaller units. In parallel, congestion from multiple frames can be reduced by combining consecutive frames (that have the same destination) into one long frame. The two forms should be interchangeable depending on the current network condition in a busy network. Frame transmissions are divided by their application protocol, destination and its critical time window. This level of division removes any hidden obstacles within the network, and re-orders network traffic to suit the network availability. Thus the NTO increases frame transmission punctuality by dynamically allocating transmission space based on the network arrangement.

The NTO has two model components, a buffer and a flow controller. At the input to the device, the payload per packet is R_{IN} and at its output, this becomes R_{OUT}. The difference arises from the action of the flow controller (R_F) in proportion to the rate of change of output packets per second.

$$R_{\mathrm{OUT}} = R_{\mathrm{IN}} - R_{\mathrm{F}} \tag{40.1}$$

Without the flow controller, we have a buffer that accepts an input stream of packets per second and divides it into packets of size B

$$R_{\mathrm{IN}} = R_{\mathrm{B}} = \frac{1}{B}\int P_{\mathrm{IN}}(t)dt. \tag{40.2}$$

Using the NTO:

$$R_{\mathrm{OUT}} = \frac{1}{B}\int P_{\mathrm{OUT}}(t)dt. \tag{40.3}$$

Now, for the flow controller:

$$R_{\mathrm{F}} = F\frac{\mathrm{d}P_{\mathrm{OUT}}(t)}{\mathrm{dt}}. \tag{40.4}$$

So from (40.1), (40.2), (40.3), and (40.4) we can obtain the Laplace domain transfer function of the NTO by recognising the resonant frequency $\omega_n^2 = (BF)^{-1}$:

$$H_{\mathrm{NTO}}(s) = \frac{\tilde{P}_{\mathrm{OUT}}(s)}{\tilde{P}_{\mathrm{IN}}(s)} = \frac{\omega_n^2}{s^2 + \omega_n^2} \tag{40.5}$$

The resonant frequency is low compared to the rate of packet arrival at the device and so the response observed will be of that to a step input of size $\overline{P}$, the mean arrival rate in packets per second, thus:

$$P_{\mathrm{OUT}} = \overline{P}\left\{1 - \cos\left(\omega_n t\right)\right\}. \tag{40.6}$$

The buffer collects a multitude of frames in a period of instances into a longer payload per packet R. The flow controller divides a long payload per packet into multiple packets per second depending on the network workload. The directly opposing nature of the functions of these two components creates traffic oscillation patterns within packet transmission. The buffer factor B is the collection of frames and measures the level of traffic load in the network, while the flow controller increases or decreases the frame flow rate F based on the traffic load. These varying transmission patterns create a critical time response window between workload (frame size) and channel division management (frames per second). Oscillating traffic is managed within network switches and physical packet queuing buffers to deliver deterministic arrival rate transmission.

3 Methodology

A simulation was conducted using the setup seen in Fig. 40.1. The two router switches and multiple radar servers are connected in a star topology network. Each server transmitted and processed data for their respective time critical communications and applications. This simulation had seven servers; five radar servers, main SWIM pool server and a backup server. All the servers in this simulation used NTOs for their time critical transmission with their own oscillation frequencies fixed on a per application specification to produce a traffic load for link capacity planning. Radar information, air traffic control and a general server were being fed into a NTO, while other commercial communications (non-time critical applications) were also transmitting into the same network on an ad-hoc basis (non NTO). The

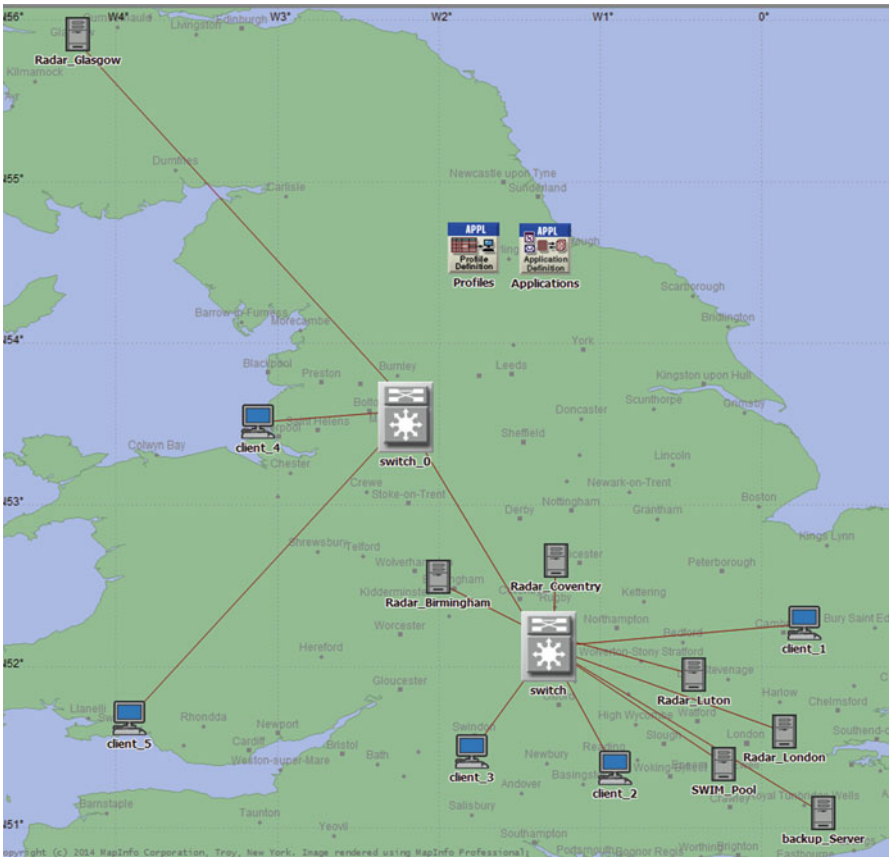


Fig. 40.1 Real-time critical Ethernet network simulation designed to mimic the air traffic SWIM infrastructure across multiple airports

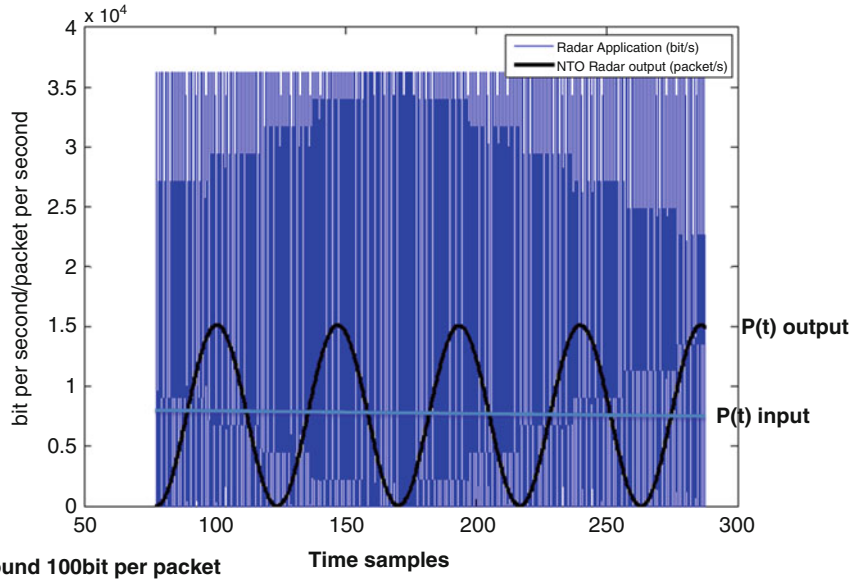


Fig. 40.2 NTO oscillation, dividing between frame sizes and number of frames per second transmission

measurements were conducted from the server client perspective. These non-time critical communications were added in later by client terminals (PCs).

3.1 Air Traffic Radar Information Transmission

The radar information in the network maintained consistent payloads per second, D(t). This payload per second is then interchanged by oscillating R(t) and P(t) in the NTO. Thus, these two could vary without changing the application network load, so long as their product remained constant. The buffer controlled the level of packets in the device by adjusting the level of the packet buffering input rate (B) which is a ratio between payload and packets. The flow controller exercised network congestion control by adjusting the packet flow rate (F), the ratio of packet per payload. The total radar transmission system always adhered to the total payload transmission rate, D(t), designated in the transmission specification (Fig. 40.2).

3.2 Oscillation Controller

The oscillation frequency was designed to meet the application specification required for transmission update. In general, application transmissions that have a low payload transmission rate, D(t), and low response time window should

preferably oscillate faster since higher oscillation cycles produce response updates in the system more readily. The system factors of buffering rate (B) and flow rate (F) could be adjusted to cater for these specifications in the simulation.

3.3 Network Payload Distance Metric Calculation

An NTO based network requires strict network resource management. Each network link must update its resource usage report before assigning transmission. Dynamic network link capacity updating is crucial to maintain a high Quality of Service amongst other real time critical applications. This strategy monitors the traffic resource usage rather than relying upon another arbitrary network management protocol to keep records (which is inaccurate) or using a feedback loop (requiring trial and error), which are both are costly in this network. Rather than using overheads (packet network information) to direct network traffic, flow can be directed using application frequency analysis to predict and prepare for network resource management. Frame overhead not only increases buffering delay by encumbering each frame with a larger payload, but it also restricts the level of freedom for network switching to manage traffic.

The NTO guarantees oscillatory periodic transmission rates of both frames P(t) and frame sizes R(t). The two quantities P(t) and R(t) are ninety degrees out of phase using the NTO because of their sinusoidal nature and this forms the relationship between them. The NTO parameters combine to deliver a deterministic payloads per second D(t) = R(t)P(t) as a sine-wave. Thus a switching arrangement is achieved to remove network delay completely by payload matching at each checkpoint per node. The transmission rate, D(t), is maintained consistently in a network to ensure the lowest minimum distance between the source and the destination.

Non-NTO traffic can also undertake critical network switching using this concept of dividing payloads per second into the two forms of R(t) and P(t). A large payload per packet should also use less of the packets per second resource to maintain the same payloads per second footprint. These two parameters, R(t) and P(t), create the payload distance of the link. To illustrate the fundamental concept, we consider demonstrating critical network switching of the three nodes, N_A and N_B via an intermediate node N_C as shown in Fig. 40.3. N_A has the (packet, payload) coordinates (P_A, R_A) and N_B the coordinates (P_B, R_B). Node N_A received payload divided as a set of payloads per packet (Y-axis) P_A and packets per second (X-axis) R_A which is fixed. We can thus define a payload distance for link AC by the square root of the sum of the squares of P_A and R_A. This may be converted to a time T_{AC} by dividing by the payloads per second value for the link N_A-N_C, which we denote by d_1:

$$T_{AC} = \frac{\sqrt{P_A^2 + R_A^2}}{d_1} \tag{40.7}$$

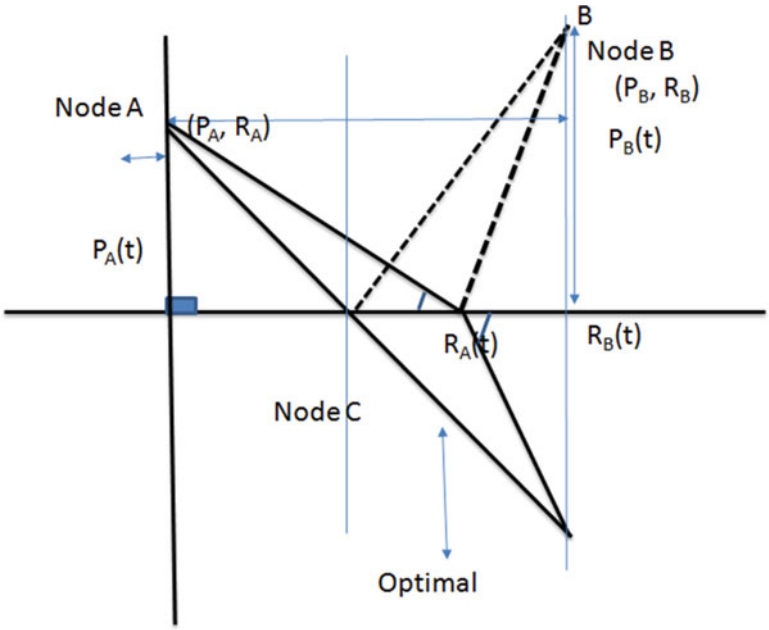


Fig. 40.3 The communication payload distance is the hypotenuse of the triangle formed by P(t) and R(t); critical networking minimises the transmission time across the two links

Considering the transmission from N_C-N_B, a similar argument may be made to give a time over the link BC, with payloads per second d_2, of:

$$T_{CB} = \frac{\sqrt{P_B^2 + (R_B - R_A)^2}}{d_2} \tag{40.8}$$

The two source rates P_A and P_B are also fixed in N_B because of the expected payload rate. Therefore, the R_A can be optimised by finding the minimum time from AC and CB by differentiating the total time $T_{AC} + T_{CB}$ with respect to R_A.

Critical Networking encourages *application* payloads per second to be deterministic and periodic such as the one produced by a NTO. NTO allows advanced *link capacity* budget projection. This method allows the network switching process the freedom to delegate other link resources for other real-time critical applications and transmit non-critical application when the network becomes available.

4 Results and Discussion

Prior to this research, the Ethernet system used discrete frame transmission, which refers to small payloads over a shorter period of time. Small payloads create the illusion that the frame exchange is faster in a network operating with a

high bandwidth, however this ignores the fact there are possible queuing and congestion problems in a busy network. The NTO alters this perspective by creating a continuous transmission by organising large payload transmission over a long period of time. The NTO enables the time-critical accommodation of continuous transmission applications such as live radar feeds needing a specified transmission window time, which were treated as a series of discrete communications in the past proving ineffective. The NTO allows continuous real-time transmission to be more flexible (interchanging between P(t) and R(t) transmission) for link capacity management, while operating within the time response design of the application (the inverse of the payloads per second). NTO continuous real-time transmissions fix the link capacity usage of the network. As a real-time application transmission is oscillating, other real-time applications can use other available link capacity resources. Other NTO continuous application transmissions are encouraged to transmit data out of phase with all existing continuous transmissions; this reduces the knock-on effects of increasing payload distance from payloads per packet, R(t) and packets per second, (P(t)).

In Fig. 40.4, the simulation results show that continuous real-time application transmissions can co-exist with regular non-time critical commercial applications such as e-mail, database access and server access created by user terminals. These do not require NTOs as they can use frame overhead to direct their transmissions, even though it is ineffective in a network. Link resources are allocated to NTO transmissions first and thus the response time of these is deterministic as the link always allocates the same resources to these communications in projection. The clear result is that servers deliver deterministic service to the real-time applications despite the presence of variable traffic loads in the network, including random data bursts from the background applications. This demonstrates that Ethernet has the capability to deliver the required service to the radar traffic in an airport scenario without dedicated links.

Although critical networking could in principle be achieved without an NTO, since it concerns maintaining the minimum payload distances in a link, the level of optimisation required would need advanced knowledge of the behaviour of each real-time continuous transmission. This is impossible when any application can use the link by adding in the correct frame overheads.

Often, sudden discrete communications that are non-time critical unintentionally offset the payload distance (P(t) and R(t)) of time critical communications. Priority queuing [8] is difficult to achieve given the unknown frame arrival rate of each real-time critical communication session. Links that have a large bandwidth (in bps) can transmit payloads faster, however increasing the application payload, D(t), per second due to this extra link capacity also creates application payload distortion effects. This is not noticeable when the application transmits discretely and not measured directly (not treated by NTO), but is noticeable when the continuous transmission is managed by the NTO as it is monitored consistently from node to node. The level of distortion is directly proportionate to the different payloads per second between the two link bandwidth technologies. This distortion can be repaired by adding phase shifts to the NTO frame transmission switching from payload size

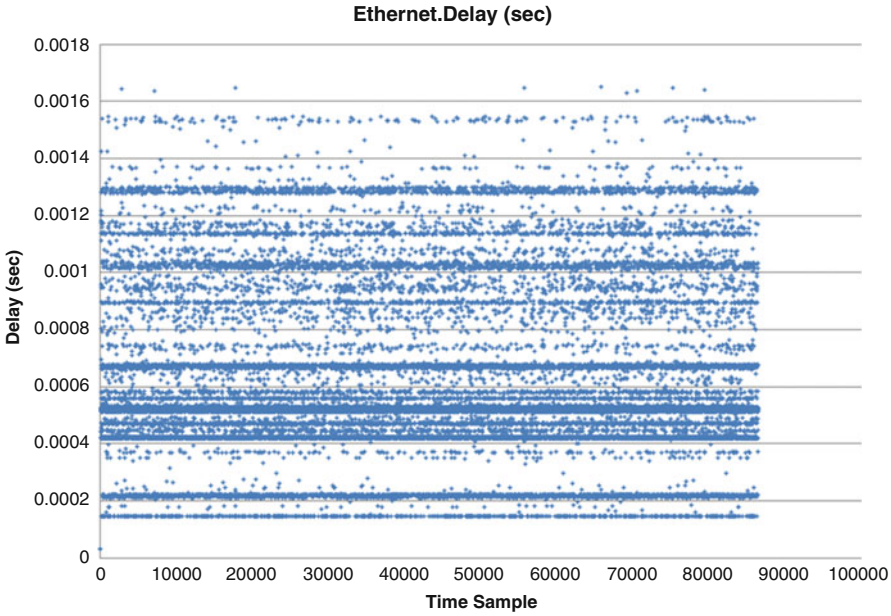


Fig. 40.4 The servers show deterministic response times for time critical communication using NTO (*solid line*) non-time critical transmissions are handled using the critical networking switching method using ad-hoc response time utilisation (*dotted lines*)

to frames per second. Link capacity management can be easily simplified by just maintaining the same payloads per second regardless of the underlying switching bandwidth.

5 Conclusion and Future Work

As the size of a network increases, resource planning is increasingly difficult when the network relies solely on overheads to direct traffic. Worse still, an overhead also increases larger payloads per frame, which overloads the links remaining capacity. Larger networks naturally need bigger frame overheads i.e. larger addresses, greater frame padding, more information for additional multilayer network services [3] and support protocols for frame diagnostics. Thus, overheads and unnecessary supporting protocols increase the pressure on bandwidth and lower Quality of Service. The current network paradigm has additional failure conditions in frame transmissions via overhead errors and misinterpretation of support protocol operation. Frame overhead is useful given large dedicated bandwidths and link capacities over a very dynamic application network but offers diminishing returns when the bandwidth and link capacity are low. Although each frame can be identified by using a frame capturing tool, this level of frame quality assurance only measures the quality of the information presented; it has nothing to say concerning the reason for the

delay of the frame. This network weakness has given rise to many sophisticated technical middle-man cyber-attacks. It is infeasible to operate real-time critical communications in an environment where they may be delayed by discrete, non-time critical commercial messages, especially when those were generated to attack this network. Important payload should be given priority especially when it is also time-critical. Here, we have illustrated the utility of the NTO concept to address this problem and offer real-time deterministic performance in an Ethernet network also carrying other non-real time applications. The use of payload distance to quantify the performance is facilitated by the NTO, which delivers a controlled traffic stream into the network. The real-time traffic is shielded from the effects of users sending large non-time critical payloads by the NTO. By managing the transmission rate at the input to the network, flow management is also simplified since the uncertainty in frame arrival times is removed and reducing any congestion. An uncertain arrival rate transmission discourages ay effective advanced route planning. Network resource wastage such as low data utilisation occurs because low payload frames are kept in a buffer even when the link capacity is perfectly able to handle them – this problem is removed using the NTO. In short, we have shown that Ethernet can deliver deterministic service to critical real-time applications without dedicated links and in the presence of random traffic from other applications.

Acknowledgment This work is funded by a Collaborative Awards in Science and Engineering (CASE) studentship from the Engineering and Physical Sciences Research Council (EPSRC) and FTI Communication Systems Ltd.

References

1. Talledo, J.P.: Design and implementation of an ethernet frame analyzer for high speed networks. In: Electronics, Communications and Computers. Proceedings. 15th International Conference on, 2005, pp. 171–176. Pueblo, Mexico, 28 Feb – 2 Mar 2005
2. Carvajal, G., Chun Wah, W., Fischmeister, S.: Evaluation of communication architectures for switched real-time ethernet. Comput. IEEE Trans. **63**, 218–229 (2014)
3. Day, J.D., Zimmermann, H.: The OSI reference model. Proc. IEEE **71**, 1334–1340 (1983)
4. Obermaisser, R.: Reuse of CAN-based legacy applications in time-triggered architectures. Ind. Inform. IEEE Trans. **2**, 255–268 (2006)
5. Mifdaoui, A., Frances, F., Fraboul, C.: Performance analysis of a master/slave switched ethernet for military embedded applications. Ind. Inform. IEEE Trans. **6**, 534–547 (2010)
6. Di Crescenzo, D., Strano, A., Trausmuth, G.: System wide information management: the SWIM-SUIT prototype. In: Integrated Communications Navigation and Surveillance Conference 2010, pp. C2-1–C2-13
7. Alese, B.K., et al.: Game-based analysis of the network attack-defense interaction. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 481–487. WCE, London, 2–4 July 2014
8. Tobagi, F.A.: Carrier sense multiple access with message-based priority functions. IEEE Trans. Commun. **30**, 185–200 (1982)
9. Mo, Y.K., Leeson, M.S., Green, R.J.: Effective frame switching using a network traffic oscillator. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 461–465. WCE, London, 2–4 July 2014

Chapter 41
Cryptographic Adversary Model: Timing and Power Attacks

Mohd Anuar Mat Isa, Habibah Hashim, Amir Hamzah Abd Ghafar, Jamalul-lail Ab Manan, Syed Farid Syed Adnan, and Ramlan Mahmod

Abstract In this work, we present an adversary model that incorporates side channel attacks in the Indistinguishability Experiment for Adaptive Chosen Ciphertext Attack (CCA2). We propose security assumptions and an attack model for a secure SSW-ARQ protocol with an integration of TFTP protocol. We also present the security reduction of SSW-ARQ protocol from Cramer-Shoup encryption scheme, timing and power attacks as side channel security for the SSW-ARQ protocol. We suggest using a lightweight symmetric encryption for data encryption and asymmetric encryption for key exchange protocols in the TFTP. The target implementation of secure TFTP is for embedded devices such as Wi-Fi Access Points (AP) and remote Base Stations (BS). In this paper we present the security proofs based on an attack model (IND-CCA2) for securing TFTP protocol. We have also introduce a novel adversary model in IND-CCA2-(TA, PA, TPA) and it is considered a practical model because the model incorporates the *timing attack* and *power attack.*

Keywords Adversary model • Adaptive chosen ciphertext attack • Cryptography • Embedded Raspberry Pi • Indistinguishability • Power analysis attack • Provable security • Random oracle model • Trivial file transfer protocol • Timing attack

1 Introduction

This chapter describes an extension of our previous work [1] that is related to a security for TFTP protocol. Formal research works with regard to improvements

M.A. Mat Isa (✉) • H. Hashim • S.F. Syed Adnan
Universiti Teknologi MARA, Shah Alam, Selangor 40450, Malaysia
e-mail: anuarls@hotmail.com; habib350@salam.uitm.edu.my; syed_farid@salam.uitm.edu.my

A.H. Abd Ghafar • R. Mahmod
Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia
e-mail: amirghafar87@gmail.com; ramlan@upm.edu.my

J.A. Manan
MIMOS Berhad, Technology Park Malaysia, Kuala Lumpur 57000, Malaysia
e-mail: jamalul.lail@mimos.my

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_41

in the TFTP protocol had been quiet for almost 11 years since the most recent publication on RFC 3617 (2003) [2]. The RFC 3617 mentioned that "there is no mechanism for access control within the protocol, and there is no protection from a man in the middle attack". Our publication in 2013 [3] proposed an implementation of a lightweight and secure TFTP protocol for embedded systems. We proposed a new packet header for RRQ, WRQ and OACK. These headers provide security information for TFTP's data payload encryption. Our most recent publication in 2014 [1] expended the security discussion on TFTP [3] with an additional security proof through reduction of Cramer-Shoup [4] encryption scheme security to TFTP security. This effort expands the work in [1] with regards to an Adversary Model for timing and power attacks.

Within this chapter we have purposely written using general information security terminology with simple mathematical notation (semi-formal). Our intention is to assist more for information security practitioners as the main audience rather than for mathematicians or cryptographers. In doing so we hope that it will give a commendable understanding of cryptographic scheme and its security arguments. We also note that it is quite difficult for a non-mathematical person to grasp the *reductionist style*. Therefore, we will take a simplistic approach and we purposely skip the math intensive parts in the *Adversary Model* and *Security Analysis* discussion sections which otherwise can be obtained from references [4–6]. We hope that, using this approach, the reader can easily understand the security assumptions and arguments for the secure TFTP.

2 Motivation

The purpose of this research work is to offer security in TFTP protocol in terms of Adversary Model and security reduction. We decided to use Cramer-Shoup [4] encryption scheme and side channel security as underlying security protocol for a new secure TFTP. Referring to our previous work [1, 3], we mentioned the need of a secure TFTP protocol, particularly in various network administrative tasks, such as for monitoring and upgrading of remote embedded device's firmware, where a lightweight protocol such as TFTP is usually employed. The security risks associated with such situations have also been discussed with emphasis on concerns due to physical attacks, wherein attackers illegally access and modify Wi-Fi AP hardware and software [3, 7, 8]. In our preceding work [3], we proposed an enhanced data communication package for DENX-UBOOT [9] firmware which also included a secure TFTP protocol. However, our proposal did not suggest a specific cryptographic protocol for the successful implementation of the secure TFTP protocol. In our effort to further augment the work, a proven secure and practical asymmetric cryptographic scheme, i.e. the Cramer-Shoup (CS) protocol was proposed to be deployed as the underlying cryptographic protocol [4] in the overall scheme. In the latter part, the CS provides a secure asymmetric key exchange, wherein CS is used to encrypt symmetric key (e.g., AES 512) for a secure TFTP data communication.

3 Related Work

3.1 Trivial File Transfer Protocol (TFTP)

TFTP is a simple protocol that has been widely used for transmitting files albeit with limited functionalities [10]. It provides upload and download operations using UDP protocol. The actual transmission protocol that is used to control file transfer is *"Simplex Stop and Wait with Automatic Repeat reQuest"* (SSW-ARQ). TFTP was designed as an application for the Internet Protocol (IP) [11] because at that moment, computers or embedded systems did not have sufficient memory or lack enough disk space to provide full FTP support. Nowadays, TFTP become quite popular and it is used by network administrators to upgrade router firmware and to distribute software within a corporate network (e.g., DENXU-Boot [9] firmware). Thus, it is beneficial for booting embedded devices (e.g., sensor nodes) that may not have sufficient volatile memory to store OS kernel and applications.

Recently, there were some research works that addressed new environments such as usage of TFTP protocol for Radio Frequency (RF) [12], remote attestation for Trusted Computing [11] (e.g., Trusted Platform Modules (TPM)), lightweight protocol for remote accessing the cloud infrastructure [13, 14], Wide Area Network (WAN) surveillance system [8], secure database [15] and etc. However, their suggestions to use TFTP as the solution in their research works were impractical and insecure because TFTP exposes all data packet in plaintext. The authors have assumed that TFTP can provide secure communication (confidentiality, integrity and authenticity) for data transfer which is not necessarily true.

3.2 Simplex Stop and Wait Automatic Repeat Request (SSW-ARQ)

SSW-ARQ is a simple network protocol used by network applications (e.g., TFTP) to enable stop and wait flow control in frame transmission when using unreliable UDP/IP stacks [10, 16]. It allows retransmission of frames in the event of frame loss or corrupted frame [17, 10]. Figure 41.1 shows an example of frame transmission using SSW-ARQ. We believe that the best way to have better security in this protocol, is to integrate it with Cramer-Shoup [4] encryption scheme in the frame data payload. We explain the security enhancement mechanism as follows.

From Fig. 41.1, ***A*** wants to transmit data or file to ***B*** in a secure manner. Therefore, both parties need to establish a secure key exchange for symmetric encryption (e.g., share AES512's secret keys). This entails the AES512's secret keys to be shared in a secure communication protocol which is best accomplished using Cramer-Shoup [4] encryption scheme. During communication setup, both parties are pre-installed with Cramer-Shoup's asymmetric keys by the network administrator. Thereafter, it is assumed that both communication parties communicate with each other in full knowledge of the recipient's public key.

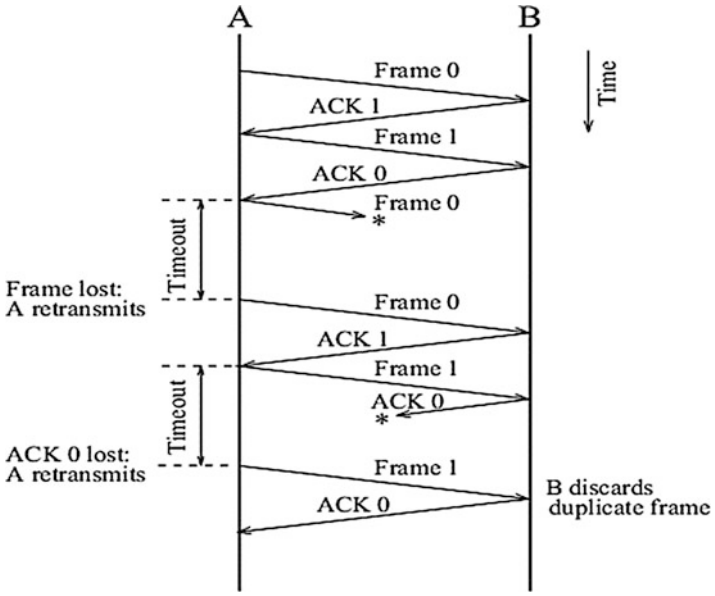


Fig. 41.1 SSW-ARQ protocol [18]

Communication commences with ***B*** generating the AES512's secret keys. Then, the AES512's secret keys is wrapped (encrypted) using ***B***'s public key. Due to limitation of SSW-ARQ's frame size, a ciphertext generated using ***B***'s public key must be divided into chunks that fit into the frame. Next, ***A*** transmits multiple frame segments containing B's public key chunks of ciphertext. However, the SSW-ARQ communication protocol allows only one frame to be sent at a time. The next frame will only be transmitted after receiving a correct acknowledgement (ACK) from ***B***. At this stage, all transmitted frame must be verified as free from data corruption (e.g., bit-error) using the checksum function. After all frames have been successfully transmitted, ***B*** assembles all frame segments into the complete ciphertext string. Next, ***B*** calls Cramer-Shoup [4] decryption function to decrypts the ciphertext and then retrieves the AES512's secret keys. Finally, ***A*** encrypts the file using the AES512's secret keys and sends the encrypted file using standard TFTP protocol. B decrypts the file using the AES512's secret keys. However, due to limited space in this chapter, we will not discuss the usage of symmetric encryption scheme and its security.

3.3 Cramer-Shoup Encryption Scheme

Cramer-Shoup [4] protocol has been proven to be secure against IND-CCA2. The protocol provides an improvement of El-Gamal [19] wherein the El-Gamal is vulnerable to chosen-ciphertext attack (CCA). However, the Cramer-Shoup

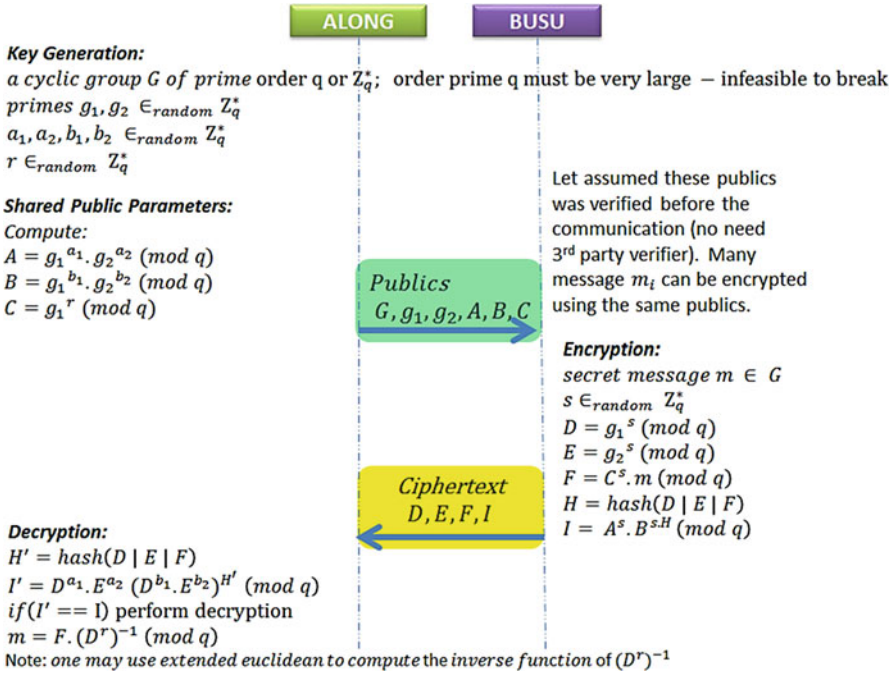


Fig. 41.2 A simplified Cramer-Shoup encryption scheme

is slower than the El-Gamal (approximately twice) in performing cryptographic computation [20]. As compared to RSA, Cramer-Shoup is slower in the encryption process but it is nearly equal in decryption process [20]. We illustrate the Cramer-Shoup protocol in Fig. 41.2.

4 Adversary Model

4.1 IND-CCA2

Rackoff-Simon (1991) [21] argued that an adversary in CCA1 may get access to a decryption oracle even after the challenge's ciphertext c^* was issued. This attack is a practical security problem because it happens in real-world. A security property for this kind of attack is that it can prevent the adversary from getting *any useful information from other ciphertext* c_i that can help to get a non-negligible *advantage* to distinguish the challenge's ciphertext c^* in a polynomial time. The authors in [21] stressed that it is important to protect our system against this attack because a digital signature scheme (a practical scheme in our real-world) is vulnerable to this attack. The digital signature scheme is secure "*if any such attacker succeeds*

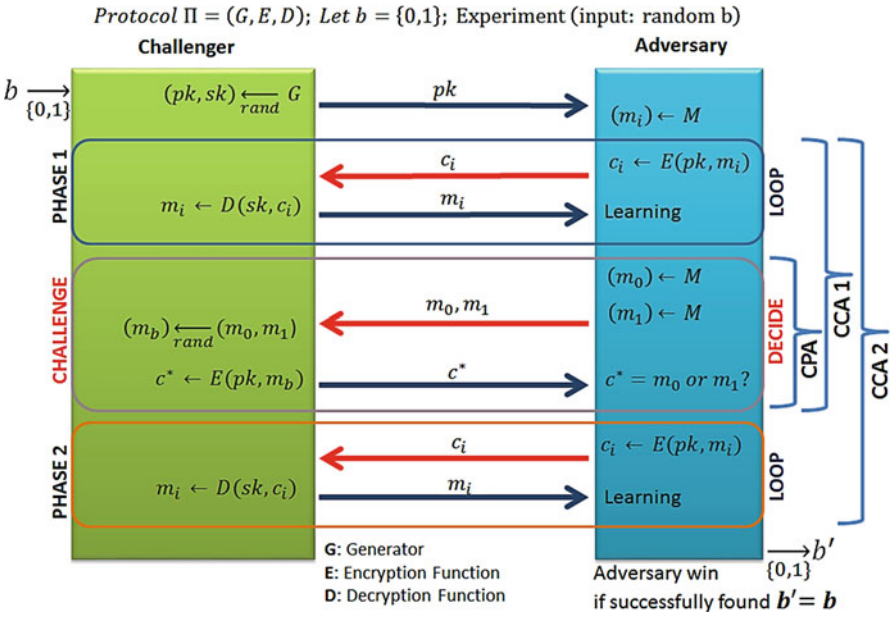


Fig. 41.3 CPA, CCA1, and CCA2 experiments

in generating a valid signature for this last document with negligible probability" [21].

Indistinguishability-Adaptive Chosen-Ciphertext Attack [21] is an attack that allows an adversary to access a decryption function through the decryption oracle. The adversary can ask the oracle to decrypt any ciphertext except the one that is being used for indistinguishability test or experiment. The IND-CCA2 allows the adversary to get a decryption of ciphertext from the oracle in Phase 1 (before) and Phase 2 (after) the challenge messages $(m_0,\ m_1\ where\ |m_0| = |m_1|)$ are issued to Challenger (refer to Fig. 41.3). For the indistinguishability test, the adversary sends two plaintext messages $(m_0,\ m_1)$ to the challenger. In place of a fair indistinguishability experiment, both plaintext messages must never be used for decryption using the oracle. This means that the adversary could never know the ciphertext of both messages after the encryption function has been applied. Referring to Fig. 41.3, the challenger chooses randomly either m_0 *or* m_1 to be encrypted using encryption scheme $\Pi = (\mathcal{K},\ \mathcal{E},\ \mathcal{D})$. Ciphertext c^* of the encrypted message either m_0 *or* m_1 is sent to the adversary. The adversary needs to distinguish the ciphertext c^* which is either m_0 *or* m_1 with probability of $\frac{1}{2}$ If the probability of guessing the correct ciphertext c^* is greater than $\frac{1}{2}$ we can conclude that the adversary has an "advantage" and the given protocol is considered not secure in terms of indistinguishabili.

Let $p(n)$ denote the set of prime in size of n, for all sufficiently large n in IND − CCA2

$$|Pr[success] - Pr[failure]| < \frac{1}{p(n)}$$

4.2 IND-CCA2-(TA)

Indistinguishability-Adaptive Chosen-Ciphertext Attack – (also known as Timing Attack) is an attack that allows an adversary to access identical computing resources in terms of computing speeds (e.g., CPU). The adversary is given knowledge of time to perform cryptographic computations (e.g., primitive computation and protocol execution). Sometimes the adversary is given knowledge of the delay of network transmission for all transactions in Phase 1, Phase 2 and Challenge phase (refer to Fig. 41.4). In both messages m_0, m_1 will be computed for encryption by a challenger. Then, the challenger will return the computation time for both messages as t_{m0}, t_{m1} the adversary with a random ciphertext c^* Either the encryption time of t_{m0} *or* t_{m1} is the actual time for the ciphertext c^* was computed. Given the knowledge of t_{m0}, t_{m1}

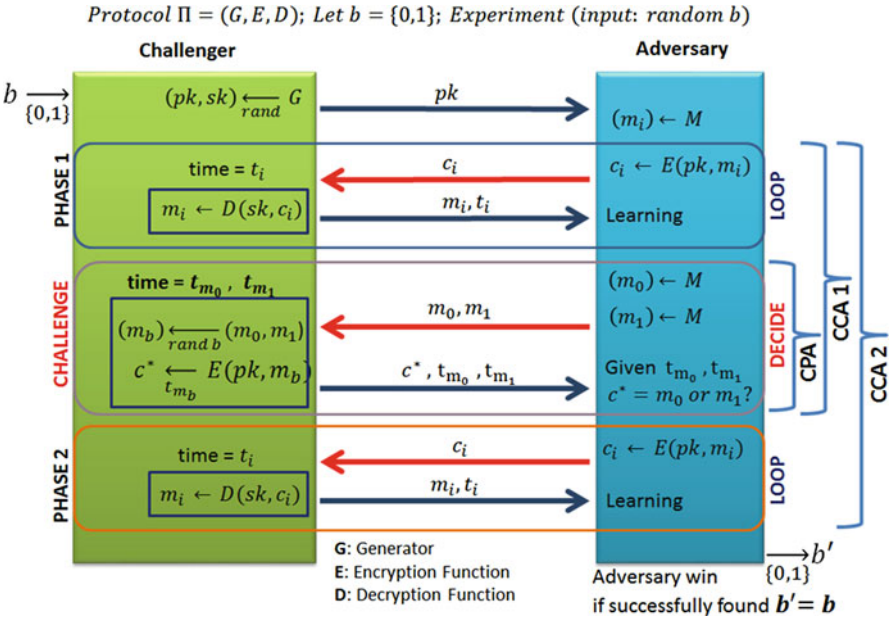


Fig. 41.4 IND-CCA2-(TA) experiment

can it assist the adversary to gain a non-negligible *advantage* to distinguish the challenge's ciphertext c^* in a polynomial time?

Adversary Model: IND (CPA, CCA1, CCA2) with Timing Attack
Adversary Knowledge: E, D, c_i, m_i, t_i, t_{m_0}, t_{m_1}, c^*
Adversary Limitation: Outsider Attack, Protocol Π is protected/sealed using temper resistance device (but adversary may have managed to capture all network frames/packets (wired or wireless) with precision of time in/out to the sealed box of Protocol Π during runtime or idle)
Adversary Goal: successfully distinguish challenge ciphertext c^* with higher ε probability

4.3 IND-CCA2-(PA)

Indistinguishability-Adaptive Chosen-Ciphertext Attack-(Power Attack) is an attack that allows an adversary to access identical computing resources in terms of computing power (e.g., electrical current–ampere and voltage). The adversary is given knowledge of electrical current to perform cryptographic computations (e.g., primitive computation and protocol execution). Adversary is also be given knowledge of the total amount of current usage for all models in Phase 1, Phase 2 and Challenge phase. Based on Fig. 41.4, Power Attack was conducted using similar experiment as shown in the Fig. 41.4 except that we now replace it with the notation where $power = p_{m_0}, p_{m_1}$ and $\mathrm{c} \underset{p_{\mathrm{m_b}}}{\leftarrow} \mathrm{E}(\mathrm{pk}, \mathrm{m_b})$ in the Challenger section. Given the knowledge of p_{m0}, p_{m1} can it help the adversary to gain a non-negligible *advantage* to distinguish the challenge's ciphertext c^* in a polynomial time?

Adversary Model: IND (CPA, CCA1, CCA2) with Power Attack
Adversary Knowledge: E, D, c_i, m_i, p_i, p_{m_0}, p_{m_1}, c^*.
Adversary Limitation: Outsider Attack, Protocol Π is protected/sealed using temper resistance device (but adversary manage to clamp the power cord to capture current and voltage during runtime Protocol Π or idle).
Adversary Goal: successful in distinguishing challenge ciphertext c^* with higher ε probability

4.4 IND-CCA2-(TPA)

Indistinguishability-Adaptive Chosen-Ciphertext Attack- (also known as Timing and Power Attacks) is a combination of IND-CCA2-(TA) and IND-CCA2-(PA) with **Adversary Knowledge:** E, D, c_i, m_i, t_i, t_{m_0}, t_{m_1}, p_i, p_{m_0}, p_{m_1}, c^*. Given the knowledge of power p_{m0}, p_{m1} and timing t_{m_0}, t_{m_1} can it help the adversary to gain a non-negligible *advantage* to distinguish the challenge's ciphertext c^* in a polynomial time?

5 Security Analysis

5.1 Cramer-Shoup with IND-CCA2

Adversary Model: IND-CCA2
Security Assumption:

1. Decision Diffie-Hellman Problem (DDHP) problem is hard [5] in a cyclic group G;
2. Hash function is a universal one-way hash function with strong collision-resistant [4, 22];
   If assumptions 1 & 2 are true; then the Cramer-Shoup encryption scheme is secure against CCA2 using indistinguishability test.

Reductionist Security Claim: Anyone who can read message m from a ciphertext c^* must also be able to solve Decisional Diffe-Hellman (DDH) problem and also be able to reverse the on-way hash function in terms of collusion resistance property.

Security Reduction Experiment: An adversary claims that he can break Cramer-Shoup protocol using an efficient algorithm A in a program A To test the adversary claim, we conduct an experiment by taking the program A and put a simple "wrapper" into it, and we call it program $A^{'}$ The program $A^{'}$ will use the program A is a sub-routine in the experiment. Then, the program $A^{'}$ will run the IND-CCA2 experiment with random input b and with expected output $b^{'}$ in indistinguishability test. The adversary is considered a winner in the experiment, if the probabilities to guess for all correct messages are non-negligible with an advantage of $\left(\frac{1}{2}\right) + \varepsilon(n)$ where $\varepsilon(n)$ is the adversary's success probability. Due to the non-negligible advantage, the program $A^{'}$ can break the Cramer-Shoup protocol. However, if there are no other efficient programs (including program $A^{'}$ that can win in the experiment with non-negligible advantage, the Cramer-Shoup protocol wins the experiment with negligible advantage of program $A^{'}$ Since the **Security Assumptions 1** and **2** in the previous paragraph use an acceptable primitive assumption (DDHP is hard and collision-resistance of hash function), the program $A^{'}$ advantage over probabilistic polynomial-time[1] is negligible. Therefore, the program $A^{'}$ loses in the experiment by indistinguishability test with a negligible advantage and the adversary claim is invalid (false) in that it "can break Cramer-Shoup protocol using all efficient algorithm A in a program A".

[1] "polynomial-time" is a term used for measuring an algorithm's running time as a function, wherein it is measured by length of its input into the function [6]. E.g. function $f(x)$ take $x = 1024$ input string during execution, then the running time is x.

5.2 SSW-ARQ: Cramer-Shoup with IND-CCA2-(TA)

Adversary Model: IND-CCA2-(TA)
Security Assumption:

3. SSW-ARQ inherits all security strength from the Cramer-Shoup encryption scheme and the Cramer-Shoup encryption scheme has been proven secure in the IND-CCA2;
4. SSW-ARQ is secure against *Timing Attack* (refer to Sect. 4.2) if and only if an adversary is not be able to distinguish ciphertext c^* in IND-CCA2-(TA) experiment with negligible advantage. The knowledge of time t_i gives a negligible advantage for all runtime in fixed input length in a polynomial time.

If assumptions 1 & 2 are true; then the Cramer-Shoup encryption scheme is secure against CCA2-(TA) using indistinguishability test.

Security Argument: For the **Security Assumption 3**, it is easy to observe the security proof because Cramer-Shoup encryption scheme has been embedded into SSW-ARQ protocol. All strings (e.g., ciphertext, public key) that are generated by Cramer-Shoup encryption scheme are divided into chunks that are fitted into the SSW-ARQ's frame. Any modification (even a single bit error) in the SSW-ARQ's frame will result in a failure in Message Authentication Codes (MAC) and digital signature in the Cramer-Shoup encryption scheme. This good security property has been derived from the collision-resistant hash function. Therefore, *"Given that **Security Assumption 3** is true, the SSW-ARQ is secure against IND-CCA2"*.

For the **Security Assumption 4:** we can use a similar experiment that has been used for Cramer-Shoup encryption scheme except that an adversary is given knowledge of runtime to perform cryptographic computation and network transmission delay (as mentioned in section adversary model for IND-CCA2-(TA)). If and only if the SSW-ARQ is secure against *Timing Attack* using *fixed-time* of runtime for all fixed input length in the function in a *polynomial time*; in simplified words: for any function that receives any valid input with the same length (e.g., $f(1001)\ and\ f(0001),\ where\ |f(1001)| = |f(0001)|$ will have identical runtime or execution for all conditions; Then, SSW-ARQ protocol is secure against CCA2-(CS-TA) for the *implementation and deployment of the protocol*.

However, it is impossible to attain the same fixed time for the encryption and decryption processes of different input strings of ciphertext (with same length ciphertext and different key) using specific encryption functions or decryption functions. Running time to compute an exponential such as g^x and g^{x+1} in DLP/CDH/DDH is different because of the different computer machine capabilities in performing addition to representing multiplication as well as the different limitations of hardware data bus. It might be similar for small inputs of 32-bits or 64-bits length, but it is not so for crypto numbers with extensive lengths such as 2048-bits length of public key. From a practical point of view, we can use a

subset of the *Security Assumption 4*; **Security Assumption 4.1:** a fixed-time is based on worst-case scenario to do *decryption process in the indistinguishability test* for all string of plaintext that has the same length and within the same cyclic group G of prime order q; or **Security Assumption 4.2:** an implementation of *blinding and masking to the decryption processes* [23] *in the indistinguishability test*. However, this assumption is not strong as well as the *Security Assumption 4.1*.

5.3 SSW-ARQ: Cramer-Shoup with IND-CCA2-(PA)

Adversary Model: IND-CCA2-(PA)
Security Assumption:

5. SSW-ARQ inherits all security strength from the Cramer-Shoup encryption scheme and the Cramer-Shoup encryption scheme has been proven secure in the IND-CCA2;
6. SSW-ARQ is secure against *power analysis attack* (refer to Sect. 4.3) if and only if an adversary is not be able to distinguish ciphertext c^* in IND-CCA2-(PA) experiment with negligible advantage. The knowledge of time p_i gives a negligible advantage for all runtime in fixed input length in a polynomial time.

   If assumptions 1 & 2 are true; then the Cramer-Shoup encryption scheme is secure against CCA2-(PA) using indistinguishability test.

Security Argument: For the **Security Assumption 5:** it is similar to security argument for *Security Assumption 3*. For the **Security Assumption 6:** we can use a similar experiment that has been used for Cramer-Shoup encryption scheme except that an adversary is given a knowledge of electric-current to perform cryptographic computation such as decryption (as mentioned in section adversary model for IND-CCA2-(PA)). If and only if the SSW-ARQ is secure against *Power Analysis Attack* using (1) *fixed-electric-current consumption* for all fixed input length in the function in a *polynomial time*; (2) *no biases in the power consumption during decryption process in the indistinguishability test*. One may considered the *fixed-current consumption over a time* in worst-case scenario, blinding and masking as mentioned in the timing attack section. This assumption can be true if the adversary is acting in a passive mode (he just observe the current flows and he will not try to vary the current or voltage to make some bit errors or flipping).

5.4 SSW-ARQ: Cramer-Shoup with IND-CCA2-(TPA)

We can directly use the CCA2-(TA) and CCA2-(PA) in the previous sections as security assumption for a combination of timing and power attacks. This can

be true if and only if the *timing and power attacks do not correlate to provide additional information to an adversary* to have a non-negligible advantage in the indistinguishability experiment.

6 Discussion

In this chapter we propose to implement security in the TFTP protocol. In previous sections we have discussed the security properties and security assumptions of an adversary Model for timing and power attacks. Past sections have only shown the security of SSW-ARQ protocol against IND-CCA2-(TA, PA, TPA) but not the TFTP protocol wherein the SSW-ARQ protocol is a subset of the TFTP protocol. In our case, TFTP protocol has been designed to just manage file transfer and key management. The TFTP invokes the file transfer using SSW-ARQ protocol and passes a security related key that is needed by SSW-ARQ protocol to perform cryptographic computation (e.g. Cramer-Shoup protocol). Therefore, to prove that the TFTP application is secure, the TFTP must be programmed to follow the standard [24] and practice [25] for a secure application. However, this is beyond the scope of this chapter.

A secure key management protocol in the TFTP application plays an important role in making sure that all cryptographic schemes are secure. Bad implementation of key management will expose the cryptographic scheme through many side-channel attacks such as timing attacks, power monitoring attacks, differential fault analysis, acoustic cryptanalysis and etc. These security vulnerabilities can be exploited in generating, distributing and managing cryptographic keys for embedded devices (e.g., RaspberryPi board) and DENX-UBOOT's TFTP application. Tamper resistant devices or Physical Unclonable Function (PUF) [26] can be integrated into embedded hardware for protecting the cryptographic keys such as TPM chip [27]. To limit our research scope, we have not included the physical security attacks (active adversary) and the side-channel attacks; rather we focused on timing and power attacks in TFTP.

We have introduced a novel adversary model in IND-CCA2- (TA, PA, TPA). This adversary model includes knowledge of time and electric-current to perform cryptographic computation. This has assumed that an adversary become more powerful than the base adversary model in IND-CCA2. For example, if the *timing attack* is mounted into the IND-CCA2, the adversary has a significant non-negligible advantage. The adversary can build a timing dictionary for every request of decryption of ciphertext c_i with time t_i in Phase 1 and Phase 2. The timing dictionary will give a non-negligible advantage to the adversary to choose a correct encrypted message by a given challenge ciphertext c^* in the Challenge process.

However, the timing dictionary for the IND-CCA2-(TA) is unable to choose the correct encrypted message with better probability because of *fixed-time* constraint in the encryption function for the indistinguishability test. We believe that, the IND-CCA2-(TA)'s Adversary Model will provide a sufficient proof to assert that

SSW-ARQ protocol is secure in the indistinguishability test and secure in timing attack. The *fixed-time* using "*worst-case scenario"* is a practical solution to be implemented in the DENX-UBOOT's TFTP application. One may think that using *"worst-case scenario"* slows down the security computation but based on observations in our laboratory, to transmit a file (e.g., Linux Kernel *"wheezy-raspbian"* 2.8 MB size) using DENX-UBOOT's TFTP application; the required Estimated Time of Completion (ETC) is around 15–30 s. Adding an extra 3–7 s to implement the security protocol in the DENX-UBOOT's TFTP application can be considered quite negligible.

7 Conclusion

We have presented the security assumptions as well as attack models for a secure TFTP protocol. We have also presented the security reduction of SSW-ARQ protocol from Cramer-Shoup encryption scheme, timing and power attacks as side channel security for SSW-ARQ protocol. The proposed secure TFTP protocol would overcome security problems (confidentiality, integrity and authenticity) in controlling, monitoring and upgrading embedded infrastructure in a pervasive computing environment. The target implementation of secure TFTP is for embedded devices such as Wi-Fi Access Points (AP), remote Base Stations (BS) and wireless sensor nodes. In the next stage of our research work, we want to conduct in a laboratory experiment for the given Adversary Models for the DENX-UBOOT's firmware with Raspberry Pi's radio frequency (RF) module.

Acknowledgments The authors would like to acknowledge the Ministry of Education (MOE) Malaysia for providing the grant 600-RMI/NRGS 5/3 (5/2013), and Universiti Teknologi MARA (UiTM) for supporting this research work.

References

1. Mohd Anuar Mat Isa, Habibah Hashim, Syed Farid Syed Adnan, Jamalul-lail Ab Manan, Ramlan Mahmod: A secure TFTP protocol with security proofs. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, vol. 1, pp. 443–448. WCE, London, 2–4 July 2014
2. Lear, E.: Uniform Resource Identifier (URI) Scheme and Applicability Statement for the Trivial File Transfer Protocol (TFTP). In: RFC 3617 (2003)
3. Mohd Anuar Mat Isa, Nur Nabila Mohamed, Habibah Hashim, Syed Farid Syed Adnan, R.M., Jamalul-lail Ab Manan: A lightweight and secure TFTP protocol in the embedded system. In: 2012 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE 2012) (2012). Kota Kinabalu Sabah, Malaysia. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6482117&tag=1
4. Cramer, R., Shoup, V.: A practical public key cryptosystem provably secure against adaptive chosen ciphertext attack. In: Lecture Notes in Computer Science: Advances in Cryptology – CRYPTO'98, pp. 1–18. http://link.springer.com/chapter/10.1007/BFb0055717 (1998).

5. Boneh, D.: The decision Diffie-Hellman problem. In: Algorithmic Number Theory. http://link.springer.com/chapter/10.1007/BFb0054851 (1998).
6. Katz, J., Lindell, Y.: Introduction to Modern Cryptography. Chapman and Hall, Boca Raton (2008)
7. Mohamed, N.N., Hashim, H., Yussoff, Y.M., Isa, M.A.M.: Securing TFTP packet: a preliminary study. Control and System Graduate Research Colloquium. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6653295 (2013).
8. Mohd Anuar Mat Isa, Habibah Hashim, Jamalul-lail Ab Manan, Ramlan Mahmod, Mohd Saufy Rohmad, Abdul Hafiz Hamzah, Meor Mohd Azreen Meor Hamzah, Lucyantie Mazalan, Hanunah Othman, Lukman Adnan: Secure system architecture for wide area surveillance using security, trust and privacy (STP) framework. J. Proc. Eng. 41. International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), pp. 480–485 (2012)
9. DENX Software Engineering: DENX U-Boot, 2014. http://www.denx.de/wiki/U-Boot/WebHome
10. Sollins, K.R.: The TFTP Protocol (Revision 2) RFC 1350. In IAB Official Protocol Standards, pp. 1–11 (1992)
11. Schiffman, J., Moyer, T., Jaeger, T., McDaniel, P.: Network-based root of trust for installation. IEEE Secur. Priv. **9**(1), 40–48 (2011)
12. Kao, K., Liao, I., Lyu, J.: An indoor location-based service using access points as signal strength data collectors. Indoor Positioning and Indoor Navigation (IPIN) (2010)
13. Doelitzscher, F., Sulistio, A., Reich, C., Kuijs, H., Wolf, D.: Private cloud for collaboration and e-Learning services: from IaaS to SaaS. Computing **91**, 23–42 (2011). http://link.springer.com/article/10.1007%2Fs00607-010-0106-z
14. Machidon, O., Sandu, F., Zaharia, C., Cotfas, P., Cotfas, D.: Remote SoC/FPGA platform configuration for cloud applications. In: 2014 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), pp. 827–832. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6850986 (2014).
15. Liu, C., Tan, K.: Research and Implementation of the Secure Database-Update Mechanism. Appl. Mech. Mater. **517**, 1752–1755 (2014)
16. Malkin, G., Harkin, A.: TFTP option extension (RFC 2347). In: The Internet Society, pp. 1–7 (1998)
17. Fairhurst, G., Wood, L.: Advice to link designers on link Automatic Repeat reQuest (ARQ). In: RFC 3366, pp. 1–28 (2002)
18. Chen, S.: Simplex stop and wait with ARQ (2007). http://users.ecs.soton.ac.uk/sqc/EL336/CNL-5.pdf
19. Elgamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans. Inf. Theory **31**(4), 469–472 (1985)
20. Shoup, V.: Research Report Why Chosen Ciphertext Security Matters. http://www.shoup.net/papers/expo.pdf (1998).
21. Rackoff, C., Simon, D.R.: Non-interactive zero-knowledge proof of knowledge and chosen ciphertext attack. CRYPTO' 91, vol. LNCS 576, pp. 433–444 (1992)
22. Tsudik, G.: Message authentication with one-way hash functions. ACM SIGCOMM Comput. Commun. Rev. **22**(5), 29–38 (1992)
23. Kocher, P.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Advances in Cryptology – CRYPTO'96 (1996)
24. Common Criteria Members: Common Criteria for Information Technology Security Evaluation Part 1□: Introduction and general model July 2009 Revision 3 Final (2009)
25. Raymond, J., Stiglic, A.: Security issues in the Diffie-Hellman key agreement protocol. In: McGill University Technical Manuscript (2002)
26. Suh, G.E., O'Donnell, C.W., Devadas, S.: AEGIS: a single-chip secure processor. Inf. Secur. Tech. Rep. **10**(2), 63–73 (2005)
27. Mohd Anuar Mat Isa, Azhar Abu Talib, Jamalul-lail Ab Manan, Siti Hamimah Rasidi: Establishing trusted process in trusted computing platform. In: Conference on Engineering and Technology Education, World Engineering Congress (2010)

Chapter 42
Using Elliptic Curve Encryption and Decryption for Securing Audio Messages

Artan Luma, Besnik Selimi, and Lirim Ameti

Abstract The security of message transmission is usually a challenge for its participants. Many available programs that work with audio data claim to enable secured communication, but usually do not show the details of the methods used for data encryption. For end users to be confident, it is essential to be aware of the methods and techniques used for data encryption and decryption. Elliptic curve cryptography, an approach to public key cryptography, is now commonly used in cryptosystems. Hence, in this paper we present a method for using elliptic curve cryptography in order to secure audio data communications. Furthermore, we present a tool that implements this method for encrypting an audio file, transmitting it through the network and decrypting the file at the other end.

Keywords Audio message • Cryptography • Decryption • Elliptic curves • Encryption • Privacy • Secure transmission

1 Introduction

Digital audio transmission is omnipresent nowadays and there is a rising concern about the privacy of communication between parties. A lot of applications in existence today claim they enable a secure audio communication, without divulging the underlying technology, and this makes final users more suspicious about the level of security. For end users to be confident, it is essential to be able to assess the degree of confidentiality of their communication. We present here, a method of encryption/decryption using elliptic curves in order to secure the transmission of voice messages.

Public-key cryptosystems [1] as a concept, implemented by a big number of different algorithms, are widely used in modern cryptography. But, known algorithms such as RSA [2] are not suitable for use when dealing with large amounts of data. When dealing with such amount of data, along the privacy concerns, one

A. Luma (✉) • B. Selimi • L. Ameti
South East European University, Ilindenska no. 335, Tetovo 1200, FYR. Macedonia
e-mail: a.luma@seeu.edu.mk; b.selimi@seeu.edu.mk; la13258@seeu.edu.mk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_42

should also consider the practical performance of algorithms, and thus consider a more efficient cryptographic system.

Compared to the encryption of text messages [3], voice messages contain a large amount of data and therefore the aforementioned algorithms are not efficient enough. We propose a cryptosystem that addresses this issue [4].

The elliptic curve cryptography [5] is more than appropriate for achieving the goal. Compared to RSA algorithm, the efficiency of elliptic curve cryptography is stated as follows: "Safety of elliptical curves is based on elliptic curve discrete logarithm problem (ECDLP) which enables ECC to reach the same level of security with RSA for smaller keys and greater computation efficiency. ECC-160 provides security compared with RSA-1024 and ECC-224 provides security compared with RSA-2048 [6]". This fact is sufficient to build our cryptographic system based on elliptic curves, which is the main purpose of this chapter.

The rest of this chapter is organized as follows: Sect. 2 describes elliptic curve operations, Sect. 3 describes the usage of these mathematical operations to perform encryption and decryption, Sect. 4 describes the particular audio format (.wav) that we use in our system, Sect. 5 explains the implementation of our system. Finally, Sect. 6 concludes this chapter and gives future directions.

2 Elliptic Curve Operations

Elliptic curve operations which are relevant to this chapter are: point generation, point addition, point subtraction, point doubling and point multiplication. For these operations to be faster, more accurate and more efficient, an elliptic curve is defined over two finite fields:

- Prime field F_p, where p is a prime and
- Binary field $F_2{}^m$, where m is a positive integer.

We use the prime field F_p. In order to illustrate the point operations, we consider as a use case the following elliptic curve ($p = 277$):

$$y^2 = x^3 + x + 1 \; mod \; 277, \; or \; E_{277}(1, 1)$$

2.1 Point Generation

Elliptic curve operations are defined over the points of the elliptic curve. Therefore we need to generate those points for a selected curve.

To generate the points of the elliptic curve we need to perform the following steps:

1. Calculate $y^2 \; mod \; 277$ for every value of $x \in [0, 277)$
2. Calculate $a^2 \; mod \; 277$ for each value of $a \in [0, 277)$,

Table 42.1 Point generation

x	y^2	a	$a^2\ mod\ 277$	$a^2\ mod\ 277 == y^2$	$(x,a) \iff (x,y)$
0	1	0	0	F	/
0	1	1	1	T	(0,1)
0	1	2	4	F	/
...	...	...	...	...	
0	1	276	1	T	(0,276)
1	3	0	0	F	/
...	...	...	...	...	...
1	3	130	3	T	(1,130)
1	3	131	264	F	/
...	...	...	...	...	...
1	3	147	3	T	(1,147)
...	...	...	...	...	...
4	69	30	69	T	(4,30)
...	...	...		...	...

3. Each time the condition $a^2\ mod\ 277 = y^2\ mod\ 277$ is satisfied, we register the point (x, y).

In the following table, for illustration purposes, we show the generation of points for the values of $x = 0,\ 1,\ 2,\ 3\ and\ 4$.

Let us take $x = 0$, then:

$$y^2 = 0^3 + 0 + 1\ mod\ 277$$

$$y^2 = 1$$

Now we need to calculate the square of every number between 0 and $p - 1$, i.e. 276, and then we modulate the result with p, i.e. 277. In the end, we compare the result with y^2, in our case with 1. The result of these calculations is better depicted in Table 42.1 [7]:

Our curve is specially selected for illustration purposes so it has 256 points including the point at infinity, and with them we can represent every character of the ASCII table.

Elliptic curve operations like point addition, point subtraction, point doubling and point multiplication in the prime field (F_p) are defined as follows [8]:

2.1.1 Point Addition

Consider two distinct points J and K such that $J = (x_J, y_J)$ and $K = (x_K, y_K)$.

Let $L = J + K$, where $L = (x_L, y_L)$, then s is the slope of the line through J and K. The slope s is calculated as follows:

$$x_L = \left(s^2 - x_J - x_K\right) \ mod \ p$$

$$y_L = \left(s\left(x_J - x_L\right) - y_J\right) \ mod \ p$$

$$s = \left(\frac{y_K - y_J}{x_K - x_J}\right) mod \ p$$

If $K = -J$, i.e. $K = (x_J, -y_J) \ mod \ p$ then $J + K = O$, where O is the point at infinity.

If $K = J$ then $J + K = 2 \cdot J$, then point doubling operations are used. Also:

$$J + K = K + J$$

In the following, we illustrate the addition of the points $J = (1, 130)$ and $K = (4, 30)$. Then, the point $L(x_L, y_L)$ can be calculated as:

First s is calculated:

$$s = \frac{30 - 130}{4 - 1} \ mod \ 277$$

$$s = \frac{-100}{3} \ mod \ 277$$

$$s = -100 \cdot \frac{1}{3} \ mod \ 277$$

$$s = (-100) \cdot (-92) \ mod \ 277$$

$$s = 9200 \, mod \ 277$$

$$s = 59 \, mod \ 277$$

then

$$x_L = \left(59^2 - 1 - 4\right) \ mod \ 277$$

$$x_L = (3481 - 5) \ mod \ 277$$

$$x_L = 3476 \, mod \ 277$$

$$x_L = 152 \, mod \ 277$$

and finally

$$y_L = (59 \cdot (1 - 152) - 130) \, mod \; 277$$

$$y_L = (59 \cdot (-151) - 130) \, mod \; 277$$

$$y_L = -9039 \, mod \; 277$$

$$y_L = 102 \, mod \; 277$$

Hence the result of point addition of $(1, 130)$ and $(4, 30)$ for the elliptic group $E_{277}(1, 1)$ is $(152, 102)$.

2.1.2 Point Subtraction

Consider two distinct points J and K such that $J = (x_J, y_J)$ and $K = (x_K, y_K)$, then $J - K = J + (-K)$, where $-K = (x_K, -y_K) \; mod \; p$.

As an example, let consider the points $J = (1, 130)$, and $K = (4, 30)$. Then,

$$-K = (4, -30) \; mod \; 277 = (4, 247) \, mod \; 277$$

$$L = J - K = J + (-K)$$

$$L = (1, 130) + (4, 247) = (131, 63)$$

Hence the subtraction of $J = (1, 130)$ and $K = (4, 30)$, i.e. $L = J - K$, gives the point $L = (131, 63)$ which also lies in our elliptic curve.

2.1.3 Point Doubling

Consider a point J such that $J = (x_J, y_J)$, where $y_J \neq 0$.

Let $L = 2 \cdot J$, where $L = (x_L, y_L)$, then:

$$x_L = \left(s^2 - 2 \cdot x_J\right) \; mod \; p$$

$$y_L = (s \cdot (x_J - x_L) - y_J) \; mod \; p$$

$$s = \frac{3 \cdot x_J^2 + a}{2 \cdot y_J} mod \; p$$

s is the tangent at point J and a is one of the parameters that generates the point J.

If $y_J = 0, 2 \cdot J = O$, where O is the point at infinity.

As an illustration, we calculate the point $L = 2 \cdot J$ for $J = (1, 130)$.

Calculation of s:

$$s = \frac{3 \cdot 1^2 + 1}{2 \cdot 130} mod\ 277$$

$$s = \frac{4}{260} mod\ 277$$

$$s = 4 \cdot \frac{1}{260} mod\ 277$$

$$s = (4 \cdot 114)\ mod\ 277$$

$$s = 456\, mod\ 277$$

$$s = 179\, mod\ 277$$

then

$$x_L = \left(179^2 - 2 \cdot 1\right)\ mod\ 277$$

$$x_L = (32041 - 2)\ mod\ 277$$

$$x_L = 32039\, mod\ 277$$

$$x_L = 184\, mod\ 277$$

and finally

$$y_L = (179 \cdot (1 - 184) - 130)\, mod\ 277$$

$$y_L = (179 \cdot (-183) - 130)\, mod\ 277$$

$$y_L = (-32757 - 130)\, mod\ 277$$

$$y_L = -32887\, mod\ 277$$

$$y_L = 76\, mod\ 277$$

Hence the result of doubling of the point $(1, 130)$ for the group $E_{277}(1, 1)$ is the point $(184, 76)$.

2.1.4 Point Multiplication

We calculate point multiplication by combining point addition and point doubling. The following *double-and-add* algorithm functions like following:

$$
\begin{aligned}
&T \leftarrow P \\
&\textbf{for } i \leftarrow 1 \textbf{ upto } t-1 \\
&\quad T \leftarrow (T+T)\ mod\ n \\
&\quad \textbf{if } d_i = 1 \textbf{ then} \\
&\qquad T \leftarrow (T+P)\ mod\ n \\
&\textbf{return } t
\end{aligned}
$$

where P is a point in the elliptic curve, T is the variable where the result is stored, t is the binary width of the scalar which multiplies the point and d_i is the bit with the index i [9].

Consider the point $P(1, 130)$ that lies in the curve. Let us take a scalar $d = 47$, i.e. $d = (101111)_2$ and $d_i = [1, 0, 1, 1, 1, 1]$, then $T = d \cdot P$, i.e. $T = 47 \cdot (1, 130)$. The execution of the algorithm will produce the following calculations:

$$
\begin{aligned}
& T = (1, 130) \\
i = 1, \quad & T = ((1, 130) + (1, 130))\ mod\ 277 \\
& d_1 = 0, / \\
i = 1 + 1 = 2, \ & T = ((184, 76) + (184, 76))\ mod\ 277 \\
& d_2 = 1, \ T = ((15, 90) + (1, 130))\ mod\ 277 \\
i = 2 + 1 = 3, \ & T = ((60, 276) + (60, 276))\ mod\ 277 \\
& d_3 = 1, \ T = ((184, 86) + (1, 130))\ mod\ 277 \\
i = 3 + 1 = 4, \ & T = ((67, 3) + (67, 3))\ mod\ 277 \\
& d_4 = 1, \ T = ((103, 73) + (1, 130))\ mod\ 277 \\
i = 4 + 1 = 5, \ & T = ((244, 128) + (244, 128))\ mod\ 277 \\
& d_5 = 1, \ T = ((227, 102) + (1, 130))\ mod\ 277
\end{aligned}
$$

$$\mathbf{T = (46,106)}$$

Hence the multiplication of point $P = (1, 130)$ with the scalar $d = 47$, in the elliptic group $E_{277}(1, 1)$ gives as a result the point $T = d \cdot P = (46, 106)$.

3 ECC Encryption and Decryption

Elliptic curve cryptography can be used to encrypt plaintext messages into ciphertexts. The plaintext message, say M, is encoded into a point P_M from the finite set of points in the elliptic group, $E_p(a, b)$. The first step consists of choosing a generator point, $G \in E_p\,(a, b)$ such that the smaller value of n for which $n \cdot G = O$ is a very large prime number. The elliptic group $E_p(a, b)$ and the generator point G are made public.

Each user selects a private key, $n_A < n$ and computes the public key $P_A = n_A \cdot G$. To encrypt the message point P_M for B, A chooses a random integer k and computes the ciphertext pair of points P_C using B's public key P_B:

$$P_C = \left[(k \cdot G), (P_M + k \cdot P_B)\right]$$

After receiving the ciphertext pair of points P_C, B multiplies the first point, $(k \cdot G)$ with his private key n_B and then adds the result to the second point in the ciphertext pair of points, $(P_M + k \cdot P_B)$:

$$(P_M + k \cdot P_B) - (n_B \cdot k \cdot G) = (P_M + k \cdot n_B \cdot G) - (n_B \cdot k \cdot G) = P_M$$

which is the plaintext point, corresponding to the plaintext message M. Only B, knowing the private key n_B, can remove $n_B \cdot (k \cdot G)$ from the second point of the ciphertext pair of point, i.e. $(P_M + k \cdot P_B)$, and hence retrieve the plaintext information P_M [10].

Consider our elliptic curve:

$$y^2 = \left(x^3 + x + 1\right) \ mod \ 277$$

That is $a = 1, b = 1, p = 277$. The elliptic curve group generated by the above elliptic curve is $E_p(a, b) = E_{277}(1, 1)$.

Let $G = (0, 276)$ be the generator point, so the multiples $k \cdot G$ of the generator point G are (for $1 \leq k \leq 277$):

$G = (0, 276)$; $2G = (208, 105)$; $3G = (72, 220)$; $4G = (274, 91)$;
$5G = (174, 74)$; $6G = (84, 3)$; $7G = (117, 35)$; $8G = (47, 157)$;
$9G = (146, 241)$; $10G = (122, 201) \ldots\ldots\ldots 274G = (121, 180)$;
$275G = (258, 34)$; $276G = (149, 188)$; $277G = (175, 175)$;

If A wants to send to B the message M which is encoded as the plaintext point $P_M = (18, 158) \in E_{277}(1, 1)$, A must use B's public key to encrypt it. Suppose that B's secret key is $n_B = 85$, then B's public key will be:

$$P_B = n_B \cdot G = 85 \cdot (0, 276)$$

$$P_B = (237, 15)$$

A selects a random number k, say k = 113, and uses B's public key $\mathrm{P_B} = (237, 15)$ to encrypt the message point into the ciphertext pair of points:

$$\mathrm{P_C} = [(\mathrm{k} \cdot \mathrm{G}), (\mathrm{P_M} + \mathrm{k} \cdot \mathrm{P_B})]$$

$$\mathrm{P_C} = [113 \cdot (0, 276), (18, 158) + 113 \cdot (237, 15)]$$

$$P_C = [(260, 67), (18, 158) + (253, 130)]$$

$$P_C = [(260, 67), (68, 178)]$$

Upon receiving the ciphertext pair of points $P_C = [(260, 67), (68, 178)]$, B uses his private key $n_B = 85$, to compute the plaintext point P_M as follows:

$$(P_M + k \cdot P_B) - [n_B \cdot (k \cdot G)] = (68, 178) - [85 \cdot (260, 67)]$$

$$(P_M + k \cdot P_B) - [n_B \cdot (k \cdot G)] = (68, 178) - (253, 130)$$

$$(P_M + k \cdot P_B) - [n_B \cdot (k \cdot G)] = (68, 178) + (253, -130)$$

Because $-P = (x_P, -y_P)$

$$(P_M + k \cdot P_B) - [n_B \cdot (k \cdot G)] = (68, 178) + (253, 147)$$

Because $-130 \equiv 147 \mod 277$

$$(P_M + k \cdot P_B) - [n_B \cdot (k \cdot G)] = (18, 158)$$

and then maps the plaintext point $P_M = (18, 158)$ back into the original plaintext message M [10].

4 Audio File Format

The purpose of this chapter is to build a cryptosystem based on elliptic curves to secure the transmission of voice messages. We use a common format of audio file, the Waveform Audio File Format (WAVE) [11, 12]. The WAVE file format is a subset of Microsoft's RIFF specification for the storage of multimedia files. In this section we examine the integral structure of the RIFF file upon which will be applied encryption and decryption.

A RIFF file starts out with a file header followed by a sequence of data chunks. A WAVE file is often just a RIFF file with a single "WAVE" chunk which consists of two sub-chunks – a "fmt" chunk specifying the data format and a "data" chunk containing the actual sample data.

Figure 42.1 clearly shows that the actual data is stored after the 44-th byte and for illustrative purposes in our implementation we will encrypt only the part of actual data which is stored from the 45-th byte till the end of the file. Although the headers are left in clear, playing such a file produces a meaningless noise.

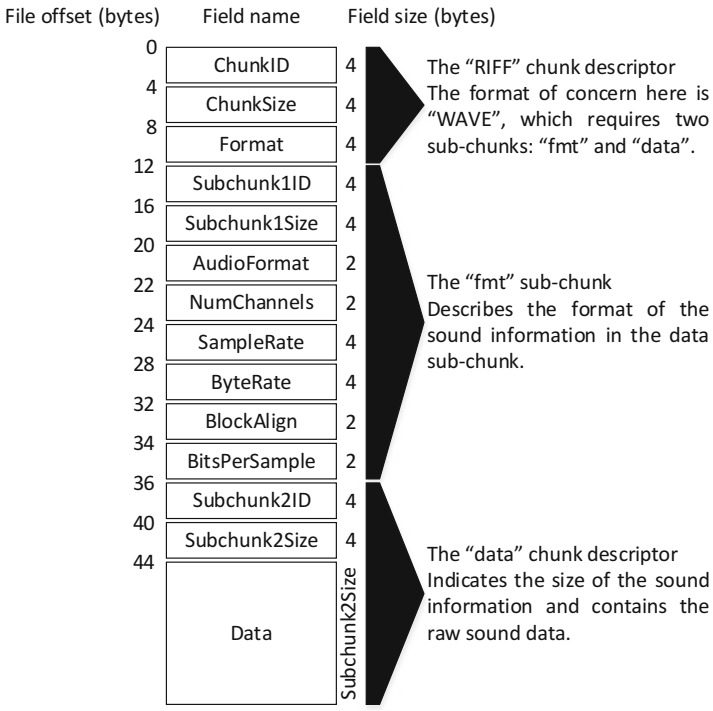


Fig. 42.1 The structure of .wav file format

5 Implementation

This section presents a tool that implements elliptic curve cryptography for encrypting an audio file, transmitting it through the network and decrypting the file at the other end. Figure 42.2 shows the interface that allows to experiment with the tool. It allows the user to define a private key, calculates the corresponding public key and allows sending the public key to the other communicating end. The user may record audio sequences, encrypt them and send the encrypted file. At the receiving end, the application decrypts and plays the audio data.

5.1 Encryption of Voice Messages

The voice message is read and its bytes are stored in the array **Ori[]** (Fig. 42.3). The array then is split into two other arrays where the first array **Header[]** contains the first 44 bytes which represent the header bytes, while the second array **WavData[]** contains the following bytes which represent the actual data of the voice.

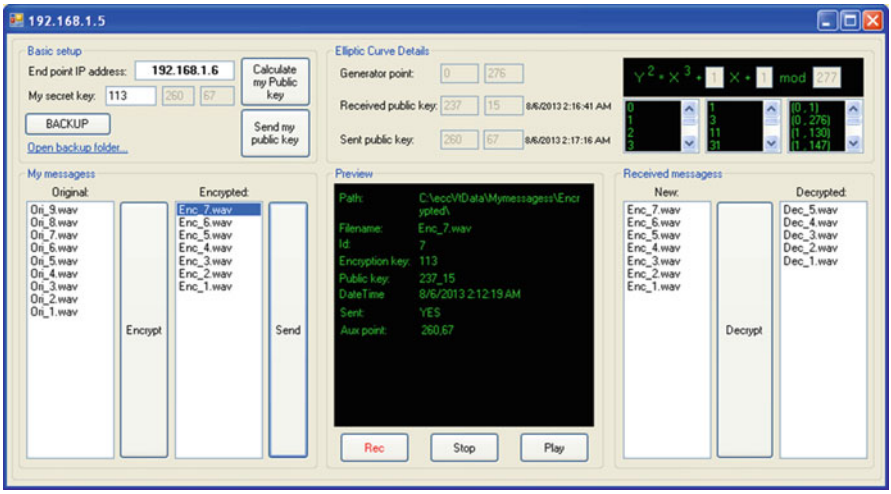


Fig. 42.2 System interface

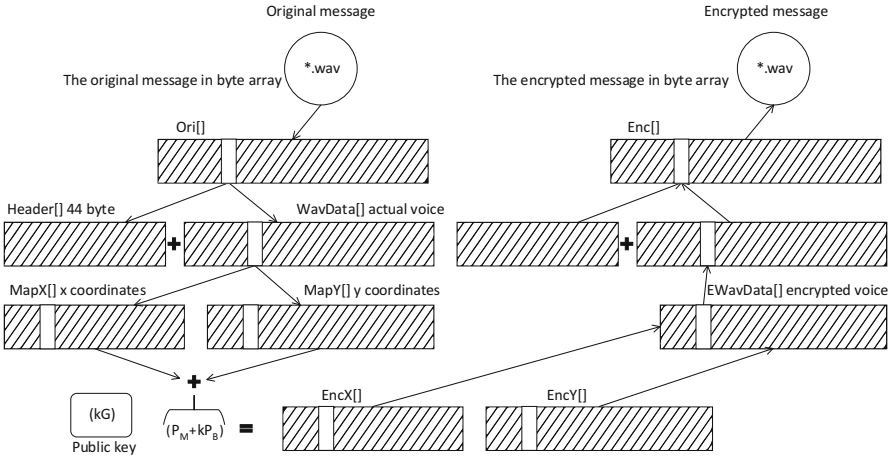


Fig. 42.3 Encryption process

Header[] won't be encrypted in order to enable playing the encrypted file. The bytes from **WavData[]** will be mapped into corresponding points of the elliptic curve and stored in the arrays **MapX[]** and **MapY[]**.

The encryption algorithm described in Sect. 3 is applied upon the stored points in the arrays **MapX[]** and **MapY[]**. The encrypted points are stored in the arrays **EncX[]** and **EncY[]**.

The points from **EncX[]** and **EncY[]** are mapped back into the corresponding bytes and stored into the array **EWavData[]. Header[]** and **EWavData[]** are merged into **Enc[]** and the content is written in a .wav file. The file represents the encrypted message which plays a meaningless noise, and this way could be securely transmitted through the network.

5.2 Decryption of Voice Messages

The encrypted voice message is read and its bytes are stored in the array **Enc[]** (Fig. 42.4). The array then is split into two other arrays where the first array **Header[]** contains the first 44 bytes which represent the header bytes, while the second array **EWavData[]** contains the following bytes which represent the actual data of the encrypted voice.

Header[] won't be decrypted since it represents the original header. The bytes from **EWavData[]** are mapped into corresponding points of the elliptic curve and stored in the arrays **MapX[]** and **MapY[]**.

The decryption algorithm described in Sect. 3 is applied upon the stored points in the arrays **MapX[]** and **MapY[]**. The decrypted points are stored in the arrays **DecX[]** and **DecY[]**.

The points from **DecX[]** and **DecY[]** are mapped back into the corresponding bytes and stored into the array **WavData[]**.

Header[] and **WavData[]** are merged into **Dec[]** and the content is written in a .wav file. The file represents the decrypted message which plays the original audio.

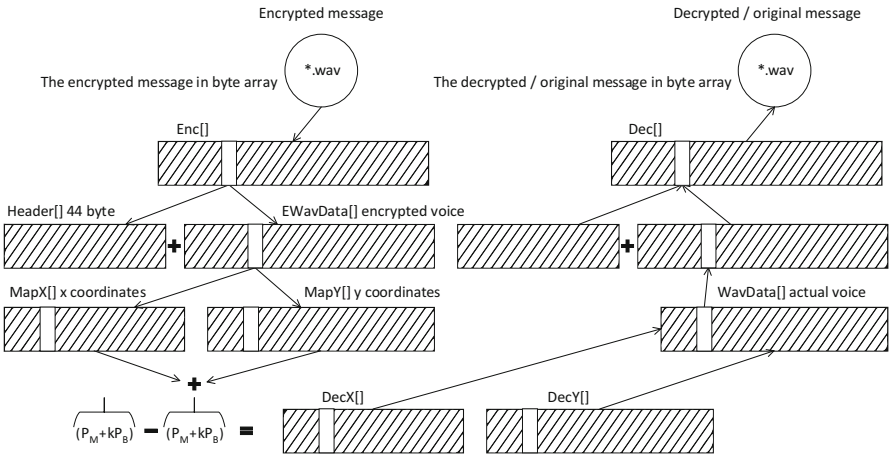


Fig. 42.4 Decryption process

5.3 Transmission of Voice Messages

Transmission of messages is done using standard network protocols. The main path of the system operation is illustrated as follows (Fig. 42.5):

To make the system more stable, during transmission of the encrypted voice message, public data related to the message can also be transmitted, which gives flexibility in changing and exchanging keys.

The transmission is done in three instances (Fig. 42.6). Every instance consists of two parts: header (the size of the instance being transmitted) and the actual data. The size of the message is merged to ensure that all the data is received by the receiver.

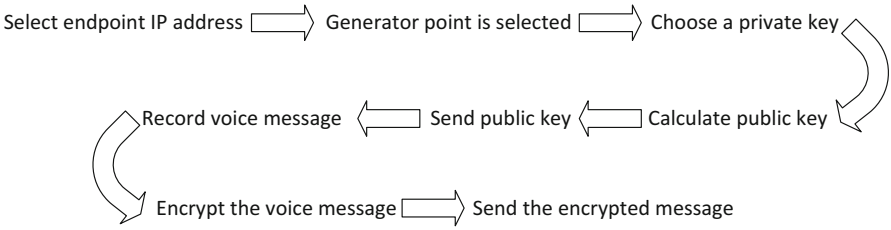


Fig. 42.5 System main path operation

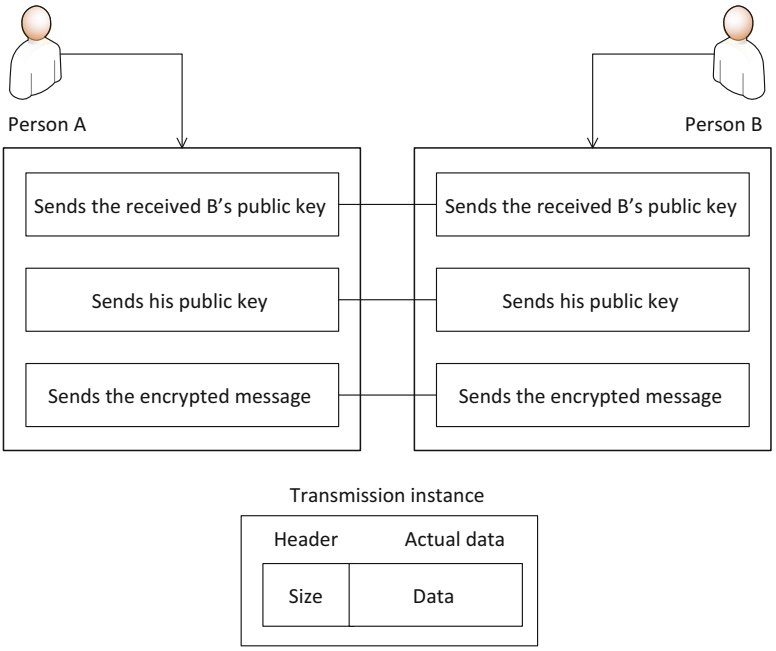


Fig. 42.6 Message transmission

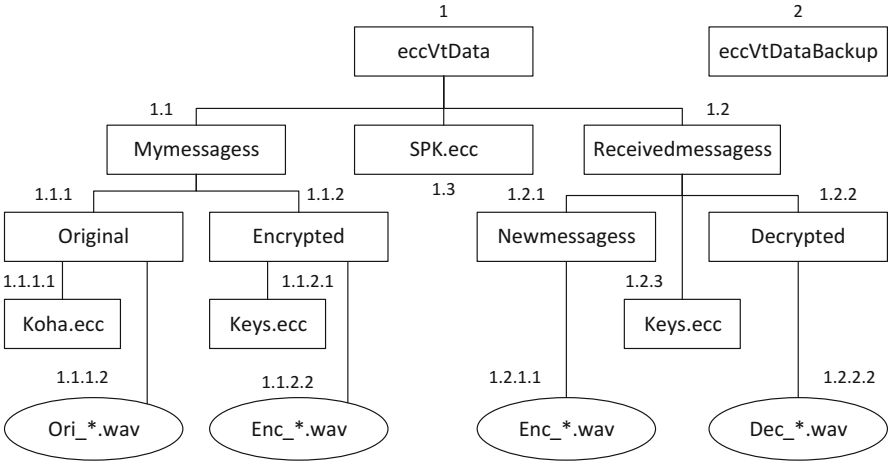


Fig. 42.7 Data organization

5.4 Data Organization

The implemented system stores the audio messages in separate files. Figure 42.7 shows the overall organization of data used by our system.

Files of type .ecc are files in which are stored voice message details such as: recording time, encryption and decryption keys, transmission time, and similar information.

6 Conclusion and Future Work

In this paper we have proposed and implemented a cryptosystem based on elliptic curve cryptography, intended to provide secure transmission of audio messages between communicating parties. This method of encryption is faster than RSA and thus it is more suitable for encrypting bigger amounts of data, as is the case with audio files. The advantage of elliptic curves relies in the fact that using a smaller-length key results in a stronger encryption compared to RSA encryption.

Using a similar approach, one can use elliptic curves for encryption of other types of data like image, video, text.

Considering the prospect of elliptic curves in terms of cryptosystems, there remain to work on optimizing the provided solution and adapt it for an implementation which will enable secure real time mobile communication with dynamic exchange of user-generated private keys.

References

1. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. Inform. Theory IT-22 **6**, 644–654 (1976)
2. Rivest,R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
3. Shoewu, O., Olatinwo, S.O.: Securing text messages using elliptic curve cryptography and orthogonal frequency division multiplexing. Pac. J. Sci. Technol. **14**, 220–227 (2013)
4. Luma, A., Ameti, L.: ECC secured voice transmitter. Lecture notes in engineering and computer science: Proceedings of The World Congress on Engineering 2014, WCE 2014, London, pp. 488–491. 2–4 July 2014
5. Koblitz, N., Menezes, A., Vanstone, S.: The state of elliptic curve cryptography. In: Towards a Quarter-Century of Public Key Cryptography, pp. 103–123. Springer (2000)
6. Gura, N., Patel, A., Wander, A., Eberle, H., Shantz, S.C.: Comparing elliptic curve cryptography and RSA on 8-bit CPUs. 6th International Workshop on Cryptographic Hardware and Embedded Systems, Boston, August 2004
7. Blake, I., Seroussi, G., Smart, N.: Advances in elliptic curve cryptography. London Mathematical Society Lecture Note Series, vol. 317, Cambridge University Press, Cambridge (2005)
8. Hankerson, D., Menezes, A., Vanstone, S.: Guide to Elliptic Curve Cryptograph. Springer-Verlag, New York (2004). ISBN 0-387-95273-X
9. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. In: Advances in Cryptology – CRYPTO 2001, pp. 213–229. Springer, Berlin/Heidelberg (2001)
10. Stallings, W.: Cryptography and network security (4th Edition). Prentice-Hall, Inc., Upper Saddle River, New Jersey (2005)
11. Microsoft Corporation: WAVE and AVI Codec registries – RFC 2361. IETF. http://tools.ietf.org/html/rfc2361 (1998). Retrieved 15 Dec 2013
12. Library of Congress: (2008-09-12). WAVE. Waveform audio file format. http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml (2008). Retrieved 15 Dec 2013

Chapter 43
A Series of Secret Keys in a Key Distribution Protocol

Mohd Anuar Mat Isa, Habibah Hashim, Jamalul-lail Ab Manan, Syed Farid Syed Adnan, and Ramlan Mahmod

Abstract In this chapter, we present a series of secret keys distribution in a key exchange protocol that incorporates protection against side channel attacks using Indistinguishability Experiment (modified) for Adaptive Chosen Ciphertext Attack (CCA2). We also present a security analysis and a new attack model for a secure Chain Key Exchange Protocol with an integration of TFTP protocol in the UBOOT firmware. To enable RasberberryPi *"system on chip"* (SoC) to perform cryptographic computation, we modified the GNU GMP Bignum library to support a simple primitive cryptographic computation in the UBOOT firmware. We suggest using our key exchange protocol for a secure key distribution in the UBOOT's TFTP protocol. Latter, the TFTP protocol can use the secure key which has been distributed by our key exchange protocol to encrypt the TFTP's data using another symmetric encryption scheme such as AES256. Lastly, we introduce a variance of adversary model in IND-CCA2-(TA, PA, TPA) which may be considered as a more realistic and practical model because it incorporates *timing attack* and *power attack.*

Keywords Adversary model • Adaptive chosen ciphertext attack • Cryptography • Embedded Raspberry Pi • Indistinguishability • Power attack • Provable security • Random oracle model • Trivial file transfer protocol • Timing attack

1 Introduction

This chapter describes an extension of our previous work [1] that is related to a security for a new key exchange protocol against side-channel attack. The state

M.A. Mat Isa (✉) • H. Hashim • S.F. Syed Adnan
Universiti Teknologi MARA, Shah Alam, Selangor 40450, Malaysia
e-mail: anuarls@hotmail.com; habib350@salam.uitm.edu.my; syed_farid@salam.uitm.edu.my

J.A. Manan
MIMOS Berhad, Technology Park Malaysia, Kuala Lumpur 57000, Malaysia
e-mail: jamalul.lail@mimos.my

R. Mahmod
Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia
e-mail: ramlan@upm.edu.my

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_43

of art for key exchange protocol is based on 1976 paper "New Directions in Cryptography" [2], Diffie and Hellman Key Exchange (DHKE) present a secure key agreement protocol that can be carried out over unsecure public communication channels. This protocol seems quite simple to be implemented; but it can be vulnerable to many types of attacks that are based on Number Theory. In this work, we propose a series of keys distribution in a Chain Key Exchange Protocol which is intended to be used as a case study to explore cryptographic computation capability of embedded system on chip (SoC) and its protection against timing and power analysis attack.

2 Motivation

The objective of this chapter is to explore cryptographic computation and security assessment for a series of keys distribution in Chain Key Exchange [1] scheme. The proposed protocol was intended for an implementation in a system on chip (SoC) with constrained environment consideration. Our main motivation in proposing the Chain Key Exchange scheme is to explore the computation capability of embedded microcontrollers such as ARM6 RaspberryPi board in performing cryptographic computation. To explore the possible constraints in the theoretical and experimental designs, we have decided to only use the RaspberryPi board and a USB debug/console cable as experimental setup for the experiment. The RaspberryPi board can support extra I/O functions (add-on card) including sensors, Wi-Fi, camera, sub controllers (e.g., random number generator and customized FPGA with cryptographic functions) and etc. However, we omitted these extra I/O features because we want to study a plain embedded board to perform cryptographic functions.

2.1 Target Application

This study will attempt to establish a secure and trust based key exchange protocol in the embedded controller. The term of "trust" is based on our previous work in Trusted Computing wherein *"How can we be assured device(s) and system(s) are trusted if we use trusted computing platform (*e.g.*, TPM) as root of trust?"* [3]. For this experiment, we do not use Trusted Platform Modules (TPM), but rather, we explore the concept of *"chain of trust"* in the cryptographic scheme, i.e. chain of trust of secret keys. The *"chain of trust of series secret keys"* allows our protocol to verify that new communication with third parties is the same as previous communication through secure transitive sessions. The proposed protocol would be useful for lightweight or smart embedded device to identify whether an adversary is trying to intrude into the confidential communication. Energy usage becomes major factor for operational consideration by lightweight devices especially for

deployment without compromising on security. In our proposal, we will use minimal I/O peripheral to reduce energy consumption, and at the same time yield high cryptographic computation in the autonomous environment. We implement this key exchange protocol in a UBOOT firmware; which gives advantage that key exchange protocol has a fairly quick boot (less than 5 s) to activate and perform key exchange process.

In the long term run, our scheme will provide an implementation of Secure Trivial File Transfer Protocol (TFTP) application in the UBOOT firmware. It will ensure remote system updates and patching (e.g., firmware, kernel or application) processes are secure from attacks which aim to eavesdrop and modify the TFTP packet. The target employment of Secure TFTP protocol is in the Wi-Fi Access Points, remote base stations, wireless sensor nodes and etc.

3 Experiment Setup

3.1 Embedded System

We decided to use RasberberryPi Model B (Fig. 43.1) with specifications: BCM2835 (ARMv7) 700 MHz, 512 MB RAM, 16GB SD memory card, 10/100 Ethernet port. This board is widely used for system prototyping or experiment, system controller, surveillance system, cluster nodes, embedded programming etc. We have done literature review on past works and we found that it is not well explored yet. From here, we decided to conduct cryptographic primitive computation using this board. Among the major issues need to be considered when

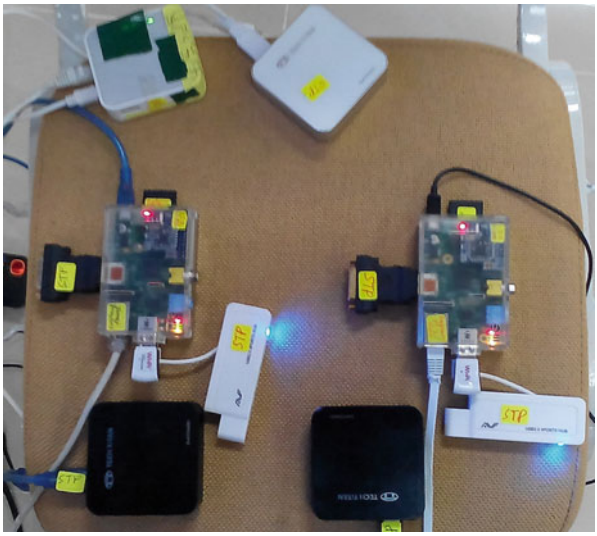


Fig. 43.1 Experimental testbed

using this board are GCC ARM compiler and GMP Bignum [4] library to compute numbers beyond 32-bit integers (e.g., exponential, modular, etc.).

We conducted an experiment to evaluate the Chained Key Exchange scheme. The first test group was conducted in application layer (user space) through Linux Raspbian "wheezy" Kernel using precompiled image "2013-07-26-wheezy-raspbian.zip" [5]. The second test group was conducted in firmware layer (bare metal) using Denx U-Boot [6] as platform for bare metal execution of our scheme. U-Boot provides cross platform execution because it supports multiple embedded architecture such as ARM, MIPS, PPC, x86, 68 k, Nios and etc. Therefore, we are confident that with a very minimal configuration, our protocol can be deployed in multiple embedded systems.

To enable RasberberryPi "system on chip" (SoC) to perform cryptographic computation, we modified the GMP Bignum version gmp-5.1.0 [4] library for a simple primitive cryptographic library. However, major modification is required in a bare metal system because of missing C library and its dependencies in the U-Boot. We noted that most of standard C libraries are meant for application and kernel layers, but not in firmware layer. This means that most of C libraries in firmware programming are minimal for the purpose of startup for the device and loading an operating system kernel for a system to boot up. To reduce the complexity, we modified the "mini-gmp" section to diminish the dependency problems. The modified "mini-gmp" is encoded in the first and second group experiments for fairness of execution and timing measurements. Our new "mini-gmp" library support the major functions for cryptographic computations such as mpz_init(), mpz_clear(), mpz_t, mpz_set_str(), mpz_powm(), mpz_get_str(), mpz_cmp(), mpz_sub(), mpz_add(), mpz_ui_pow_ui(), mpz_gcdext(), mpz_invert() clock(), SHA512(), and etc. Based on our previous work, we work on the communication protocol for two sets of RasberberryPi board using a secure TFTP protocol for smart environment [7]. The previous work [7] discussed the modification of U-Boot's TFTP protocol to support a secure key exchange and data encryption.

4 Chain Key Exchange Scheme

We divided the Chain Key Exchange into three major stages (refer Fig. 43.2).

4.1 Stage 1: Initialization of Pre-shared Knowledge Between Two Parties

This pre-shared knowledge must happen during production, physical exchange or through a trusted communication. Let assume these two parties names are Along and Busu. Let observe a Fig. 43.2 for a visualization of this stage. This initialization of crt key is less likely to be computed compared to the series of chain session key. We assumed Key_{crt} computation happens only in safe environments (e.g., during

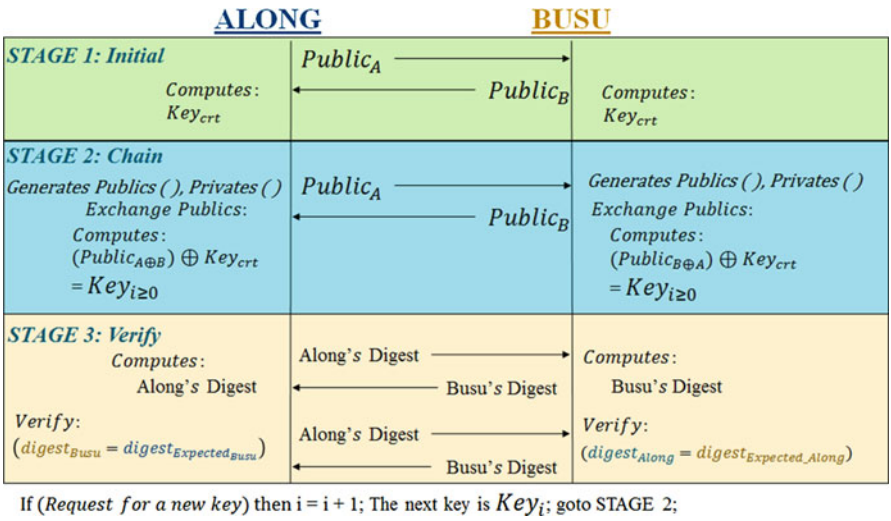


Fig. 43.2 A simplified protocol of chain key exchange scheme

production of embedded device) and no integrity verification of the messages is required. Furthermore, an adversary would not be able to eavesdrop this information because it happens in close environments. This initialization scheme has been originated from DHKE [2] scheme.

4.2 Stage 2: Initialization and Series of "Chained of Series Session Key" Between Two Parties

In this scenario, the *"chain of series session key"* occurs in open communication channel; hence it is still vulnerable to adversary's attacks. Figure 43.3 shows the *first chain of series session key; such that the initial session chain number is i = 0.* Figure 43.4 shows the *next series session key* (e.g. *let i = 1).* The series session key i = 1 can be generated after the previous key (let the previous key is i = 0) has successfully been verified in Stage 3. For the next session of key computation, we use key derivative function to derive $\overline{Key_{i=0}}$ from $Key_{i=0}$. Therefore, we conclude that the $Key_{i=1}$ as follow:

$$Key_1 \equiv {g_1}^{a_1} \cdot {g_1}^{b_1} \cdot {g_1}^{\overline{Key_0}} \pmod{p_1}$$

However, there is no guarantee that Key_i in Figs. 43.4 and 43.5 will be the size of *large n-key size* after the successful key exchange process. In worst case scenario, it produces a weak key with a short length. Therefore, Key_i needs to be

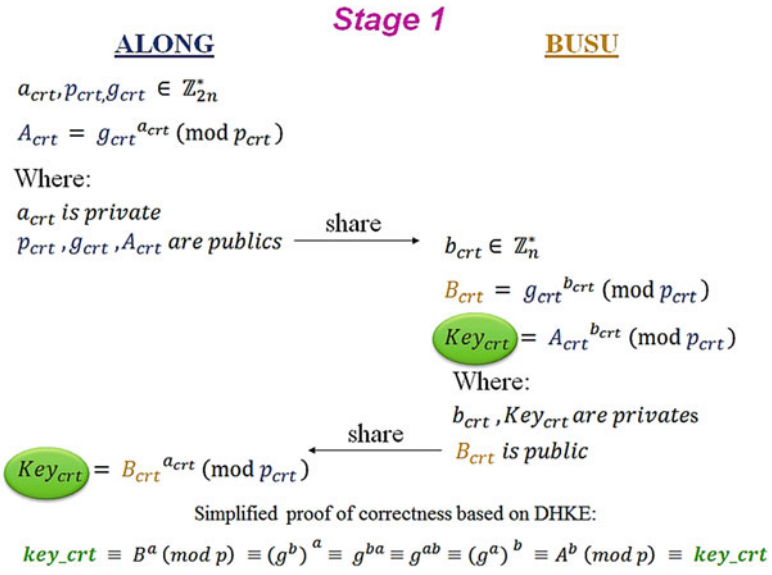


Fig. 43.3 Stage 1: generation of a "chain root trust" (CRT) key

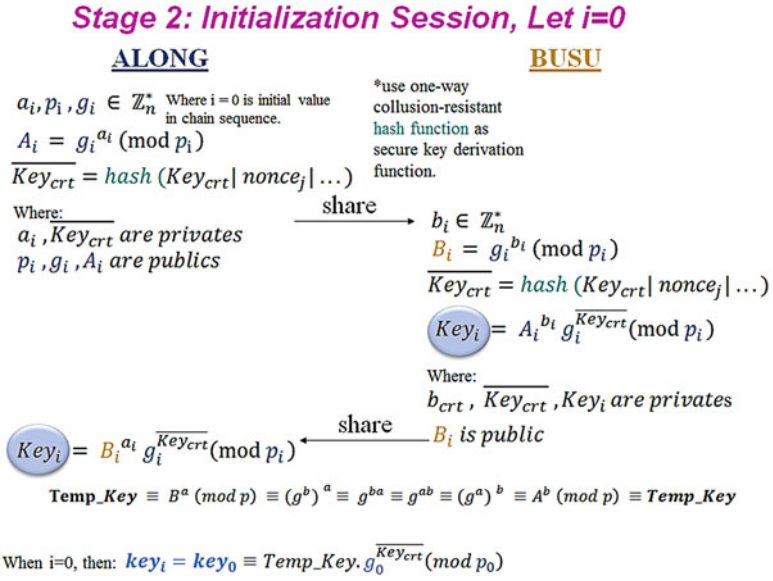


Fig. 43.4 Stage 2: initialization of "chained of series session key", i = 0

checked[1]/discarded before we can proceed to the verification process as it will be explained next.

[1]We can use generated key with a key length less than n (e.g. $(n-2)$ length); but we need to use a secure one way key expander/derivation function to fill-up (or padding) the less significant part of number in $(n-2)$ length. However this is very risky when the $\left(n - \left(\frac{n}{2}\right)\right)$ length is too short.

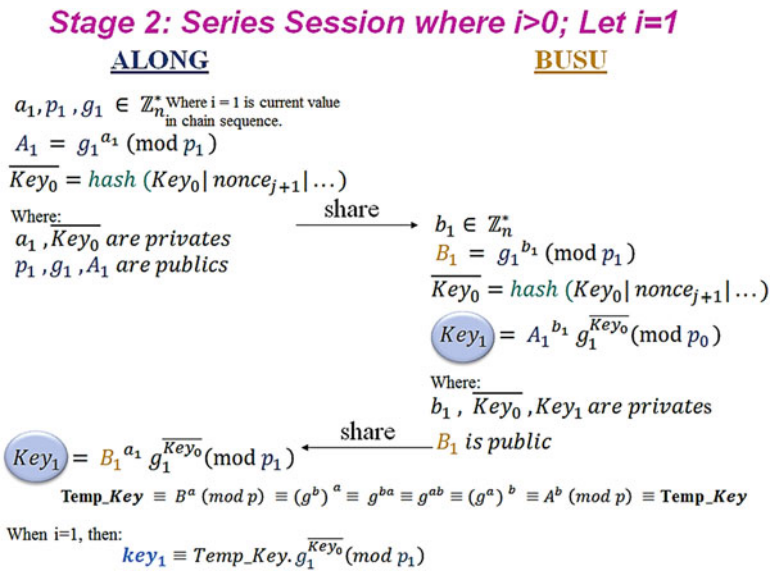


Fig. 43.5 Stage 2: next series session key, i = 1

4.3 Verification of "Chain of Series Session Key"

The purpose of verification is to ensure that both parties will synchronize a correct secret key with a Message Authentication Code (MAC). This will ensure that both parties are communicating with the right intended one. If successful, both parties will store the matching secret Key_i in the non-volatile memory equipped with physical tamper resistant technology wherein the secret key is protected using user authentication. The previous Key_{i-1} (if it still exists) is safely wiped out from nonvolatile memory. Figure 43.5 shown the verification in initialization of chain series session key (this is for i = 0). Figure 43.6 shown the verification in next of chain series session key (this is for i = 1).

Observe that in hashing function between Along and Busu, the sequence of hashing function is different in the first parameter and the second parameter for the hashing input. This will guarantee that the hashing digests of Along and Busu are different for the MAC authentication process. A good hashing function provides random values for blinding and masking [8] process to a key exchange process. A *strong collusion resistance hashing function* is can also help to secure the key exchange process against CCA2 [9] and timing attack if we use it correctly. To protect from an attack based on Number Theory, such as "degenerate message attack" [10], we need to ensure random secrets, public parameters and Key_i are not recycled numbers. For the next session, we must use a secure one way key derivation function (e.g. hashing) to derive $\overline{Key_i}$ from Key_i to avoid using previous key.

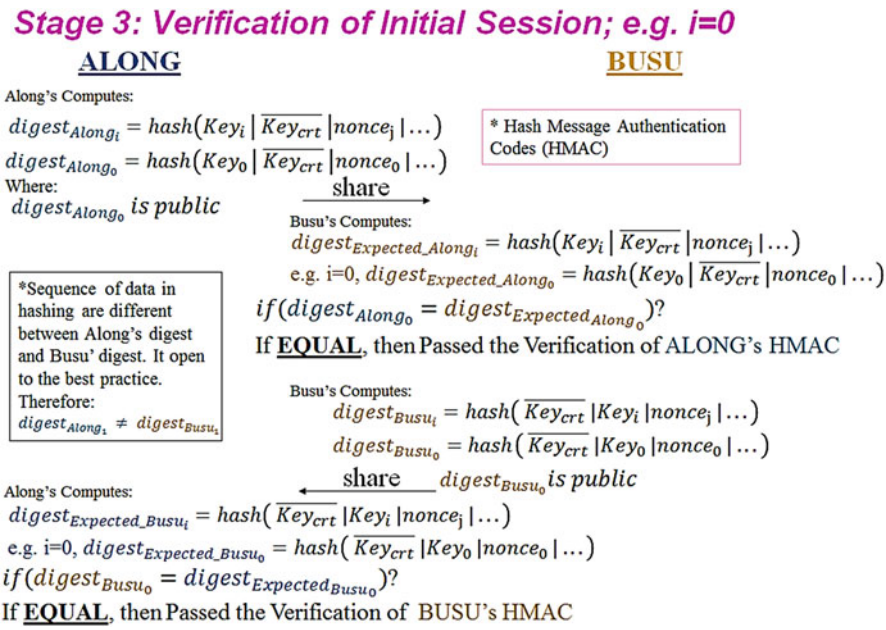


Fig. 43.6 Stage 3: verification for Initialization of "chained of series session key", i = 0

After that, both parties store Key_i that has been successfully verified. In case of failure, the digest need to be retransmitted (retry) because errors may happen in communication medium when using non-reliable network. Failure to do correction and verification within the allowed number of retries, the verification process is considered invalid and the chain of session *i* must be dropped. All temporary data in *Stage 2* must also be safely wiped out from volatile memory. If this problem happens, we can consider that there are problems: *(i) an error in the communication channel, (ii) an active adversary is impersonating either parties, or (iii) an active adversary has tampered the digest or data stream in the Stage 2* (Fig. 43.7).

5 Security Analysis

5.1 Previous Work

The underlying security for a Series Keys Distribution in key exchange protocol are based on two principles; (i) *Decisional Diffie-Hellman Problem (DDHP)* [11] and (ii) *a secure one way hash function* [12]. We have previously discussed some security analysis in our earlier publications [1, 13] such as session state reveal attack,

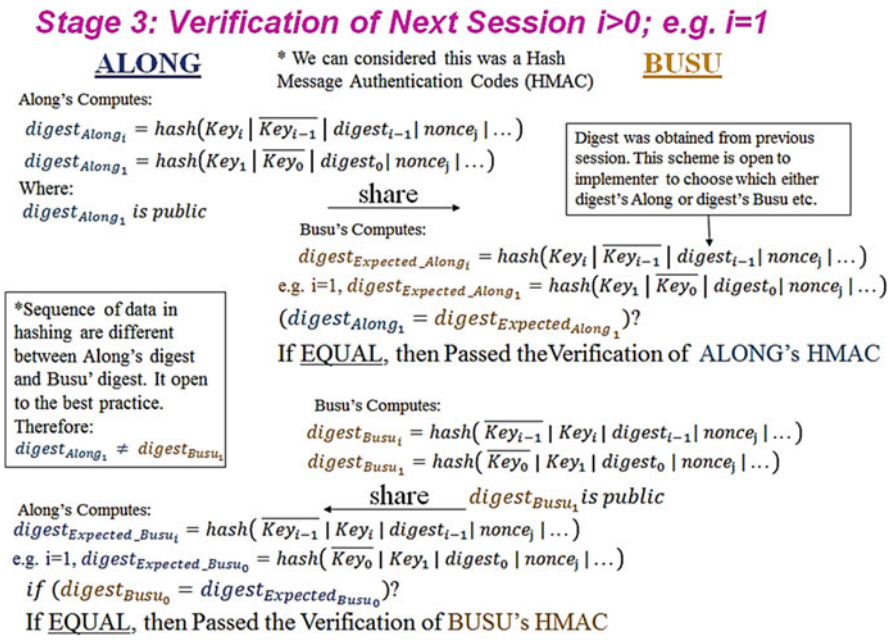


Fig. 43.7 Stage 3: verification for the next series session key, i = 1

forward secrecy, key independence and Burmester triangle attack. Readers may refer to our previous publications for further details.

5.2 *Side Channel Adversary Model: Timing and Power Attacks*

We have also introduced an adversary model Indistinguishability-Adaptive Chosen-Ciphertext (IND-CCA2) with timing attack in our previous work [13]. It uses the same indistinguishability challenge[2] in our general model [14]. However, in this chapter we use a different adversary model from the original one because we have not done any implementation of challenge ciphertext c^* in the indistinguishability experiment that is corresponding to Goldwasser-Micali [15], Naor-Yung [16] and Rackoff-Simon [17]. Furthermore, our new adversary model incorporates protection against *timing attacks* and *power attacks* in IND-CCA2. We believe that, it is essential to change the adversary model to allow for adaptation and protection against timing attack (TA) and power analysis attack (PA) in the indistinguishability experiment.

[2]We also present the latest work on adversary model in another chapter of this book.

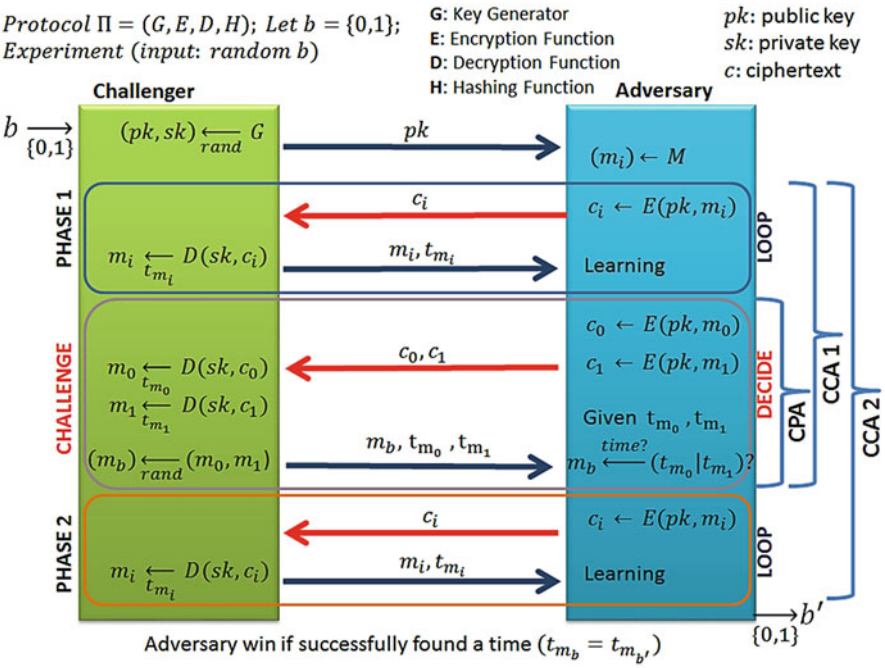


Fig. 43.8 IND-CCA2-(TA) experiment

To simulate the new proposed adversary model (TA, PA), we use indistinguishability test by letting an adversary to choose two ciphertexts (c_0, c_1) ε C where ($c_0 \neq c_1$) and length $|c_0| = |c_1|$. Both ciphertext (c_0, c_1) are freely chosen by an adversary with an intelligence of the best choices. The adversary is given the public key to encrypt any message for both chosen ciphertexts such that $c_0 = \mathcal{E}(pk, m_0)$ and $c_1 = \mathcal{E}(pk, m_1)$ in an encryption scheme $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{D}, \mathcal{H})$ as shown in Fig. 43.8.

The encryption/signature scheme Π is *CCA2-TA secure,* if and only if, any *probabilistic polynomial-time (PPT)* algorithm that was used in the adversary model to determine a correct timing from message m_b with a negligible probability such that:

$$|Pr[success] - Pr[failure]| < \frac{1}{p(n)}$$

$p(n)$ denotes the set of prime in size of n, for all sufficiently large n.

The adversary needs to distinguish whether m_b, t_{m_0}, t_{m_1} are either timing for c_0 *or* c_1 with probability of $\frac{1}{2} + \varepsilon$. If the probability to guess a correct decrypted ciphertext with the right timing is greater than $\frac{1}{2}$ with non-negligible advantage in PPT; we conclude that the adversary has an *"advantage"* and the given protocol Π is consider not secure in terms of indistinguishability experiment. The timing

attack can be represented using the same diagram as Fig. 43.8 with small changes in the notation such that: we replace the notation with $power = p_{m_0}, p_{m_1}$ and $m_b \underset{p_{m_b}}{\leftarrow} D(sk, c_b)$ in the challenger section. We assume that, the adversary is given knowledge of electrical current to perform cryptographic computations (e.g., primitive computation and protocol execution). The adversary is also given opportunity to use both timing and power attacks to gain more advantages in the experiment. The protocol Π can be secure, if and only if, the *timing and power attacks do not correlate to provide additional information to the adversary* to have a non-negligible advantage in the indistinguishability experiment.

6 Discussion

6.1 Hardware Security

Physical access to RaspberryPi board is a major problem in the product deployment. An attacker can easily remove the SD card memory from the board and use memory card reader to access all data in the SD card. We can encrypt the memory region in SD card using common techniques such as disk encryption.[3] At the moment, there is no temper resistant circuit in the devices, it can be easily broken if an attacker has physical access to the device. We need extra I/O devices that are connected to board with implementation of physical temper resistant (e.g., memory), physically unclonable function (PUF) [18], and etc.

6.2 Protocol Security: Side-Channel Attack

We have tested the Chain Session Key in Key Exchange Protocol in the RaspberryPi ARM7 (in bare-metal and with OS) and HP Elitebook 8440w i7 ×64 laptop in our previous work [13]. The result in [13] showed that the performance in RaspberryPi was slower by 11–12 times compared to laptop performance which can be considered as a performance issue. However, the main issue in the previous experiment is that the runtime of the RaspberryPi in UBOOT's bare-metal is extremely vulnerable to the timing and power analysis attacks. Based on our observation in the bare-metal environment, the computation for the same session key (e.g. session $i = 1$) is almost similar with differences of 30–60 Hz (CPU cycles) for the same public and private parameters. An adversary can easily re-compute the best predicted private parameters to sieve (similar to quadratic sieve)

[3]The size of UBOOT firmware with implementation of our scheme is 185,584 bytes. We can add user authentication module in the UBOOT firmware if user wants to access our scheme (encrypted memory regions of secrets and keys). We consider this as a future work.

to sieve the Computational Diffie-Hellman Problem (CDHP) in the protocol. The adversary begins the sieve attack through searching the lower bound of secret key and upper bound secret key in a computation of Discreet Log Problem (DLP) and CDHP. The adversary knows that the lower bound secret key is smaller than the upper bound secret key; and to compute the lower bound secret key is much faster than to compute upper bound secret key in $g^{lower} < g^{upper} \pmod p$. Given the DLP/CDHP computation timing, the adversary can reduce the searching space for the secret parameters e.g. $find\ x\ when\ given\ h,\ such\ that\ h = g^x \pmod p$; $when\ x\ is\ smaller,\ then\ it\ is\ faster\ to\ compute\ h.$

There is also a note from a manual in GNU's GMP library [4] that reminds about the risk of mpz_powm() function being vulnerable to timing attack.

To overcome the above mentioned vulnerability, we can use key derivation function, hashing secret parameters with unique strings or random nonce, unique sequence data input in a hashing function, partially shared secret and public parameters and fixed-time runtime (e.g. worst cases for all decryption time) that have been implemented in our key exchange protocol. We have already discussed these security properties in our previous work [1]. Another way to protect against timing attack is using blinding and masking [8] to provide randomness of runtime in the cryptographic computation. For power analysis attack, the prevention techniques discussed above is not enough to protect our key exchange protocol while running in bare-metal (RaspberryPi board). This happens because our key exchange protocol runs as single program (single threaded) in the UBOOT's firmware (without OS). Therefore, an Adversary can observe the entire running program with the exact current consumption. A method proposed by Molnar et al. [19] perhaps can be used for eliminating control-flow side channel in the C source codes. However, we have no idea whether it can work in a bare-metal with constrained environment. For a combined timing and power analysis attacks, as discussed in the side-channel adversary model section above, we believed that there is a correlation between timing attack and power attack for our key exchange protocol. This correlation will increase the probability of breaking the key exchange protocol compared to a standalone side-channel attack only.

7 Conclusion

Various research and developments works must be done to improve the proposed Chain Key Exchange Protocol. We need to further explore the implementation security or side-channel security. Our major contribution is the integration of our key exchange protocol for a secure TFTP implementation in the UBOOT firmware. The proposed secure TFTP protocol would overcome security problems (confidentiality, integrity and authenticity) in controlling, monitoring and upgrading embedded infrastructure in a pervasive computing environment. The target implementation of secure TFTP is for embedded devices such as Wi-Fi Access Points (AP), remote Base Stations (BS) and wireless sensor nodes. For the future work, we want to

investigate Molnar et al. [19] suggestion for a secure an implementation of C source codes for the side-channel prevention in a bare-metal programming.

Acknowledgment The authors would like to acknowledge the Ministry of Education (MOE) Malaysia for providing the grant 600-RMI/ERGS 5/3 (12/2013), and Universiti Teknologi MARA (UiTM) for supporting this research work.

References

1. Mohd Anuar Mat Isa, Habibah Hashim, Syed Farid Syed Adnan, Jamalul-lail Ab Manan, Ramlan Mahmod: An experimental study of cryptography capability using chained key exchange scheme for embedded devices. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, pp. 510–515. WCE, London, 02–04 July 2014
2. Diffie, W., Hellman, M.E.: New directions in cryptography. IEEE Trans. Inf. Theory **22**, 644–654 (1976)
3. Mohd Anuar Isa Mat, Azhar Abu Talib, Jamalul-lail Ab Manan, Siti Hamimah Rasidi: Establishing trusted process in trusted computing platform. In: Conference on Engineering and Technology Education, World Engineering Congress 2010, Kuching Sarawak, Malaysia. (2010).
4. The GNU Multiple Precision Arithmetic Library [Online]. Available: http://gmplib.org/ (2014). Kota Kinabalu Sabah, Malaysia.
5. Raspberry Pi Foundation: Raspberry Pi [Online]. Available: http://www.raspberrypi.org/downloads (2014)
6. DENX Software Engineering: DENX U-boot [Online]. Available: http://www.denx.de/wiki/U-Boot/WebHome (2014)
7. Mohd Anuar Mat Isa, Nur Nabila Mohamed, Habibah Hashim, Syed Farid Syed Adnan, R.M., Jamalul-lail Ab Manan: A lightweight and secure TFTP protocol in the embedded system. In: 2012 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE 2012), Kota Kinabalu Sabah, Malaysia (2012)
8. Kocher, P.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Advances in Cryptology – CRYPTO'96. http://link.springer.com/chapter/10.1007/3-540-68697-5_9 (1996).
9. Cramer, R., Shoup, V.: A practical public key cryptosystem provably secure against adaptive chosen ciphertext attack. In: Lecture Notes in Computer Science: Advances in Cryptology – CRYPTO'98, pp. 1–18 (1998)
10. Raymond, J., Stiglic, A.: Security issues in the Diffie-Hellman key agreement protocol. In: McGill University Technical Manuscript. http://crypto.cs.mcgill.ca/~stiglic/Papers/dhfull.pdf (2002).
11. Boneh, D.: The decision Diffie-Hellman problem. Algorithm. Number Theory **1423**, 1–14 (1998)
12. Tsudik, G.: Message authentication with one-way hash functions. ACM SIGCOMM Comput. Commun. Rev. **22**(5), 29–38 (1992)
13. Mohd Anuar Mat Isa, Habibah Hashim, Syed Farid Syed Adnan, Jamalul-lail Ab Manan, Ramlan Mahmod. A secure TFTP protocol with security proofs. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2014, vol. 1, pp. 443–448. WCE, London, 02–04 July 2014
14. Anuar, M., Isa, M., Hashim, H.: Adversary Model□: Adaptive Chosen Ciphertext Attack with Timing Attack. In: arXiv e-print (arXiv:1409.6556), pp. 1–3 (2014)

15. Goldwasser, S., Micali, S.: Probabilistic encryption & how to play mental poker keeping secret all partial information. In: STOC'82 Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing, pp. 365–377. http://dl.acm.org/citation.cfm?id=802212 (1982).
16. Naor, M., Yung, M.: Public-key cryptosystems provably secure against chosen ciphertext attacks. In: Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing – STOC'90, pp. 427–437. http://portal.acm.org/citation.cfm?doid=10 (1990).
17. Rackoff, C., Simon, D.R.: Non-interactive zero-knowledge proof of knowledge and chosen ciphertext attack. Adv. Cryptol. – CRYPTO'91, vol. LNCS 576, pp. 433–444. http://link.springer.com/chapter/10.1007%2F3-540-46766-1_35 (1992).
18. Suh, G.E., O'Donnell, C.W., Devadas, S.: AEGIS: a single-chip secure processor. Inf. Secur. Tech. Rep. **10**(2), 63–73 (2005)
19. Molnar, D., Piotrowski, M., Schultz, D., Wagner, D.: The program counter security model: automatic detection and removal of control-flow side channel attacks. In: Information Security and Cryptology – ICISC 2005. http://link.springer.com/chapter/10.1007/11734727_14 (2005).

Chapter 44
Forensics Issues in Cloud Usage

William R. Simpson

Abstract Forensics is undertaken to find out exactly what happened on a computing system and who or what was responsible for it. This is done by a structured investigation while maintaining a documented chain of evidence. Cloud computing is emerging as an attractive, cost effective computing paradigm. The early offerings of cloud capabilities have not provided security, monitoring or attribution that would allow an effective forensics investigation. The high assurance requirement presents many challenges to normal computing and some rather precise requirements that have developed from high assurance issues for web service applications and forensics applications of cloud systems. The challenges of high assurance and the maintenance of a documented chain of evidence associated with cloud computing are primarily in four areas. The first is virtualization and the loss of attribution that accompanies a highly virtualized environment. The second is the loss of ability to perform end-to-end communications. The third is the extent to which encryption is needed and the need for a comprehensive key management process for public key infrastructure, as well as session and other cryptologic keys. The fourth is in the availability of monitoring and logging for attribution, compliance and data forensics. Our view of high assurance and the issues associated with web services is shaped by our work with DoD and the Air Force, but applies to a broader range of applications.

Keywords Attribution • Architecture for monitoring data • Cloud computing • Forensics • IT security • Standards • Monitoring • Virtualization

1 Introduction

This paper is based in part on a paper published by WCE [1]. Cloud computing has come to mean many different things. To some, it is simply putting one's data on a remote server. However, in this paper, we utilize the definition provided by NIST [2]. They define five essential characteristics of any cloud computing environment:

W.R. Simpson (✉)
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_44

1. On demand self-service,
2. Broad network access,
3. Resource pooling,
4. Rapid elasticity, and
5. Measured service.

It is important to note that multi-tenancy and virtualization are not essential characteristics for cloud computing.

Cloud computing is, at its core, a service. There are three primary models of this service. In the lowest level Infrastructure as a Service (IaaS), storage, computation, and networking are provided by the cloud provider to the cloud consumer. In level two of the cloud models, Platform as a Service (PaaS), all of the trappings of IaaS plus an operating system and perhaps some application programming interfaces (APIs) are provided and managed by the cloud provider. The highest service model is Software as a Service (SaaS), in which the cloud provider provides an end-user service such as webmail. The higher the service model, the more control the cloud provider has as compared to the cloud consumer. There are four different models for deploying cloud services. Primarily, they are public or private clouds. In a public cloud, the infrastructure – although generally not the data on it – may be used by anyone willing to agree to its terms of use. Public clouds exist off the premises of the cloud consumer. Private cloud infrastructure is used only by one organization. It may exist either on or off the organization's premises. There are two twists to these infrastructures. In a community cloud, a group of organizations with similar interests or needs share a cloud infrastructure. That infrastructure is not open to the general public. In a hybrid cloud, two or more cloud deployment models are connected in a way that allows data or services to move between them. An example of this would be an organization's private cloud that makes use of a community cloud during periods of high utilization.

2 Cloud Benefits

Cloud computing benefits emerge from economies of scale [3]. Large cloud environments with multiple users are better able to balance heavy loads, since it is unlikely that a large proportion of cloud consumers will simultaneously have high utilization needs. The cloud environment can therefore run at a higher overall utilization, this may result in better cost effectiveness. In many cloud environments this balancing of resources is done by virtualization and the use of a hypervisor, offering resiliency and agility. In a large cloud computing environment, rather than having a number of information technology generalists, the staff has the ability to specialize and become the experts of their technical areas. With regard to information security, the staff can become even more specialized and spend more time hardening platforms to secure them from attacks. In the homogeneous cloud environment, patches can be rolled out quickly to the nearly identical hosts. Identically configured hardware elements are not a cloud requirement but do facilitate large-scale administration and focusing of expertise.

3 Cloud Weaknesses

Cloud computing is not without its negatives. In cases where services are outsourced, there can be a loss of control. This can affect compliance with laws, regulations, and organizational policies. Cloud systems have additional levels of complexity to handle intra-cloud communications, scalability, elasticity, data abstraction, and more. To be available to cloud consumers, cloud providers may need to make their services available via the Internet, opening interfaces that are subject to attack. And critically, many clouds allow multiple organizations simultaneous use of a single host and virtualization. If one tenant organization is compromised or malicious, it may be able to compromise the data or applications of the other organizations on the same host. The load balancing may use a single identity for all instances of a service whether it is virtual or real.

3.1 Changes in the Threat Model

There are clear differences in many of the threat scenarios as detailed below [4]:

- Loss of governance (or visibility and/or control of the governance process)
- Lock-in (threats may be present and locked into the cloud environment and shared among all tenants)
- Isolation failure (e.g., hypervisor attack, lack of accountability)
- Lack of distinction between virtualized instances of services)
- Compliance risks (if provider cannot provide compliance evidence or will not permit audit by customer, lack of accountability)
- Management interface compromise (and or inheritance of threats and/or malicious code from other users of the cloud).
- Data protection (how does customer verify protection, lack of accountability)
- Insecure or incomplete data protection and/or data deletion
- Malicious insider (often the cloud insider is not vetted as well as the organizational insider, and insiders from other customers could bring in contagious viruses – see 5 above.)
- Unprotected or ineffective key management for cryptography

3.2 Traditional Data Centers Versus Cloud Computing

Cloud computing relies on much of the same technical infrastructure (e.g., routers, switches, operating systems, databases, web servers) as traditional data centers and as a result, many of the security issues are similar in the two environments. The notable exception in some cases is the addition of a hypervisor for managing virtual machines. The Cloud Security Alliance's security guidance states "Security

controls in cloud computing are, for the most part, no different than security controls in any IT environment. Cloud computing is about gracefully losing control while maintaining accountability even if the operational responsibility falls upon one or more third parties." While many of the controls are similar, there are two factors at work that make cloud computing different: perimeter removal and trust. With cloud computing, the concept of a network or information perimeter changes radically. Data and applications flow from cloud to cloud via gateways along the cloud perimeters. However, since the data may be stored in clouds outside the organization's premises or control, perimeter controls become less useful. In exchange for the lack of a single perimeter around one's data and applications, cloud consumers must be able to trust their cloud providers. A lack of trust in a cloud provider does not necessarily imply a lack of security in the provider's service. A cloud provider may be acceptably secure, but the novelty of cloud computing means that many providers have not had the opportunity to satisfactorily demonstrate their security in a way that earns the trust of cloud consumers. Trust must be managed through detailed Service Level Agreements (SLAs), with clear metrics and monitoring mechanisms, and clear delineation of security mechanisms [5].

4 A Paradigm for High Assurance

While the current implementations of cloud computing provide efficient and operationally friendly solutions to data computing and content distribution, they are not up to the challenge of high assurance.

In certain enterprises, the network is continually under attack. Examples might be:

- Banking industry enterprise.
- Defense industry applications,
- Credit card consolidation processes.
- Commercial point-of-sale processes.
- Medical – privacy and statutory requirements,
- Content Distributor's – rights in data, theft of content.

The attacks have been pervasive and often include previously unseen attack vectors and they continue to the point that nefarious code may be present, even when regular monitoring and system sweeps clean up readily apparent malware. This omnipresent threat leads to a healthy paranoia of many threats including resistance to observation, intercept and masquerading. The web interface is the best way to provide access to many of its users despite this highly active threat environment. One way to maintain capability in this type of environment is to not only know and vet your users, but also your software and devices. Even that has limitations when dealing with the voluminous threat environment. Today we regularly construct seamless encrypted communications between machines through SSL or other TLS. These do not cover the "last mile" between the machine and the user (or service)

on one end, and the machine and the service on the other end. This last mile is particularly important when we assume that malware may exist on either machine, opening the transactions to exploits for eaves dropping, ex-filtration, session high-jacking, data corruption, man-in-the-middle, repeat replay, masquerade, blocking or termination of service, and other nefarious behavior. Before we examine the challenges of cloud computing systems, let us first examine what high assurance architecture might look like.

4.1 A Comprehensive Set of Tenets

We have implemented 12 tenets that guide decisions in an architectural formulation for high assurance approaches [6]. These tenets are distinct from the functional requirements normally associated with specific software component(s) (e.g., a name(s) need to be unique or identities need to be established); they relate more to the goals of the solution that guide its implementation.

The zeroth tenet is that the malicious entities are present and can look at all network traffic and may attempt to modify that traffic by sending virus software to network assets. In other words, rogue agents (including insider threats) may be present and to the extent possible, we should be able to operate in their presence, although this does not exclude their ability to view some activity. Assets are constantly monitored and cleaned, however new attacks may be successful at any time and nefarious code may be present at any given time.

The first tenet is simplicity. This seems obvious, but it is notable how often this principle is ignored in the quest to design elegant solutions with more and more features. That being said, there is a level of complexity that must be handled for security purposes and implementations should not overly simplify the problem for simplicity's sake.

The second tenet, and closely related to the first, is extensibility. Any construct we put in place for an enclave should be extensible to the domain and the enterprise, and ultimately to cross-enterprise and coalition. It is undesirable to work a point solution or custom approach for any of these levels.

The third tenet is information hiding. Essentially, information hiding involves only revealing the minimum set of information to the outside world needed for making effective, authorized use of a capability. It also involves implementation and process hiding so that this information cannot be farmed for information or used for mischief. Its corollary in software design provides only information that is needed to a software segment or process.

The fourth tenet is accountability. In this context, accountability means being able to unambiguously identify and track what active entity in the enterprise performed any particular operation (e.g., accessed a file or IP address, invoked a service). Active entities include people, machines, and software process, all of which are named registered and credentialed. By accountability we mean attribution with supporting evidence. Identity is a key attribute here and virtual elements must

have unique identities. Without such an identity process and detailed logging, it is impossible to establish a chain of custody or do effective forensic analysis to investigate security incidents.

The fifth tenet is minimal detail (to only add detail to the solution to the required level). This combines the principles of simplicity and information hiding, and preserves flexibility of implementation at lower levels. For example, adding too much detail to the access solution while all of the other IA components are still being elaborated may result in wasted work when the solution has to be adapted or retrofitted later.

The sixth is the emphasis on a service-driven rather than a product-driven solution whenever possible. Using services makes possible the flexibility, modularity, and composition of more powerful capabilities. Product-driven solutions tend to be more closely tied to specific vendors and proprietary products. That said, commercial off-the-shelf (COTS) products that are as open as possible will be emphasized and should produce cost efficiencies. This means that for acquisition, functionality and compatibility are specified as opposed to specific reference to current implementations such as "must operate in a Microsoft forest" [7] environment.

The seventh tenet is that lines of authority should be preserved and IA decisions should be made by policy and/or agreement at the appropriate level. An example here is that data owners should implement sharing requirements even when the requirements come from "higher authority."

The eighth tenet is need-to-share as overriding the need-to-know. Often effective health, defense, and finance rely upon and are ineffective without shared information.

The ninth tenet is separation of function, this makes updates easier, isolates vulnerabilities and aids in forensics.

The tenth tenet is reliability; it needs to work even if adversaries know how the process works. In setting up a large scale enterprise we need to publish exactly how things work and this should not create additional vulnerabilities.

The eleventh tenet is to trust but verify (and validate). This essentially precludes the use of identity by self-attestation which is unverified.

4.2 Element of the High Assurance Architecture

In order to build an architecture that conforms to these tenets, there must be elements that ensure that they are built into the systems. In the architecture we describe, the basic formulation follows a web 2.0 approach and uses Organization for the Advancement of Structured Information Standards (OASIS) standards of security [8]. These elements are listed below:

Identity derives from accountability. Identity will be established by the requesting agency. To avoid collisions, the identity used by all federated exchanges shall be the distinguished name as it appears on the primary credential provided by the

certificate authority. The distinguished name must be unique over time and space which means that retired names are not reused and ambiguities are eliminated. Naming must be applied to all active entities (persons, machines, and software).

Credentials derive from identity, reliability, trust but verify, and malicious entities. Credentials are an integral part of the federation schema. Each identity (all active entities) requiring access shall be credentialed by a trusted credentialing authority. Further, a Security Token Server (STS) must be used for storing attributes associated with access control. The primary exchange medium for setting up authentication of identities and setting up cryptographic flows is the Public Key Infrastructure (PKI) embodied in an X.509 certificate.

Authentication derives from accountability and malicious entities. The requestor will not only authenticate to the service (not the server), but the service will authenticate to the requestor. The preferred method of communication is secure messaging, contained in Simple Object Access Profile (SOAP) envelopes.

Confidentiality in transit derives from malicious entities and reliability. All messages are encrypted for delivery to the recipient of the message.

Authorization derives from accountability, malicious entities, lines of authority, trust but verify, and extensibility. Authorizations will be through the use of SAML packages in accordance with the SAML 2.0 specification provided by OASIS [9].

5 Cloud and High Assurance

Despite the obvious advantages of cloud computing, the large amount of virtualization and redirection poses a number of problems for high assurance. In order to understand this, let's examine a security flow in a high assurance system.

The application system consists of a web application (for communication with the user), one or more aggregation services that invoke one or more exposure services and combines their information for return to the web application and the user, As a prerequisite to end-to-end communication an SSL, or other suitable TLS is set up between each of the machines.

The exposure services retrieve information from one or more Authoritative Data Sources (ADSs). Each communication link in Fig. 44.1 will be authenticated end-to-end with the use of public keys in the X.509 certificates provided for each of the active entities.

This two-way authentication avoids a number of threat vulnerabilities. The requestor initially authenticates to the service provider. Once the authentication is completed, an TLS connection is established between the requestor and the service provider, within which a WS-Security package will be sent to the service. The WS-Security [10, 11] package contains a SAML token generated by the Security Token Server (STS) in the requestor domain. The primary method of authentication will be through the use of public keys in the X.509 certificate, which can then be used to set up encrypted communications, (either by X.509 keys or a generated session key). Session keys and certificate keys need to be robust and sufficiently protected

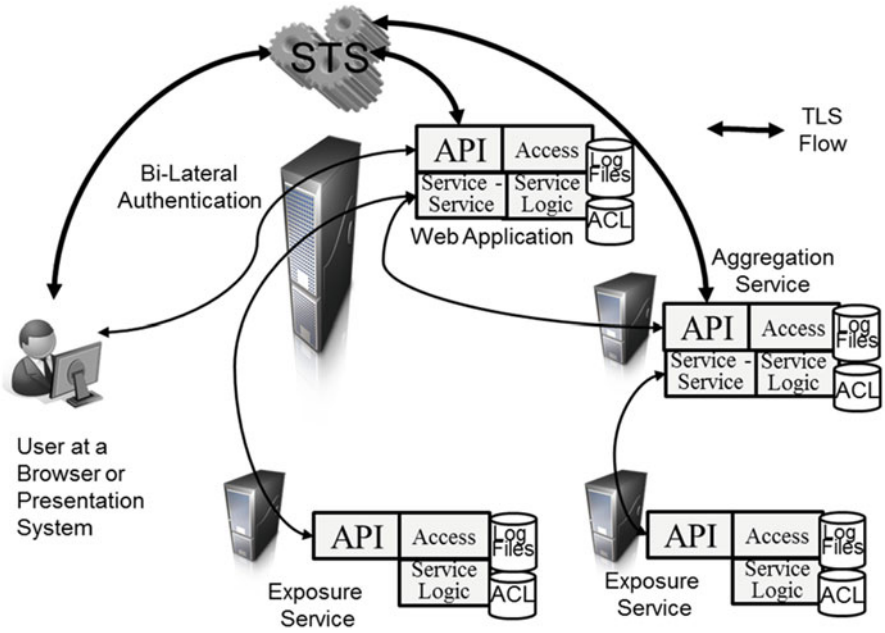


Fig. 44.1 High assurance security flows

to prevent malware exploitation. The preferred method of communication is secure messaging using WS Security, contained in SOAP envelopes. The encryption key used is the public key of the target (or a mutually derived session key), ensuring only the target can interpret the communication.

6 Accountability, Monitoring, Forensics

The goal of computer forensics is to perform a structured investigation while maintaining a documented chain of evidence to find out exactly what happened on a computing system and who or what was responsible for it [12]. There are several steps that must be taken to preserve the state, if we are interested in a forensics reconstruction of the computing.

6.1 Accountability

In order to ensure accountability, the number of independent instances must be anticipated. Names, credentials and end points must be assigned for their use. The attribute stores and HSMs must be provisioned with properties and key to be used.

The simple re-direct must be changed to a re-post loop. The requester will then have a credentialed application to authenticate with bi-laterally and an end point for end-to-end message encryption. Key management is complex and essential. When a new independent instance is required, it must be built and activated (credentials and properties in the attribute store, as well as end point assignment). All of these activities must be logged in a standard format. When a current independent instance is retired, it must be disassembled, and de-activated.

6.2 Monitoring

All of these activities must be logged in a standard format with reference values that make it easy to reassemble the chain of events for forensics. The same threats exist, and the same safeguards must be taken. There is a need for trusted and verified software to do the virtualization, and protection of the resources while they are in the resource pool.

6.3 Knowledge Repository System

The knowledge repository system is an integrated source of all information on the operation of the enterprise. It will be updated by the Enterprise Support Desk (ESD) and accessible to all the ESD and other entities with appropriate access credentials. Instrumented agents feed the data base on a schedule or on demand (Fig. 44.2).

6.3.1 Agents for Service Monitoring

There are two specific agents that are support the ESD.

Embedded Agents

The embedded agent sitsx in stacks and monitors the performance of the server and its threads. It should be configured to provide performance, connectivity and anomaly data.

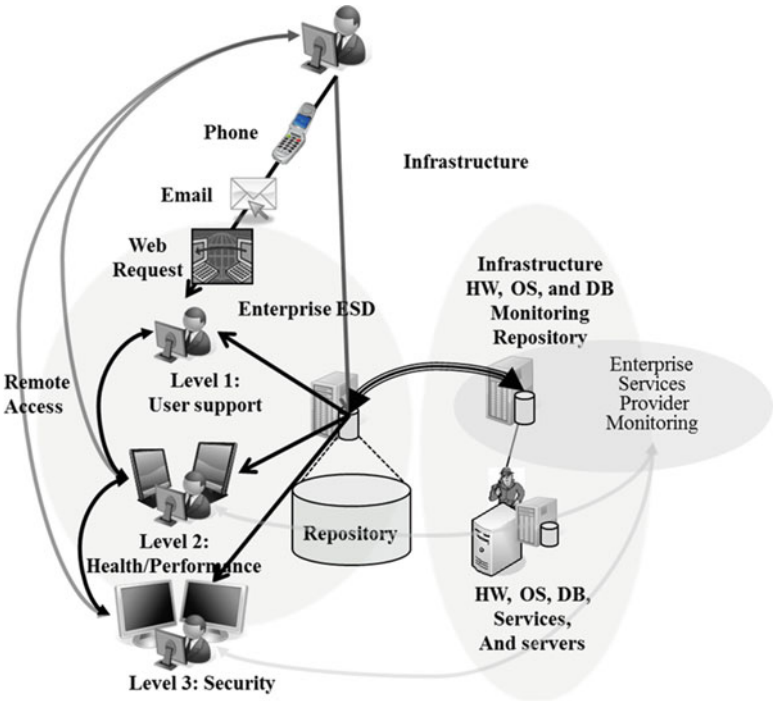


Fig. 44.2 Support desk and the knowledge repository

Monitor Sweep Agents

Figure 44.3 shows the placement of monitor sweep agents in the confines of the cloud, even though they may reach outside the cloud on external calls. It is the job of the monitor sweep agents to read, translate, and submit monitor records to the centralized knowledge repository. They may be extensive translation for some types of monitor data in external sources.

6.3.2 Knowledge Repository

The knowledge repository (KR) is a single integrated source of all information on the operation of the cloud. Instrumented agents feed the data base on a schedule or

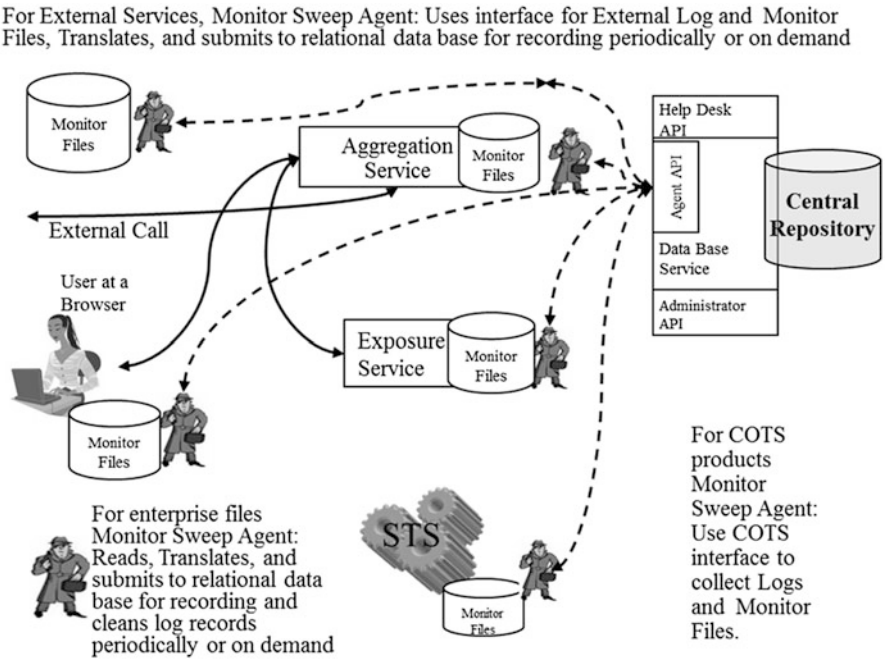


Fig. 44.3 Web service monitoring sweep model

Fig. 44.4 Knowledge repository

on demand. The knowledge base is where all information related to the enterprise SOA is stored. This will include the following (Fig. 44.4):

- Hardware/software current status from cloud or enterprise
- Current reports on test activities including response times, frequency of test, etc.
- Current reports of usage data from service agent monitors and service logs, including number of users, session times, response times, etc.
- Hardware/software historical data
- A list of current alerts for the entire enterprise
- Historical data on alerts
- Current monitoring records

Many of these feed status displays for the network monitoring and the enterprise support desk. They can provide a basis for real-time management and network defenses as well as forensics.

6.4 Forensic Tools

Tools are needed for KB query, correlation, and anomaly detection. Each tool should be functionally specified together with outputs and based upon the standard defined records described in the next section.

7 Standards Requirements for Cloud Forensics

Standards provide a basis for commercial applications as well as a market for developed tools and provide an interoperability process for generated files. Standards need to be developed, and are required for a number of cloud operations:

1. Standards for identity issuance of virtual objects. Current practice of overloading the identity makes the development of forensic strategies difficult or problematic. The standard should include credentialing and establishment of identity attributes.
2. Standards are required for a number of details associated with cloud monitoring. These requirements are based upon monitoring requirements derived for enterprise application and shown in Fig. 44.5.
   - Standard lists of events that must be logged and alerted including event specific data (these may be separated into minimal, rich, and robust categories). The events should include all of the agility events such as stand-up and tear-down of services as well as credentialing and establishment of identity attributes.

All Records: Record Number, Thread Number, Sequence Number, Active Entity Name, Event Name, ID of Service Requestor, Date/time, Other pertinent data (OP_i)	
Event	**Content**
Start-up of the audit function the service	Event Name = " Start Up Audit", OP1= ID of Service Requestor, OP2 = Level of Audit
Shutdown of the audit within the service	Event Name = "Shut Down Audit", OP1= ID of Service Requestor

Fig. 44.5 Monitoring records

- A standard process for automatic escalation of monitoring under pre-defined circumstances.
- Standard monitoring record content (that is separate from event specific content) including; Date/time, Record Number, Thread Number, Sequence Number, Active Entity Name, Event Name, ID of Service Requestor.
- Standard identifiers for monitoring content (to assist upload to a centralized KR).
- Standard processes for protection, access and integrity of monitor records.
- Standards for continuous operations and recovery of the monitoring system and its records.

3. Standards for KR construction, KR update process, forensics tools and analysis processes.
4. Standards for protection, access and integrity of stored resources that will be used in agile virtualization.

8 Summary

We have reviewed the basic approaches to clouds and their potentials for savings in computing environments. We have also discussed high assurance architectures and their requirements which provide direct challenges to the way cloud computing environments are organized. Notably the extensive use of virtualization and re-direction is severe enough that many customers who need high assurance and forensics capabilities have moved away from the concept of cloud computing [13–15]. We believe, however, that a precise statement of the high assurance and forensics requirements will lend themselves to solutions in the cloud computing environment, and expand the potential use of this technology. This work is part of a body of work for high assurance enterprise computing using web services. Elements of this work are described in [16–20].

References

1. Simpson, W.R., Coimbatore Chandersekaran: Cloud forensics issues. Lecture Notes in Engineering and Computer Science: Proceedings World Congress on Engineering 2012, pp. 475–480. London, 2–4 July 2014
2. Mell, P., Grance, T.: NIST SP 800–145 Draft: Cloud Computing, Computer Security Division, Information Technology Laboratory. National Institute of Standards and Technology, Gaithersburg, Jan 2011. http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf
3. Jansen, W., Grance, T.: NIST SP 800–144 Draft: Guidelines on Security and Privacy in Public Cloud Computing, Security Division, Information Technology Laboratory. National Institute of Standards and Technology, Gaithersburg, Jan 2011. http://csrc.nist.gov/publications/drafts/800-144/Draft-SP-800-144_cloud-computing.pdf

4. Catteddu, D., Hogben, G.: European Network Information Security Agency (ENISA), Cloud Computing Risk Assessment, Nov 2009. http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment
5. Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing V2.1, Dec 2009. https://cloudsecurityalliance.org/csaguide.pdf
6. Air Force Information Assurance Strategy Team, Air Force Information Assurance Enterprise Architecture, Version 1.70, SAF/XC, 15 Mar 2009 [Not available to all]
7. Windows Server 2003: Active Directory Infrastructure. Microsoft Press, pp. 1–8 to 1–9 (2003). ISBN: 0-7356-1438-5
8. OASIS Identity Federation, Liberty Alliance Project. Available at http://projectliberty.org/resources/specifications.php
9. OASIS Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0. Available at http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security
10. Web Service Security: Scenarios, Patterns, and Implementation Guidance for Web Services Enhancements (WSE) 3.0. Microsoft Corporation (2005)
11. WSE 3.0 and WS-ReliableMessaging. Microsoft White Paper, June 2005. Available at http://msdn2.microsoft.com/en-us/library/ms996942(d=printer).aspx
12. SearchSecurity, Security Resources, computer forensics (cyber forensics). http://searchsecurity.techtarget.com/definition/computer-forensics
13. Remarks-Debra Chrapaty, Corporate Vice President, Global Foundation Services, Microsoft Mgt Summit, Las Vegas, May 2008. http://www.microsoft.com/Presspass/exec/debrac/mms2008.mspx
14. Bobbie Johnson, technology correspondent, guardian.co.uk, Cloud computing is a trap, warns GNU founder Richard Stallman, 29 Sept 2008. http://www.guardian.co.uk/technology/2008/sep/29/cloud.computing.richard.stallman
15. Andy Plesser, Executive Producer, Beet.tv, Cloud Computing is Hyped and Overblown, Forrester's Frank Gillett.....Big Tech Companies Have "Cloud Envy" http://www.beet.tv/2008/09/cloud-computing.html, 26 Sept 2008. Accessed 19 Feb 2011
16. Simpson, W.R., Coimbatore Chandersekaran: The 2nd International Multi-Conference on Engineering and Technological Innovation: IMETI2009, vol. I, pp. 300–305. Information Sharing and Federation, Orlando, July 2009
17. Simpson, W.R., Coimbatore Chandersekaran: International Journal of Computer Technology and Application (IJCTA). An Agent-Based Web-Services Monitoring System, vol. 2, no. 9, pp. 675–685, Sept 2011
18. Coimbatore Chandersekaran, Simpson, W.R.: Claims-based enterprise-wide access control. Lecture Notes in Engineering and Computer Science: Proceedings World Congress on Engineering 2012, pp. 524–529, London, 4–6 July 2012
19. Coimbatore Chandersekaran, Simpson, W.R.: Assured content delivery in the enterprise. Lecture Notes in Engineering and Computer Science: Proceedings World Congress on Engineering 2012, pp. 555–560, London, 4–6 July 2012
20. Simpson, W.R., and Coimbatore Chandersekaran: Claims-based authentication for a web-based enterprise. Lecture Notes in Engineering and Computer Science: Proceedings World Congress on Engineering 2013, pp. pp. 1204–1209, London, 3–5 July 2012

Chapter 45
Undergraduate Student Retention Prediction Using Wavelet Decomposition

Ji-Wu Jia and Manohar Mareboyana

Abstract In this paper, we have presented some results on undergraduate student retention using signal processing techniques for classification of the student data. The experiments revealed that the main factor that influences student retention in the Historically Black Colleges and Universities (HBCU) is the cumulative grade point average (GPA). The linear smoothing of the data helped remove the noise spikes in data thereby improving the retention results. The data is decomposed into Haar coefficients that helped accurate classification. The results showed that the HBCU undergraduate student retention corresponds to an average GPA of 2.8597 and the difference of −0.023307. Using this approach we obtained more accurate retention results on training data.

Keywords Haar transform • Linear smoothing • Machine learning • Signal processing • Student retention • Wavelet transform

1 Introduction

In this paper we study the HBCU undergraduate student retention using signal processing techniques to obtain the HBCU undergraduate student retention criterion for average GPA [1–5].

We started collecting data from the HBCU Fall 2006 full-time and first-time undergraduate students. We tracked these students' activities in the following 6 years from Fall 2006 to Fall 2011. The data was queried from the Campus Solution database. The 6-year training data set size is 771 instances with two attributes shown in Table 45.1. The HBCU undergraduate 6 years retention rate 44.9 % was derived from the 6-year training data set [6]. The HBCU 6-year training data set numeric attributes and statistics are shown in Table 45.2 [2].

J.-W. Jia (✉) • M. Mareboyana
Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA
e-mail: jjia@bowiestate.edu; MMareboyana@bowiestate.edu

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_45

Table 45.1 List of data set attributes

Number	Name	Description	Type
1	GPA	The last cumulative GPA while student enrolled	Number
2	Retention	If student graduated or enrolled in Fall 2011 then yes, else no	Text

Table 45.2 Training data set numeric attributes

Naive Bayes	No retention		Retention	
Attribute name	Mean	Std. Dev.	Mean	Std. Dev.
GPA	1.9371	±0.8913	2.8864	±0.4276

We classified the data under two groups using well known technique such as machine learning [7], decision tree [8], neural network [9], support vector machines [9], and the training set and test set [10]; "Retention" – students who were retained in the HBCU and "No Retention" – students who were not retained in the HBCU [6, 11–24].

The most basic wavelet transform is the Haar transform described by Alfred Haar in 1910. It serves as the prototypical wavelet transform. We will describe the (discrete) Haar transform, as it encapsulates the basic concepts of wavelet transforms used today. We will also see its limitation, which the newer wavelet transform (in 1998 by Ingrid Daubechies) resolves [4].

The algorithm to calculate the Haar transform of an array of n (number of years) samples is below [4]:

1. Treat the array as n/2 pairs called (a, b)
2. Calculate (a + b)/sqrt(2) for each pair, these values will be the first half of the output array.
3. Calculate (a–b)/sqrt(2) for each pair, these values will be the second half.
4. Repeat the process on the first half of the array (the array length should be a power of two)

First, we pre-processed the student data, added missing data, and grouped the student data into two files. They are retention and no-retention files.

Second, we applied linear smoothing to the discrete GPA signals for removing noise.

Finally, we applied Haar transform to the GPA data, and calculated the average and difference from the retention data, and discussed the average GPA for the HBCU undergraduate student retention. The framework of the study is shown as Fig. 45.1.

In the following sections, we describe the methodology and algorithms.

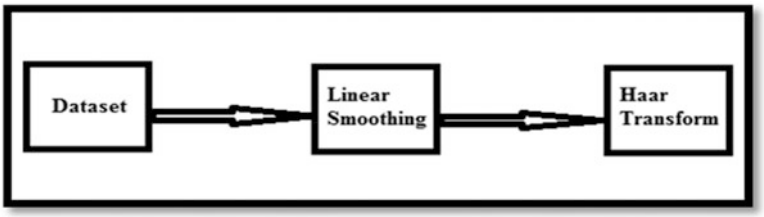


Fig. 45.1 The framework for the study

```
clf
hold
b=[1 2 3 4 5 6]
r=load('F:\Octa\used\retention.txt')
plot(b,r,'*')
```

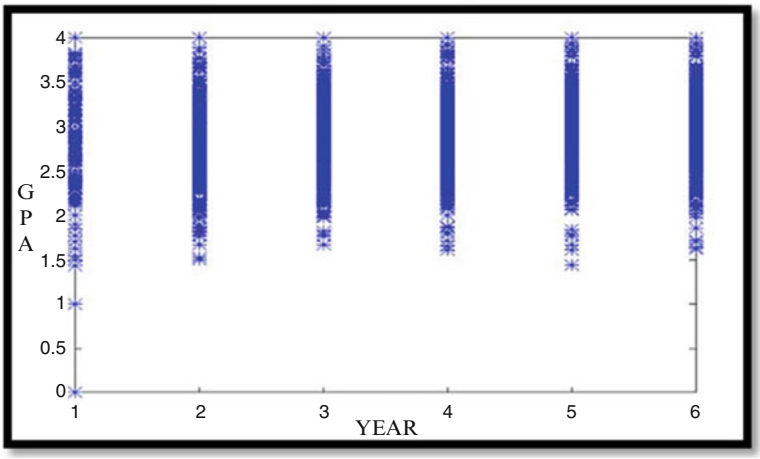


Fig. 45.2 Retention student data

2 Methodology

2.1 Retention Student Data

The retention student data are shown in Fig. 45.2 as star. The data are concentrative distributed.

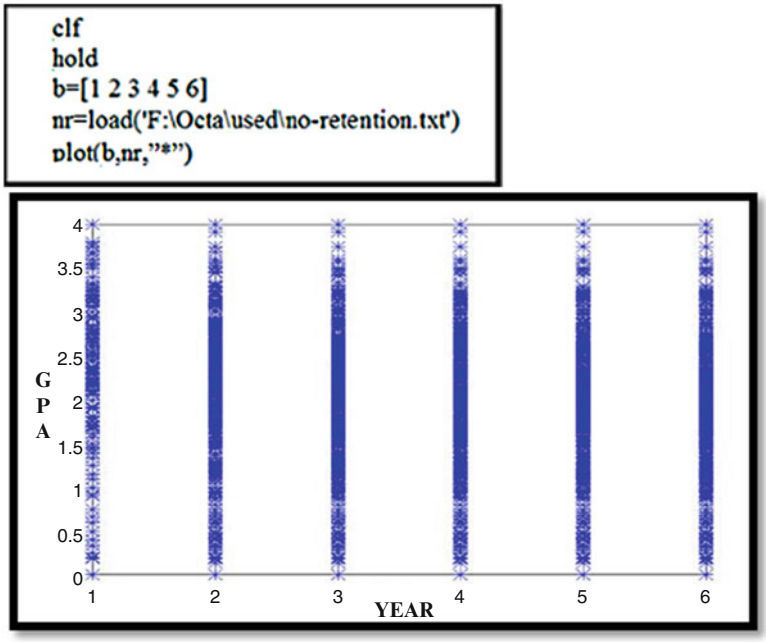


Fig. 45.3 No-retention student data

2.2 *No-Retention Student Data*

No-retention student data are shown in Fig. 45.3 as star. The data are not concentrative distributed.

2.3 *Linear Smoothing Retention GPA Data*

We applied linear smoothing to the retention GPA data. The linear smoothing is applied as shown below.

$$\begin{aligned} y_1 &= \frac{x_1 + x_2 + x_3}{3} \\ y_2 &= \frac{x_2 + x_3 + x_4}{3} \\ &\ldots\ldots \end{aligned} \tag{45.1}$$

Where $x_1, x_2, x_3 \ldots$ are the original GPA data, and $y_1, y_2, y_3 \ldots$ are the new GPA data.

```
clf
hold
a=[1 2 3 4}
r3=r(:,1:4)
[m n]=size(r)
for i=1:m
r3(i, 1)=(r(i, 1)+r(i, 2)+r(i, 3))/3;
r3(i, 2)=(r(i, 2)+r(i, 3)+r(i, 4))/3;
r3(i, 3)=(r(i, 3)+r(i, 4)+r(i, 5))/3;
r3(i, 4)=(r(i, 4)+r(i, 5)+r(i, 6))/3;
end
for k=1 m
plot(a, r3(k, :),'*');
end
```

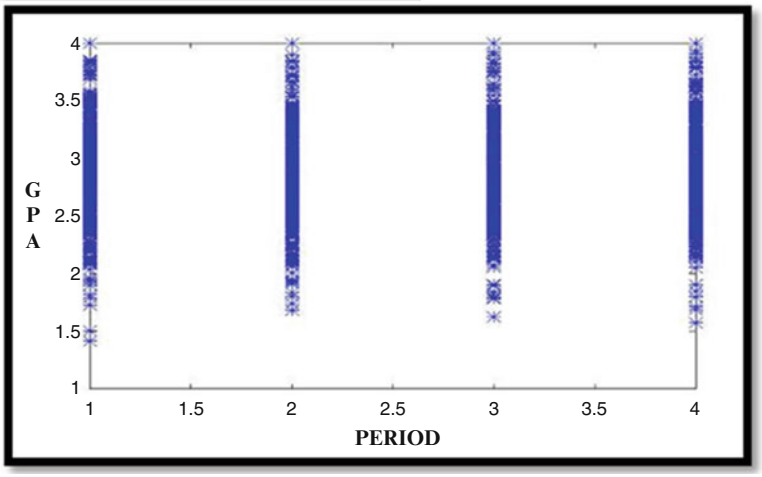


Fig. 45.4 Linear smoothing retention discrete GPA data

The new retention student data are shown in Fig. 45.4 as star. Compared the smoothing retention data to the original data (Fig. 45.2), we can see the lowest and highest values are removed, and 6 years data have been smoothed into three periods (four data points).

2.4 Linear Smoothing No-Retention GPA Data

Used the same algorithms as retention, we did linear smoothing for the no-retention data. Compared the smoothing no-retention data to the original data (Fig. 45.3),

```
clf
hold
a=[1 2 3 4}
[m n]=size(nr)
s=r(:,1:4)
for i=1:m
s(i, 1)=(nr(i, 1)+nr(i, 2)+nr(i, 3))/3;
s(i, 2)=(nr(i, 2)+nr(i, 3)+nr(i, 4))/3;
s(i, 3)=(nr(i, 3)+nr(i, 4)+nr(i, 5))/3;
s(i, 4)=(nr(i, 4)+nr(i, 5)+nr(i, 6))/3;
end
for k=1:m
plot(a, s(k, :),"*");
end
```

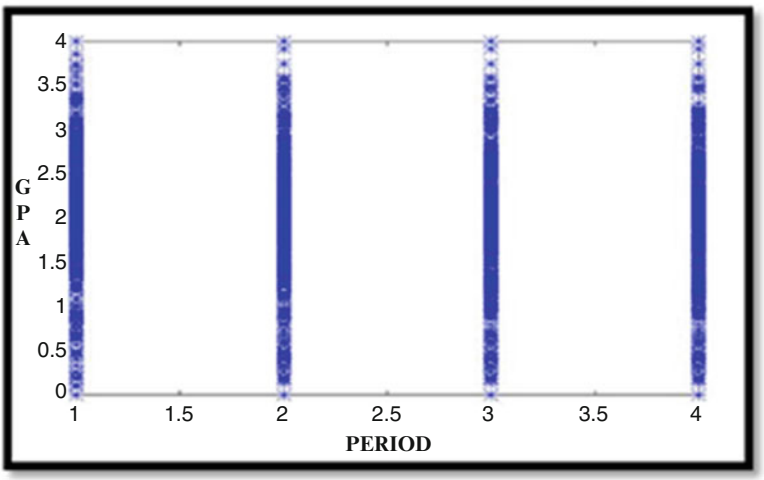


Fig. 45.5 Linear smoothing no-retention data

we can see the lowest and highest values are removed, and 6 years data have been smoothed into three periods (four data points).

The no-retention data dimension is bigger than retention data, and the no-retention linear smoothing graph is shown as Fig. 45.5 as star.

2.5 Wavelet Transform

After the retention data have been smoothed by linear filters, we used Haar transform processing. Haar transform's decompositions can be written as below [4].

```
hc1= r3(:,1:2)
[m n]=size(r)
for i=1:m
hc1(i,1)=0.5*r3(i,1)+0.5*r3(i,2);
hc1(i,2)=0.5*r3(i,3)+0.5*r3(i,4);
end
clf
hold
p2=[1 2]
for i=1:m
plot(p2,hc1(i,:),'*');
end
```

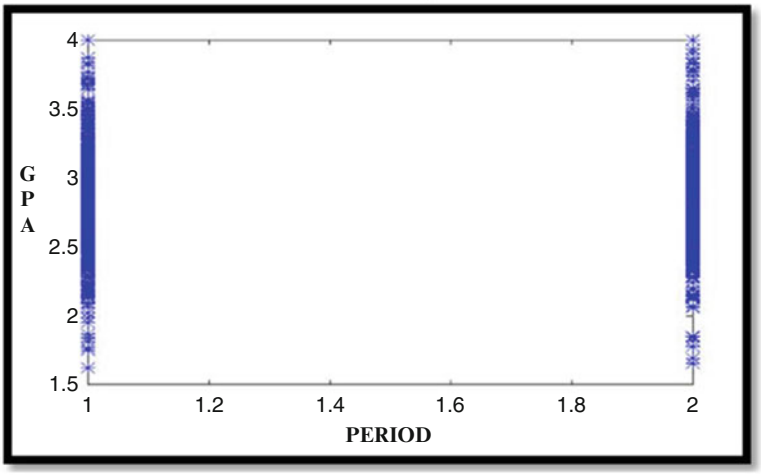


Fig. 45.6 Haar average 1

$$\begin{aligned} c(n) &= 0.5 * y(2n) + 0.5 * y\,(2n+1) \\ d(n) &= 0.5 * y(2n) - 0.5 * y\,(2n+1) \end{aligned} \tag{45.2}$$

Where c(n) is average of the pairs of data, and d(n) is their differences.

1. The First Level Decomposition

Used the Haar algorithm above applied to the retention data and smoothed data, and the new data for the first level decomposition of the average GPA are shown below as star in Fig. 45.6.

Using the Haar algorithm above applied to the retention data and smoothed data, and the new data for the first level decomposition of the difference are shown as Fig. 45.7 as rhombus.

```
hd1= r3(:,1:2)
for i=1:m
hd1(i,1)=0.5*r3(i,1)-0.5*r3(i,2);
hd1(i,2)=0.5*r3(i,3)-0.5*r3(i,4);
end
clf
hold
p2=[1 2]
for i=1:m
plot(p2,hd1(i,:),"d");
end
```

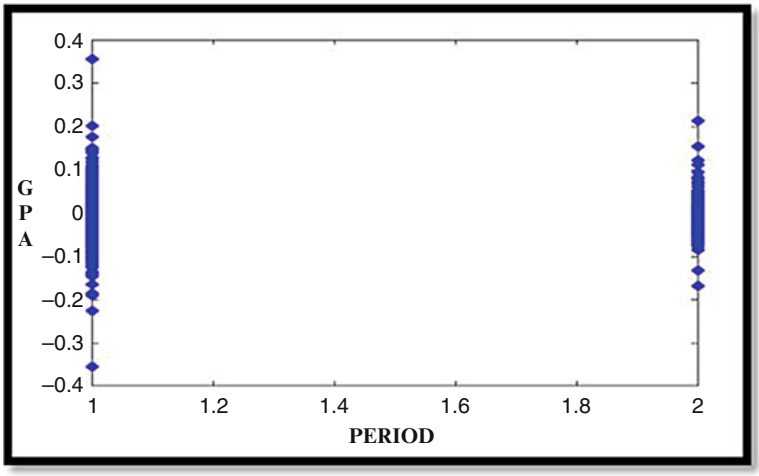


Fig. 45.7 Haar difference 1

2. The Second Level Decomposition

We applied the Haar transform above to the first level decomposition retention data, and the second level decomposition retention average, and difference are shown below.

The average points are shown as star on the top, and the difference points are shown as rhombus on the bottom in Fig. 45.8.

3 The Retention Data Representation

The retention data Haar processes are shown in Fig. 45.9.

```
hc2=r3(:,1:1)
hd2= r3(:,1:1)
for i=1:m
hc2(i)=0.5*hc1(i,1)+0.5*hc1(i,2);
hd2(i)=0.5*hc1(i,1)-0.5*hc1(i,2);
end
clf
hold
for i=1:m
plot(1,hc2(i),"*");
plot(1,hd2(i),"d");
end
```

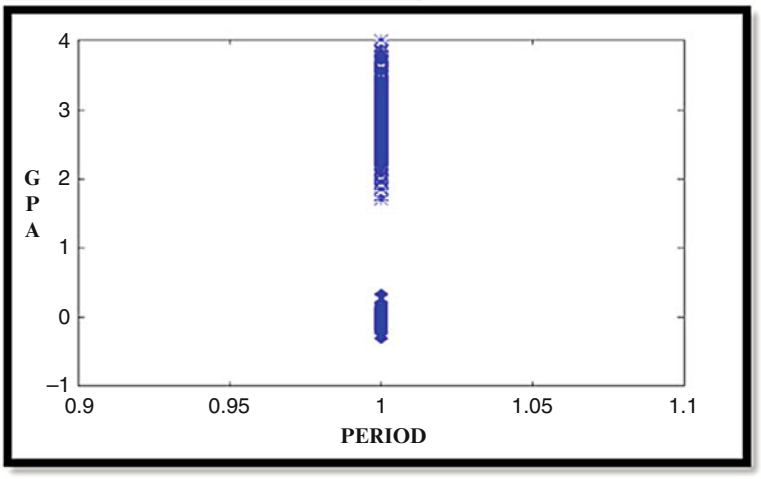


Fig. 45.8 Haar average and difference 2

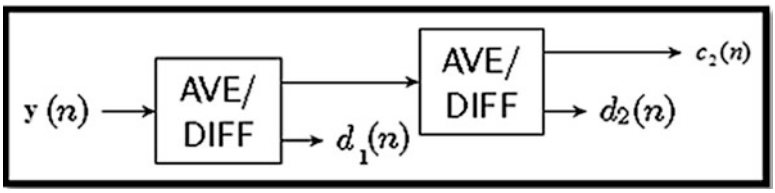


Fig. 45.9 The Haar retention processes

```
clf
hold
for i=1:m
plot(p2,hd1(i,:),"d");
end
for i=1:m
plot(3,hc2(i),"*");
plot(3,hd2(i),"d");
end
```

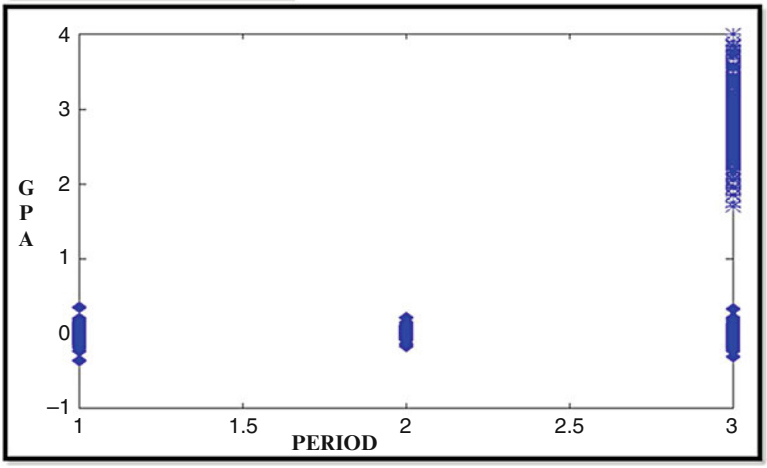


Fig. 45.10 The Haar retention representation

Where $d_1(n)$ is the first level decomposition GPA difference, $d_2(n)$ is the second level decomposition GPA difference, and $c_2(n)$ is the second level decomposition GPA average, and the retention representation is shown in Fig. 45.10. The average points are shown as star on the top, and the difference points are shown as rhombus on the bottom [4].

4 Results

By using the results from the second level Haar transform over the entire student population, we computed the average GPA and the difference for retention students.

In Fig. 45.11, the average point is shown as star on the top, and the difference point is shown as rhombus on the bottom.

```
tgpa=0
tdiff=0
for i=1:m
tgpa=tgpa+hc2(i);
tdiff=tdiff+hd2(i);
end
atgpa=0
atdiff=0
clf
hold
atgpa=tgpa/m
atdiff=tdiff/m
plot(1,atgpa,"*")
plot(1,atdiff,"d")
atgpa
atdiff
```

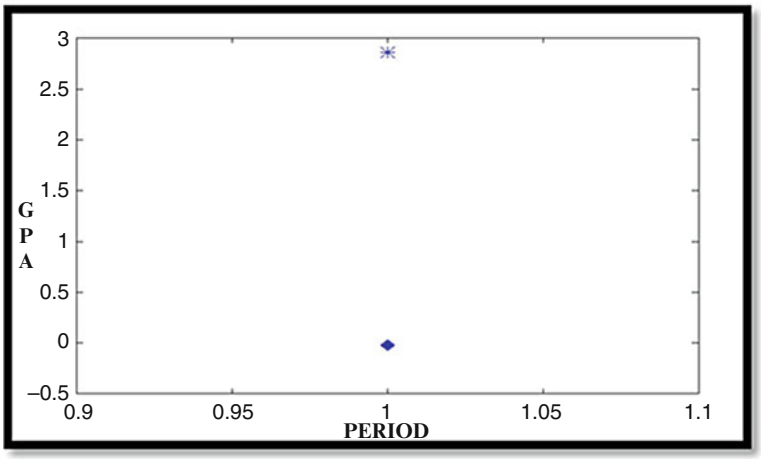


Fig. 45.11 The average GPA and difference

Table 45.3 Tested result

	Average GPA	Retention accuracy (%)
Naïve Bayes	2.8864	74.8
Haar	2.8597	75.6

We tested the Haar average retention GPA using test data set (the HBCU Fall 2007 to Fall 2012 student test data set with 820 instances) and compared the result to Naïve Bayes mean value. The results are given in Table 45.3. Haar based classification is better than Naïve Bayes.

5 Conclusions

Smoothing the data removed the highest and lowest GPA values for both of retention and no-retention data. The algorithm filtered out the noise and made the data more pure.

From the Haar transform's result we can say that the average GPA for the HBCU undergraduate student retention should be 2.8597, and the average difference should be −0.023307.

References

1. Stearns, S.D., Hush, D.R.: Digital Signal Processing with Examples in MATLAB, 2nd edn, pp. 335–356. CRC Press (2011)
2. Jia, J.W., Mareboyana, M.: Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013, pp. 222–227. San Francisco 23–25 Oct 2013
3. Gao, R.X., Yan, R.: Wavelets: Theory and Applications for Manufacturing, ISBN 978-1-4419-1544-3, pp. 1–220. Springer (2011)
4. Selesnick, I.W.: Wavelet Transforms – A Quick Study, Physics Today Magazine, pp. 1–11 (2007)
5. Jia, J.W., Mareboyana, M.: Undergraduate Student Retention Using Wavelet Decomposition, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 562–566. London 2–4 July 2014
6. Hagedorn, S.L.: How to define retention: a new look at an old problem. In: Seidman, A. (ed.) College Student Retention: Formula for Student Success, pp. 1–26. Praeger Publishers, Westport (2005)
7. Mitchell, T.: Machine Learning, pp. 52–122. McGraw Hill, New York (1997). ISBN 0070428077
8. Dunham, M.H.: Data Mining Introductory and Advanced, pp. 92–103. Prentice Hall, Upper Saddle River (2003). doi:0-13-088892-3
9. Russell, S., Norvig, P.: Artificial intelligence, a modern approach, 3rd edn., Pearson, ISBN-13: 978-0-13-604259-4, ISBN-10: 0-13-604259-7, pp. 727–748 (2010)
10. Abu-Mostafa, Y.S., Magdon-Lsmail, M., Lin, H.: Learning from Data. ISBN 10:1-60049-006-9, AMLbook, pp. 39–75 (2012)
11. Alkhasawneh, R., Hobson, R.: Modeling student retention in science and engineering disciplines using neural networks. In: IEEE Global Engineering Education Conference (EDUCON), Amman, Jordan, pp. 660–663 (2011)
12. Stone, D.B.: African-American males in computer science – examining the pipeline for clogs, The School of Engineering and Applied Science of the George Washington University, Thesis, pp. 48–60 (2008)
13. Frank, E.: Pruning Decision Trees and Lists. Department of Computer Science, University of Waikato, Thesis, pp. 29–74 (2000)
14. Yu, C.H., Digangi, S., Jannasch-pennell, A., Kaprolet, C.: A data mining approach for identifying predictors of student retention from sophomore to junior year. J. Data Sci. **8**, 307–325 (2010)
15. Yadav, S.K., Bharadwaj, B., Pal, S.: Mining educational data to predict student's retention: a comparative study. Int. J. Comput. Sci. Inf. Sec. **10**(2), 113–117 (2012)

16. Nandeshwara, A., Menziesb, T., Nelson, A.: Learning patterns of University student retention. Expert Syst. Appl. **38**(12), 14984–14996 (2011)
17. Kumar, S.A., Vijayalakshmi, M.N.: Implication of classification techniques in predicting student's recital. Int. J. Data Mining Knowl. Manage. Process (IJDKP) **1**(5), 41–51 (2011)
18. Kabakchieva, D.: Student performance prediction by using data mining classification algorithms. Int. J. Comput. Sci. Manage. Res. **1**(4), 686–690 (2012)
19. Lin, S.: Data mining for student retention management. Consort. Comput. Sci. Coll. **27**(4), 1–6 (2012)
20. Singh, S., Kumar, V.: Classification of student's data using data mining techniques for training & placement department in technical education. Int. J. Comput. Sci. Netw **1**(4), 49–56 (2012)
21. Esposito, F., Malerba, D., Semeraro, G.: A comparative analysis of methods for pruning decision tree. IEEE Trans. Patt. Anal. Mach. Intell. **19**(5), 476–491 (1997)
22. Patil, D.D., Wadhai, V.M., Gokhale, J.A.: Evaluation of decision tree pruning algorithms for complexity and classification accuracy. Int. J. Comput. Appl. (0975–8887) **11**(2), 23–30 (2010)
23. Stanevski, N., Tsvetkov, D.: Using support vector machines as a binary classifier. International Conference Computer System Technology. CompSys Tech., 1–5 (2005)
24. Sembiring, S., Zarlis, M., Hartama, D., Wani, E.: Prediction of student academic performance by an application of data mining techniques. In: International Conference on Management and Artificial Intelligence IPEDR 6, Bali, Indonesia, pp. 110–114 (2011)

Chapter 46
Construction of Radar SAR Images from Digital Terrain Model and Geometric Corrections

Philippe Durand, Luan Jaupi, and Dariush Ghorbanzadeh

Abstract We propose in this paper an original method to correct the geometric distortions of a radar image. The comparison of satellite data, reveals specific problems. Data can be noisy, but especially the geometry of their acquisition requires corrections for comparaisons between them. In this paper we show how highly deformed radar images can be geometrically corrected and compared to map data coming from digital terrain models and also with data coming from SPOT satellite. Radar images used, are from the sensor airborne radar Varan, which is used for data acquisition campaign in the South-East of France. Applications include both structural geology, land cover or study of coastline. We propose a solution to rectify radar image in the geometry of a numerical terrain model. The method adopted here, is to produce a synthesis radar image by encoding all flight parameters of aircraft or satellite from a digital terrain model; radar image can then be compared to the synthetic image because points of landmarks can be clearly identified. Finally, we obtain a correspondence between the points of real radar image distorted, and those in the land or map.

Keywords Digital terrain model • ERS1 • Geometric corrections • Radar imagery • SAR images • SPOT

1 Introduction

The images discussed in this article come from a synthetic aperture radar imaging system: (SAR system) [5]. It was used during a mission in the south of France in order to collect data images for geology, geomorphology and land use [3, 11].

P. Durand (✉) • D. Ghorbanzadeh
Department of Mathematics (M2N), Conservatoire National des Arts et Métiers, 292 rue Saint martin, Paris 75141, France
e-mail: philippe.durand@cnam.fr; dariush.ghorbanzadeh@cnam.fr

L. Jaupi
Department of Mathematics (CEDRIC), Conservatoire National des Arts et Métiers, 292 rue Saint martin, Paris 75141, France
e-mail: leon.jaupi@cnam.fr

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_46

Radar imagery, show artifacts. Firstly, it has a multiplicative noise known as speckle, secondly it is deformed geometrically because of its acquisition. In the first part of this article, we recall these facts. We not seek here the correction of speckle noise, in fact, its removal can cost the loss of precious radiometric informations. We focus on geometric corrections. These corrections will enable us to make comparisons of satellites data and map data. We give an application on the location of oyster beds invisible on a SPOT image. We give another application in geology: it is the extraction of directs line accidents with mathematical morphology tools. From this method, it is possible to simulate radar image of aircraft or satellite missions as ERS1 or ERS2 and propose some geometrical corrections. Finally, we can consider other comparisons: we then apply the superposition of radar data and SPOT images of terrains with or without landforms. Other interesting works focus in texture analysis [8].

2 Characteristics of Airborne Radar Varan

We give here the main characteristics of radar VARAN which allowed the acquisition image data southeast of France referred to here. This is an X-band radar (3 cm). It offers better resolution than L-band radar (25 cm). The slant-range resolution is $\delta_r = 3$ m. This is a theoretical resolution. In SAR mode, azimuth resolution depend of b, the real antenna length: $\delta_{az} = b/2$, here $\delta_{az} = 3$ m also. The altitude of flight is 5,800 m. At last, there are two kinds of polarization, horizontal and vertical. For example: In the HH configuration we have emission and reception of horizontally polarized wave. For More technical details about this captor you can see [12].

3 Nature of the Radar Imaging

3.1 Speckle Noise

Consider a target with a size at least equal to the resolving power of the sensor. It is further assumed that the wavelength is large compared to the roughness, but less than that of the target. This assumption ensures that the phases of the diffuser elements are independent and uniformly distributed in the interval $[0, 2\pi]$. This target can then be analyzed as consisting of several elementary reflectors. The energy reflected from the target is the sum returned by each of the elementary reflectors energy. According to the law of large numbers, the resulting energy is a discrete random number [7]. We can then show that the radar signal follows a Rayleigh distribution.

$$F_z(r) = \frac{r}{\sigma^2} exp(-\frac{r^2}{2\sigma^2}) \tag{46.1}$$

3.1.1 Multiplicative or Additive Noise

Assuming multiplicative noise is obtain as the product of the ideal image (with pixel denoted by y). The pixel of the random image is denoted u. Under these conditions, we deduce statisticals parameters:

$$\sigma_y^2 = \frac{\sigma_z^2 - \sigma_u^2 E(z)^2}{\sigma_u^2 + 1} \tag{46.2}$$

In our study, radar image come with a pretreatment: a logarithmic calibration: statistical noise become additive and we obtain a new relation:

$$\sigma_z^2 = \sigma_y^2 + \sigma_u^2 \tag{46.3}$$

In the remainder of this paper we will not consider a study of speckle noise. We will consider only the geometric corrections that may be made to the radar image. This is why in the following, we will not add noise to our synthetic images created.

3.2 Geometric Deformations

Let's see what can be said of the deformation of a flat surface. Take four points A, B, C, D ($AB = BC = CD$) equally spaced in a perpendicular direction to the flight axis of the plane or the satellite. In the radar image, Near the vertical of the plane, the points A', B', C', D' satisfies ($A'B' < B'C' < C'D'$). A second type of deformations is caused by mountainous landscapes. When the sensor takes pictures near the top of mountains, it may happen that the reflected response from the top of the relief arrives before the response of his foot. We speak of a phenomenon of reversal on the image [10].

4 Creating a Synthetic Image in the Geometry of the Flight of the Aircraft or Satellite

We propose to correct the radar image (Fig. 46.7) on the geometry of a digital terrain model: mnt (Fig. 46.1).

4.1 Pretreatment of the Digital Terrain Model

To create a synthetic image having the same geometry as the actual scene, it must first rotate the mnt so that scanning lines of the plane corresponding to the row of the matrix containing numerical datas of digital terrain model.

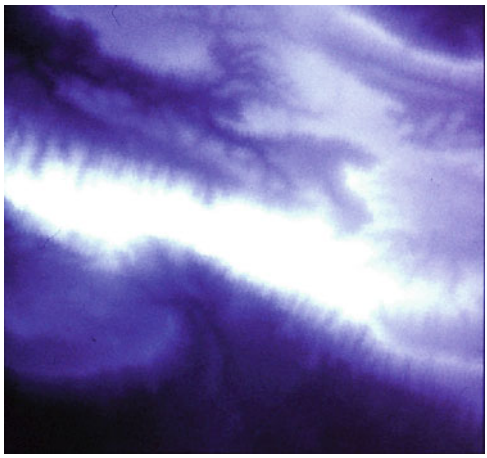


Fig. 46.1 Sainte Victoire mnt

The second step is to improve the resolution of mnt. The matrix of mnt, is small (300×400 points). We increase the mnt size, with a bicubic resampling of the matrix. Size is multiplied by two. The method is based on a spline interpolation of degree 2 [1]. The bicubic even kernel is given by:

$$H(x) = \begin{cases} 1 - 2x^2 + x^3 & 0 \leq x \leq 1 \\ 4 - 8x + 5x^2 - x^3 & 1 \leq x \leq 2 \\ 0 & otherwise \end{cases} . \tag{46.4}$$

We can then resample the digital function by writing the following convolution product:

$$f(x) = \sum f(n) * H(x - n) \tag{46.5}$$

The result is the image of the digital terrain model resampled below:

Third step: it can be useful to implement a routine rotation of the digital terrain model in order to find the exact geometry of the acquisition of the radar image (Fig. 46.2).

4.2 Synthesis of a Radar Image

Then a radar image is synthesized using the parameters of the plane or satellite: We need first of the following parameters:

- Altitude of the aircraft: A_{av}
- Aperture of the antenna: A_{pant}
- Angle to the nearest terrain feature: Φ_{min}

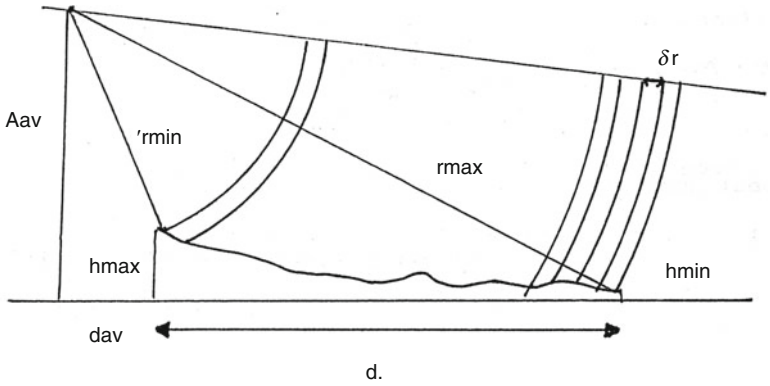


Fig. 46.2 Geometry of radar image

– Angle the more distant land: Φ_{max}
– The slope of the antenna: Φ_{moy}
– Distance of the aircraft D_{av}

Then, it is then possible to calculate other parameters:

– Φ, angle between the scan trace, and the vertical of the plane:

$$\Phi = arctan(\frac{D_{av} + x}{A_{av} - z}) \tag{46.6}$$

r-distance between the plane and the mnt-point:

$$r = \sqrt{(D_{av} + x)^2 + (A_{av} - z)^2} \tag{46.7}$$

Θ: angle of incidence:

$$\Theta = arcos\frac{derx.sin(\Phi) + cos(\Phi)}{\sqrt{derx^2 + dery^2 + 1}} \tag{46.8}$$

Where *derx*, *dery* are partial derivatives associate to mnt. From these parameters, we may, at any point of a scan line, know the total backscattered energy. This energy is given by the following equation:

$$E = \sigma_0 cos(\Theta)\sqrt{1 + derx^2} \tag{46.9}$$

4.2.1 Airborne SAR Image, Satellite SAR Image

In this section, we consider a rotation of the digital terrain model of Fig. 46.1, we obtain Fig. 46.3. Suppose that we fly over from the top down of Fig. 46.3.

Fig. 46.3 mnt rotated

Fig. 46.4 Synthesis

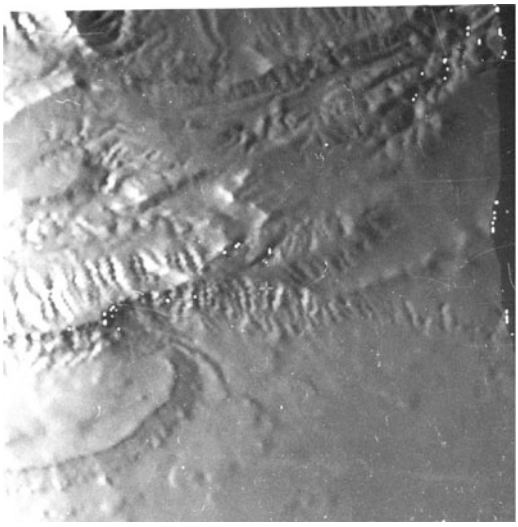


Figure 46.4 is obtained for the altitude 5,700 m (case imaging radar Varan).

Figure 46.5 for an altitude of 780 km (Case of a satellite ERS-1 or ERS-2).

We do not study here the correction of satellite images, because we do not have satellite images of ERS-1 and ERS-2 (Fig. 46.5).

If the resolution r is less than the mesh of the mnt, energy is recognized as the energy returned by several pixels. During the synthesis, we share the energy received on these pixels. We will now make a readjustment, between the real scene and the synthesized one from the digital terrain model and flight parameters of the aircraft picture.

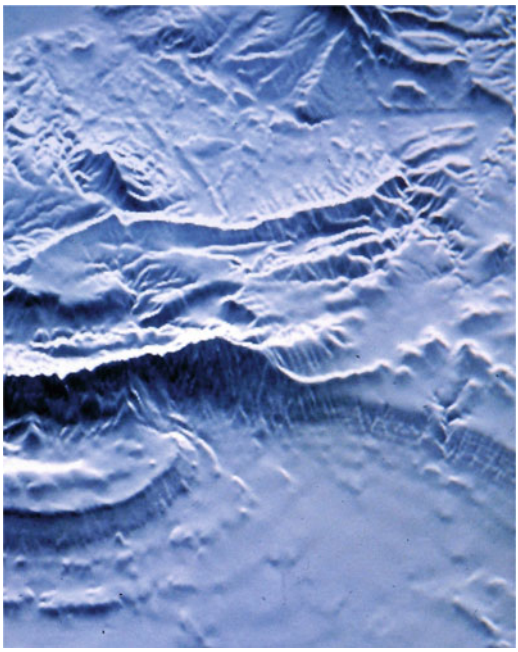


Fig. 46.5 Synthesized image at the altitude of satellite ERS

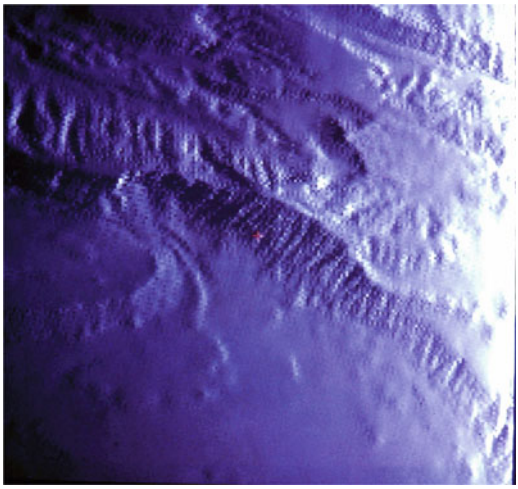


Fig. 46.6 Synthesized image

4.3 Readjustment of the Real Image on the Synthesized Image

From the digital terrain model (Fig. 46.1), we obtain the synthesis image (Fig. 46.6). The synthetic image is the ideal image that would be obtained without speckle noise. The readjustment of the real image VARAN on this image synthesis should help rectify errors flight of the aircraft. The polynomial model of degree 2 is well suited for

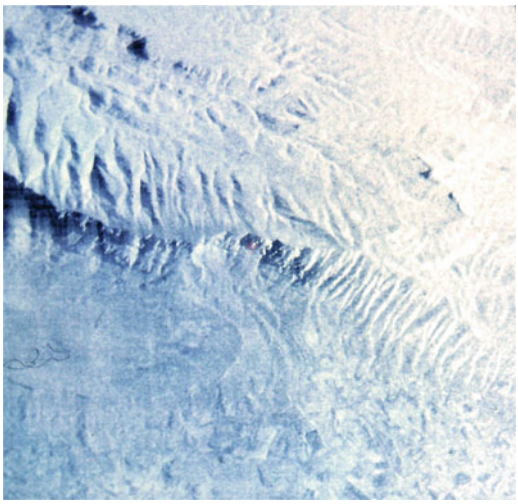


Fig. 46.7 Real scene

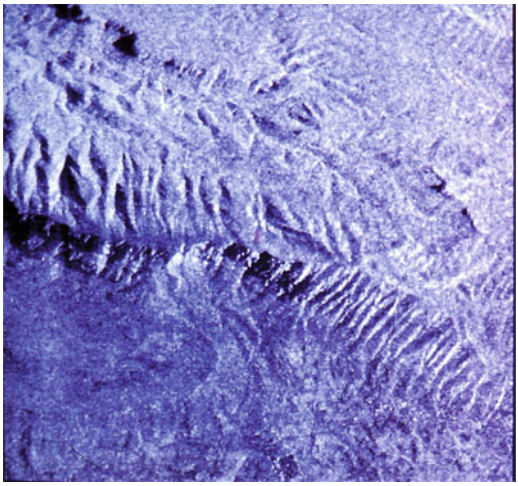


Fig. 46.8 Readjustment of the real image on the synthesized

this first correction (Fig. 46.7). The average differences in x and y are approximately one pixel. A difference of one pixel from the exact point is a mandatory approximation [6]. Therefore, the readjustment can be considered very good (Fig. 46.8).

However, the mnt resampled, the accuracy is 25 m, for a more precise result, We would need to have more accurate digital terrain models.

4.4 Readjustment of the Real Image on mnt

The correspondence between the synthetic image and the digital terrain model is exact (Fig. 46.9).

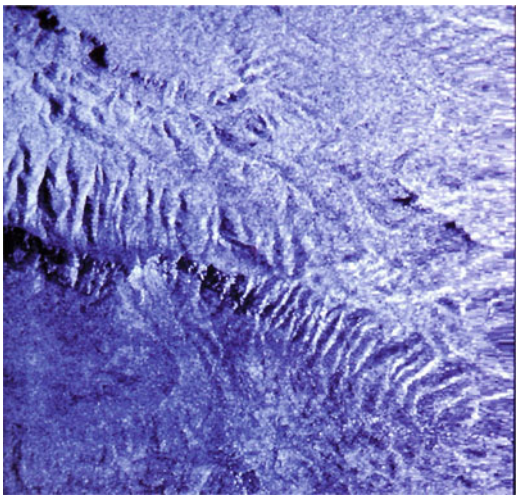


Fig. 46.9 Readjustment of the real image on the mnt

With the readjustment of the previous section, can be matched without new errors, the real scene and the digital terrain model. The method just described is interesting. However, we do not have mnt accurate enough to allow a systematic readjustment.

5 Combinations Between SPOT and VARAN Data With or Without Relief

The overlay data from multiple sensors, allows a better analysis in geology and geomorphology. It should provide a best understanding of the respective contributions of the various sensors.

5.1 Example 1

The first stage, Fig. 46.10, represents the ponds (Tau area), it shows a flat region without much distortions. Only remain, those from the radar geometry in the SAR mode, but not reversals due to reliefs. In this case, the radar and SPOT geometries are similar, in addition, the site has special features points that can be selected as landmarks: Cape border of cultivated plots, contour of the pond. A possible application could be followed in real time by satellite, drift oyster beds. Indeed, the image that we present shows the overlap between radar data and SPOT data. The park oysters, submerged do not appear on the SPOT image, however, are clearly visible on the radar image.

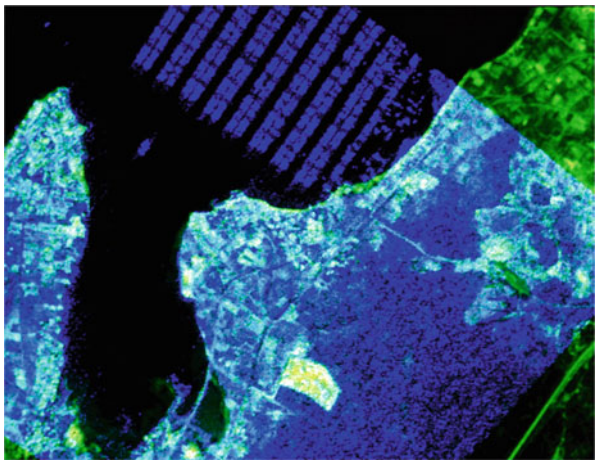


Fig. 46.10 Comparison radar-spot on "Etangs de Tau"

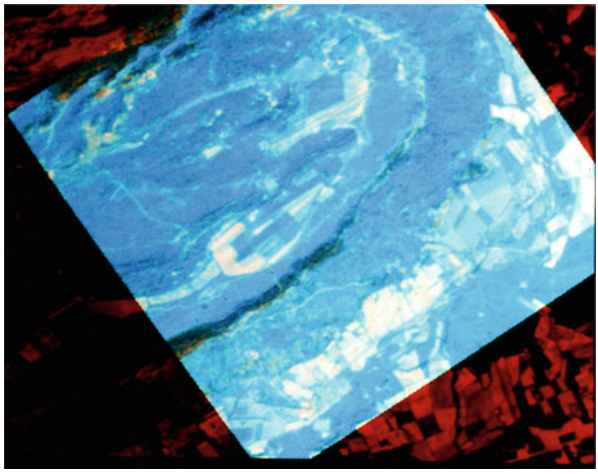


Fig. 46.11 Comparison radar-spot Sainte Victoire mountain

5.2 Example 2

We also propose a methodology to reset the radar image on a SPOT image in relief area. At first, the distortions of the radar image is corrected by restoring the geometry of the digital terrain model (see previous section). Then, the overlap between the corrected radar image and SPOT Image is performed. This is a delicate operation, with the first readjustment errors, add those from the SPOT image (not exactly in the geometry of the mnt). The acquisition of SPOT mnt on this region would allow us to solve this last source of errors. The relief features visible on the radar image are superposed on the visible areas of SPOT imagery (see Fig. 46.11).

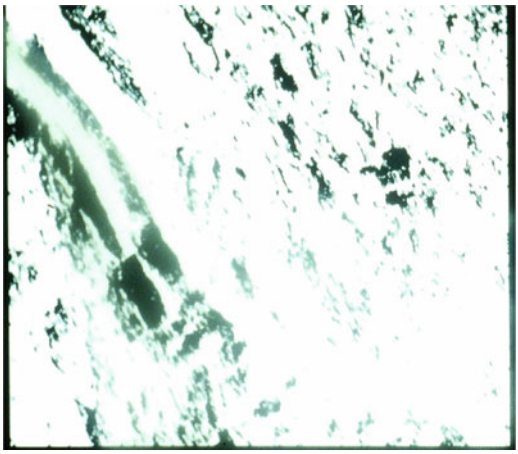


Fig. 46.12 Direction of fracturation north 160

6 Application of Mathematical Morphology Extraction Directions Geological Accidents

The extraction of networks is a central topic of a great importance in remote sensing. From the radar scene acquired on the Massif de Bras (Fig. 46.12) (Luc southern France) were extracted three directions of fractures affecting the Lias and partly Triassic. They correspond to real families accidents. Two of these families were known. The radar scene confirmed their existence, in addition, the third direction (North 160) is to be highlighted. We can strengthen these directions, by manufacturing of structuring elements, and the application of morphological operations in each direction [2, 4, 9].

7 Conclusion

The geometric rectification a radar image on the digital terrain model is the starting point that should be used to compare images with different geometries. Spot imaging is less sensitive to geomorphologic accidents. Instead, the radar image, can solve this problem. On the other hand climatic conditions do not affect collect of radar data. It is therefore essential to have more effective tools to compare the data between them. We can only encourage research to efficient pretreatments. we need to develop tools that would help to improve the fusion of different satellite data, and better understand and use data from different sensor remote sensing.

References

1. Bernstein, R.: Image geometry and rectification manual of remote sensing manual of remote sensing. Am. Soc. Photogr. **21**, 873–921 (1983)
2. Cuq, F., Durand, P., Hamidou, S., Simonin, A.: Utilisation de filtres géostatistiques pour l"analyse de changement d'échelle à partir d'images satellitaires. Photointerprétation Eska Paris **6**, 33–38 (1990)
3. Durand, P., Mekarnia, A., Chorowicz, J.: Télé-analyse géologique du radar aéroporté VARAN sur la montagne Sainte Victoire. Photointerprétation Eska Paris **27**(6), 1–10 (1988)
4. Durand, P., Hakdaoui, M., Chorowicz, J., Rudant, J.P., Simonin, A.: Caractérisation des textures urbaines sur image radar Varan par approche morphologique et statistique. Application à la ville du Luc. Int. J. Remote Sens. **15**(5), 1065–1078 (1994)
5. Durand, P., Jaupi, L., Ghorbanzadeh, D.: Geometric correction of airborne radar SAR image on a digital terrain model, and overlay with satellite SPOT data. In: Proceedings of the World Congress on Engineering (WCE 2014), London. Lecture Notes in Engineering and Computer Science, pp. 572–576 (2014)
6. Garinet, J.Y.: Techniques de recalage d'images de télédetection et transparence. Thesis, Toulouse (1985)
7. Goodman, J.W.: Some fundamental properties of speckle, J. Opt. Soc. Am. **66**(11), 1145–1150 (1976)
8. LLoyd, C.D., Berberoglu, S., Curran, P.J., Atkinson, P.M.: A comparison of texture measures for per-field classification of mediterranean land cover. Int. J. Remote Sens. **25**(19), 3943–3965 (2004)
9. Matheron, G.: Eléments pour une théorie des milieux poreux. Masson, Paris (1967)
10. Rudant, J.P., Chorowicz, J., Durand, P.: Problèmes d'interprétation géomorphologique et modélisation géométrique d'images radar à partir d'un modèle numérique de térrain, C.R.A.S,, Gauthier Villars, **306**(1), 15–20, Paris (1988)
11. Rudant, J.P., Cervelle, B., Chorowicz, J., Durand, P., Kamoun, P., Louhala, S., Polidori, L., Riazanoff, S., Simonin, A., Tannous, I.: Evaluation des données VARAN en géologie, géomorphologie sur le sud est de la France. In: Proceedings of 4th International Colloquium on Spectral Signatures of Objects in Remote Sensing (ESA), Aussois (1988)
12. Vaillant, D., Varan, S.: An airborne synthetic aperture radar for research in microwave remote sensing. In: Proceedings of EARSEL/ESA Symposium, Strasbourg, pp. 167–172 (1985)

Chapter 47
Traffic Light Control Based on the Road Image Analysis

Obada M.A. Bani Issa, Venus W. Samawi, and Jehad Q. Alnihoud

Abstract The automatic analysis of traffic scene is interesting subject in the context of traffic planning and monitoring. The main objective of this research is to construct an advanced traffic control system according to the changes of road traffic circumstances. The road traffic changes are specified through analyzing the road's image and identify the traffic load. Two approaches are suggested to specify traffic load. The first approach is measuring the length of car-queues on road based on edge detection using Sobel operator then estimate number of vehicles. The second approach is counting-vehicles on the road based on region growing segmentation algorithm. An equation to specify the estimated time is suggested to determine the time for green-light. The two developed techniques are compared from two points of views: estimate time for green-light period, and the estimated number of vehicles on the road. The impact of density of vehicles on road (low/high) is taken in consideration in comparison process. As a result, it was found that car-queue technique suits high density road, while counting-vehicles technique suits both cases (high and low density) and shows better performance, which is comparable to actual traffic results.

Keywords Automatic traffic light control • Dynamic traffic control • Edge detection • Image processing • Image segmentation • Vehicles queue length • Vehicles counting

O.M.A. Bani Issa
Department of Computer Science, Al-Albayt University, P.O. Box 130040, Mafraq 25113, Jordan
e-mail: alaishe@yahoo.com

V.W. Samawi (✉)
Department of Computer Information System, Amman Arab University, P.O. Box 2234, Amman 11953, Jordan
e-mail: venus@aau.edu.jo

J.Q. Alnihoud
Department of Computer Science, Al-Albayt University, P.O. Box 130040, Mafraq 25113, Jordan
e-mail: jehad@aabu.edu.jo

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_47

1 Introduction

As driving around town on daily travel, one may find himself stuck in traffic and receiving poor gas mileage. One of the main reasons could be the poor design of the traffic light system. Traffic signals must be instructed when to change phase. They can also be coordinated so that the phase changes occur with respect to traffic monitoring, and nearby signals. Mainly, there are two types of traffic control: fixed time control (phase changing in specified period of time), and dynamic time control (phase changing based on traffic monitoring). One of the major problems concerning traffic control is to provide a dynamic system that makes decision when to change the traffic signal phase through specifying the jam points in the road [1].

Due to the importance of real time (dynamic) traffic control, many researchers investigated the real time vision based transportation surveillance system.

The dynamic traffic control systems should analysis the traffic on urban road, detect the objects (cars), and then count number of cars. After that, extrapolate the transportation information of the main urban road [2–4]. Alvaro Soto and Aldo Cipriano [5] used a computer vision system for the real time measurement of traffic flow. The traffic images are captured by a video camera and digitized into a computer. The measuring algorithms are based on edges detection and comparison between a calculated reference without vehicles and the current image of traffic lanes. Tests under real traffic conditions were satisfactory, with over 90 % of accuracy and error below 5 %. Murphey et al. [6] present an intelligent system, Dyta (dynamic target analysis), for moving target detection. Dyta consists of two levels of processes. The first level, Dyta attempts to identify possible moving objects and compute the texture features of the moving objects. At the second level, Dyta inputs the texture features of each moving object to a fuzzy intelligent system to produce the probability of moving targets. In Dyta, three algorithms were used, moving target tracking algorithm, Gabor multi-channel filtering, and fuzzy learning and inference. In 2005, Lawrence Y. Deng et al. [7], integrated and performed vision based methodologies that include the object segmentation. Classify and tracking methodologies were used to know well the real time measurements in urban road. According to the real time traffic measurement, the adaptive traffic signal control algorithm to settle the red–green switching of traffic lights both in “go straight or turn right” and “turn left” situations is derived. The experimental result confirms the efficiency of vision based adaptive TSC approach. In the experiment results, they diminished approximately 20 % the degradation of infrastructure capacities. In 2008, Richard Lipka, Pavel Herout [8], implement light signalization in urban traffic simulator JUTS. They use JUTS in experiments dealing with impact of time plans to traffic situation. In 2011, Choudekar and Banerjee [9] detect vehicles using image processing. To do so, Prewitt edge detection operator has been carried out. Traffic light durations are controlled based on the percentage of image matching. In 2014, Bani Issa et al. developed automatic traffic control system based on image processing. Estimated time for traffic light duration is decided based on finding queue length in two lane road. The experimental results of the developed system

are compared with the performance of the actual traffic light control system. The developed queue length approach proved its efficiency [10].

The main objective of the proposed research is to construct a fast dynamic decision traffic-light model by analyzing road image, and identify the traffic load on the road. Two different approaches are developed and their performances are compared. The first approach is based on finding queue length of each lane on the road (find car-queue length) using edge detection. The second approach is counting number of vehicles after performing image segmentation. For each approach, the estimated time needed for cars to pass the traffic light region is calculated (i.e. time for car starting-up, moving from the current point to the next, and passing the traffic region). The time estimation equation is constructed based on experiment and consultation of the traffic engineer office in Amman municipality. To verify the efficiency of the developed approaches, the experimental results of each of the two approaches are compared with the performance of the actual traffic-light control system. Comparison points are number of vehicles, and estimated time for green-light period. Empty level, cars beside other, and distance between cars are also considered.

2 Methodology

Various monitoring systems are used to specify the load on a road (number of vehicles on road). To determine number of vehicles on a road, these systems could use infrared detectors, cable to count wheels, or analyze road image (which is the core of this research). In this research, traffic jam is identified through analyzing the road image taken by a digital camera (with specific resolution and from definite distance). Two approaches are developed to find out traffic jam. The first approach is finding the length of vehicles-queue on a lane, while the second approach counts number of vehicles on a lane based on object discrimination. At first, noise is removed from the road image (image pre-processing), then analyze the pre-processed image to determine the current traffic situation. Based on road image analysis, a signal will be sent to the traffic-lights control center to change the traffic condition. The developed model mainly consists of three phases (see Fig. 47.1): preprocessing phase, image analysis phase, and timing decision phase.

3 Preprocessing Phase

In this research, the empty road image is needed (background image) in addition to the image taken every period of time (traffic image). Before starting with image analysis, preprocessing steps are applied (see Fig. 47.2):

1. Read image data for base image and traffic image. The system will accept as an input image of any size. The image is resized to 897×830 to get standard image size.

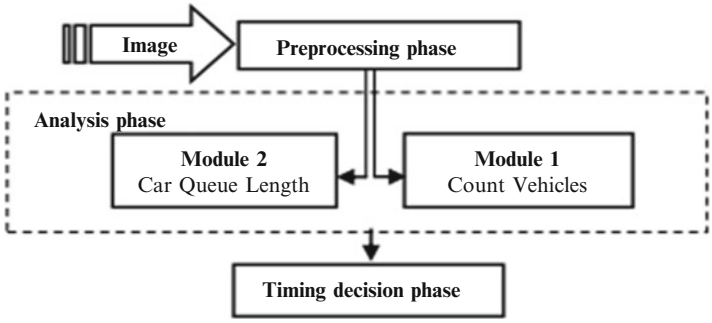


Fig. 47.1 System model

2. Convert both background image and traffic-image to gray-scale form using Eq. 47.1 [11, 12].

$$\text{Grayscale} = 0.2989 \times \text{R} + 0.5870 \times \text{G} + 0.1140 \times \text{B} \tag{47.1}$$

3. Create mask to extract the *region of interest* (ROI). This method is intended to separate the part of the road where vehicles are moving in one direction. This action is essential because it simplifies the processing of information extracted from more than one image. Masking algorithm is given by Eq. 47.2 [12]

$$\text{N}(\text{p}) = \text{M}(\text{p}) \times \text{V}(\text{p}) \tag{47.2}$$

where M(p) is an image point value in primary frame; N(p) is a new image point in the output image, and V(p) is mask value for point p: V(p) = 0 if the corresponding pixel is eliminated, otherwise V(p) = 1.

4. Image Subtraction: used to calculate the difference between the input image and the background image (subtract background image from road image). The resulted image contains only vehicles (no background road image). The output pixel values are given by Eq. 47.3 [12]:

$$\text{D}(\text{X}, \text{Y}) = \text{C}(\text{X}, \text{Y}) - \text{B}(\text{X}, \text{Y}) \tag{47.3}$$

where D(X,Y) is the difference image, C(X,Y) is the traffic image, and B(X,Y) is the background image.

5. Convert resulted image D to binary image using threshold value. Apply image thickening algorithm on the binary image R to prevent total erasure, and to ensure connectivity of edges. The background difference image D needs to be changed into binary image using Eq. 47.4 [11, 12].

$$Ri\,(x, y) = \begin{cases} 0, \text{ if } |D_i\,(x, y)| < T \\ 255 \qquad\quad \text{otherwise} \end{cases} \tag{47.4}$$

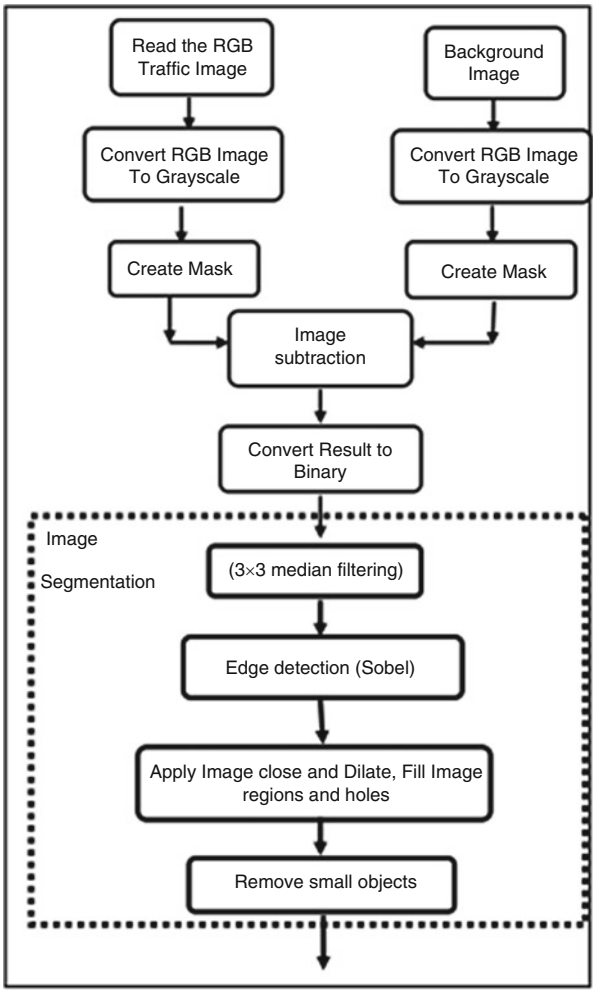


Fig. 47.2 Preprocessing phase

where $D_i(X,Y)$ is the difference image, $R_i(x,y)$ is a binary image, and T is a threshold. In this research, by experiments, it was found that the best T value is 20.

6. Noise Filtration: After image subtraction process, the resulted image has a lot of speckles caused by noise, which could be removed only by means of filtration. To perform noise filtration, median filter with 3×3 window is used.
7. Edge detection: Sobel edge detection filter is used [13].
8. Use morphological operations (dilation and erosion). Dilation and erosion are related operations, although they produce very different results. Dilation adds pixels to the boundaries of objects (i.e., changes them from off to on). Erosion removes pixels on object boundaries (changes them from on to off).

9. Filling small holes in objects and closing short gaps in strokes using the majority black operator.
10. Remove small objects <3,000 pixels (3,000 was chosen based on the used image resolution). Smaller objects could be human, animals, rock or any other thing but vehicle.

4 Analysis Phase

In this phase, the preprocessed road image will be analyzed using two modules: car-queue length module and counting vehicles module. The output of each module is the estimated number of cars on the road.

4.1 Counting Vehicles Module

To make the segmentation method effective and useful, at first separate vehicles from background, then count out number of objects (vehicles) in the image. To do so, the following conditions will be considered:

- Let SL be segment length, SW be segment width, and SIZE be segment size = SL × SW, N = number of vehicles (initially N = 0), C = number of vehicles next to each other, and NV is the total number of vehicles.
- If SL = 3 multiples of SW or SIZE >15,000 then the object is considered as a vehicle; N = N + 1
- If 22,000 > size ≥ 15,000, then the object is considered as a long or two attached vehicles; N = N + 2
- If 27,000 > SIZE ≥ 22,000, then the object is considered as a long or three attached vehicles; N = N + 3
- If SIZE ≥ 27,000, then the object is considered as a long or four attached vehicles; N = N + 4

Choosing the threshold numbers (22,000, 27,000, and 15,000) depend on the image resolution.

The important point is that what is needed is the number of vehicles of the longest lane not all objects on all lanes of the road. To solve this point, find-out the center-point of each object; compare this point with the start and end points of the other vehicles to find out which vehicles are almost in the same level of this vehicle. In this case, count them all as one object since almost all vehicles at the same level will need the same time to move. Figure 47.3 illustrates the flow-graph of counting vehicle module.

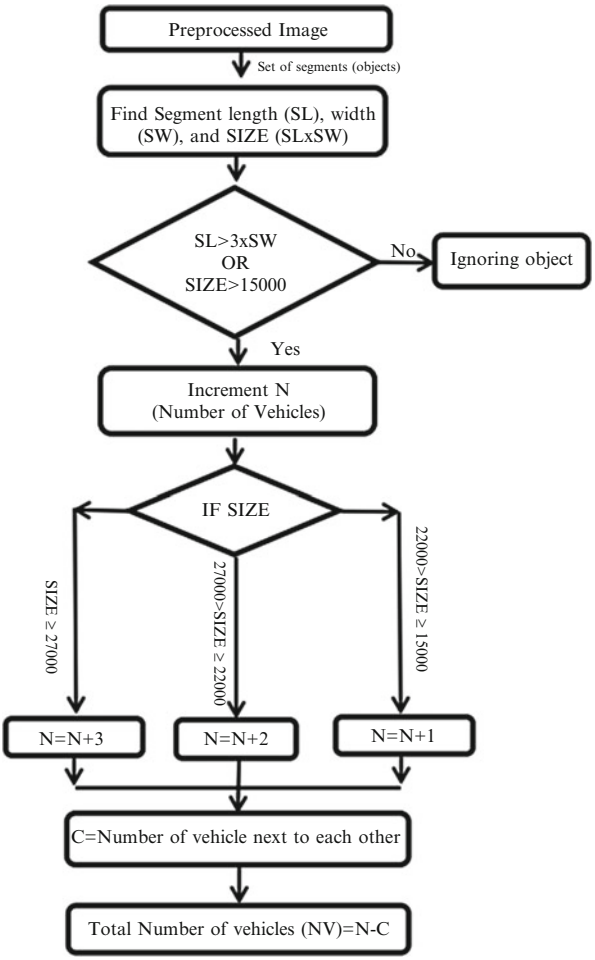


Fig. 47.3 Counting vehicle module

4.2 Queue Length Module

Traffic queuing length could be found by dividing the road into regions depending on number of road lanes, and lanes width. In this work, the road consists of two lanes (right and left). Each part is scanned top-down and bottom-up to calculate the queue length. The main steps of this model are [10]:

1. Find the first pixel in first object (scan bottom-up and register the first white dot), and last pixel in the last object (scan image top-down and register the first white dot).

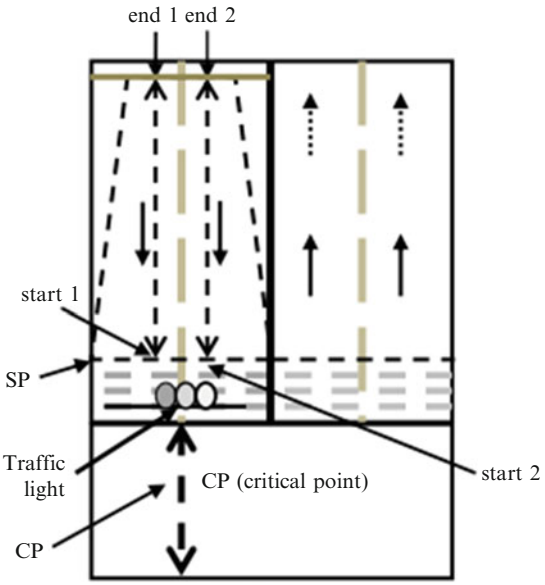


Fig. 47.4 Road map

2. Find the distance between first pixel and last pixel of the object, which represents the queue length (as will be shown later).
3. Compute how many vehicles in the road by dividing the queue length by the assumed vehicle length. The assumed vehicle length is found out by calculating different objects sizes in the given images after segmentation. Different image resolutions will affect the assumed object size.

The length of the queue has been calculated in six ways (for street that has two lanes as shown in Fig. 47.4).

The distance between two points of the XY-plane can be found using Eq. 47.5, which calculates the distance between (x_1, y_1) and (x_2, y_2) [14].

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{47.5}$$

Let the first column be Start1; the start of the second column be Start2; the end of the first column End1; the end of the second column End2; the start point of the street (SP); the beginning of the queue from SP is SR. The empty space between the queues is EL, and the safety zone CP. In this work, six cases is considered as queues categories, these are:

$$\left.\begin{aligned} &\text{Case 1}: \text{End}_1 < \text{End}_2 \text{ and } \text{Start}_1 > \text{Start}_2 \\ &\quad \text{D} = |\text{End}_2 - \text{Start}_2| \\ &\quad \text{SR} = |\text{SP} - \text{End}_2| \end{aligned}\right\} \tag{47.6}$$

$$\left.\begin{array}{l}\text{Case 2}: \text{End}_1 > \text{End}_2 \text{ and } \text{Start}_1 < \text{Start}_2 \\ \quad \text{D} = \left|\text{End}_1 - \text{Start}_1\right| \\ \quad \text{SR} = |\text{SP} - \text{End}_1|\end{array}\right\} \tag{47.7}$$

$$\left.\begin{array}{l}\text{Case 3}: \text{Start}_2 < \text{End}_1 \text{ and } \text{End}_1 < \text{End}_2 \text{ and } \text{Start}_1 < \text{Start}_2 \\ \quad \text{D} = |\text{End}_2 - \text{Start}_1| \\ \quad \text{SR} = |\text{SP} - \text{End}_2|\end{array}\right\} \tag{47.8}$$

$$\left.\begin{array}{l}\text{Case 4}: \text{End}_1 > \text{End}_2 \text{ and } \text{Start}_1 < \text{End}_2 \text{ and } \text{Start}_1 > \text{Start}_2 \\ \quad \text{D} = |\text{End}_1 - \text{Start}_2| \\ \quad \text{SR} = |\text{SP} - \text{End}_1|\end{array}\right\} \tag{47.9}$$

$$\left.\begin{array}{l}\text{Case 5}: \text{End}_1 < \text{Start}_2 \\ \quad \text{D} = (\text{End}_1 - \text{Start}_1) + (\text{End}_2 - \text{Start}_2) \\ \quad \text{EL} = |\text{Start}_2 - \text{End}_1| \\ \quad \text{SR} = |\text{SP} - \text{End}_2|\end{array}\right\} \tag{47.10}$$

$$\left.\begin{array}{l}\text{Case 6}: \text{End}_2 < \text{Start}_1 \\ \quad \text{D} = (\text{End}_1 - \text{Start}_1) + (\text{End}_2 - \text{Start}_2) \\ \quad \text{EL} = |\text{Start}_1 - \text{End}_2| \\ \quad \text{SR} = |\text{SP} - \text{End}_1|\end{array}\right\} \tag{47.11}$$

In Eqs. 47.10 and 47.11, we know that there is space between queues. We need to calculate the distance and deduce the time needed by the vehicle to cross through. After obtaining the queue length of the two lanes, the queue lengths are divided by the assumed car length to find the expected number of cares (NV) using Eq. 47.12.

$$\text{NV} = \text{Q\_length}/100 \tag{47.12}$$

The average length of a car is ∼100 pixels with respect to the used image resolution. Figure 47.5 shows the flow-graph of the queue length module [10].

5 Timing Decision Phase

The estimated time for each technique (depending on the computed number of vehicles) is calculated and compared with the actual time from control unit. To find out the estimated time needed for green traffic light, the following factors should be calculated:

- Number of Vehicles (NV).
- The distance of the first vehicle from the Start Point (SR). The distance is set to one vehicle length. Since it is open area, it was given half the time needed for vehicle movement.
- The time needed to move to the next point is set to two seconds (MT).

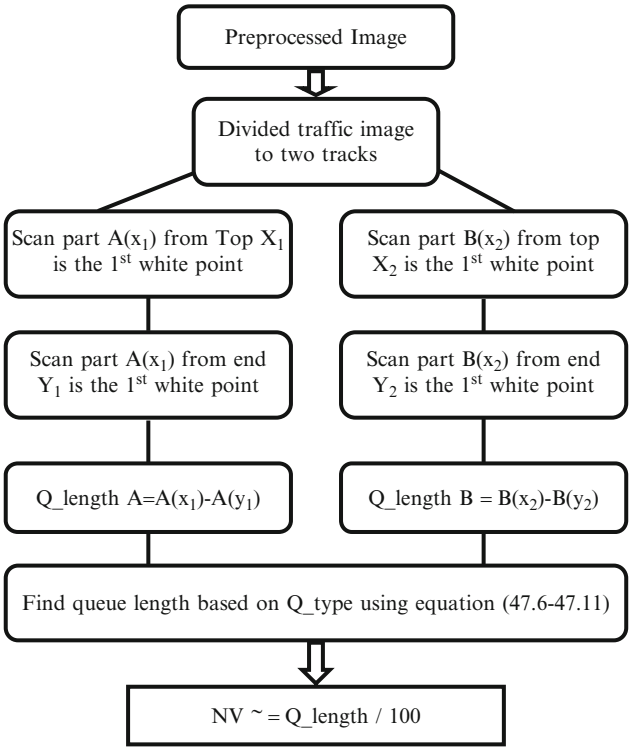


Fig. 47.5 Queue length module

- Distance between vehicles (EL) which represents empty levels. Estimated time for EL is one second.

After calculating the above factors, the total number of vehicles (NV) that affect the estimated time is calculated using Eq. 47.12. The estimated time (in seconds) is calculated using Eq. 47.13.

$$\text{Time} = \text{NV} \times \text{MT} + \text{SR} + \text{EL} \tag{47.13}$$

6 Experimental Results

To evaluate the performance of the proposed approach, 32 images were taken with different sizes but with the same resolution. These images are partitioned into two datasets, according to the number of cars in the image. Crowd images (high density), and few cars with a lot of space (low density) as shown in Figs. 47.11 and 47.12

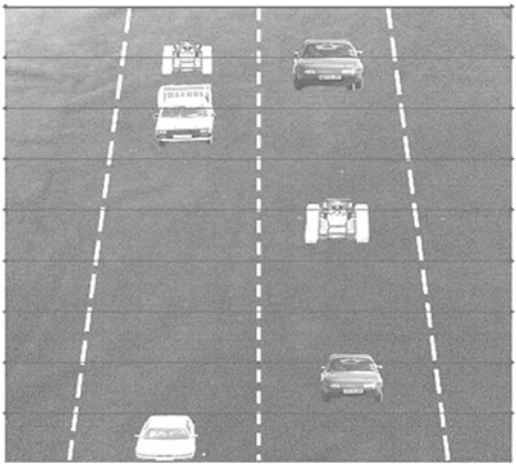


Fig. 47.6 Image is divided into levels

in Appendix. Each of the two sets will be analyzed separately. The behavior of the suggested approach is evaluated by comparing the predicted number of cars (NV_final) compared with the actual number of cars in each picture. The actual time was calculated as follows (after dividing the road into levels as in Fig. 47.6):

Let SV be number of small vehicles. Assume BV represents number of large vehicle. Each large vehicle is counted as two small vehicles.

1. Calculate number of vehicles

$$\mathrm{NV} = \mathrm{SV} + 2 \times \mathrm{BV} \tag{47.14}$$

2. Calculate number of vehicles on the same level (SL).
3. Calculating the empty level. (EL)
4. Calculate the final number of vehicles (NV_final).

$$\mathrm{NV\_final} = \mathrm{NV} - \mathrm{SL} \tag{47.15}$$

5. Find the distance from the starting point. (SR).
6. Calculate the estimated final time (Time).

$$\mathrm{Time} = (\mathrm{NV\_final} \times 2) + \mathrm{EL} + \mathrm{SR} \tag{47.16}$$

Figures 47.7, 47.8, 47.9, and 47.10 show the comparison between the actual and the developed two modules. The comparison is from number of vehicle, and estimated time point of views (considering both low-density images and high density images). Figures 47.7 and 47.8 show the estimated time and estimated number of vehicles in images with low densities. One could notice that the

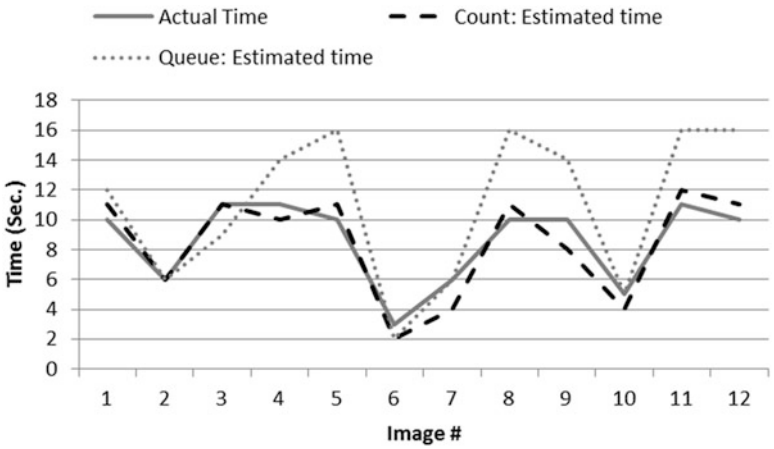


Fig. 47.7 Estimated time in low density

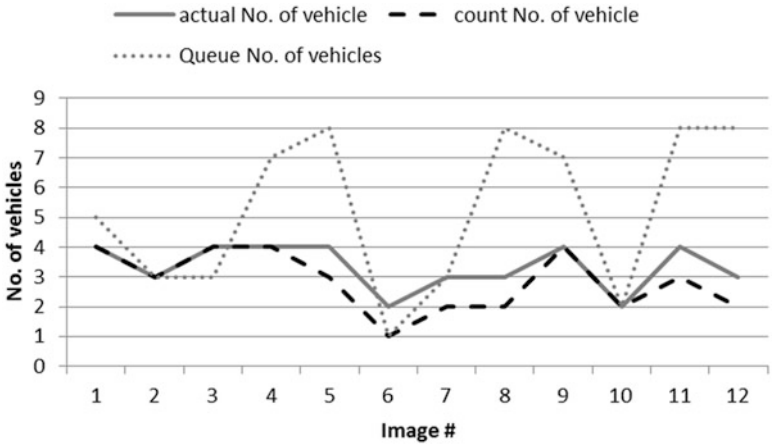


Fig. 47.8 Number of vehicle in low density

difference between estimated time and actual time is relatively high when car-queue is used. Also, the difference between estimated number of vehicles and actual number of vehicles is relatively high. This is because counting vehicles in the queue is based on the deference between first and last vehicle divided by vehicle size, regardless if the vehicles exist or not. Thus, the calculation time on this basis lead to increase in error rate. On the other hand, the number of vehicles detected and

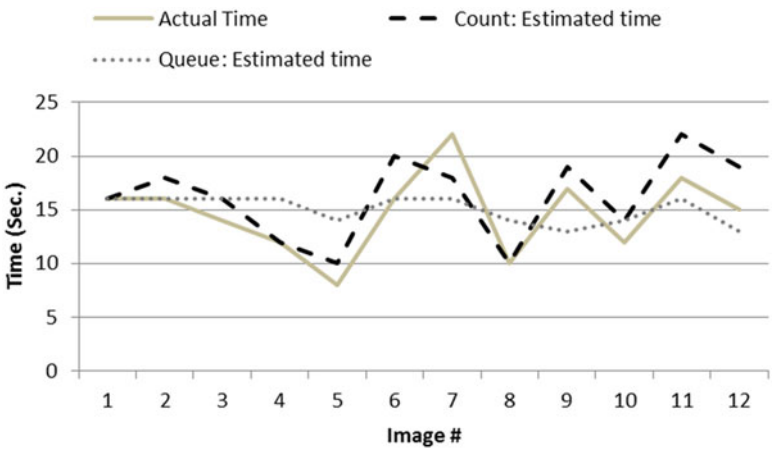


Fig. 47.9 Estimated time high density

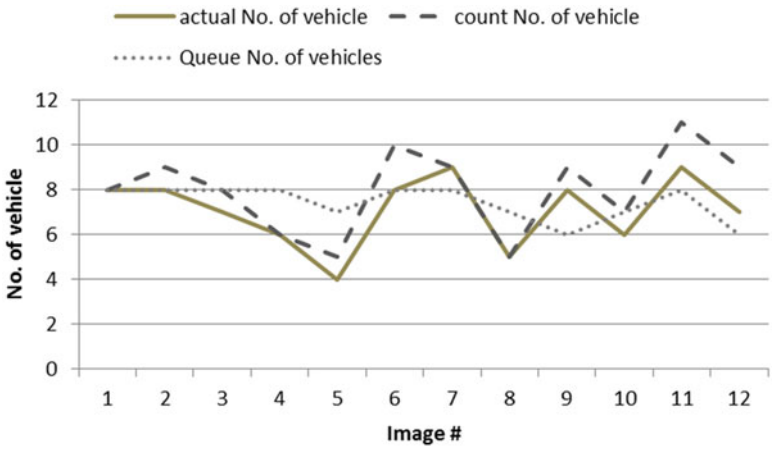


Fig. 47.10 Number of vehicle for high density

the estimated time when counting vehicle module is used is almost the same to actual number of vehicles and the actual time. One can deduce that vehicle counting module is more suitable to be used with low density than car-queuing module.

Figures 47.9 and 47.10 show the estimated time and estimated number of vehicles in images with high densities. It is noticeable that the car-queuing module shows better performance with high density pictures. But, still counting vehicle module shows better performance than car-queue module from both number of cars and estimated time point of views.

The mean square error between the estimated time and the actual time using car-queue module, for both high and low density images, is ∼17.4069. The mean square error between estimated number of vehicles and the actual number of vehicles, for both low and high density images, is 13. Number of vehicles is relatively larger since the suggested approach will consider the space between vehicles as cars. This consideration will affect the calculation of estimated time. Therefore, the error rate increases with low density roads, where vehicles are scattered. When counting vehicles module is used, the mean square error between the estimated time and the actual time (for both high and low density images) is ∼10. The mean square error between estimated number of vehicles and the actual number of vehicles (for both low and high density images) is ∼4. This will clearly indicates that count vehicle module gives approximation results comparable to the actual traffic.

7 Conclusion and Future Work

In this paper, vision based traffic control is developed. In the developed model, edge detection and segmentation is used to analyze the road image and find traffic load. The analysis is done by two different modules. Find the length of queue in two lane road, and counting vehicles on road. The estimated number of vehicles and the estimated time needed for green light period is calculated. From time execution point of view, counting vehicle module is better since it needs less processing time to make decision than queue module. From estimated-time of green-light period and number-of-vehicles point of views, we concluded that the queue length shows better performance in high density situations. But counting-vehicles module is recommended to be used since it performs better than queue length module when compared with the actual results. As future work, weather and night vision problem need to be solved. Also, choosing proper threshold when converting images to binary form partially solves day light shadows problem. Finally, counting-vehicles module should be developed to control more than one cross point on the road (i.e. manage all traffic lights at a cross point).

Appendix

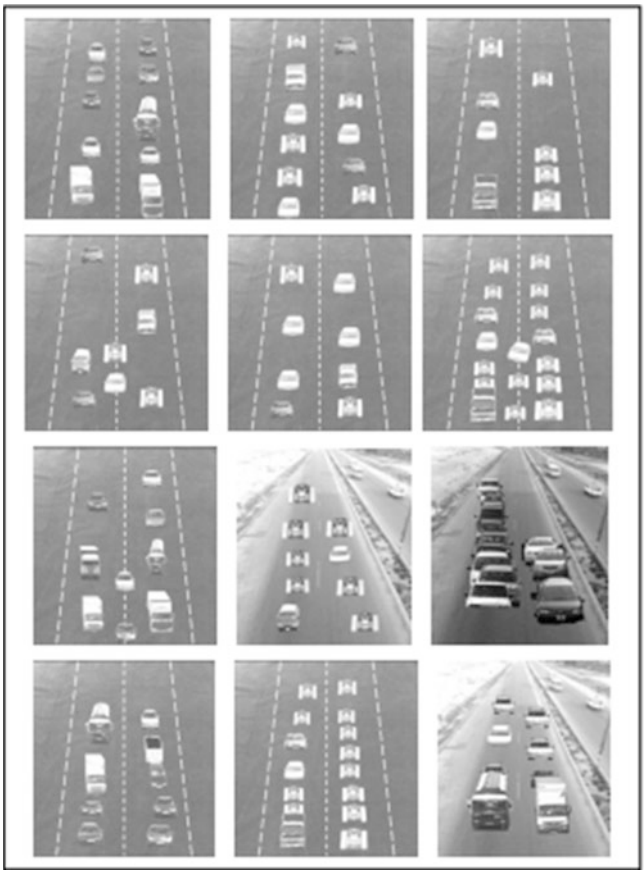


Fig. 47.11 High density images

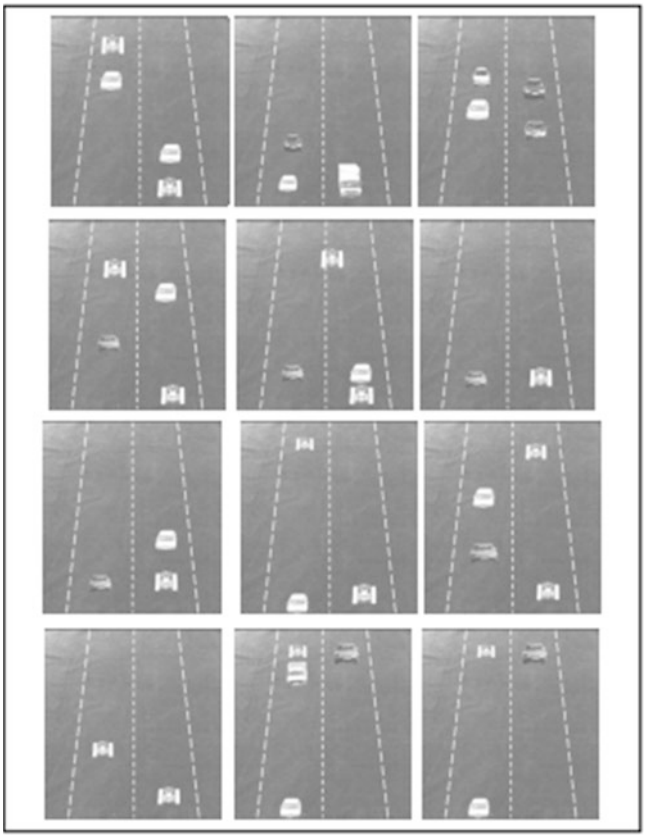


Fig. 47.12 Low density images

References

1. Papageoggius, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., Yibing Wang: Review of road traffic control strategies. Proc. IEEE **91**(12), 2043–2067 (2003)
2. Hsu, W.L., Liao, H.Y.M., Jeng, B.S., Fan, K.C.: Real-time vehicle tracking on highway. J. Inf. Sci. Eng. *21*, 733–752 (2005)
3. Haritaoglu, I., Harwood, D., Davis, L.S.: W4: real-time surveillance of people and their activities. Pattern Anal. Mach. Intell. IEEE Tran. **22**(8), 809–830 (2000)
4. Giusto, D.D., Massidda, F., Perra, C.: A fast algorithm for video segmentation and object 286 tracking. In: Digital Signal Processing, 14th International Conference, vol. 2, pp. 697–700 287, Santorini, Hellas (Greece) (2002)
5. Soto, A., Cipriano, A.: Image processing applied to real time measurement of traffic flow. In: Proceedings of the 28th Southeastern Symposium on System Theory, pp. 312–316, Baton Rouge (1996)
6. Murphey, Y.L., Lu, H., Lakshmanan, S., Karlsen, R., Gerhart, G., Meitzler, T.: Dyta: An intelligent system for moving target detection. In: Proceedings of the 10th International Conference of Image Analysis and Processing, Venice, pp. 1116–1121 (1999)

7. Lawrence, Y. D., Nick, C. T., Dong-liang Lee, Chin Thin Wang, Ming Chih Lu: Vision based adaptive traffic signal control system development. In: Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05), vol. 2, pp. 385–388. IEEE Press (2005)
8. Lipka, R., Herout. P.: Implementation of traffic lights in JUTS. In: The 10th International Conference on Computer Modeling and Simulation 2008, Uksim, pp. 422–427. 1–3 Apr 2008
9. Choudekar, P., Banerjee, S., Muju, M.K.: Implementation of image processing in real time traffic light control. Electronics Computer Technology (ICECT), the 3rd International Conference, vol. 2, pp. 94–98. IEEE Press (2011)
10. Bani Issa, O.M.A., Venus W. Samawi, Jehad Q. Alnihoud.: Traffic light control utilizing queue length, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, WCE 2014, pp. 590–594. London, 2–4 July 2014
11. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn (kindle). Pearson Prentice Hall, Upper Saddle River (2012)
12. Mathworks website: http://www.mathworks.com/
13. Vincent, R., Folorunso, O.: A descriptive algorithm for sobel image edge detection. In: 309 Proceedings of Informing Science & IT Education Conference (InSITE), pp. 97–107, Macon (2009)
14. Anton, H., Bivens, I., Davis, S.: Calculus: Early Transcendental, 10th edn. Houghton Mifflin, Boston (2011)

Chapter 48
On Cellular Automata Framework for Image Processing Hardware

Abdul Raouf Khan and Md. Al-Amin Bhuiyan

Abstract Image processing employing software routines are time consuming in compared to their counterfeits of hardware strategies. Cellular automata appear as abstract models of systems in image manipulation due to their local nature, inherently parallel and computationally simple implementation. This paper addresses a hardware framework based on two dimensional cellular automata, for manipulating images in certain specific positions and orientations. The proposed architecture can be easily implemented using VLSI technology.

Keywords 2DCA • Cellular automata • Image processing • Image translation • Neighborhood function • Real time processing

1 Introduction

Cellular automata have attracted significant attention in many applications in a wide variety of divergent fields of image processing, computer graphics, pattern recognition, computer vision, and so on [1, 2].

Image processing, the operations for improving image quality, intelligibility, and visual appearance, is formulated in either spatial domain or spectral domain. One of the commonly used spatial domain techniques is that of the convolution masks. In spectral domain, the widely used techniques are Fourier transform and wavelet transform. These procedures are almost based on calculations of directional derivatives that result in computationally intensive tasks or previous knowledge of

A.R. Khan (✉)
Department of Computer Sciences, King Faisal University, P. O. Box 400, Alhassa 31982, Saudi Arabia
e-mail: raoufkhan@kfu.edu.sa

M.A. Bhuiyan
Department of Computer Engineering, King Faisal University, P. O. Box 400, Alhassa 31982, Saudi Arabia
e-mail: mbhuiyan@kfu.edu.sa

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_48

the image nature. The processing time of images is very important, and usually depends on the software routines used to manipulate the images. Cellular automata play a momentous role as alternate to enhance the processing speed.

A fair amount of research works have been reported in literature to augment the processing speed of images employing various hardware solutions. In [3, 4] the hardware proposed is only for image rotation. In [5] a highly parallel machine CLIP4 was developed. In CLIP4 an array of identical cells corresponding to each pixel consists of a logical unit and four memories, where three of them are single bit buffers and one is a 32 bit RAM. Also there is a Boolean processor in each cell, which provides 16 Boolean functions. In [6] a similar kind of work was proposed and dealt with the implementation of 2D transform operators as a sequence of window shift operations. In [17] we proposed the architecture, however, the experimentation results reported were based on certain binary shapes and not real images.

In this paper we propose hardware architecture for manipulating binary images in specific directions and angles. The proposed architecture is based on the theory of Cellular Automata (CA) and is highly suitable for VLSI implementation. For reducing the complexity on the silicon floor, designers look for simple and regular blocks which can be cascaded and reused. Cellular automata have the advantage to fulfill these objectives. It has been demonstrated by us [7] that parallel processing architecture built around the CA machine is highly suitable for variety of applications. Further, various VLSI applications of Cellular Automata have been reported [8–12]. In addition, various theoretical developments of Cellular Automata have been reported in [13–16].

2 Cellular Automata Framework

The CA structure investigated by Wolfram [18] can be viewed as discrete lattice of sites (cells) where each cell can assume either the value 0 or 1. The next state of a cell is assumed to depend on itself and on its two neighboring cells for a three neighborhood dependency. The cells evolve in discrete time steps according to some deterministic rule that depends only on local neighborhood. In effect, each cell as shown in Fig. 48.1, consists of a storage element (D- Flip Flop) and a combinational logic implementing the next state.

Mathematically, the next state q of the i^{th} cell of a CA is given by

$$q_i\,(t+1) = f\,[q_{i-1}(t), q_i(t), q_{i+1}(t)]$$

In 2D CA the cells are arranged in two dimensional grid with connections among the neighborhood cells. The state of CA at any time instant can be represented by an $(m \times n)$ binary matrix. The neighborhood function specifying the next state of a

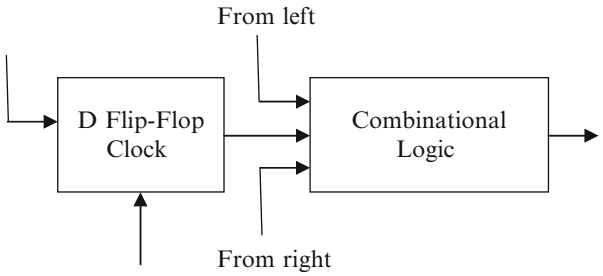


Fig. 48.1 Basic CA cell

particular cell of the CA is affected by the current state of itself and eight cells in its nearest neighborhood. Mathematically, the next state q of the $(i,j)^{th}$ cell of a 2D CA is given by

$$q_{i,j}\,(t+1) = f\Big[q_{i,j}(t), q_{i,j-1}(t), q_{i,j+1}(t), q_{i+1,j-1}(t), q_{i+1,j}(t), q_{i+1,j+1}(t), q_{i-1,j-1}(t), q_{i-1,j}(t), q_{i-1,j+1}(t)\Big]$$

Where, f is the Boolean function of nine variables.

To express a transition rule of 2D CA, a specific rule convention proposed in [5] is noted below

64	128	256
32	1	2
16	8	4

The central box represents the current cell (that is the cell being considered) and all other boxes represent the eight nearest neighbors of that cell. The number within each box represents the rule number associated with that particular neighbor of the current cell – that is, if the next state of a cell is dependent only on its present state, it is referred to as rule 1. If the next state depends on the present state of itself and its right neighbor, it is referred to as rule 3 (=1 + 2). If the next state depends on the present state of itself and its right, bottom, left, and top neighbors, it is referred to as rule 171 (=1 + 2 + 8 + 32 + 128) and so on. The minimized expression for Rule 171 is given by

$$q_{i,j}\,(t+1) = \Big[q_{i,j}(t) \oplus q_{i,j-1}(t) \oplus q_{i,j+1}(t) \oplus q_{i+1,j}(t) \oplus q_{i-1,j}(t)\Big]$$

3 Mathematical Model

The 2D CA behavior can be analyzed with the help of an elegant mathematical model [5], where two fundamental matrices are used to obtain row and column dependencies of the cells. Let the two dimensional binary information matrix be denoted as X_t that represents the current state of a 2D CA configured with a specific rule. The next state of any cell will be obtained by XOR operation of the states of its relevant neighbors associated with the rule. The global transformation associated with different rules are made effective with following fundamental matrices referred to as T_1 & T_2.

$$T_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \& \; T_2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Next we define CA under various possible basic conditions. If same rule is applied to all the cells in a CA, then the CA is said to be Uniform or Regular CA. If different rules are applied to different cells in a CA, then the CA is said to be Hybrid CA. The CA is said to be a Periodic CA if the extreme cells are adjacent to each other. The CA is said to be a Null boundary CA if the extreme cells are connected to logic 0-state. If in a CA the neighborhood dependence is on XOR or XNOR only, then the CA is called additive CA, specifically, a linear CA employs XOR rules only. A CA whose transformation is invertible (i.e. all the states in the state transition diagram lie in some cycle) is called a Group CA, otherwise it is called a Non Group CA.

The following theorems [5], specifies the value of the next state of a 2D CA referred to as X_{t+1} given that its current state is X_t. The CA is assumed to be configured with primary rule only – that is it depends on only one of its nine neighbors.

Theorem 1 *The next state transition of all the primary rules (1, 2, 4, 8, 16, 32, 64, 128, 256) with null boundary condition, can be represented as*

$$\begin{aligned}
Rule1 &\rightarrow [X_{t+1}] = [X_t] \\
Rule2 &\rightarrow [X_{t+1}] = [X_t]\ [T_2] \\
Rule4 &\rightarrow [X_{t+1}] = [T_1]\ [X_t]\ [T_2] \\
Rule8 &\rightarrow [X_{t+1}] = [T_1]\ [X_t] \\
Rule16 &\rightarrow [X_{t+1}] = [T_1]\ [X_t]\ [T_1] \\
Rule32 &\rightarrow [X_{t+1}] = [X_t]\ [T_1] \\
Rule64 &\rightarrow [X_{t+1}] = [T_2]\ [X_t]\ [T_1] \\
Rule128 &\rightarrow [X_{t+1}] = [T_2]\ [X_t] \\
Rule256 &\rightarrow [X_{t+1}] = [T_2]\ [X_t]\ [T_2]
\end{aligned}$$

Theorem 2 *The next state transition of all the primary rules with periodic boundary condition can be represented by*

$$
\begin{aligned}
Rule1 &\rightarrow [X_{t+1}] = [X_t] \\
Rule2 &\rightarrow [X_{t+1}] = [X_t]\ [T_{2c}] \\
Rule4 &\rightarrow [X_{t+1}] = [T_{1c}]\ [X_t]\ [T_{2c}] \\
Rule8 &\rightarrow [X_{t+1}] = [T_{1c}]\ [X_t] \\
Rule16 &\rightarrow [X_{t+1}] = [T_{1c}]\ [X_t]\ [T_{1c}] \\
Rule32 &\rightarrow [X_{t+1}] = [X_t]\ [T_{1c}] \\
Rule64 &\rightarrow [X_{t+1}] = [T_{2c}]\ [X_t]\ [T_{1c}] \\
Rule128 &\rightarrow [X_{t+1}] = [T_{2c}]\ [X_t] \\
Rule256 &\rightarrow [X_{t+1}] = [T_{2c}]\ [X_t]\ [T_{2c}]
\end{aligned}
$$

Where

$$
T_{1c} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \ \& \ T_{2c} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}
$$

4 Experimental Results

Experiments have been carried out to investigate the effectiveness of training cellular automata to perform several image processing tasks using Intel® Core™ i5 CPU with 2.70 GHz PC. The algorithm has been implemented using Visual C++. Although the gray scale images have been employed, these were being converted into various binary matrices, considering binary "0" as white pixel and a binary "1" as a black pixel. Various rules were applied to different $(m \times n)$ matrices. This study investigated the use of cellular automata with Rule 2, 4, 8, 16, 32, 64, 128 and 256, respectively for image manipulation. The snapshots of the graphical interfaces for some image manipulation processes are illustrated in Figs. 48.2, 48.3, 48.4, and 48.5, respectively.

Almost all primary rules, except rule 1, can be used to translate the images, in different directions including corner directions, both with null and periodic boundary conditions. However, under null boundary condition, once the image has been translated up to the screen boundary, further translation means that images are lost and cannot be retrieved. For similar translation under periodic boundary conditions, the images appear again on the other side of the screen boundary. Rotation of images about x- axis and y- axis, can be implemented with the help of secondary rules i.e. rule 136 $(128 + 8)$ and rule 34 $(32 + 2)$ under null boundary conditions, as shown in Fig. 48.6. To achieve the required translation or rotation the CA is run for several clock periods and depends on the cycle length of a particular

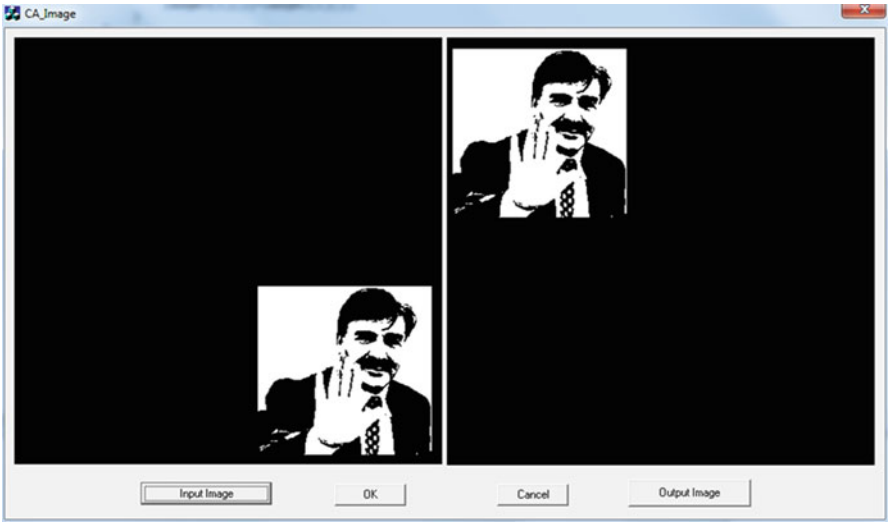


Fig. 48.2 Application of rule 4

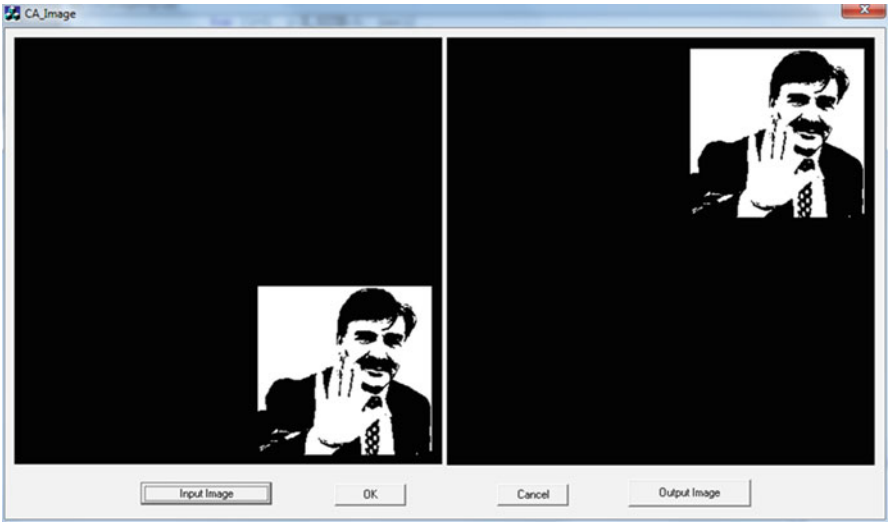


Fig. 48.3 Application of rule 8

rule applied as well as the number of rows and columns of the CA ($m \times n$). To create special effects, like duplication of image, as shown in Fig. 48.7, rule 3, rule 9, rule 33 and rule 129 can be used. It is possible to produce multiple images of an object with other rules.

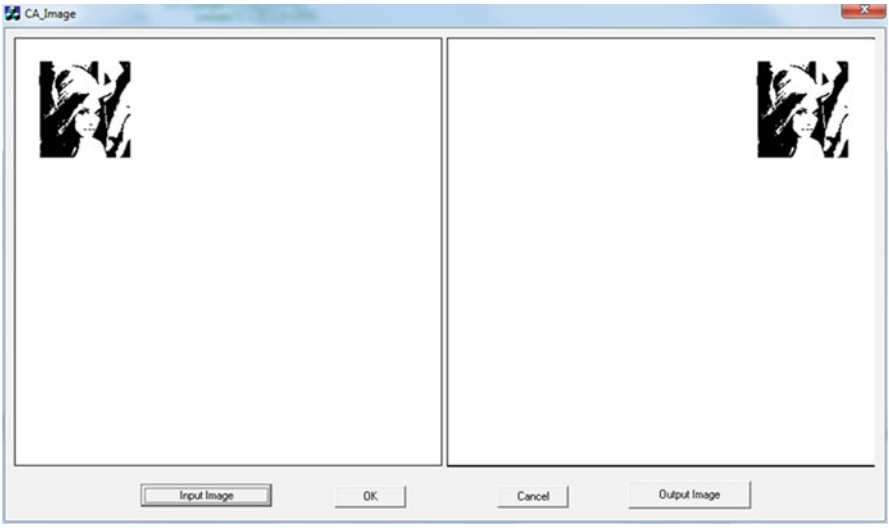


Fig. 48.4 Application of rule 32

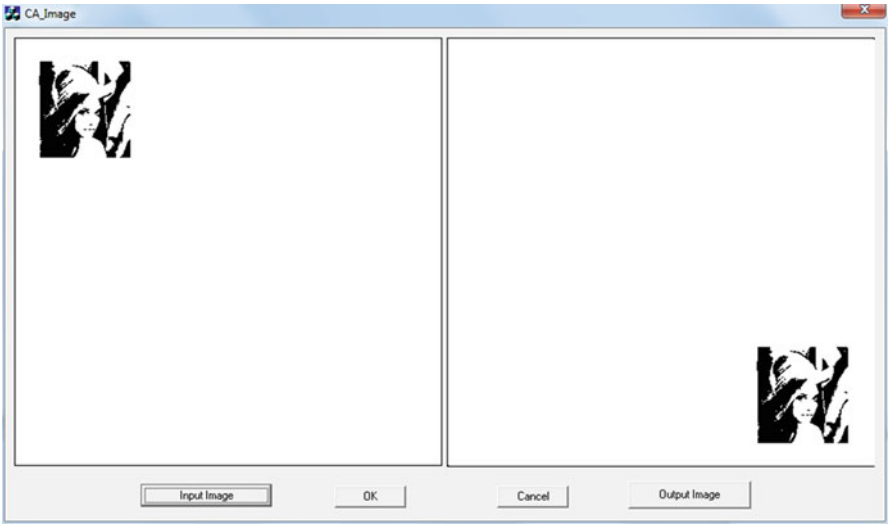


Fig. 48.5 Application of rule 32

5 Proposed Framework

The architecture of the proposed hardware consists of an $(m \times n)$ matrix of cells. Each cell (as shown in Fig. 48.8) consists of a memory element (D-flip Flop) and a Multiplexer. Eight inputs of the MUX (Fig. 48.9) are labeled as

Fig. 48.6 Rotation of an image (180°)

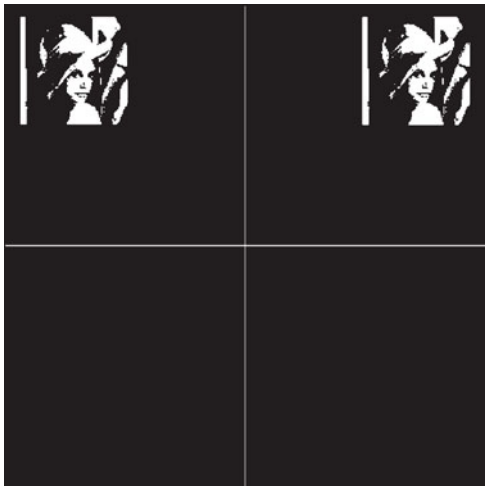


Fig. 48.7 Duplication of an image

$L, R, T, B, LT, RT, LB, RB$. These stand for the eight outputs of the memory elements of the neighboring cells i.e. left, right, top, bottom, left top, right top, left bottom, right bottom respectively. The other two inputs R_x and R_y of the MUX are obtained from $T \oplus B$ and $L \oplus R$ respectively.

Inputs for duplication take the values $D_{xp} = Q \oplus R, D_{xn} = Q + L, D_{yp} = Q + T, D_{yn} = Q + B$ respectively.

For Zoom Out and Zoom in $Z_o = T \oplus B \oplus R \oplus L$ and $Z_i = Q \oplus T \oplus B \oplus R \oplus L$ inputs are applied to the MUX. The select lines of the MUX and the clock are same for all the cells.

For applying any of the transformations, the select lines of each MUX selects the particular input and the CA is evolved for some no. of cycles depending on the

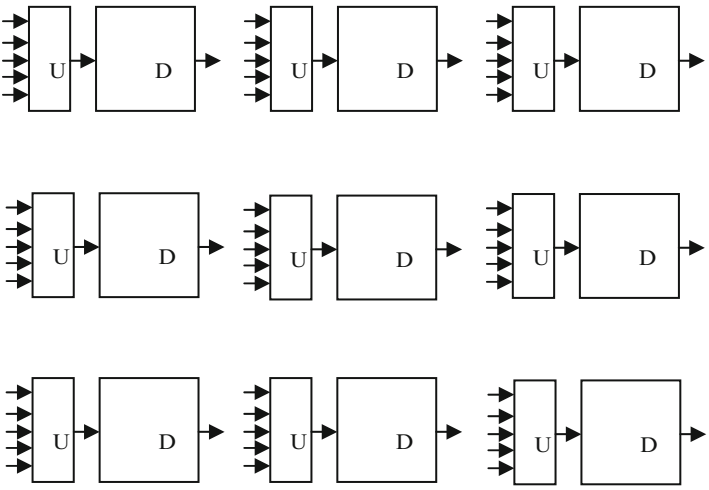
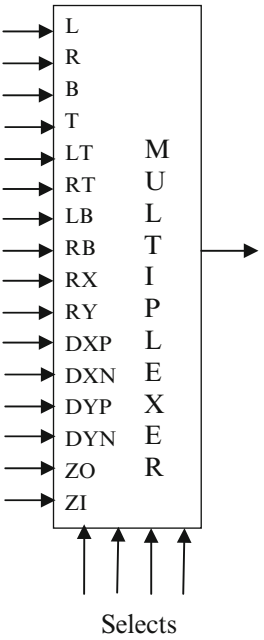


Fig. 48.8 Details of each cell

Fig. 48.9 Details of the multiplexer



operation. For example if any binary image located at $x_1 \times y_1$ is to be translated to new location $x_2 \times y_2$ towards left in the x direction, the select lines select L input of the MUX and the CA is evolved for $|x_1 \times x_2|$ times. Similarly, other transformations are carried.

6 Conclusion and Future Work

This paper presents the framework of an application specific image processing hardware. It explores the application of two dimensional cellular automata to the shifting and rotation of digital images with neighborhood based rules approach. The local dependency of the cells or pixels aids for real time operations. The proposed architecture can be extended to build hardware based video games and animation.

References

1. Hernandez, G., Herrmann, H.J.: Automata for elementary image enhancement". Graph. Models Image Process. **58**(1), 82–89 (1996)
2. Rosin, P.L.: Image processing using 3-state cellular automata. Comput. Vis. Image Underst. **114**, 790–802 (2010)
3. Ghosh, I., Mujamdar, B.: Design of an application specific VLSI chip for image rotation. In: Proceedings of 7th International Conference on VLSI Design, pp. 275–278 (1994)
4. Tsuchida, N., Yamada, Y., Weda, M.: Hardware for image rotation by twice skew transformation. IEEE Trans. ASSP **34**(4), 527–532 (1987)
5. Duff, M.J.B., et al.: A cellular logic array for image processing. Pattern Recognit. **5**, 229–247 (1973)
6. Biswas, G.P., et al.: Cellular architecture for affine transformation on raster images. IEE Proc. Comput. Digit. Tech. **143**(2), 103–110 (1996)
7. Khan, A.R., Choudhury, P.P., et al.: VLSI architecture of a cellular automata machine". Int. J. Comput. Math. Appl. **33**(5), 79–94 (1997)
8. Choudhuri, P.P., Choudhuri, D.R., Nandi, S., Chatterjee, S.: Additive Cellular Automata, Theory and Applications, vol. 1. IEEE Computer Society Press, Los Alamitos (1997)
9. Chowdhury, D.R., Gupta, I.S., Choudhuri, P.P.: A class of two dimensional cellular automata and applications in random pattern testing". J. Electron. Test. Theory Appl. **5**, 65–80 (1994)
10. Chowdhury, D.R., et al.: A low cost high capacity associative memory design using cellular automata. IEEE Trans. Comput. **44**(10), 1260–1264 (1995)
11. Khan, A.R.: Architecture of Image Encryption Hardware Using Two Dimensional Cellular automata. In: Proceedings of International Conference on IT Convergence and Security. Lecture notes in Electrical Engineering, vol. 120. Springer (2011)
12. Sarkar, P.: A brief history of cellular automata. ACM Comput. Surv. **32**(1), 80–107 (2000)
13. Khan, A.R.: Classification of 2D cellular automata uniform group rules. Eur. J. Sci. Res. **64**(1), 51–57 (2011)
14. Munshi, S., et al.: An analytical framework for characterizing restricted two dimensional cellular automata evolution. J. Cell. Autom. **3**(2), 313–335 (2008)
15. Nayak, K., et al.: Colour Graph: An efficient model for two dimensional cellular automata. Orissa Mathematical Society Conference (2008)
16. Gosh A. K., Choudhury, P.P., Basuray, A.: Chotis fractals with multivalued logic in cellular automata. In: Innovation and Advanced Techniques in Computer and Information Sciences and Engineering, pp. 77–82. University of Bridgeport, Springer (2007)
17. Raouf Khan, A., Md. Al-Amin Bhuiyan.: Real time application specific image processing hardware. Lecture Notes in Engineering and Computer Sciences: Proceedings of the World Congress on Engineering 2014, WCE 2014, pp. 595–598. London 2–4 July 2014
18. Wolfram, S.: Statistical mechanics of cellular automata. Rev. Mod. Phys. **55**(3), 601–644 (1983)

Chapter 49
Dynamic Simulation of the Temporomandibular Joint

Janith Muhandiram, Bin Wang, and Mahmoud Chizari

Abstract This research is aiming to investigate the mechanical behaviour of the temporomandibular joint (TMJ), in response to cyclical loading caused through actions of speech and mastication. A set of in-vitro experimental tests has been performed in three different groups on a fresh sheep jaw bone to examine the hypothesis of the study. No failure was monitored during the cyclical test. The study was concluded that the amount of loading is effective on the displacement of the TMJ.

Keywords Cyclic loading • Experimental test • Mastication • Sheep jaw bone • Speech • Temporomandibular joint

1 Introduction

1.1 Temporomandibular Joint (TMJ)

The temporomandibular joint (TMJ) is the region which connects the bone of lower jaw, mandible, to the upper temporal bone of the skull as shown in Fig. 49.1. The TMJ is the most constantly used joint in the human body [2]. The flexibility of these joints allows the movement of the lower jaw, enabling the actions of speech, chewing and yawn [1].

J. Muhandiram (✉) • B. Wang
College of Engineering, Design and Physical Sciences, Mechanical Engineering Department, Brunel University, West London, UK
e-mail: janith.muhandiram@gmail.com; bin.wang@brunel.ac.uk

M. Chizari
College of Engineering, Design and Physical Sciences, Mechanical Engineering Department, Brunel University, West London, UK

Mechanical Engineering Department, Sharif University of Technology, Tehran, Iran

Orthopaedic Research and Learning Centre, College of Engineering, Design and Physical Sciences, Brunel University, Uxbridge, London UB8 3PH, UK
e-mail: mahmoudchizari@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_49

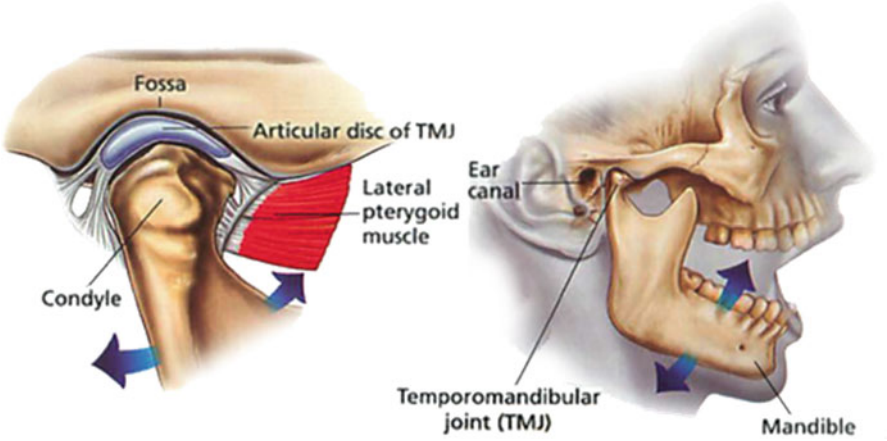


Fig. 49.1 Schematic of a temporomandibular joint. The image has been reproduced from Ref. [3]

The TMJ is composed of a mobile segment known as condyle, the round upper end of the lower jaw and a joint socket of the temporal bone called the articular fossa [2] (Fig. 49.1). The condyle glide along the articular fossa, when the mouth opens and returns to the original position as the mouth is closed [1]. A disc composed of cartilage is located between the condyle and the articular fossa to ensure a smooth motion of the condyle. The disc functions through the absorption of stress to the jaw caused by chewing and other actions [2].

The significant character of hinge and slide movements along with the unique composition of the TMJ imposes many challenges to the study of this complex [1]. The function of the TMJ maybe obstructed by a variety of conditions and those conditions may lead to a joint disorder of the TMJ [3].

1.2 Temporomandibular Disorders (TMD)

Due to the dynamic nature of the mandible and its constant movement during functions, issues may occur at the temporomandibular joint. Varying degrees of force is applied on the mandible during functions of mastication and speech, and force overload leads to the rise of temporomandibular disorders (TMD). Temporomandibular disorders are also termed as temporomandibular joint dysfunction (TMJD) and temporomandibular joint pain dysfunction syndrome (TMJPDS) [3, 4]. TMD is a group of conditions in the temporomandibular joint and the muscles of mastication [4].

Temporomandibular disorder is a benign condition and has a variety of symptoms including the ongoing pain in the jaw, pain on jaw movement, restricted jaw movement, earache and headache [3]. These conditions are also characterized by clicking or grating sounds in the temporomandibular joint and pain on the opening and closing of the mouth [1].

TMD falls into key subtypes including myofacial pain, arthralgia and internal derangement of the joint involving replacement of the disc [4]. A key cause of the TMD is arthritis as joint disorders affects the TMJ [1]. TMD may also be caused by injury to the TMJ due to grinding teeth at night, clenching teeth together or habitual movements that cause stress to muscles of the jaw [3]. An additional cause of TMD is the disc displacement in the TMJ due to internal derangement which may result in the clicking and popping sound, pain on opening and closing the mouth and restricted jaw movement. A grating sound on movement of the jaw maybe produced due to perforation of the disc. Rheumatoid Arthritis and trauma is capable of fusing components of the TMJ preventing jaw movement which also leads to TMD [2].

There has been no cure found for TMD and the condition of majority of patients with the TMD have been reported to be improved with simple and effective treatments such as avoiding extreme jaw movements, avoiding chewing of hard food, application of hot or cold packs on the face, jaw and face exercises [1]. Also the use of the hard bite guard is a treatment of the TMD, which reduces tension of jaw muscles, preventing further injury to the TMJ as a result of teeth grinding or clenching of the jaw at night [3]. Anti-inflammatory medication is used at times to reduce the swelling of the jaw and aid the relaxation of the jaw [2]. Jaw joint surgery is an unusual treatment of the TMD [3]. Further investigation of the complex and unique nature of the temporomandibular joint is required to discover a cure for temporomandibular disorders.

1.3 Jaw Movement in Mastication and Speech

The jaw movements are involved in primitive motor functions such as mastication and speech. A communal muscle architecture and histochemistry as well as sensory, motor and reflex components are shared between the functions of mastication and speech to a certain degree [5]. During these functions, varying amounts of force is applied on the temporomandibular joint, depending on the behaviour of the jaw movement.

The movement of jaw has been investigated through the measurements of movement, rate and amplitude in speech and the compliance of the bolus in mastication. In 1989, it was reported by Ostry and Flangan [5] that in mastication, the movement, rate and amplitude of jaw movement was greater in comparison to that in speech, as reported in Table 49.1. However, the normalized functions were similar in jaw-closing movements in contrast to the movement, rate and amplitude values in both mastication and speech. The functions for mastication and speech were similar in jaw-opening movements although the amplitude and compliance values were significantly different. The duration of acceleration and deceleration of fast movements were approximately equivalent, whereas the deceleration duration was significantly longer than that of acceleration duration of slower movements.

Table 49.1 Kinematics of mandible at mastication and speech

	Normal		Normal closing	
	Mastication	Speech	Mastication	Speech
Amplitude, cm	0.63	0.26	0.64	0.26
Duration, ms	318	234	356	267
Acceleration, ms	110	69	99	94
Deceleration, ms	208	163	257	133
*V_{max}, cm/s	6.82	2.93	5.38	2.82
	Fast opening		Fast closing	
	Mastication	Speech	Mastication	Speech
Amplitude, cm	0.45	0.24	0.44	0.24
Duration, ms	132	110	143	342
Acceleration, ms	74	49	145	148
Deceleration, ms	58	61	95	64
*V_{max}, cm/s	7.21	4.06	6.76	4.00

*V_{max} is the maximum instantaneous velocity

Due to the dynamic nature of the mandible and its constant movement during functions, complications may occur at the temporomandibular joint. Varying degrees of force is applied on to the mandible during functions of mastication and speech and force overload leads to the rise of temporomandibular disorders.

1.4 Objective of the Study

Simulation of the loading applied on the TMJ of a healthy human jaw during functions of speech and mastication is the key focus of this research. The study is aiming to model a cyclical load applied on TMJ and assesses its mechanical behaviour. Investigating possible modes of failure, while the loading is applied, is also an interest of the study.

2 Methods

2.1 Input Data

The study uses fresh sheep jaw bones to simulate the mechanical behaviour of the jaw under daily mechanical loading on the TMJ. In order to simplify the experiment while accommodating available testing facility, an assumption was made and the input data was defined [7]. The assumptions made for the use of sheep jaw specimens are reported in Table 49.2.

Table 49.2 The assumed data for sheep's jaw specimen

Chews per day	3,750 ± 50 cycles
Frequency	1.57 Hz
Amplitude at the mouth	22.4 mm
Amplitude at TMJ	1.3 mm

Fig. 49.2 The sheep jaw bone sample was mounted into the testing machine

2.2 Testing Setup

Fresh sheep jaw bones were used for the experimental examination of this study. Sheep head samples were obtained from an authorized farm and the mandible jaw bones of the heads were carefully harvested by an qualified individual. The jaw bone was then cleaned off from all the meat, muscles and ligaments, only the hard bone as shown in Fig. 49.2 was kept for experimental investigation of the study. The harvested mandible jaw bones were packed in a zip plastic bag and then stored in a freezer at temperature of −20 °C. On the day of the experiment, the specimen was removed from the freezer and kept for 3 h at room temperature to be thawed. The specimen was kept moist by spraying pure water during thawing and the experimental setup.

A custom made fixture was utilized to mount the specimen into the testing machine as shown in Fig. 49.3. The specimen was adjusted on the testing rig to make sure the TMJ of the jaw bone is placed in front of the loading bar which was attached to the crosshead. The size of contact zone at the loading bar was approximately the size of TMJ. Before applying the load, the distance between the loading bar and

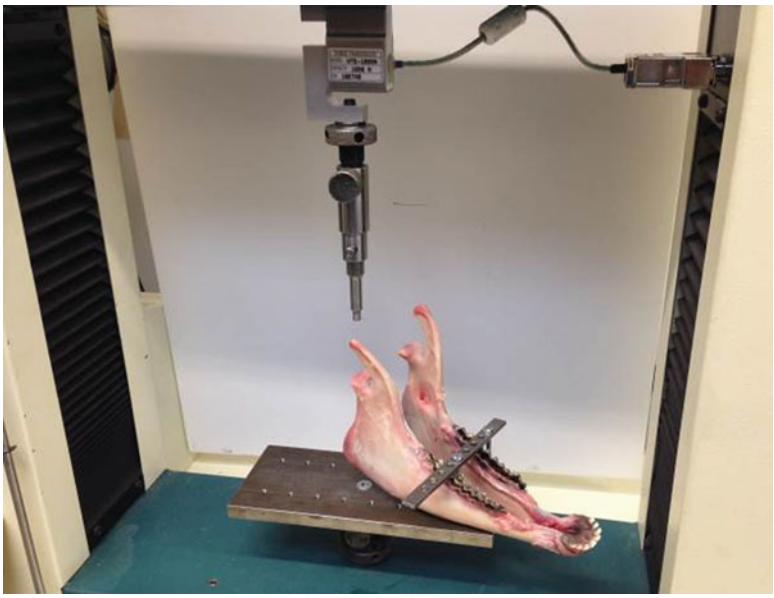


Fig. 49.3 The mounted specimen in a Hounsfield testing machine. The specimen was mounted into the machine using a custom made fixture

the TMJ was controlled. The specimen was then fixed using metal bridges on the metal rig.

A computer drive Hounsfield hydraulic testing machine was used to carry out the test. The machine was armed with a 1,000 N load-cell on its crosshead. The machine was controlled with the QMat V5.3 software (Tinius Olsen, UK). The software enabled the user to introduce the input data to the testing machine. A compression-cyclic routine from the QMat database was used to carry on the test.

2.3 Biomechanical Testing

In order to determine the mechanical behaviour of the TMJ under a cyclical load that was applied on the specimen, the data was recorded and possible mode of failure was visually monitored [6].

The frequency of daily chewing by the number of days of a given period of time, determines the amount of mechanical load applied on to the jaw during a specific life time. Figure 49.3 shows the experimental setup and the position of the specimen in relation to the tensile machine.

The specimens were tested in three different groups. Table 49.3 shows the specification of each testing groups. The specimen was kept moist whilst it was attached to the testing machine.

Table 49.3 Specification of the three testing groups

	Group 1	Group 2	Group 3
Load range, N	10	20	50
Extension range, mm	0.5	10	10
Speed, mm/min	1,000	1,000	1,000
Sample height, mm	5	5	5
Load target, N	10	10	10
Number of cycle	100	100	100
Preload, N	2	5	5

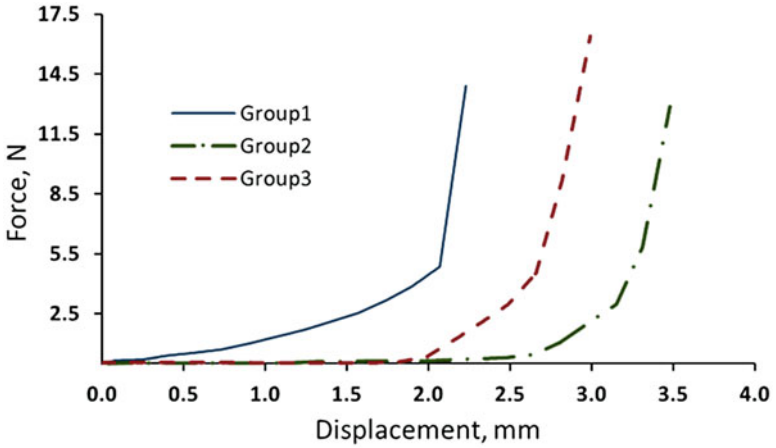


Fig. 49.4 Load and displacement results of the three groups applied on a sheep jaw bone sample

3 Results and Discussion

Experimental measurements were carried out using Hounsfield testing machine and QMat V5.3 software on a sheep jaw bone specimen. The cyclical load and its correspondence displacement were recorded for all three groups. The load displacement results of the three groups are illustrated in Fig. 49.4. No significant difference was monitored between the three groups. However the effect of the applied load on TMJ was evident. The results show that increase of the load on TMJ, would increase the displacement of the bone. However, the behaviour of the load-displacement curve was not linear.

The maximum displacement after a full cyclical test for group 1 was 2.1 mm. This displacement was in correspondence with 14.6 N loads which was applied onto the TMJ. For group 2 and 3, maximum displacement was 2.9 and 3.4 mm which were in correspondence with 16.2 N and 13.6 N loads respectively.

A difficulty was experienced on the fixation of sheep jaw bone on the testing rig which may have caused minor movement while the specimen was under loading. This unexpected movement may impact accuracy of the recorded data.

The study was investigating that the cyclic loading on the TMJ is effective on its mechanical behaviour however the result obtained here are preliminary and further investigation is required to complete the hypothesis of the study.

4 Conclusion and Future Work

This research focuses on the kinematics of the temporomandibular joint during speech and mastication. A set of in-vitro tests were performed to examine the hypothesis of the study. A fresh sheep jaw bone was used to investigate the mechanical properties of the TMJ in response to cyclical loading caused through actions of speech and mastication. The data was recorded and possible mode of failure was screened. The study examined three different methods of input loading and concluded that the amount of kinematic loading may be effective on the deformation of the TMJ. No significant difference was monitored between the three groups and there was no failure during the test.

This study is still under research and outcome has not yet finalized. However, the concept of the study may be used to improve the treatment of TMJ.

The hypothesis of the study may need further investigation with more jaw bone specimens. The utilization of an improved testing rig to mount the specimen into the testing machine may increase the accuracy of the future experiments. In addition, the replacement of the contact point of the loading bar with an sheep TMJ may enhance the results of further investigations.

Acknowledgment The authors would like to thank the Biosurgical Lab at the Brunel University London for providing the animal materials and test facilities.

References

1. TMJ Disorders.: U.S. Department of Health and Human Services, National Institute of Health Services, MD (2013) [Online]. Available http://www.nidcr.nih.gov/oralhealth/Topics/TMJ/Documents/TMJDisorders_100113_508C.pdf
2. LAWSON.: Oxford University Hospitals NHS Trust, Specialist Surgery, TMJ Exercises [Online] (2011) Available http://www.ouh.nhs.uk/patientguide/leaflets/files%5C120417tmjexercises.pdf
3. Temporomandibular joint disease treatment-North Shore Oral and Maxillofacial Surgery [Online]. Available http://nsoms.com.au/procedures/temporomandibular-joint-disease-treatment/
4. LERESCHE, L.: Epidemiology of temporomandibular disorders: implications for the investigation of etiologic factors. SAGE J **8**(3), 291–305 (1997) [Online]. Available http://cro.sagepub.com/content/8/3/291.full.pdf+html
5. Ostry, D.J., Flanagan, J.R.: Human jaw movement in mastication and speech. Arch. Oral Biol. **34**(9),685–693 (1989) [Online]. Available http://www.ncbi.nlm.nih.gov/pubmed/2624559

6. Manfredini, D., Guarda-Nardini, L., Winocur, E., Piccotti F, Ahlberg J, Lobbezoo F.: Research diagnostic criteria for temporomandibular disorders: a systematic review of axis I epidemiologic findings. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodont. **112**(4), 453–62 (2011). doi:10.1016/j.tripleo.2011.04.021. PMID 21835653
7. Muhandiram, J., Wang, B., Chizari, M.: Dynamic simulation of the Temporomandibular Joint. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2014, vol I, WCE 2014, pp. 655–658 [Online]. London 2–4 July 2014. Available http://www.iaeng.org/publication/WCE2014/WCE2014_pp655-658.pdf

Chapter 50
Using Information Gain to Compare the Effeciency of Machine Learning Techniques When Classifying Influenza Based on Viral Hosts

Nermin Shaltout, Ahmed Rafea, Ahmed Moustafa, and Mahmoud ElHefnawi

Abstract The paper compares the performance of two classical machine learning techniques when features selection is used to improve Influenza-A host classification. The impact of using the most informative positions on both the classifier efficiency and performance of decision trees (DTs) and neural networks (NNs) was measured. The experiments were conducted on cDNA sequences belonging to all the viral segments of the subtype H1 to ensure authenticity of results. Sequences belonging to each viral segment were further divided into viruses infecting human and non-human hosts prior to classification analysis. The performance measures, accuracy, sensitivity, specificity, precision, and time, were used. Extracting the best hundred informative positions with the information gain *(IG)* algorithm increased classification efficiency for both classifiers by more than 80 % for all viral segments. The change in performance was insignificant. The overall results of statistical significant tests showed that NNs classified viral hosts more accurately than DTs for subtype H1. The tests also showed that the DTs are significantly faster than NNs in classifying Influenza hosts despite the slight decrease inperformance.

The surname of the fourth chapter co-author was corrected in this chapter. An erratum to this chapter can be found at: [DOI 10.1007/978-94-017-9804-4_52]

N. Shaltout (✉) • A. Rafea
Computer Science and Engineering Department, School of Sciences and Engineering, The American University of Cairo, New Cairo, Cairo 11835, Egypt
e-mail: nermeena@gmail.com; rafea@aucegypt.edu

A. Moustafa
Biology Department, School of Sciences and Engineering, The American University of Cairo, New Cairo, Cairo 11835, Egypt
e-mail: amoustafa@aucegypt.edu

M. ElHefnawi
Informatics and Systems Department and Biomedical Informatics and Chemoinformatics group, Division of Engineering Research and Centre of Excellence for Advanced Sciences, National Research Centre, Tahrir Street, Cairo 12311, Egypt
e-mail: mahef111@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_50

Keywords Antiviral classification • Decision tree • Feature selection • Host classification • Influenza-A • Information gain • Machine learning • Neural network • Segment • Subtype

1 Introduction

During epidemics and pandemics, rapid analysis is crucial to identifying the causative infectious virus and its cure. Virus analysis using laboratory techniques is time costly, usually posing a health risk to the handlers. Classical Bioinformatics computation techniques, in addition to being time consuming, are too coarse for detailed virus analysis. Improving the efficiency of machine learning techniques intends to remedy this by cutting down the computational time of virus analysis further without drastically sacrificing performance. Of the many infectious diseases, Influenza-A was selected for analysis due to its rapid mutation, yearly seasonal outbreaks, and its high morbidity and mortality rates. Influenza-A's constantly growing and changing data additionally renders it more challenging and time consuming to analyze than other viruses.

1.1 Overview of Influenza-A's Properties

Influenza-A's genetic data is divided into eight RNA segments [1] that can be swapped between different virus strains to produce new viruses. Of these segments, the Hemagglutinin (HA) and Neuraminidase (NA) code for the virus's surface antigens. The target host's antibodies recognize and destroy the virus via these antigens. Outbreaks occur when the mutation of the virus leads to the production of unrecognizable HA and NA proteins. Due to Influenza-A's rapid mutation rate, there are currently at least 16 HA and 9 NA known Influenza-A subtypes. The combination of the different HA (H) and NA (N) subtypes determines the virus's strain [2]. E.g. H1N1 is a virus strain which mutated, resulting in the swine flu pandemic. Figure 50.1 summarizes the most important features of the Influenza-A virus.

The severity of Influenza-A outbreaks is determined by its virulence or mortality rate. Influenza-A's virulence is increased by its capacity to infect a variety of hosts [3]. Based on previous pandemic and epidemic history, the most problematic hosts are human, avian, and swine in nature. Some strains pertaining to these hosts have gained the capability of infecting multiple hosts simultaneously via mutation [3]. To aid in the virus' analysis, decoded Influenza-A sequences are stored in online databases as RNA/cDNA and protein sequences. These sequences can be utilized for predicting the target host(s) of the virus in future outbreaks.

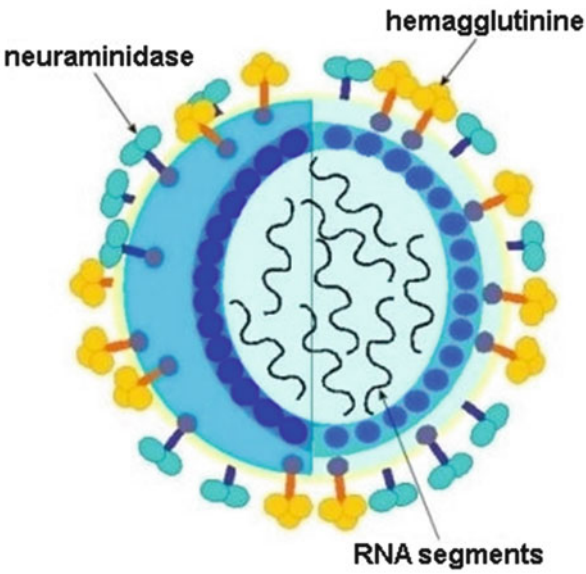


Fig. 50.1 An overview of the structure of Influenza-A virus

1.2 Motivation for Improving Influenza-A Classification

Influenza-A's mutation, through a process known as antigenic drift, causes it to infect its host seasonally [1]. Although the process alters the virus' surface antigens minimally, it still renders them unrecognizable by the host's antibodies. Typically, Influenza-A strains do not intercross species; i.e. Influenza-A strains infecting avian and swine species are not transmittable to human hosts and vice versa. In cases where virus strains with interspecies transmission exist, the strain's virulence is low. However, in some rare cases the virus undergoes massive changes by a mutation process known as antigenic shift [1]. This in turn leads to the formation of a virulent strain with interspecies transmission ability. During outbreaks, these mutated strains pose a significant health risk to the human population due to their accelerated spread using swine and avian hosts as intermediaries [4]. By determining the most efficient machine learning methods for classifying Influenza-A by host, the infectious ability of new virus strains can swiftly be assessed in an outbreak's early stages. Determining the efficient classification technique will also open way for the identification of other Influenza-A viral properties such as antiviral resistance.

1.3 Related Works

The following techniques are used classically in Bioinformatics. Hemagglutinin inhibition (HI) and Neuraminidase inhibition (NI) assay are classical laboratory

analysis methods for determination of Influenza-A host and subtype [5, 6]. In addition to the technique posing a risk to handlers, the processes are time consuming taking several days to generate results. Classical Bioinformatics computational techniques such as sequence alignment, phylogenic trees, and the blast algorithm [7, 8] are also frequently used for virus identification. This is done by comparison against known sequences. Although suitable for differentiating between different species, these methods are unsuitable for detecting fine grained individual differences within the same species, e.g. Influenza-A hosts. The methods are additionally time costly rendering them inapplicable to high dimensionality datasets such as that of Influenza-A.

In order for DNA and protein sequences to be suitable for machine learning analysis, they have to be encoded in a format recognizable by the classifiers at hand. Sami A. & Takahashi developed techniques to map DNA to common data mining methods and introduced a technique that suits all classification methods [9]. Salzber et al. created an interpolated Markov Model system to locate genes in DNA [10].

Neural network (NN) encoding schemes for DNA analyses were discussed by the following researches. Brunak et al. described a direct input encoding scheme, utilizing four units to represent the nucleotides [11]. Demeler & Zhou presented a dense representation using two units for four nucleotides [12]. They demonstrated how this direct encoding scheme performs better than using four units. Wu & Shivakumar discovered an indirect encoding scheme that calculates frequencies of nucleotide bases either individually or in k-tuples. These frequencies are fed to the classifier as input [13]. Farber et al. contrasted the pros and cons of direct versus indirect encoding methods [14]. He proved that indirect 2-Mer frequency representation schemes performed better than direct encoding methods. Attaluri [15] compared and contrasted the effect of different NN encoding schemes on the classification performance of Influenza-A. He discovered that including the gaps in direct encoding schemes generated more accurate classification results. He additionally determined the k-frequencies at which indirect encoding schemes generated the best results.

Yuan et al. [16] compared and contrasted decision trees (DTs) and support vector machines (SVMs) in classifying gene expressions. They did so by building a SVM bank holding all the possible encoding schemes. They opted for using nucleotide sequence data. An example scheme they used was (00 01 10 11) to code for the four different nucleotides: ‘A’, ‘C’, ‘G’, and ‘T’.

To enhance machine learning efficiency in Bioinformatics further, feature selection is usually applied to datasets with high dimensionality. Saeys et al. [17] describe the pros and cons of three types of feature selection techniques when applied to Bioinformatics problems: filter, embedded, and wrapper methods. They emphasized that speedy, univariate, filter methods are favored when analyzing high dimensionality data. Leung et al. [18] utilized a filter method known as information gain to indentify the biomarkers in Hepatitis-B DNA sequences responsible for liver cancer. They utilized Rule Learning and Nonlinear Integrals to classify their data after the feature selection step. The feature selection technique has not been tried on other viruses.

The following key researches on the classification of Influenza-A via machine learning were conducted by ElHefnawi et al. [19, 20] and Attaluri [15]. ElHefnawi et al. [19] classified Influenza-A hosts and subtypes using hidden Markov models (HMM). Protein sequences from the HA pertaining to subtypes H1, H2, H3, H4, and H5 segment were used. To classify hosts, the virus sequences were further divided into human and non-human hosts. The research yielded an accuracy of 100 % for subtype classification and accuracies ranging from 50 % to 100 % for host classification, depending on subtype. Although sequences pertaining to subtypes H1 and H2 yielded higher host classification accuracies, the performance was reduced when using sequences pertaining to subtypes H3 and H5. This drop in performance can be explained by the shortage of documented protein sequence targeting human hosts in the latter subtypes. The method didn't analyze classification performance at the RNA level or the effect of applying of feature selection prior to classification.

In another experiment, El Hefnawi et al. [20] used HMM and DTs to find host associated signatures in order to increase classification accuracy. The experiments were conducted on the HA protein of various subtypes. DTs yielded higher host classification accuracies, ranging from 92 % to 100 %, as opposed to HMMs. The research did not explore extracting host associated signatures at the RNA level.

Attaluri [15] analyzed the use of NNs, DTs, and SVMs on the classification performance of Influenza-A hosts and subtypes. He conducted the experiments using cDNA and protein sequences. A subset of sequences belonging to the H1N1 strain was selected for host classification. Sequences belonging to the H1, H2, H3, N1, and N2 subtypes were selected for subtype classification. The overall classification accuracies, for both subtype and host classification, were 96.5 %, 96.2 %, and 95.1 % when using DTs, SVMs, and NNs respectively.

Attaluri [15] additionally integrated DTs and HMMs to classify the Influenza-A data. He used DTs to extract informative positions from the cDNA sequences then converted these positions into their protein equivalent. The attained protein sequences were then fed as input to a HMM classifier. Both viral hosts and subtypes classification were analyzed. The technique yielded an overall accuracy of 97 %.

The results attained by Attaluri, however, are not suitable for comparing the performances of the different classifiers used to each other. This was due to the use of varying data along different classifiers. Different classifiers did not always classify the same type of virus hosts. E.g. The DTs and SVMs were trained with sequences infecting swine and human hosts, while the NN was trained with sequences infecting human, avian, and swine hosts. The type of data used for classification was additionally not consistent. For instance, while DTs and NNs were trained using DNA data, SVMs and HMMs were trained using protein data. Thus, a fair comparison between classifier performances cannot be made.

While the integrated model using DTs and HMM performed feature selection using cDNA data, the model did not consider feeding the results directly to the HMM classifier. It instead requires the informative positions to be converted first into protein format. In a situation where an epidemic or pandemic suddenly arises, the protein data on that particular virus won't be initially available and the aforementioned classification method cannot be used. Lastly, while the performance

of the integrated model was assessed, the effect of the feature selection on classifier efficiency or speed was not measured. The study additionally did not explore using feature selection to improve classifiers' performances other than HMM.

In a previous conference paper [21], the performance of NNs and DTs using information gain as a feature selection technique were compared for two subtypes and two segments of the Influenza-A virus. In this research we hope to improve on all the aforementioned points mentioned by analyzing the improvement of classification more thoroughly. This will be achieved by: analyzing the classification results of more than one viral segment; concentrating on the H1 subtype since it yielded better results with human-nonhuman binary classification of viral hosts; and last but not least measuring the statistical significance of the results of two classifiers in order to compare them and select the more efficient classifier.

1.4 Aim of the Research

The aim of the research is to find which classifier generates more efficient results when coupled with information gain algorithm, and to enhance Influenza-A host detection using cDNA. This should be done without deteriorating classifier performance. Two classical machine learning methods neural networks and decision trees are contrasted and compared to measure their potential in host identification.

The same techniques employed in [21] are used but with more viral segments to ensure result authenticity. The objective will be achieved as follows: Using information gain as a feature selection technique to extract the most important positions for host classification; feeding the informative positions directly from several segments as inputs to NNs and DTs to contrast their performance; comparing the results attained using feature selection technique to the base classification performance when no informative positions are used; recording time taken for classifiers construction as an extra performance parameter to measure classifier efficiency; evaluating the results using accuracy, precision, specificity, sensitivity, and time; measuring the statistical significance of the results generated by both classifiers in comparison to each other.

2 Information Gain as a Feature Selection Technique for Influenza-A Data

Information gain (IG) is the feature selection method of choice to be applied prior to the classification step. It was selected due to being a filter technique; thus it can scale well with the high dimensionality data of Influenza-A; it is applicable on several classifiers due to being classifier independent. The latter is useful in this research's scope to compare the effect of feature selection on the efficiency of the classifiers under study.

IG is able to detect features(s) possessing the most information based on a specific class. It is derived from an attribute known as entropy using Eqs. 50.2, 50.3, 50.4, and 50.5. Details on derivation of IG from entropy can be found in [18]. Entropy is a measure of the uncertainty of class using the probability of a certain event or attribute. It is inversely proportional to IG. When defined over a binary training set *S*, it is calculated using the following equation:

$$E(S) = -[P(+) \times log\ P(+)] - [P(-) \times log\ P(-)] \tag{50.1}$$

Where *P(+)* represents the probability of the positive class in the training set, while *P(−)* represents the probability of the negative class. When more than two classes are classified the entropy is measured using:

$$E(X) = \sum_{i=1}^{m} -P(x_i) \log_2 P(x_i) \tag{50.2}$$

Where $m =$ number of classes being analyzed. In the scope of this research, the entropy is calculated based on the virus hosts. Since cDNA data is used, the value of the nucleotide positions are consequently used as the classification attributes or features. By deriving the IG values from the entropies of the nucleotide positions, the most informative positions in the sequence can be identified. The amount of features or nucleotide positions used for classification can thus be reduced. The entropies of the nucleotide positions are calculated using the probabilities of the four nucleotides that make up the sequence, 'A', 'G', 'C', or 'T'. This is done relative to the Influenza classes in consideration: avian, human and swine. Further details of how this is achieved can be found in [21]. The probabilities are then used to determine the entropy as follows:

$$R(X) = \sum_{i=0}^{m} \frac{\sum_{k=1}^{K} C_{ki}}{\sum_{k=1}^{K} C_k} E(P(c_{li}), \ldots, P(c_{ki})) \tag{50.3}$$

A weighted average of the entropies is then calculated by multiplying the overall probability of each nucleotide in a certain position with the entropy of that nucleotide [21]. The sum of weighted averages of the entropies of all four nucleotides will be named the remainder, *R(X),* as shown below:

$$R(X) = \sum_{i=0}^{m} \frac{\sum_{k=1}^{K} C_{ki}}{\sum_{k=1}^{K} C_k} E(P(c_{li}), \ldots, P(c_{ki})) \tag{50.4}$$

The equation above can be explained as follows: $i = 1, \ldots, m$ represents the nucleotide at a certain aligned site whose sequence is 'A', 'G', 'C', or 'T'. Assuming

the data has K classes $C = C1, C2, \ldots, C_k$, then $|C_k|$, where $k = 1,2,\ldots,K$, is the number of sequences in class C_k. $|C_{ki}|$ as a consequence is the number of sequences in class C_k, whose character at the aligned site is i. $\frac{|C_{ki}|}{|C_k|}$ is thus the probability of nucleotide i, in a certain position and $E(P(c_{li}), \ldots, P(c_{ki}))$ is the entropy of the nucleotides relative to the classes we want to identify. $R(X)$ is calculated, once per nucleotide position, over a group of sequences. After determining the remainder, IG is calculated using:

$$IG_J = E(C) - R(J) \tag{50.5}$$

Where $E(C)$ is the class entropy and J is the nucleotide position. $E(C)$ is calculated, once, using (50.2). Taking host classification as an example, to calculate $E(C)$, the entropies of the three main classes; swine, avian, and human, are calculated using the number of sequences belonging to each class. Like the remainder, IG_J is calculated once per nucleotide position by subtracting the remainder, $R(J)$, of each nucleotide position from the value obtained from $E(C)$. Once the IG is calculated for all nucleotide positions, over a group of sequences, the nucleotide positions are arranged in descending order according to their IG value. The best hundred positions with highest IG are then used to build and train a classifier of choice. As a rule, calculating IG should be done after the genetic data has been preprocessed by multiple sequence alignment (MSA). This is in order to group similar sequences together and unify sequences' length.

The equations are implemented from scratch using *Matlab*. The informative positions extracted using IG will represent conserved regions between the viruses, at the cDNA level, within each host. Although the exact function of most of these regions are not known they can prove to be medically significant if analyzed more closely.

3 Methodology Summary

The following chapter explains the experimentation steps. The main steps are: Virus subtype and segment specification, data collection and preprocessing, feature selection using IG, classifier construction, classification based on hosts, and classifier evaluation and comparison.

– *Subtype and Segment Specification:* Sequences belonging to the H1 dataset were selected. Unlike the previous experiment in [21] only the H1 subtype was selected, as it yielded better results for human vs. non-human classification. Details of the sequence frequencies per host before preprocessing are available in Table 50.1. Contrary to the previous experiment in [21] where only two segments were selected, all eight viral cDNA/RNA segments were used. This was done in

Table 50.1 Frequency of cDNA sequences contained per segment, for subtype H1, before and after data preprocessing

	Before preprocessing H1 cDNA		*After preprocessing H1 cDNA*	
Segment	Human	Non-human	Human	Non-human
PB2	1,338	1,236	1,338	1,236
PB1	1,505	1,342	1,505	1,342
PA	754	913	754	913
HA	1,395	1,342	1,395	1,342
NP	1,389	1,189	1,389	1,189
NA	1,409	1,268	1,409	1,268
M1	1,277	1,529	1,277	1,529
NS1	1,348	1,135	1,348	1,135

order to: observe if the results differ significantly from segment to segment; in order to better compare the performance of NNs to DTs.

- *Data Collection*: Complete and near complete cDNA sequences were selected from the online Influenza database: http://flu.lanl.gov. The most important Influenza hosts were selected: Avian, Human, and Swine. This was repeated for all virus segments pertaining to subtype H1.
- *Data Alignment*: The collected data was aligned using a MSA program to unify the sequences' length. An online MSA program *Mafft*, was used due to its capability to align high dimensionality data rapidly. The FFT-NS-2 function, which contains a modified version of Fast Fourier Transform (FFT), was used to achieve this.
- *Feature Selection:* For each segment in subtype H1, the best hundred informative positions of the cDNA sequences were determined. This was done by measuring IG across the selected hosts: Avian, Human, and Swine. Although the binary classifier built in the scope of this paper will differentiate only between human and non-human hosts, the IG algorithm is run on all three hosts to allow for future expansion of the project.
- *Classifier Construction:* Binary classifiers that differentiate human viral strains from non-human strains were constructed. This was implemented once with NNs and once with DTs. The two aforementioned classifiers were selected for the following reasons: NNs were chosen for their ability to classify patterns in the presence of noisy or unbalanced data; a common place phenomenon in the Influenza-A dataset. DTs were selected due to their ability to generate, easy to understand, classification rules. This is important in identifying RNA signatures in Influenza-A, that are specific to each viral host.
- *Classifier Evaluation:* Classifier performance was evaluated by running the set of experiments detailed below:

  (a) *Creating Benchmarks*: The base classification performance of the DTs and NNs using raw cDNA, unprocessed by IG, is obtained.

(b) *Measuring the Impact of Feature Selection*: IG is used to extract the best 100 informative positions from the cDNA sequence. These positions are then directly fed to the NNs and DTs as input. The classification performance is then analyzed.

(c) *Comparing the Performance*: After running both experiments in the first two steps, the classification performance is compared based on accuracy, sensitivity, specificity, and precision. The time taken for the classifier to test 200 sequences is recorded and used for measuring efficiency. In this way the benchmark performance is compared to the performance attained when using feature selection. The method is also used to compare the efficiency of both classification techniques.

Figure 50.2 provides a summary of the overall methodology described above. It shows how the most important positions in the cDNA data from each segment and subtype are used in order to classify the Influenza-A hosts.

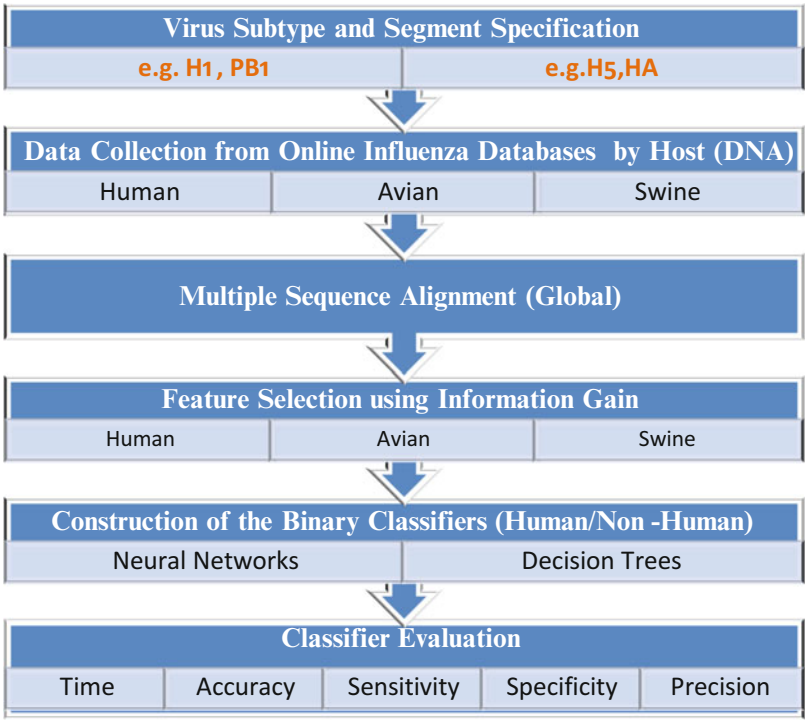


Fig. 50.2 Summary of the methodology

4 Results

The following section details the setting used for classification and the corresponding results attained from running the experiment described in the Methodology section. The results for DTs are shown in Sect. 4.1, while the results for NNs are shown in Sect. 4.2.

4.1 Experiment-1: Measuring the Impact of Using the Most Informative Positions on Classification Performance

- *Objective:* To measure the impact of using the most informative positions more thoroughly on the accuracy or speed of classification; C4.5 DTs are used to demonstrate this.
- *Method:* The section below explains additional settings and steps used during the classification process.

  (a) *Classifier Settings*: The training dataset of the DTs is divided into sequences infecting humans and non-humans respectively. To prevent overtraining, data preprocessing is performed to balance the dataset as shown in Table 50.1. The *Weka* program is used to train, build, and test the DTs. Details of how the DT was configured can be found in Sect. 4.1 in [21]. The DT was constructed so that: the nucleotide position represented the DT's nodes; the nucleotide value of the positions represented the attributes on the node; and the host classes represented the DT's leaf nodes. The classifier is used to classify the viral segments based on host, once with aligned DNA data, and once with data where the best 100 positions were selected.

- *Results*: Tables 50.2 and 50.3 demonstrate the impact of using the most informative positions on DTs as opposed to using raw aligned cDNA data. cDNA sequences belonging to the H1 dataset were used to generate the results in both tables. The 'Time' column represents the time to test 200 sequences. There is a noted speed increase when using the most informative positions as opposed to using raw cDNA data.

4.2 Experiment-2: Measuring the Impact of Using the Most Informative Positions on a Different Classifier

- *Objectives*: To measure the improvement in classification efficiency, after applying IG to NNs, more thoroughly; to compare the performance of NNs' to DTs. A typical three-layered, feed forward NN is used.

Table 50.2 DT performance, per segment, when using raw cDNA data pertaining to subtype H1 for host classification

Segment	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Time (s)
PB2	99.4	99.3	99.4	99.5	0.15
PB1/PB1_F2	98	97.5	98.6	98.7	0.18
PA/PA-X	99	98.4	99.5	99.3	0.17
HA	98.7	98.8	98.7	98.7	0.17
NP	99.5	99.5	99.5	99.6	0.11
NA	99.2	98.7	99.7	99.7	0.11
M1/M2	98.9	98.4	99.3	99.2	0.05
NS1/NS2	99.1	98.8	99.4	99.5	0.06

Table 50.3 DT performance per segment when 100 informative positions are used. The experiment was conducted on sequences pertaining to the H1 subtype

Segment	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Time (s)
PB2	99.3	99.1	99.4	99.5	0.01
PB1/PB1-F2	98.1	97.8	98.4	98.5	0
PA/PA-X	98.9	98.1	99.5	99.3	0.01
HA	97.4	97.2	97.5	96.6	0.02
NP	99.3	99.1	99.6	99.6	0.01
NA	99.2	99.4	99.1	99.2	0.01
M1/M2	99.0	98.5	99.5	99.4	0.02
NS1/NS2	98.9	98.4	99.4	99.5	0.01

– *Method*: The section below explains additional settings and steps used during the classification process.

  (a) *Classifier Settings:* The same datasets used in Experiment-1, shown in Table 50.1, are used to train the NNs. Details of the NNs' configuration are available in [21] in Sect. 4.2.
  (b) *Classifier Evaluation:* The results, attained using IG with DT and NN classification as illustrated by Tables 50.3 and 50.5, are run through the *Anova* algorithm. The *Anova* algorithm is used to generate box plots which are then used as a rough measure of statistical significance of the performance measures. This is done to better evaluate the performance of both classifiers compared to one another when IG is used. *Matlab* is used to achieve this.

– *Results*: The results of the NN host classification of both datasets are summarized in Tables 50.4 and 50.5. The tables show the impact of using the top 100 informative positions on NN network classification as opposed to the benchmark; the classification speed increased without marked deterioration in classification performance. The 'Time' column, in both tables, represents the time to test 200 sequences. Figure 50.3 shows the box plots attained from the *Anova* algorithm when comparing the performance measures of NNs & DTS that were

Table 50.4 The performance of NNs on the H1 dataset per segment when raw aligned cDNA sequences are used to classify hosts

Segment	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Time (s)
PB2	99.1	99	99.4	99.4	0.40
PB1/PB1-F2	98.9	98.3	99.6	99.6	0.62
PA/PA-X	99.2	98.7	99.7	99.6	0.26
HA	96.9	98.4	95.4	95.7	0.41
NP	99.4	99.1	99.7	99.8	0.34
NA	99.2	99	99.4	99.4	0.39
M1/M2	99.2	98.7	99.6	99.5	0.19
NS1/NS2	99.4	99.1	99.7	99.8	0.17

Table 50.5 The performance of NNs per segment when 100 informative positions are used. The experiment was conducted on sequences pertaining to the H1 subtype

Segment	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Time (s)
PB2	99.3	99	99.6	99.6	0.04
PB1/PB1-F2	99.0	98.6	99.5	99.5	0.05
PA/PA-X	99.2	98.4	99.8	99.7	0.04
HA	98.1	97.7	98.3	97.6	0.04
NP	99.5	99.7	99.2	99.4	0.04
NA	98.4	97.5	99.4	99.5	0.04
M1/M2	99.3	98.7	99.7	99.0	0.04
NS1/NS2	99.4	99.2	99.7	99.8	0.04

constructed with 100 informative positions. Group-1 represents the result of NN classification, while Group-2 represents the result of the DT classification. The segments used to generate the results belong to the H1 subtype. Accuracy, sensitivity, specificity, precision, and time are analyzed respectively.

5 Discussion

In Experiment-1, Tables 50.2 and 50.3 shows that building the DT host classifiers using the best 100 most informative positions increased the classification efficiency without performance deterioration. The accuracy, sensitivity, specificity, and precision stayed comparable to the results attained in the benchmark represented by Table 50.2. The classification speed increased by an overall of 92 % with respect to the benchmark. The four performance measures did not show a significant decrease in values when comparing the results attained from using the most informative 100 positions to the benchmark. The values of all four performance parameters remained above 90 %. This shows that using a combination of the most informative positions and DTs increases the classification efficiency of viral sequences belonging to the H1 subtype.

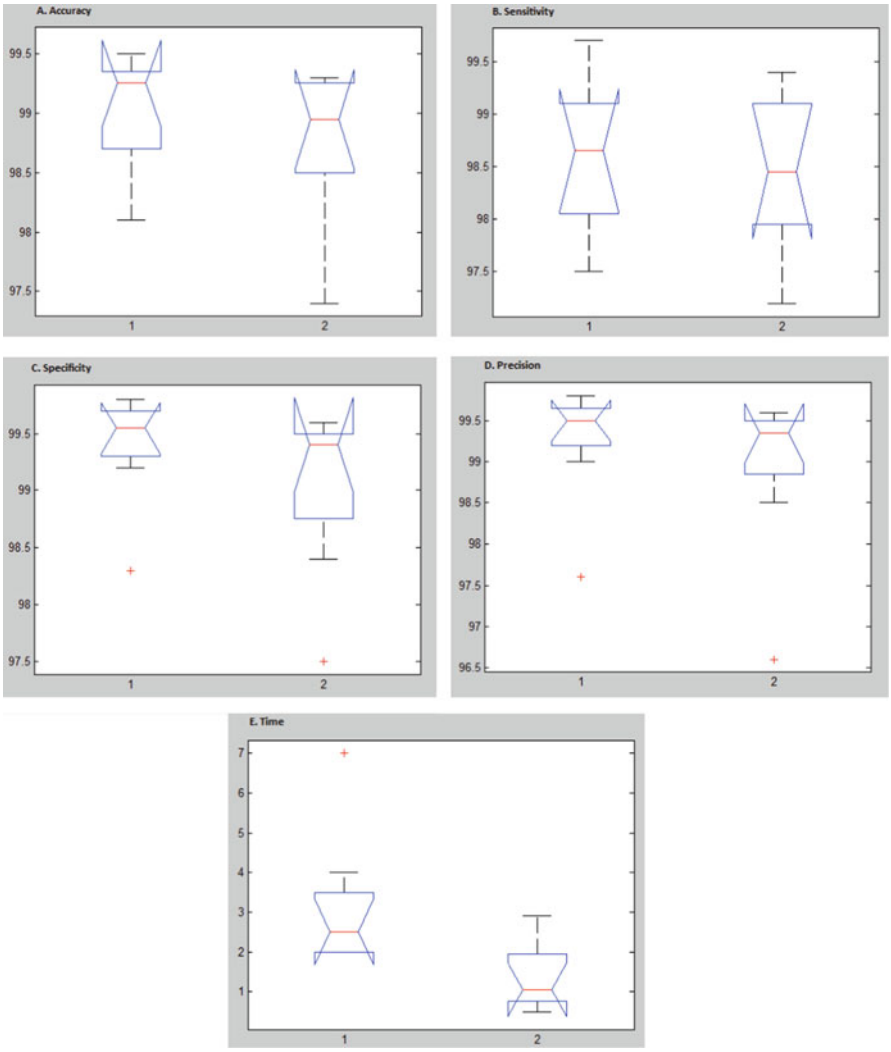


Fig. 50.3 Comparison between NNs and DT respectively using box plots generated with the Anova algorithm

In Experiment-2, Tables 50.4 and 50.5 show that using the most informative positions to train the NN also increased classification speed. A deeper analysis of the results shows an overall significant increase in testing speed of 88 % when using 100 informative positions for classification compared to the benchmark. The four performance parameters, on the other hand, show no significant decrease in value. This confirms that the overall classifier performance does not deteriorate, compared to the benchmark, when the most informative 100 positions are used for classifying the H1 dataset.

Lastly, the box-plots in Fig. 50.3 shows that NNs show an overall increased performance in terms of accuracy, sensitivity, specificity and precision over DTs. NNs however require more time for constructing and testing the classifiers, than DTs, as shown in section-E of Fig. 50.3.

6 Conclusion

The results of Tables 50.2 and 50.3 show that feature selection using IG can improve DT classification efficiency, in terms of speed, for the H1 subtype. The results of Tables 50.4 and 50.5 show that feature selection using IG greatly increases NN classifier efficiency, in terms of speed, for the same subtype.

The overall results of Tables 50.2, 50.3, 50.4, and 50.5 further verifies the claim that using the most informative positions can increase cDNA host classification efficiency or speed. This can be achieved regardless of the segment being classified as well as the classifier used, when using subtype H1. Finally, the box-plots in Fig. 50.3 show that despite NNs being more accurate, DTs are more efficient in classifying sequences pertaining to the H1 subtype. This is because the decrease in performance compared to classification speed achieved with DTs is not significant.

Based on the above conclusions, feature selection using IG can be used to improve the efficiency of cDNA host classification of various classifiers, when analyzing the H1 subtype. This opens two main future research areas. The first research area is that the most informative positions extracted during feature selection, which represent host associated signatures at the RNA level, can prove to be biologically significant with further experimentation. The second research area involves using IG in combination with classifiers to improve the cDNA classification efficiency of other influenza-A problems in the field such as virus subtype and viral anti-resistance determination. E.g. Given that the strain H1N1 is responsible for contracting swine flu, combining IG with DTs to identify viral antiresistance of the H1N1 viral strain and measuring the resulting classifier performance can be a future scope of research to consider.

References

1. Bouvier, N.M., Palese, P.: The biology of influenza viruses. Vaccine **26**, D49–D53 (2008)
2. Ghedin, E., Sengamalay, N., Shumway, M., Zaborsky, J., Feldblyum, T., et al.: Large-scale sequencing of human Influenza reveals the dynamic nature of viral genome evolution. Nature **437**, 1162–1166 (2005)
3. Fislova, T., Kostolansky, F.: The factors of virulence of Influenza-A virus. Acta Virol. **49**, 147–157 (2005)
4. Gabriel, G., Dauber, B., Wolff, T., Planz, O., Klenk, H.D., Stech, J.: The viral polymerase mediates adaptation of an Avian Influenza virus to a mammalian host. Proc. Natl. Acad. Sci. **102**, 18590–18595 (2005)

5. Pedersen, J.C.: Hemagglutination-inhibition test for Avian Influenza virus subtype identification and the detection and quantitation of serum antibodies to the Avian Influenza virus. Methods Mol. Biol. **436**, 53–66 (2008)
6. Pedersen, J.C.: Neuraminidase-Inhibition assay for the identification of Influenza-A virus Neuraminidase subtype or Neuraminidase antibody specificity. Methods Mol. Biol. **436**, 67–75 (2008)
7. Song, D.S., Lee, C.S., Jung, K., Kang, B.K., Oh, J.S., Yoon, Y.D., et al.: Isolation and phylogenetic analysis of H1N1 Swine Influenza virus isolated in Korea. Virus Res. **125**, 98–103 (2007)
8. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. **25**, 3389–3402 (1997)
9. Sami, A., Takahashi, M.: Decision tree construction for genetic applications based on association rules. IEEE TENCON 2005. pp. 21–25. Melbourne (2005)
10. Salzber, S.L., Delcher, A.L., Kasif, S., White, O.: Microbial gene identification using interpolated Markov models. Nucl. Acids Res. **26**, 544–548 (1980)
11. Brunak, S., Engelbrecht, J., Knudsen, S.: Neural network detects errors in the assignment of mRNA splice sites. Nucl. Acids Res. **18**, 4797–4801 (1990)
12. Demeler, B., Zhou, G.: Neural network optimization for E. coli promoter prediction. Nucl. Acids Res. **19**, 1593–1599 (1991)
13. Wu, C.H., Shivakumar, S.: Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences. Nucl. Acids Res. **22**, 4291–4299 (1994)
14. Farber, R., Lapedes, A., Sirotkin, K.: Determination of eukaryotic protein coding regions using neural networks and information theory. J. Mol. Biol. **226**, 471–479 (1992)
15. Attaluri, P.K.: Classifying Influenza subtypes and hosts using machine learning techniques. ProQuest. UMI Dissertation Publishing (2012)
16. Yuan, X., Xiaohui Yuan, Yang, F., Peng, J., Buckles, B.P.: Gene expression classification: decision Trees vs. SVMs, FLAIRS (2003)
17. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Brief Bioinform. **23**, 2507–2517 (2007)
18. Leung, K.S., Lee, K.H., et al.: Data mining on DNA sequences of Hepatitis B virus. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 428–440 (2011)
19. ElHefnawi, M., Kadah, Y.M., Sherif, F.: Influenza-A subtyping and host origin classification using profile hidden Markov models. J. Mech. Med. Biol. **12**, (1240009) 1–11 (2012)
20. ElHefnawi, M., Kadah, Y.M., Sherif, F.: Accurate classification and Hemagglutinin amino acid signatures for Influenza-A virus host-origin association and subtyping. Virology **449**, 328–338 (2014)
21. Shaltout, N., Rafea, A., Moustafa, A., El Hefnawi, M.: Information gain as a feature selection method for the efficient classification of Influenza based on Viral hosts. Lecture Notes in Engineering and Computer Science*:* Proceedings of the World Congress on Engineering 2014, WCE 2014, pp. 625–631. London 2–4 July 2014

Chapter 51
Process Optimization, Empirical Modeling and Characterization of Biodiesel from Cottonseed Oil

Umaru Musa, Aboje Audu Alechenu, Ibrahim Aris Mohammed, Aliyu Musa Aliyu, Muhammad Munir Sadiq, and Aminat Oladunni Olaibi

Abstract The optimization of biodiesel production from a Typical Nigerian cottonseed oil using sodium hydroxide as catalyst is presented. 2^3 factorial design was employed to study the combined effect of temperature (40–60 °C), mole ratio of methanol to oil (6:1–8:1) and catalyst concentration (0.5–1.0 wt%). The optimum yield of 96.23 wt% was obtained at 0.5% w/v potassium hydroxide at a temperature of 40 °C and Molar ratio of 6:1 for 90 min. A Model equation that shows the relationship between biodiesel yield and process variables was developed using a statistical tool. Statistical Analysis of variance (ANOVA) of the data shows that molar ratio has the most pronounced positive effect on the biodiesel yield. The methyl ester was characterized and all the properties were consistent with the ASTM standard.

Keywords Biodiesel • Cottonseed oil • Modeling • Optimization • Process variables • Transesterification

1 Introduction

Energy is globally regarded as an index for all forms of development [1]. A large percentage of the world's total energy output is generated from petroleum fossil fuels [2]. Humanity's dependence on petroleum for energy comes at high cost. Experts already suggest that the world current oil and gas reserves would only last a few more decades [3]. Besides emissions from the combustion of fossil derived fuel

U. Musa (✉) • A.A. Alechenu • I.A. Mohammed • M.M. Sadiq • A.O. Olaibi
Department of Chemical Engineering, Federal University of Technology Minna, Minna, Nigeria
e-mail: umar.musa@futminna.edu.ng; ma.ibrahim@futminna.edu.ng; mmsadiq@futminna.edu.ng; aminatolaibi@gmail.com

A.M. Aliyu
Process Systems Engineering Group, School of Engineering, Cranfield University, Cranfield, UK
e-mail: alimaliyu@futminna.edu.ng; a.m.aliyu@cranfield.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4_51

are the greater contributors to global warming. The concern for climate change and energy sustainability has stimulated the widespread search for cheap, renewable and eco-friendly alternative energy sources [4].

Bioethanol, biohydrogen and biodiesel are examples of renewable energy source derived from biomass [5]. Among these biofuels, biodiesel derived from vegetable oils is gaining acceptance and market share as Diesel fuel in Europe and the United States [6]. Therefore, in recent years several researches have been undertaken on the use of vegetable oils for biodiesel production as replacement for petroleum diesel in diesel engines [7]. Biodiesel is a mono-alkyl ester of fatty acids derived from vegetable oils or animal fats. It is usually produced from the transesterification of vegetable oils or animal fats with methanol or ethanol in the presence of a catalyst. Biodiesel is one of the most attractive biofuels because of its excellent biodegradability, higher flash point, high lubricity, reduced exhaust emissions, miscibility in all ratios with petro diesel, ability to be used in internal existing compression without modification [4]. Biodiesel production is worthy of continuous study and optimization, because of its environmentally friendly characteristics and renewable nature. A major hurdle towards widespread commercialization is the high price of biodiesel [8]. Feedstocks are reported to account for the high cost (70–95%) of total cost of biodiesel production [9, 10]. In recent times almost all commercial processes for biodiesel production around the world use edible vegetable oils such as soya bean oil, sunflower oil, rapeseed oil, palm oil, peanut oil, canola oil, and corn oil as their feedstock [11]. The high price of these edible plant oils is responsible for the inability of biodiesel to compete favourably with petroleum diesel [12]. One important approach towards the reduction of this cost is the use of relatively cheap vegetable oil (non-edible oils, animal fats and waste vegetable oil from extensive cooking or process of vegetable oil refining) [13]. Another vital way of minimizing the cost of biodiesel production is to improve on the production technology through optimizing the fundamental variables that affect the yield and purity of biodiesel [8]. According to Ma and Hanna, [14] higher yields are obtained when these fundamental reaction conditions are optimized. These variables include catalyst type and concentration, molar ratio of alcohol to vegetable oil, temperature, time, agitation speed and the purity of reactants.

In Nigeria for example, common vegetable oils available for biodiesel production include palm oil, palm kernel oil, groundnut oil, soyabean oil, Jatropha oil, castor oil, neem oil and cotton seed oil. The exploration of cottonseed oil and sesame seed for biodiesel production have not been thoroughly investigated. Cotton plant is a naturally grown perennial plant belonging to the family of *Malvaceae*, tribe of *Gossypieae* and genus of *Gossypium*. The plant bears seeds that contain about 15% of oil [15]. The seeds have been used for the production of cattle feed but the presence of a natural toxin (gossypol) is a potential risk to this animal. The oil has been reportedly used for the production of potato chips, doughnuts, biscuits, alkyl resins, lubricants and soaps. But research has shown that the seed contains high levels of saturated fat and pesticide residues making it unfit for human consumption. The response to this health concern has seriously resulted into fall in demand of this oil in the food industry [16]. This oil is characterized by with excellent stability

to oxidation. The seeds do not have practical uses except for oil production and so the anticipated usage for biodiesel production is not expected to have negative consequences on food value [17]. A few number of work have been reported on biodiesel production from cottonseed oil in different part of the globe [17–23]. There is however a dearth of literature of the optimization of the production of biodiesel from a typical Nigerian cottonseed oil using factorial designs. Factorial design is an example of technique employed in design of experiments. It is a planned approach of experiments to explore cause and effect relationships. It is a proficient means of minimizing the total number of experimental trial while simultaneously altering all variables that formalize an experimental system. This approach provides a clear choice of strategy that enables realistic solutions to be obtained after each sequence of experiments in a multivariable system [24].

This study is therefore aimed at investigating the optimization, empirical modelling and characterization of biodiesel production from a typical Nigerian cottonseed oil with a view to developing model that establishes the relationship between biodiesel yield and the process variables studied.

2 Methodology

2.1 Materials

The feedstock for this study (Refined cottonseed oil) was obtained from Lagos, Nigeria, while the chemicals such as methanol and sodium hydroxide used were manufactured by Aldrich Co. Ltd England. All the equipment and glass wares were obtained and used in Federal University of Technology Minna, Nigeria.

2.2 Equipment

The experiment was conducted in a 500 mL three-necked flat-bottom flask equipped with a reflux condenser (to reduce the loss of methanol by evaporation), thermometer, and a stopper to add the catalyst solution. The reaction mixture, 100 g by weight of the refined cottonseed oil was heated and stirred by a hot plate with a magnetic stirrer.

2.3 Design of Experiments

In this study, 2^3 factorial experimental designs were employed to determine the optimum conditions. Three variables, which includes Alcohol to Oil mole ratio; Reaction Temperature; and Catalyst Concentration were studied at both high and low

Table 51.1 Variation of variables for the 2^3 factorial design

	Lower level		Higher level	
Properties	Coded	Actual	Coded	Actual
Alcohol to oil mole ratio	−1	1:6	+1	1:8
Reaction temperature	−1	40	+1	60
Catalyst concentration	−1	0.5	+1	1

levels with the response been methyl ester yield. The high level of alcohol/oil mole ratio was 1:8 and the low level was 1:6. The high level of temperature was chosen at 60 °C and the low level at 40 °C. The high level of catalyst concentration was chosen at 1.0% and low level was 0.5% sodium hydroxide catalyst by weight of cottonseed oil. A constant reaction time of 90 min and constant agitation were maintained throughout the experiment. The design matrix is as shown above (Table 51.1).

2.4 Transesterification Procedure

100 g of refined cottonseed oil was measured using weighing balance and was pre-heated to 70 °C using a hot plate magnetic stirrer equipped with a thermometer, to remove residual water or moisture from the oil. The oil sample was then allowed to cool to about 45 °C and was maintained at this temperature. 0.5 wt% of sodium hydroxide was dissolved into the required volume of methanol (having 99.5% purity). The solution (methoxide) was then poured into the reactor containing the pre-heated oil to react. The reaction took place for 90 min at the required reaction temperature with the agitation maintained at 300 rpm.

2.5 Purification

After the reaction, the mixture was poured into the separating funnel and allowed to settle under gravity for 24 h. The mixture separated into two separate layers. The upper layer contained biodiesel, excess methanol, soap, and residual catalyst while the lower layer contained glycerol. The glycerol was gradually drained off leaving behind the upper layer.

Phosphoric acid was added to neutralize the residual catalyst, washed with warm distilled water of between 40 to 45 °C to remove methanol, residual catalyst and soaps. During the washing process, warm water was added to the biodiesel produced in the separating funnel and allowed to settle. The lower layer was drained off again until a clear lower layer with no trace of soap was observed. The biodiesel was then dried using a hot plate to remove the water content. The resultant product is an amber coloured substance known as biodiesel.

2.6 Fuel Properties Determination

The characterization of the methyl ester produced was performed in accordance with the ASTM Standard as reported by Gerpen et al. [25].

3 Results and Discussion

3.1 Effect of Individual Parameters

3.1.1 Effect of Molar Ratio of Methanol to Oil

The molar ratio of alcohol to oil is an important variable that affect the yield of biodiesel [14] (Table 51.2).

In this study the biodiesel yield was observed to decrease as the molar ratio of alcohol to oil increases from 6:1 to 8:1. This is due to the increase in the polarity of the reacting system. According to the Lalita et al. [26] increasing molar ratio of methanol to oil above 6:1 decrease ester yield because higher amount of the alcohol interferes with the separation of glycerin due to increase in solubility and so part of the glycerin remained in the biodiesel phase which in turn drive the equilibrium back to the reverse reaction. The result of investigation shows that 6:1 was the most appropriate molar ratio. This finding show appreciable consistency with the work of Freedman et al. [27]. Therefore, increasing molar ratio of methanol to oil beyond 6:1 did not really increase biodiesel yield as well as the ester content but rather complicated the ester recovery process and thereby raising the cost of methanol recovery.

Table 51.2 Design matrix for the 2^3 factorial for the transesterification of refine cottonseed oil showing experimental and predicted result

	Alcohol to oil mole ratio		Reaction temperature		Catalyst concentration			
Run	Coded	Actual	Coded	Actual	Coded	Actual	Experimental value	Predicted value
1	−1	1:6	−1	40	−1	0.5	96.23	96.19
2	+1	1:8	−1	40	−1	0.5	80.66	80.00
3	−1	1:6	+1	60	−1	0.5	92.10	91.44
4	+1	1:8	+1	60	−1	0.5	82.15	82.11
5	−1	1:6	−1	40	+1	1	92.30	92.34
6	+1	1:8	−1	40	+1	1	88.10	88.76
7	−1	1:6	+1	60	+1	1	82.02	82.68
8	+1	1:8	+1	60	+1	1	85.91	85.95

3.1.2 Effect of Catalyst Concentration

Sodium hydroxide as a catalyst was used at two levels (0.5 and 1.0 wt%). It can be observed that the methyl ester yield decreases with increase in catalyst concentration. This can be due to the formation of soap by the oil and excess sodium hydroxide used. The addition of an excess amount of catalyst beyond the optimum give rise to the formation of an emulsion increased the viscosity and led to the formation of gels. This also hinders the glycerin separation resulting into dilution of methyl ester and ester yield tend to decrease [28]. This result shows quantitative agreement with the Patil and Deng [29] who reported an optimum of 0.5 wt% for transesterification of karanja oil but differ with the 1 wt% reported for Jatropha oil. This variation could be attributed to the purity of the oil used in this study.

3.1.3 Effect of Reaction Temperature

Temperature is one of the important parameter for the production of biodiesel because the rate of reaction is strongly influenced by the reaction temperature. Transesterification can occur at different temperatures depending on the oil used [14]. The effect of temperature was investigated at 40 °C and 60 °C respectively. It was observed that the biodiesel yield decreases from 96.23 to 92.10 wt% when the temperature was increased from 40 °C to 60 °C at a molar ratio of 6:1, catalyst concentration of 0.5 wt% and reaction time of 90 min. This could be attributed to the fact that at higher temperature the solubility of the reactants increases leading to reduction in the separation of methyl ester and glycerol phase. It is important to add that higher reaction temperature of 60 °C that was close to the boiling point of methanol can lead to evaporation of this alcohol during the transesterification process. However when a higher molar ratio of alcohol to oil (8:1) was employed under the same condition; the biodiesel yield increases from 80.66 to 82.15 wt%. It was obvious that 40 °C is the optimum temperature for methyl ester synthesis from this oil. Because the less than 2% increase observed do not appear to be cost effective when compared with the cost of rising temperature from 40 °C to 60 °C.

3.2 Statistical Analysis and Model Fitting

Analysis of variance (ANOVA) was carried out on the experimental results obtained in Table 51.3, the effects of the three (3) main variables and interaction were estimated. The result shows that molar ratio, temperature and catalyst concentration have a main effects of −6.46, −3.78 and −0.70 on the biodiesel yield and with a percentage contributions of 36.4, 12.48 and 0.43 respectively. It can be deduced that catalyst concentration has the highest effect while the molar ratio of methanol to oil has the least effect. With respect to percentage contribution molar ratio of methanol to oil have the highest percentage contribution of 36.47 followed

Table 51.3 Analysis of variance and percentage contribution

Source	Sum of squares	df	Mean square	% contribution	Effect	F-value	P-valueProb > F
Model	226.95	5	45.39			51.89	0.0190
A-mole ratio	83.40	1	83.40	36.47	−6.46	95.33	0.0103
B-temperature	28.54	1	28.54	12.48	−3.78	32.62	0.0293
C- catalyst concentration	0.99	1	0.99	0.4367	−0.70		
AB	23.50	1	23.50	10.27	3.43	26.86	0.0353
AC	79.44	1	79.44	34.74	6.30	90.81	0.0108
BC	12.08	1	12.08	5.28	−2.46	13.81	0.0654
ABC	0.76	1	0.76	0.33	0.62		
Residual	1.75	2	0.87				
Cor total	*228.70*	7					

by temperature. Among the interaction AC has the highest effect (6.30) and the highest percentage contribution (34.74) while ABC have the least effect (0.62) and percentage contribution of 0.33. This followed by AB and then BC while ABC has negligible or no effect and percentage contribution. A model F-value of 45.39 implies that the model is significant. The Values of "Prob > F" less than 0.0500 (95% confidence level) indicates that the model terms are significant, In this case A, B, AB, AC and BC are significant model terms. However, when the value is greater than 0.1000, it indicates that the model terms are not significant which means that C and ABC are not significant. This leads to their elimination from the model developed. Thus the linear regression model to describe the relationship between the biodiesel yield and process variables investigated *in terms of coded factors* is given as:

$$\begin{aligned} \textit{Methyl Ester Yield } (Y) &= 87.43 - 3.23 \times A - 1.89 \times B + 1.71 \times AB \\ &\quad + 3.15 \times AC - 1.23 \times BC \end{aligned}$$

From the equation, AB, AC have positive coefficients and this means that a simultaneous increase in these variables has a direct proportionality to the biodiesel yield while A, B and AB all have negative coefficient and exhibit an inverse proportionality behavior with respect to biodiesel yield. The reasonably close agreement between the predicted R^2 and R^2 value of 0.8776 and 0.9923 indicates the correctness of model. The model equation developed to establish the relationship between variables investigated and the biodiesel yield was simulated and the percentage biodiesel yield was predicted within the range of the experimental yield. A plot of the actual yield against the predicted yield shows a straight line which pass through most of the point. This depicts of an agreement between the experimental values and the model equation values. Hence the equation can be use to obtain the optimum yield for the transesterification. Statistically the rule of thumb specifies

that the coefficient of determination (R^2) must be greater than 0.8 to typically indicate that the regression model is a good fit. Otherwise, a second order model will be required because the linear regression will not be fit enough to describe the experimental result. The R^2 of the first order regression model was 0.9923 which is far greater than 0.8 and this means the first order model obtained above was adequate enough to completely describe the system. The R^2 of 0.9923 indicates that the model was able to predict 99.23% of the variance and only 0.0077% of the total variance was not explained by the model.

3.3 Effect of Interaction Between Process Variables

In this research simultaneous increase in the reaction temperature and mole ratio of methanol to oil will lead to a reduction in methyl ester yield. A molar ratio of 6:1 and temperature of 40 °C appear to be the optimum for the synthesis of cottonseed methyl ester further increase beyond this point have negative consequence on the biodiesel yield (Fig. 51.1).

This is because higher amount of the methanol interferes with the separation of biodiesel from the glycerin leaving part of the glycerin remained in the biodiesel phase which in turn drives the reaction in the reverse reaction. The results are quantitatively similar to those of the literature Lalita et al. [26].

The Figure below (Fig. 51.2) shows the methyl ester yield as function of molar ratio and catalyst concentration. From the Figure it was obvious that higher methyl ester yields was obtained at lower molar ratio (6:1) and at lower catalyst concentration (0.5 wt%). When molar ratio values were higher (8:1), yields decreased with catalyst concentration. This can be explained on the basis of the reactant (oil) concentration in the reaction mixture. By increasing alcohol to oil molar ratio,

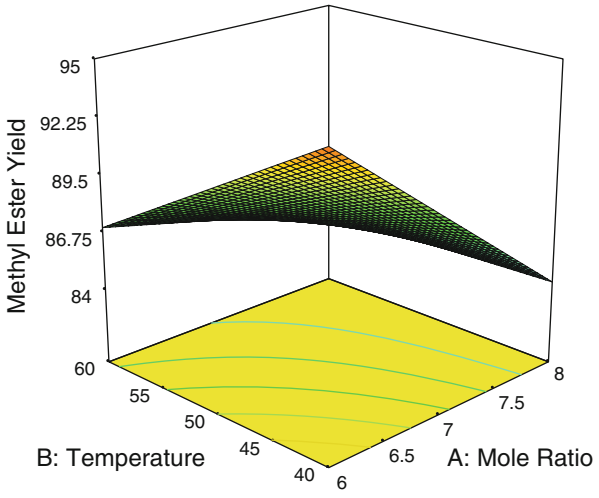


Fig. 51.1 Interaction between temperature and molar ratio

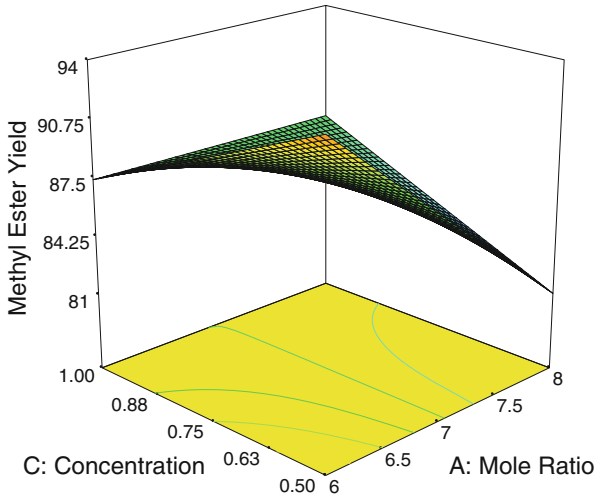


Fig. 51.2 Interactive effect between catalyst concentration and temperature

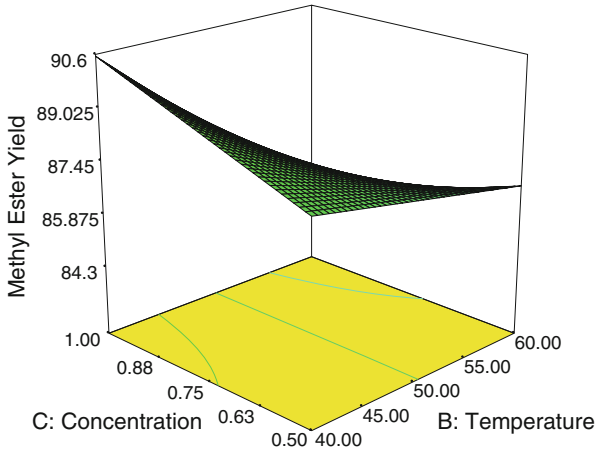


Fig. 51.3 Interactive effect between catalyst concentration and temperature

the amount of alcohol was increased, and catalyst concentrations were diminished, which lead to a reduction on the biodiesel yield. For higher molar ratio of 8:1, the glycerol separation was difficult and the apparent yield of esters decreased, because of a part of the glycerol remained in the biodiesel phase (Fig. 51.3).

When cottonseed oil and methanol react in the presence of a catalyst, two possible products are expected to be obtained. It is either biodiesel produced from transesterification or soap from saponification. The formation of these products depends on the purity of oil, quantity of catalyst and temperature employed. In

Table 51.4 Characterization of biodiesel produced

Property	ASTM D 6751	EN 14214	This work	Anya et al. [30]	Bello and Makanju [31]
Specific gravity	0.87–0.9	–	0.88	0.8732	0.886
Flash point (°C)	130 min	101 min	165	175	160
Total sulphur (% wt)	0.05 max	–	0.006	–	0.006
Pour point (°C)	–	–		10	−26
Kinematic viscosity (mm^2/s)	1.9–6.0	3.5–5	4.5		10.43
Biodiesel index	–	–		–	–
Cetane number	47 min	51 min	52.91	–	53
Free gylcerine (%mass)	0.02	–	0.02	–	–
Total gylcerine (%mass)	0.240	0.25 max	0.02	–	–
Cloud point (°C)	–	–	5.5	13	7
Water by distillation (% vol)	0.05 max	–	Trace	–	–
Bottom sediment & water (% vol)	0.05 max	–	Trace	–	–

this study increase in these variables concurrently leads to a decrease in methyl ester yield. This is because at higher catalyst concentration and temperature the catalyst becomes a reagent for soap formation and this temperature also favours saponification (Table 51.4).

3.3.1 Specific Gravity/Density

The density of a fuel helps in the estimation of the Cetane index of the fuel [32]. The specific gravity for the biodiesel synthesized was determined to be 0.88. This value agrees quite well with ASTM standard and the findings of Bajpai and Tyagi [33] who reported that the density of petroleum diesel (0.85) is lower than that of biodiesel; and reported 0.89 for soya bean oil methyl ester. The slight difference in densities of biodiesel is attributed to fatty acid composition of ester and purity of feedstock.

3.3.2 Flash Point

The flammability or non-flammability behavior of the diesel fuel is generally characterized by the flash point [25]. It is the lowest temperature at which the fuel will be flammable. The ASTM D93 standard specifies 130 °C as minimum temperature. The result of this study reveals a flash point of 165 °C. The result obtained is also in agreement with 164 °C reported by Bello and Makanju, [31]. This shows that Biodiesel from cottonseed oil is safe for handling and storage, non flammable and safe for engine. According to Bajpai and Tyagi [33] that materials with flash point of 93 °C and above are classified as non-hazardous. The Biodiesel

produced is much safer than petroleum diesel. The high flash point obtained shows that the biodiesel produced is essentially free from methanol as even small amount of methanol can reduce the flash point greatly, affects fuel pumps, seals and elastomers [25].

3.3.3 Sulphur Content

The sulphur content of the cottonseed biodiesel was 0.006. This value is very consistent with 0.05% maximum stipulated by ASTM standard and other reported work. This shows that the biodiesel produced from cottonseed oil is environmentally friendly. Higher levels of sulphur are not desirable in fuels as it contribute to the emission of green house gases and acids rain. This can also necessitate the use of catalytic diesel particulate emissions as well as the use of advance technologies such as sulphur oxide absorber to reduce tail pipe emission from petroleum diesel. The biodiesel produced from this oil is a clean fuel and a more preferable than petroleum diesel. Biodiesel as a fuel aside been renewable have been widely acknowledged as sulfur-free fuel and this reason accounts for environmentally benign characteristics over petroleum diesel the results obtained in this work is a testimony to this fact.

3.3.4 Viscosity

The measure of a fluid's resistance to flow is referred to as viscosity. The result obtained shows a kinematic viscosity of 4.5. From the results obtained, biodiesel has higher viscosity than conventional diesel fuel. The result is higher than 3.8–4.1 reported by Bajpai and Tyagi [33] for soy methyl ester and lower than 15.98 mm/s^2 reported by Bello and Makanju [31]. But fall within the ASTM specification of 1.9–6.0 mm/s^2. Fuels with low viscosity may not provide sufficient lubrication for the precision fit of fuel injection pumps, resulting in leakage or increased wear. Diesel fuels with high viscosity tend to form larger droplets on injection, which can cause poor combustion, increased exhaust smoke and emissions On the other hand high viscosity may lead to poor atomization of the fuel, incomplete combustion, choking of the injectors, ring carbonization and accumulation of the fuel in the lubricating oils [34]. The kinematic viscosity of the produced biodiesel was much lower than that of the parent oil.

3.3.5 Cetane Number

Cetane number is a measure of ignition quality of diesel fuel. The higher the cetane number, the easier the fuel ignites when it is injected into the engine [32]. The result from this work indicates a cetane number of 52.91. This value is consistent with most international standard (ASTM and EN) which set a minimum limit of 47.

The result also agrees well with other reported work. This is an indication that the biodiesel has good ignition quality and would perform optimally when used in diesel engine. The value is greater than the cetane number of petrodiesel whose minimum is 40.

3.3.6 Cloud Point

The cloud point is used to specify cold temperature usability of a fuel [24]. The cloud points for this work was 5.5 °C. The values obtained in this study differ from 13 °C and 7 °C reported by Anya et al. [30] and Bello and Makanju [31]. The low cloud points from this study is an indication that the fuel will performs satisfactory even in cold climate conditions since the tendency for gel formation are low. Higher cloud point can affect the engine performance and emission adversely under cold climate conditions. Gerpen et al. [25] stated that operating a diesel fuel at temperatures below the cloud point of the fuel can result in fuel filter clogging due to the wax crystals.

3.3.7 Free and Total Glycerol

The Free glycerin and total glycerine in the biodiesel were determined to be 0.02 % by mass respectively which is in accordance with the European standard (EN 14105) and ASTM D6584 standard which both stipulate a minimum of 0.02%. These values provide an indication of amount of unconverted or partially converted fats and by product of glycerin. The results of this study demonstrate that washing process and glycerin removal was satisfactory carried out because higher free glycerol usually results from incomplete separation of ester and glycerin product after transesterification and as a result of imperfect waste water washing. Gerpen et al. [25] reported that incomplete removal of free glycerin can result in contamination of storage tank, fuel system and the diesel engine.

3.3.8 Water Content

The water content of biodiesel was less than the minimum standard of 0.05 Max. In this study only minute quantity of water was detected in traces. This result attest to the cleanliness of the fuel which is apparently due to the effectiveness in the drying method employed after washing. Fuel contaminated with water can cause engine corrosion or react with the glyceride to produce soap [25]. The product from this study will not support microbial growth under storage because it is essentially free from water.

4 Conclusion and Future Work

From the experimental result for the transesterification of refined cottonseed oil, it was concluded that the optimal yield is attained at the following condition: oil to methanol molar ratio of 1:6; reaction temperature of 40 °C and catalyst concentration of 0.5 wt% at constant reaction time and agitation of 90 min and 300 rpm respectively. The optimum yield of 96.23% was obtained at these conditions. The statistically analysis yielded an R^2 of 0.9923 which depicts that 99.23% variability can be explained by linear regression model developed. Characterization of biodiesel shows that it compares favorably well with relevant standards and can be used as alternative to petroleum diesel in diesel engine.

Future work is expected to explore the effect of more variables using other methods of design of experiment. The performance, emission and combustion characteristics of the biodiesel and it blends should be tested in diesel engine.

Acknowledgment The authors wish to acknowledge Mr. Bulus Baba of Water Aquaculture and Fishery Technology Department, Federal University of Technology Minna, for his technical support during the course of laboratory experiments.

References

1. Bugaje, I.M., Mohammed, I.A.: Biofuel Production Technology, 1st edn, pp. 1–65. Science and Technology Forum (STF), Zaria (2008)
2. Tashtoush, G.M., Al-widyan, M.H., Al-shyoukh, A.O.: Experimental study on evaluation and optimization of waste animal fat into biodiesel. Energy Convers. Manag. **45**, 2679–711 (2004)
3. Demirbas, A.: Global biofuel strategies. Energy Edu. Sci. Technol. **17**, 27–63 (2006)
4. Umaru, M., Aboje, A.A., Mohammed Ibrahim, A., Aliyu, A.M., Munir, S.M., Aminat Olaibi, O.: The effect of process variables on the transesterification of refined cottonseed oil. Lecture Notes in Systems Biology and Bioengineering: Proceedings of the World Congress on Engineering 2014, vol. I, pp. 622–624. WCE, London, 2–4 July 2014
5. Sadiq Muhammad, M., Umaru, M., Mohammed Ibrahim. A., Aliyu Musa, A., Oyiza M.R.: Influence of process parameters on solid state fermentation of pineapple pomace for bioethanol production. 7th International Conference on Sustainable Energy and Environmental Protection, Dubai, 23–25 Nov 2014
6. Demirbas, A.: Biodiesel fuels from vegetable oils via catalytic and non-catalytic supercritical alcohol transesterifications and other methods: a survey. Energy Convers. Manag. **44**, 2093–2109 (2003)
7. Bozbas, K.: Biodiesel as an alternative motor fuel. 2006 production and polices in the European Union. Renewable and Sustainable Energy Reviews. dio:10.1016/j.reser.2005.06 (2005)
8. Refaat, A.A., Attia, N.K., Sibak, H.A., El-Sheltawy, S.T., Eldiwani, G.I.: Production optimization and quality of biodiesel from waste vegetable oil. Int. J. Environ. Sci. Technol. **5**(1), 75–82 (2008)
9. Zhang, Y., Dube, M.A., Mclean, D.D., Kates, M.: Biodiesel production from waste cooking oil. Bioresour. Technol. **90**, 299–2240 (2003)
10. Lim, S., Hoong, S.S., Teon, L.K., Bhatia, S.: Supercritical fluid extraction of *Jatropha curcas* oil L. seed with methanol: a novel biodiesel production method. Bioresour. Technol. **101**, 7169–7172 (2010)

11. Kansedo, J., Lee, K.T., Bhatia, S.: Cerbraodollam (sea mango) oil as a promising non edible feedstock for biodiesel production. Energy Fuel **88**, 1148–1150 (2009)
12. Deng, X., Fang, Z., Liu, Y.-H., Yu, C.-L.: Production of biodiesel from jatropha oil catalyzed by nano solid basic catalyst. Energy, homepage: www.elsevier.com/locate/energy (2011)
13. Umaru, M., Aberuagba, F.: Characteristics of a typical Nigerian Jatropha curcas oil seeds for biodiesel production. Res. J. Chem. Sci. **2**(10), 7–12 (2012)
14. Ma, F., Hanna, M.A.: Biodiesel production: a review. Bioresour. Technol. J. **70**, 1–15 (1999)
15. Orhevba, B.A., Efomah, A.N.: Extraction and characterization of cottonseed (*Gossypium*) Oil. Int. J. Basic Appl. Sci. **1**(2), 398–402 (2012)
16. Mohammed Ibrahim, A., Umaru, M., Sadiq Muhammad, M., Aliyu Musa, A., Bawa, A.A., Isah, A.G., Ronke.:, Effect of catalyst type and concentration on biodiesel yield and quality. 3rd Biennial Engineering Conference, School of Engineering and Engineering Technology, pp. 233–238. Federal University of Technology, Gidan Kwano- Minna, 14–16 May 2013
17. Abdulwakil, M.D., Ahmed, Z.U., Rahman, H., Arifuzzaman, M.D.: Study on fuel properties of various vegetable oil available in Bangladesh for biodiesel production. J. Mech. Eng. **2**(5), 10–17 (2012)
18. Royon, D., Daz, M., Ellenrieder, G., Locatelli, S.: Enzymatic production of biodiesel from cotton seed oil using t-butanol as a solvent. Bioresour. Technol. **98**, 648–653 (2007)
19. Demirbas, A.: Studies on cottonseed oil biodiesel prepared in non-catalytic SCF conditions. Bioresour. Technol. **99**(5), 1125–1130 (2008)
20. Qian, J., Wang, F., Liu, S., Yun, Z.: In situ alkaline transesterification of cottonseed oil for production of biodiesel and nontoxic cottonseed meal. Bio Resour. Technol. **99**(18), 9009–9012 (2008)
21. Rashid, U., Anwar, F., Knothe, G.: Evaluation of biodiesel obtained from cottonseed oil. Fuel Process. Technol. **90**(9), 1157–1163 (2009)
22. Xiaohu Fan, Xi Wang, Feng Chen: Biodiesel Production from Crude cottonseed oil: an optimization process using response surface methodology. Open Fuel. Energy Sci. J. **4**, 1–8 (2011)
23. Özçimen, D., Yücel, S.: Novel methods in biodiesel production. In: Marco Aurelio Dos Santos Bernardes (ed.) Biofuel's Engineering Process Technology. ISBN: 978-953-307-480-1, InTech, Available from: http://www.intechopen.com/books/biofuel-s-engineering-process-technology/novel-methods-in-biodiesel-production (2011)
24. Živorad, R.L.: Design of Experiments in Chemical Engineering, pp. 154–158. WILY. VCH. VerlagGmbH & Co. KGaA, Weinheim (2004)
25. Gerpen, J.V., Shanks, B., Pruszko, R., Clements, D., Knothe, G.: Biodiesel Analytical Methods. NREL/SR-510-36240, National Renewable Energy Laboratory (2004)
26. Lalita, A., Sukunya, M., Peesamai, J.: Factors affecting the synthesis of biodiesel from crude palm kernel oil. The Joint International Conference on "Sustainable Energy and Environment (SEE)", pp. 359–361. Hua Hin (2004)
27. Freedman, B., Pryde, E.H., Mounts, T.L.: Variables affecting the yields of fatty esters from transesterified vegetable oils. J. Am. Oil Chem. Soc. **61**, 1638–43 (1984)
28. Vicente, C., Coteron, A., Martinez, M., Aracil, J.: Application of the factorial design of experiments and response surface methodology to optimize biodiesel production. Ind. Crop Prod. **8**, 29–35 (1998)
29. Patil, P.D., Deng, S.: Optimization of biodiesel production from edible and non edible vegetable oils. Fuel **88**(7), 1302–1306 (2009)
30. Anya, U.A., Nwobia, C.N., Ofoegbu, O.: Optimized reduction of free fatty acid content on neem seed oil for biodiesel production. J. Basic Appl. Chem. **2**(4), 21–28 (2012)
31. Bello, E.I., Makanju, A.: Production, characterization and evaluation of castor oil biodiesel as alternative fuel for diesel engines. J. Emerg. Trends Eng. Appl. Sci. **2**(3), 525–530 (2011). Scholarlink Research Institute Journals, (ISSN: 2141–7016)
32. Demirbas, A: Characterization of biodiesel fuels. Taylor & Francis Group, LLC, Energy Sources, Part A **31,** 889–896 (2009)

33. Bajpai, D., Tyagi, V.K.: Biodiesel: sources, production, composition, properties and its benefits. J. Oleo Sci. **55**(10), 487–506 (2006). ISSN 1345 – 8957 print/ISSN 1347 – 3352 online
34. Mohammed, I.A., Musa, U., Onifade, K.R., Aberuagba, F.: Fuel properties of Nigerian castor seed oil biodiesel. 3rd Biennial Engineering Conference, School of Engineering and Engineering Technology, pp. 239–246. Federal University of Technology, Gidan Kwano-Minna, 14–16 May 2013

ERRATUM

Chapter 50
Using Information Gain to Compare the Effeciency of Machine Learning Techniques When Classifying Influenza Based on Viral Hosts

Nermin Shaltout, Ahmed Rafea, Ahmed Moustafa, and Mahmoud El-Hefnawi

G.-C. Yang et al. (eds.), Transactions on Engineering Technologies,
DOI 10.1007/978-94-017-9804-4

DOI 10.1007/978-94-017-9804-4_52

The surname of the fourth chapter co-author was spelled incorrectly in the original chapter. The correct spelling is Mahmoud ElHefnawi.

The corrections described in this erratum were also done in the original publication.

The online version of the original chapter can be found at
http://dx.doi.org/10.1007/978-94-017-9804-4_50

Author Index

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4

B

H

M

S

Subject Index

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-9804-4