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

Transactions on Engineering Technologies

Special Volume of the World Congress on Engineering 2013

Springer

Transactions on Engineering Technologies

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

Transactions on Engineering Technologies

Special Volume of the World Congress on Engineering 2013

Springer

Editors
Gi-Chul Yang
Department of Multimedia Engineering,
College of Engineering
Mokpo National University
Mokpo, Jeonnam
Republic of Korea

Len Gelman
Department of Applied Mathematics
and Computing
Cranfield University
Cranfield, Bedfordshire
UK

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

ISBN 978-94-017-8831-1 ISBN 978-94-017-8832-8 (eBook)
DOI 10.1007/978-94-017-8832-8
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013953195

Printed on acid-free paper

Springer 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., 3–5 July, 2013, under the World Congress on Engineering 2013 (WCE 2013). The WCE 2013 is organized by the International Association of Engineers (IAENG); the Congress details are available at: http://www.iaeng.org/WCE2013. IAENG is a nonprofit 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 an international meeting with a high level of participation from many countries. The response to the Congress has been excellent. There have been more than 1,100 manuscript submissions for the WCE 2013. All submitted papers have gone through the peer review process, and the overall acceptance rate is 55.12 %.

This volume contains 48 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include mechanical engineering, bioengineering, Internet engineering, image engineering, wireless networks, knowledge 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.

Gi-Chul Yang
Sio-Iong Ao
Len Gelman

Contents

Viscous Fingering of Reversible Reactive Flows in Porous Media

Hesham Alhumade and Jalel Azaiez

Abstract The dynamics of viscous fingering instability of miscible displacements in a homogeneous porous medium are examined in the case of flows that involve reversible chemical reactions between the displacing and displaced fluid. The flows are modeled using the continuity equation, Darcy's law, and volume-averaged forms of the convection-diffusion-reaction equation for mass balance of a bi-molecular reaction. Numerical simulations were carried out using a Hartley transform based pseudo-spectral method combined with semi-implicit finite-difference time-stepping algorithm. The results of the simulations allowed to analyze the mechanisms of fingering instability that result from the dependence of the fluids viscosities on the concentrations of the different species, and focused on different flow scenarios. In particular, the study examined the effects of varying important parameters namely the Damkohler number that represents the ratio of the hydrodynamic and chemical characteristic time scales, and the chemical reversibility coefficient, and analyzed the resulting changes in the finger structures. The results are presented for flows with an initially stable as well as initially unstable front between the two reactants.

Keywords Fluid mechanics · Homogeneous porous media · Hydrodynamics · Miscible displacements · Reversible chemical reaction · Stability · Viscous fingering

H. Alhumade
University of Waterloo, 200 University Avenue West Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: halhumade@uwaterloo.ca

J. Azaiez (✉)
University of Calgary, 2500 University Dr. NW Calgary, Calgary, Alberta T2N 1N4, Canada
e-mail: azaiez@ucalgary.ca

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_1,

1 Introduction

When a viscous fluid is used to displace another one of a larger viscosity, a frontal instability appears at the interface between the two fluids, which may dramatically affect the overall efficiency of the displacement process. This instability may grow to form fingers that propagate in both upstream and downstream directions and is referred to as fingering or Saffman–Taylor instability [21]. The instability can be triggered by either viscosity mismatch and is referred to as viscous fingering or density mismatch, where it is known as the Rayleigh-Taylor instability. Such instabilities are encountered in a wide variety of processes that include enhanced oil recovery, soil remediation, chromatography and CO_2 sequestration. Many experimental and theoretical studies have focused on the frontal instability of non-reactive displacement processes, where hydrodynamic interactions between the fluids result in the viscous fingering instability. In these studies the effects of different parameters were examined and most of these studies were reviewed in [11, 13].

The viscous fingering instability may develop in conjunction with chemical reactions in a wide variety of processes such as underground water treatment, tertiary heavy oil recovery, spreading of chemical pollutants chromatographic separation, polymer synthesis and processing as well as fixed bed regeneration [14]. There has been a growing interest in analyzing such reactive flow instability and a number of studies have examined the reactive flow. One of the earliest studies of reactive displacements in porous media was conducted out by [19], where the reaction leads to an interfacial tension decrease in a secondary oil recovery process. A number of subsequent experimental studies examined the effects of different parameters such as stoichiometry [16], geometry orientation [12], finger growth rate [17], chemical composition [3], external electrical field [25], variation in the physical properties of the phases [20], and precipitation [18].

Analytical and numerical Modelling of reactive flow displacements has been carried out by a limited, but growing number of studies [6–10]. These studies have considered either auto-catalytic or non-autocatalytic reactions.

All existing studies dealing with reactive viscous fingering have assumed the chemical reaction to be complete. However the reversibility of the reaction plays an important role in many phenomena studied in physics, chemistry, biology, and geology [9]. For example, in the in situ soil remediation, promising results were reported by [26], where a reactive fluid was injected to remove the pollutant from the underground water. The first study on reactive-diffusive systems with reversible reaction was carried out by [5], where the properties of a reversible reactive front with initially separated reactants were examined. It was reported that the dynamics of the reactive front can be described as a crossover between irreversible and reversible regimes at short and long times, respectively. A subsequent study [23] confirmed the existence of a crossover between short time “irreversible” and long-time “reversible” regimes. In a recent study [22], the reaction rate of a reversible reactive-diffusive process when the reactants are initially mixed with

different diffusion coefficients by using the boundary layer function method was investigated. The authors reported that the reactive-diffusive process for this case can be considered as a quasi-equilibrium process and analyzed the dependence of the reaction rate on the initial distribution of the reactants.

It should be noted that even though the above studies did examine the role of chemical reversibility, there conclusions are actually very restrictive since they accounted only for diffusive effects. It is however known that in actual flow displacements that involve the injection of chemical species; convective effects must be included in the model to analyze properly the flow. Actually, for such flows convective effects can be dominant at least at some stages of the flow and hence cannot be ignored. Motivated by this, the first linear stability analysis to understand the effects of chemical reversibility on the stability of some cases of reactive-diffusive-convective flow displacements [1]. In a recent study, the role of chemical reversibility on the stability of some cases of reactive-diffusive-convective flow displacements was investigated [2]. In this study, the nonlinear development of the flow are analyzed through numerical simulations.

2 Mathematical Model

2.1 Physical Problem

A two-dimensional displacement is considered in which both fluids are incompressible and fully miscible. The flow takes place in horizontal direction in a homogeneous medium of constant porosity ϕ and permeability K. A schematic of the two-dimensional porous medium is shown in Fig. 1. The length, width and thickness (z-direction) of the medium are L_x, L_y and b respectively.

The medium is assumed to be initially filled with a solution of a reactant (B) of viscosity μ_B. A miscible fluid (A) of viscosity μ_A is injected from the left-hand side with a uniform velocity U to displace fluid (B). The direction of the flow is along the x-axis and the y-axis is parallel to the initial plane of the interface. A reversible chemical reaction occurs between the two fluids leading to the formation of a product (C) of viscosity μ_C:

$$A + B \rightleftharpoons C \tag{1}$$

As time proceeds, the bi-molecular reaction results in the accumulation of more chemical product at the interface between the two reactants. This leads to the coexistence of the three chemical species (A) Fig. 1 shows an idealized distribution of the two reactants (A) and (B) and the product (C), with two fronts. One between the reactant (A) and the product (C); (A–C) while the other is between the reactant (B) and the product (C); (C–B), and they are referred to as the trailing and the leading front, respectively. It should be stressed that this is an idealization of the system and the three chemical species are actually present to a more or less

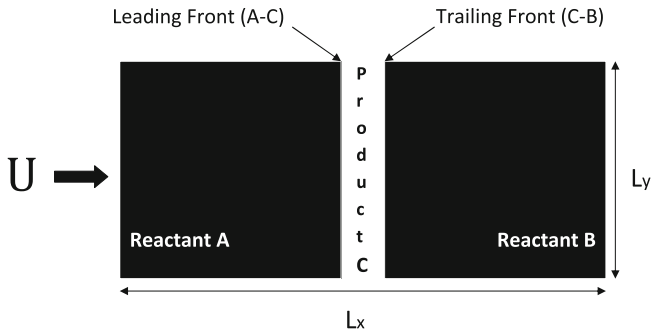


Fig. 1 Schematic of a reactive front displacement process

degree everywhere in the region where the reaction takes place. However, this concept of a leading and trailing front will be helpful in the interpretation and explanation of the results.

2.2 Governing Equations

The flow is governed by the equations for conservation of mass, momentum (Darcy's Equation) and the transport of the three chemical species.

$$\nabla \cdot \mathbf{v} = 0, \tag{2}$$

$$\nabla p = -\frac{\mu}{K}\mathbf{v}, \tag{3}$$

$$\phi\frac{\partial A}{\partial t} + u\frac{\partial A}{x} + v\frac{\partial A}{y} = \phi D_A\left[\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2}\right] - kAB + k_r C, \tag{4}$$

$$\phi\frac{\partial B}{\partial t} + u\frac{\partial B}{x} + v\frac{\partial B}{y} = \phi D_B\left[\frac{\partial^2 B}{\partial x^2} + \frac{\partial^2 B}{\partial y^2}\right] - kAB + k_r C, \tag{5}$$

$$\phi\frac{\partial C}{\partial t} + u\frac{\partial C}{x} + v\frac{\partial C}{y} = \phi D_C\left[\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2}\right] + kAB - k_r C. \tag{6}$$

In the above equation, $\mathbf{v} = ui + vj$ is the velocity vector with u and v the $x-$ and $y-$ components respectively, p the pressure, μ the viscosity, K the medium permeability and ϕ its porosity. The concentrations of the two reactants and the product are denoted by A, B and C, respectively while D_A, D_B and D_C are their corresponding diffusion coefficients. Furthermore, k is the reaction constant while k_r represents the reverse reaction constant. For simplicity, it will be assumed that all species have the same diffusion coefficient, i.e. $D_A = D_B = D_C = D$.

Since the characteristic velocity for the fluid flow through the porous medium is U/ϕ, we adopted a Lagrangian reference frame moving at a velocity U/ϕ. Furthermore, diffusive time $D\phi^2/U^2$ and diffusive length $D\phi/U$ are chosen to make the length and time dimensionless. The constant permeability K is incorporated in the expression of the viscosity by treating μ/K as μ, and we shall refer to ratios of μ as either viscosity or mobility ratios. The rest of the scaling is as follows: the velocity is scaled with U/ϕ, the viscosity and pressure with μ_A and $\mu_A D/\phi$, respectively, and the concentration with that of the pure displacing fluid, A_0. The dimensionless equations are:

$$\nabla \cdot \mathbf{v} = 0, \tag{7}$$

$$\nabla p = -\mu(\mathbf{v} + \mathbf{i}), \tag{8}$$

$$\frac{\partial A}{\partial t} + u\frac{\partial A}{x} + v\frac{\partial A}{y} = \left[\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2}\right] - D_a AB + D_r C, \tag{9}$$

$$\frac{\partial B}{\partial t} + u\frac{\partial B}{x} + v\frac{\partial B}{y} = \left[\frac{\partial^2 B}{\partial x^2} + \frac{\partial^2 B}{\partial y^2}\right] - D_a AB + D_r C, \tag{10}$$

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{x} + v\frac{\partial C}{y} = \left[\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2}\right] + D_a AB - D_r C. \tag{11}$$

In the above equations, dimensionless variables are represented with asterisk while $D_a = kA_0 D/U^2$ is the Damkohler number representing the ratio of hydrodynamic to chemical characteristic times and $D_r = k_r D/U^2$ represents a reversible Damkohler number. Two additional dimensionless groups are also involved, namely the Péclet number $P_e = UL_x/D$ and the cell aspect ratio $A_r = L_x/L_y$ that appear in the boundary conditions equations.

Following previous studies, an exponential concentration dependent viscosity model is adopted to complete the model [8, 10, 24],

$$\mu = \exp(R_b B + R_c C) \tag{12}$$

where R_b and R_c are the log-mobility ratios between the species as follows:

$$R_b = \ln\left(\frac{\mu_B}{\mu_A}\right) \quad and \quad R_c = \ln\left(\frac{\mu_C}{\mu_A}\right) \tag{13}$$

An associated mobility ratio at the chemical front between the chemical product (C) and the reactant (B), and between the reactant (A) and the product (C) can be also defined as:

$$R_{AC} = \ln\left(\frac{\mu_C}{\mu_A}\right) = \frac{R_c}{2} \quad and \quad R_{CB} = \ln\left(\frac{\mu_B}{\mu_C}\right) = R_b - \frac{R_c}{2} \tag{14}$$

It should be stressed here that the different fronts, whether it is the initial reactive front between (A) and (B) or the idealized leading and trailing ones will

be unstable whenever their mobility ratios are strictly positive, while they will be stable if the mobility ratios are negative or zero. For convenience, in all that follows, the asterisks will be dropped from all dimensionless variables.

2.3 Numerical Techniques

The above problem is formulated using a stream-function vorticity formulation, where the velocity field, the streamfunction ψ and the vorticity ω are related as:

$$u = \frac{\partial \psi}{\partial y}, \quad v = -\frac{\partial \psi}{\partial x}, \quad \nabla^2 \psi = -\omega. \tag{15}$$

where ∇^2 is the Laplacian operator.

The pressure term is eliminated by taking the curl of Eq. (8) resulting in the following relationship between the vorticity and the concentrations of the three chemical species:

$$\omega = R_b \left(\frac{\partial \psi}{\partial x}\frac{\partial B}{\partial x} + \frac{\partial \psi}{\partial y}\frac{\partial B}{\partial y} + \frac{\partial B}{\partial y} \right) + R_c \left(\frac{\partial \psi}{\partial x}\frac{\partial C}{\partial x} + \frac{\partial \psi}{\partial y}\frac{\partial C}{\partial y} + \frac{\partial C}{\partial y} \right) \tag{16}$$

Equations (9)–(11) and (16) form a closed set that can be solved for the concentration and velocity fields. The system of partial differential equation is solved by decomposing the variables as a base-state and a perturbation. The perturbation terms consist of a random noise centered at the initial interface between the reactants A and B, with the magnitude of the noise decaying rapidly away from the interface. The resulting system of three partial differential equations is solved using a highly accurate pseudo-spectral method based on the Hartley transform [4, 10]. This method allows to recast the partial differential equation in time and space into an ordinary differential equation in time. The solution for the time stepping of the reactive-diffusive-convective equations was generated by using a semi-implicit predictor-corrector method along with an operator-splitting algorithm.

3 Result

3.1 Numerical Code Validation

The numerical code has been validated by comparing the time evolution and the related viscous fingers interactions to those presented by Hejazi and Azaiez [10] for the non-reversible case ($\alpha = 0$). It has been noted that the dynamics of fingering were identical when the same parameters were used along with the same spatial resolution and time step size. In addition, the numerical convergence of the

numerical results has also been tested by varying the spatial resolution and the time step. In this study, unless mentioned otherwise, a spatial resolution of 256×256 is used along with a time step dt $= 0.005$.

3.2 Concentration ISO-Surfaces Contours

The dynamics of the flow and the development of the instability are expected to depend on the flow parameters, namely the Péclet number Pe, cell aspect ratio A_r, Damkohler number D_a, reversibility ratio coefficient α as well as the species' mobility ratios; R_b and R_c. Prior to discussing the effects of chemical reversibility, a brief explanation of the effects of chemical reaction on the viscous fingering instability is presented. More details can be found in [10].

3.2.1 Effects of the Chemical Reaction

Figure 2 shows a time sequence of contours of the displacing fluid in the case of non-reactive flow displacement for a mobility ratio of $R_b = 3$. In this case the displaced fluid is more viscous than the displacing one resulting in an unstable front. The finger structures develop and become more complex with time as a result of different mechanisms of interactions that have already been analyzed in the literature.

When chemical reactions take place, the instability is modified and will depend on both mobility ratios R_b and R_c. Figure 3 shows concentration iso-surfaces of the reactant (A) in the case $R_b = 3$, $R_c = 5$, $D_a = 0.5$ and $Pe = 1{,}000$. It is clear that the frontal instability is affected by the reaction resulting in a larger number of fingers that tend to be thinner and to have more complex structures. Figure 4 depicting the corresponding contours for the chemical product (C) shows that the contours of (C) actually allow to illustrate simultaneously the finger structures of the displacing fluid (A) (compare the trailing front in Fig. 4 with those in Fig. 3) and those of the displaced one (B) through the leading front. This indicates that plots of the chemical product contours allow us to show the development of the instability at all fronts, and hence it will be used in all subsequent figures. Note that both the leading and trailing fronts are unstable. Furthermore, as time proceeds and more product is generated, the fingers develop more and extend both upstream and downstream.

In what follows, the effects of chemical reversibility are analyzed. Given the large number of parameters and the fact that the effects of the Péclet number and Damkohler numbers have already been analyzed in the case of irreversible reactions, a number of parameters will be fixed in order to focus the analysis on the role of chemical reversibility. In all that follows, the aspect ratio, Péclet number and Damkohler number are fixed as $A_r = 2$, $P_e = 1{,}000$ and $D_a = 1$. Furthermore, since the dynamics of the flow depend on the mobility ratios, the results will be discussed first for systems that involve an initially stable front between the two

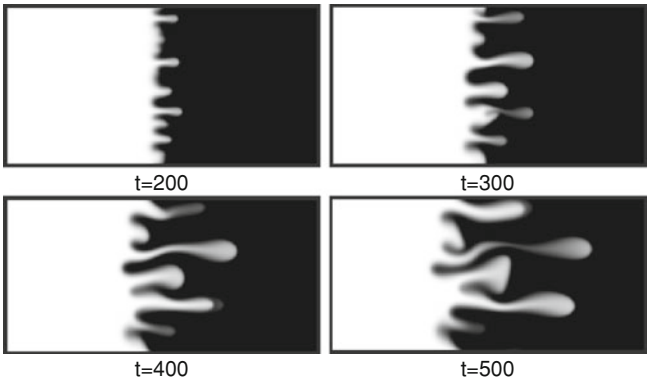


Fig. 2 Contours of (A) for a non-reactive flow: $R_b = 3$

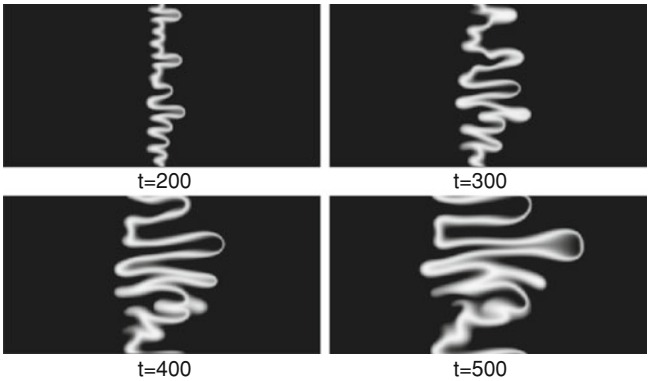


Fig. 3 Contours of (A) for a reactive flow: $R_b = 3$, $R_c = 5$, $D_a = 0.5$

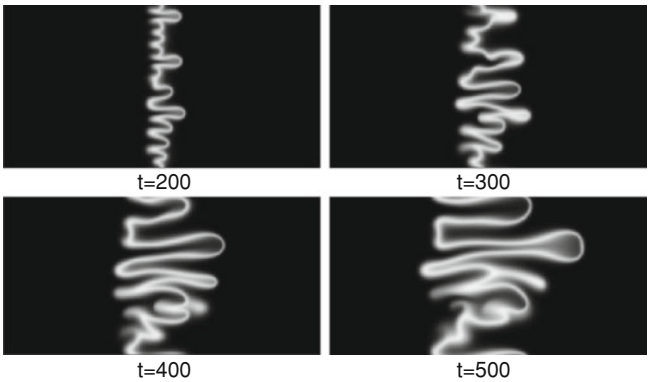


Fig. 4 Contours of the chemical product (C) for a reactive flow: $R_b = 3$, $R_c = 5$, $D_a = 0.5$

reactants ($R_b \leq 0$) followed by that where the initial front is unstable ($R_b > 0$). All qualitative results are presented in the form of concentration iso-surfaces of the chemical product (C).

3.2.2 Stable or Neutrally Stable Initial Interface ($R_b \leq 0$)

The initial interface between the two reactants is stable or neutrally stable if the viscosity of the displacing fluid is larger than or equal to that of the displaced one ($R_b \leq 0$). However, as the reaction takes place and chemical product is generated, instability may develop at either the trailing or the leading front but not at both. The case where the viscosity of (C) lies between those of (B) and (A) or equal to both or any of them ($\mu_A \geq \mu_C \geq \mu_B$) results in a stable displacement process and therefore, will not be discussed. In what follows, the two cases where instability appears at only the trailing or the leading front are examined.

An unstable trailing front and a stable leading front are observed in the case where the viscosity of (C) is grater than that of both reactants. On the other hand, a stable trailing front and an unstable leading will occur when the viscosity of the product (C) is smaller than the viscosities of both reactants. It is worth mentioning that in these cases, the mixing between the two reactants is mainly controlled by diffusion. As a result, the growth of fingers is rather slow compared to cases with unstable initial fronts.

In such cases involving stable initial reactive fronts, reversibility tends to attenuate the instability at the unstable trailing or leading front and may actually result in a completely stable system for a period of time. Figure 5 depicts the case where the instability takes place at the trailing front when the chemical reaction is complete ($\alpha = 0$), while the instability of the system decreases when the reaction reverses ($\alpha \neq 0$). This can be attributed to the fact that in the initial stages of the flow, the mixing of the different species is governed by diffusion and reversibility perpetuates this state by preventing more of the product to accumulate and to trigger instability at the unstable front. However, the amount of chemical product will still keep growing with time, though slowly, and instability will eventually appear at the unstable trailing front at later times (see Fig. 5). The time for the instability to appear depends on the rate of chemical reversibility, with smaller reversibility coefficients leading to an earlier growth of fingers. This time also depends on the mobility ratios of the different chemical species, with distributions that lead to a more unstable trailing front resulting in the instability developing earlier in time. To illustrate this, the results for ($R_b = -1.0$, $R_c = 5$, $R_{AC} = 2.5$) depicted in Fig. 6 show that the instability develops earlier in time in comparison with the case in Fig. 5 where $R_{AC} = 1.5$. Similar conclusions were reached for the case where the instability takes place at the leading front, and for brevity the corresponding contours are not shown.

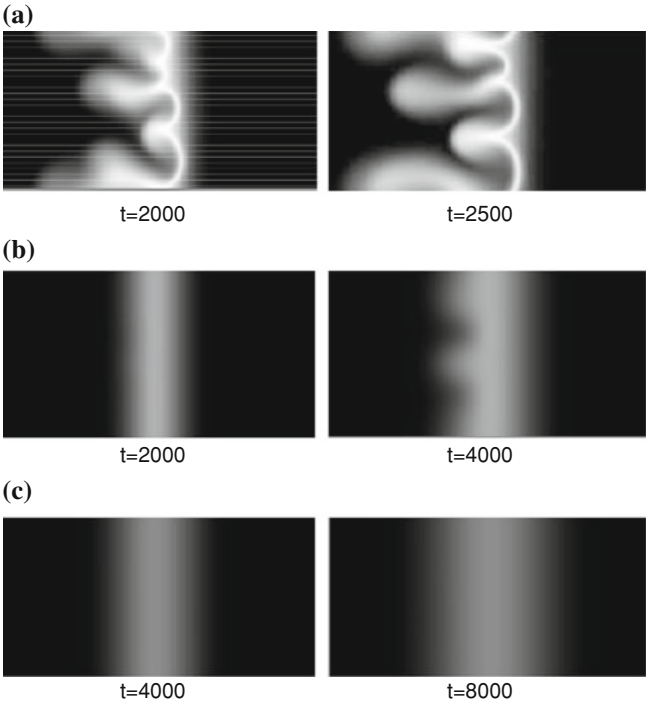


Fig. 5 Concentration iso-surfaces for $R_b = -1$, $R_c = 3$ (unstable trailing front, stable leading front): **a** $\alpha = 0.0$, **b** $\alpha = 0.3$, **c** $\alpha = 0.8$

Fig. 6 Concentration iso-surfaces for $\alpha = 0.3$, $R_b = -1$, $R_c = 5$ (unstable trailing front, stable leading front)

3.2.3 Unstable Initial Interface ($R_b > 0$)

When a less viscous fluid (A) is used to displace another one (B) with a higher viscosity, the initial interface between the two fluids is unstable ($R_b > 0$). The viscosity of the chemical product (C) can be either smaller than, larger than or in between the viscosities of (A) and (B). As a result, regardless of the viscosity of the product (C), instability will take place at least at the trailing or the leading front, if not at both. In what follows, various cases of instability are discussed.

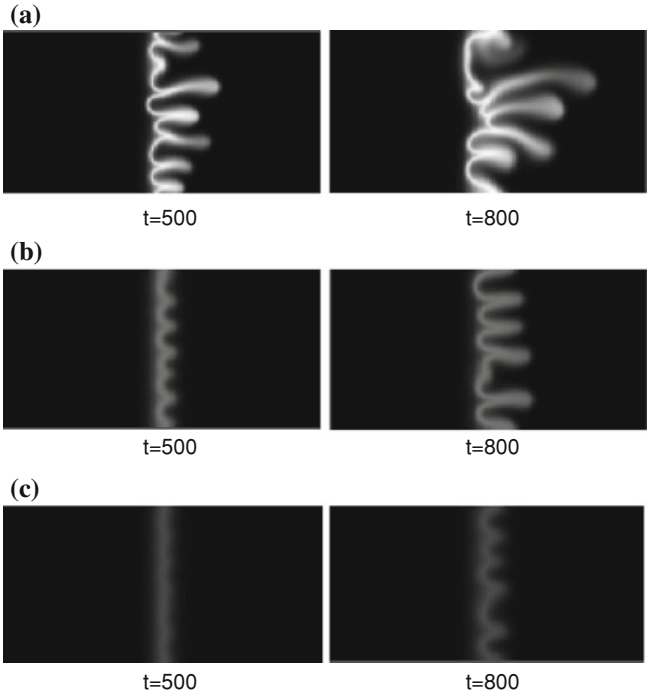


Fig. 7 Concentration iso-surfaces $R_b = 1$, $R_c = -2$ (stable trailing front and unstable leading front): **a** $\alpha = 0.0$, **b** $\alpha = 0.3$, **c** $\alpha = 0.8$

Concentration contours of the chemical product (C) are depicted in Fig. 7 for the case where the product's viscosity is smaller than those of both reactants; i.e. $R_{AC}(=R_c/2) < 0$ and $R_{CB}(=R_b - R_c/2) > 0$. The results are presented for the cases where the chemical reaction is complete ($\alpha = 0$), weakly reversible ($\alpha = 0.3$) and strongly reversible ($\alpha = 0.8$). It is worth noting that in all three cases the instability develops mainly on the leading front and the fingers extend in the downstream direction. Furthermore, it is clear that in this case reversibility tends to attenuate the instability of the flow. In particular, fingers are less developed and more diffuse than in the non-reversible case. However, reversibility also increased the number of developed fingers. Moreover, it should be noted that the distribution of the chemical product is more homogeneous and shows less gradients than in the non-reversible reaction flow. This indicates that in such reversible-reaction flows, the chemical product is more uniformly distributed in the medium. When the viscosity of the product (C) is larger than those of both reactants, the viscosity ratio will be in favour of the growth of instability at the trailing ($R_c > 0$), but not the leading front ($R_b - R_c/2 < 0$). As a result, fingers appear on the trailing front and extend in the opposite direction of the flow, while the leading front is expected to be stable. Figure 8 depicts the case $R_b = 1$ and $R_c = 3$, which corresponds to unstable trailing and stable leading fronts. It is interesting to note that in this case,

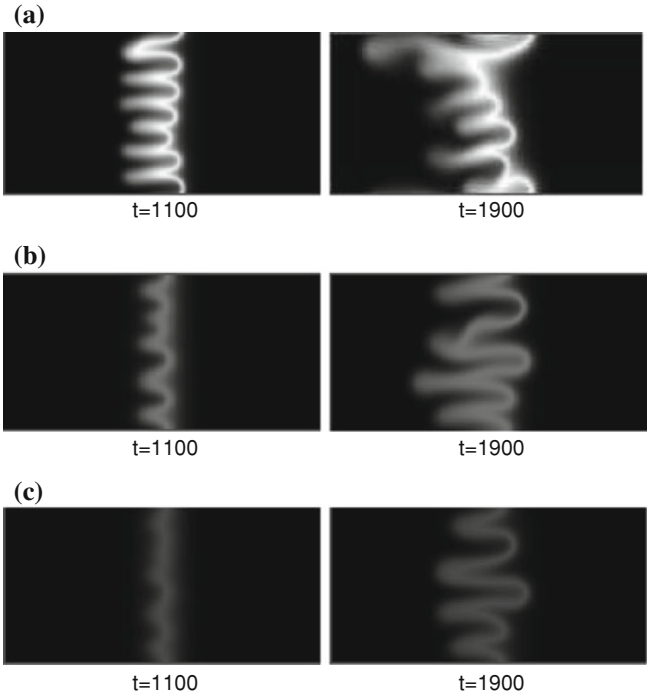


Fig. 8 Concentration iso-surfaces $R_b = 1$, $R_c = 3$ (unstable trailing front and stable leading front): **a** $\alpha = 0.0$, **b** $\alpha = 0.3$, **c** $\alpha = 0.8$

reversibility does actually enhance the instability particularly at the leading front, where the fingers become more developed and narrower with increasing α. Furthermore here too, reversibility leads to a more homogeneous distribution of the chemical product. Figure 9 depict results for reactive flow displacements where the viscosity of (C) lies between those of the two reactants. In this case both the trailing and the leading fronts are unstable ($R_c > 0$, $R_b - R_c/2 > 0$). The non-linear simulations indicate that reversibility does not actually have a major effect on the finger structure, and aside from the fact that the chemical product is more uniformly distributed in the porous medium when the reaction reverses, the number and overall structures of fingers are virtually unchanged. It should be finally noted that in all previous cases where the initial reactive front is unstable, stronger reversibility systematically leads to thinner and less diffuse fingers with a uniform distribution of the chemical product. The previous results can be explained by examining the effects of the chemical reversibility on the distribution of the viscosity on the different fronts. First, it should be noted that in a reactive displacement process, instability will not grow until a certain amount of the product (C) is generated. Furthermore, for the unstable initial reactive front case ($R_b > 0$) when the instability develops at one of the trailing or the leading fronts,

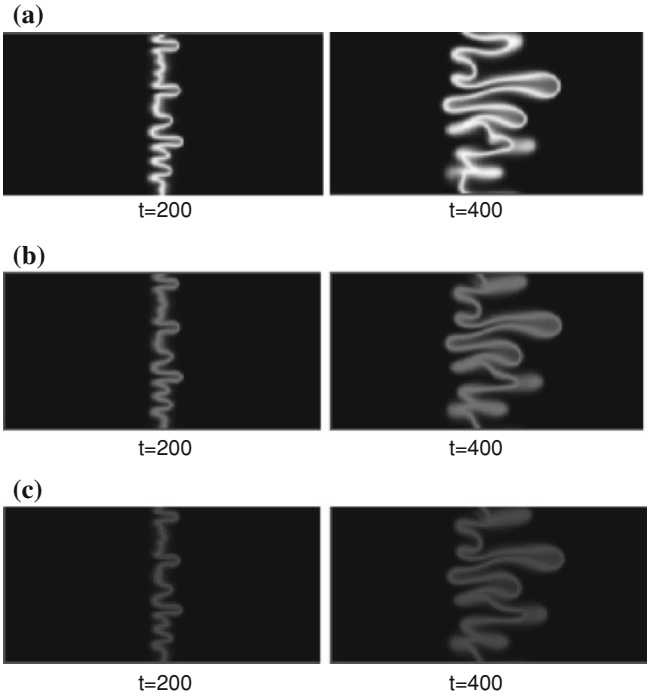


Fig. 9 Concentration iso-surfaces $R_b = 3$, $R_c = 4$ (unstable trailing and leading fronts): **a** $\alpha = 0.0$, **b** $\alpha = 0.3$, **c** $\alpha = 0.8$

the mobility ratio at that unstable front is always larger than that of the initial reactive front. As the reaction reverses and (C) is converted back into (A) and (B), the favourable mobility ratio between the reactants will increase and decrease the mobility ratios at the stable and the unstable front, respectively. Furthermore, it is known that the direction of injection is in favour of the growth of the instability at the leading (C–B), but not the trailing (A–C) front [15]. These two factors explain the influence of reversibility in attenuating or enhancing the instability of the cases where the instability developed at the leading ($R_b = 1, Rc = -2$ and $R_{CB} > R_b$) or the trailing ($R_b = 1$, $R_c = 3$ and $R_{CB} < R_b$) front, respectively.

The less noticeable effects of reversibility in Fig. 9 corresponding to ($R_b = 3$, $R_c = 4$, $R_{AC} = 2$, $R_{CB} = 1$) can be attributed to the fact that in this case the trailing and leading fronts are unstable, resulting in a stronger mixing of the chemical species. Furthermore, the favorable mobility ratio between the reactants (R_b) increases the mobility ratios at both the trailing and the leading fronts as the reaction reverses. This helps the instability to keep growing regardless of how fast the product (C) is converted back to (A) and (B).

4 Conclusion

In this study, the nonlinear development of fingering instabilities that develop in reactive flow displacements in porous media is examined. The study has in particular focused on the effects of chemical reversibility in bi-molecular reactions that affect the viscosity distributions of the three chemical species and in turn the fate of the flow. The study examined the effects of reversibility under the conditions of stable and unstable initial reactive fronts.

Analyses of concentration iso-surfaces of the chemical product revealed that in all flow situations, chemical reversibility tends to lead to a more uniform and homogeneous distribution of the product, when compared with the non-reversible reaction case. Furthermore, foe flows involving an initially stable front between the two chemical reactants, reversibility of the chemical reaction systematically induce an attenuation of the fingering instability. This attenuation was also observed in the case of an initially unstable front between the reactants that result in unstable leading and stable trailing fronts. No noticeable effects were however noted for viscosity distributions that correspond to an unstable initial reactive front with unstable trailing and leading fronts. The only exception where it was found that reversibility can actually enhance the fingering instability is for flows involving an unstable initial front between the two reactants with unstable trailing and stable leading fronts. In this particular case, it was found that the reversibility of the chemical reaction tends actually to enhance the growth of instabilities at the stable leading front. This enhancement can be attributed to the combined effect of the increase of the mobility ratio at the leading front due to the favourable mobility ratio between the initial reactants and the direction of the injection that promotes the growth of the fingers at the leading front.

Acknowledgements H. Alhumade acknowledges financial support from the Ministry of higher education in Saudi Arabia. The authors would like also to acknowledge WestGrid for providing computational resources.

References

1. H. Alhumade, J. Azaiez, In: *Reversible reactive flow displacements in homogeneous porous media*, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013, WCE 2013, 3–5 July, 2013, London, pp. 1681–1686
2. H. Alhumade, J. Azaiez, Stability analysis of reversible reactive flow displacements in porous media. Chem. Eng. Sci. **101**, 46–55 (2013)
3. T. Bansagi, D. Horvath, A. Toth, Nonlinear interactions in the density fingering of an acidity front. J. Chem. Phys. **121**, 11912–11915 (2004)
4. R.N. Bracewell, *The Fourier Transform and its Applications*, 2nd edn. (McGraw Hill, New York, 2000)
5. B. Chopard, M. Droz, T. Karapiperis, Z. Racz, Properties of the reaction front in a reversible reaction-diffusion process. Phys. Rev. E **47**, R40–R43 (1993). Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics

6. A. De Wit, G.M. Homsy, Nonlinear interactions of chemical reactions and viscous fingering in porous media. Phys. Fluids **11**, 949–951 (1999)
7. A. De Wit, G.M. Homsy, Viscous fingering in reaction-diffusion systems. J. Chem. Phys. **110**, 8663–8675 (1999)
8. K. Ghesmat, J. Azaiez, Miscible displacements of reactive and anisotropic dispersive flows in porous media. Tran. Porous Med. **77**, 489–506 (2009)
9. S. Havlin, D. Ben-Avraham, Diffusion in disordered media. Adv. Phys. **51**, 187–292 (2002)
10. S.H. Hejazi, J. Azaiez, Nonlinear interactions of dynamic reactive interfaces in porous media. Chem. Eng. Sci. **65**, 938–949 (2010)
11. G.M. Homsy, Viscous Fingering in Porous Media. Ann. Rev. Fluid Mech. **19**, 271–311 (1987)
12. D. Horvath, T. Bansagi, A. Toth, Orientation-dependent density fingering in an acidity front. J. Chem. Phys. **117**, 4399–4402 (2002)
13. M.N. Islam, J. Azaiez, Fully implicit finite difference pseudo-spectral method for simulating high mobility-ratio miscible displacements. Int. J. Num. Meth. Fluids **47**, 161–183 (2005)
14. K.V. McCloud, J.V. Maher, Experimental perturbations to Saffman–Taylor flow. Phys. Rep. **260**, 139–185 (1995)
15. M. Mishra, M. Martin, A. de Wit, Miscible viscous fingering with linear adsorption on the porous matrix. Phys. Fluids **19**, 1–9 (2007)
16. Y. Nagatsu, T. Ueda, Effects of reactant concentrations on reactive miscible viscous fingering. Fluid. Mech. Trans. Phen. **47**, 1711–1720 (2001)
17. Y. Nagatsu, T. Ueda, Effects of finger-growth velocity on reactive miscible viscous fingering. AIChE J. **49**, 789–792 (2003)
18. Y. Nagatsu, S. Bae, Y. Kato, Y. Tada, Miscible viscous fingering with a chemical reaction involving precipitation. Phys. Rev. E: Stat., Nonlin., Soft Matter Phys. **77**, 1–4 (2008)
19. H. Nasr-El-Din, K. Khulbe, V. Hornof, G. Neale, Effects of interfacial reaction on the radial displacement of oil by alkaline solutions. Revue—Institut Francais du Petrole **45**, 231–244 (1990)
20. T. Rica, D. Horvath, A. Toth, Density fingering in acidity fronts: Effect of viscosity. Chem. Phys. Lett. **408**, 422–425 (2005)
21. P. Saffman, G. Taylor, The penetration of a fluid into a porous medium or hele-shaw cell containing a more viscous liquid. Proc. R. Soc. Lond. A **245**, 312–329 (1958)
22. M. Sinder, V. Sokolovsky, J. Pelleg, Reversible reaction diffusion process with initially mixed reactants: Boundary layer function approach. Phys. B: Condens. Matter **406**, 3042–3049 (2011)
23. M. Sinder, H. Taitelbaum, J. Pelleg, Reversible and irreversible reaction fronts in two competing reactions system. Nucl. Instrum. Methods Phys. Res. Sect. B **186**, 161–165 (2002)
24. C. Tan, G. Homsy, Simulation of nonlinear viscous fingering in miscible displacement. Phys. Fluids **31**, 1330–1338 (1998)
25. A. Zadrazil, I. Kiss, J. D'Hernoncourt, H. Sevcikova, J. Merkin, A. De Wit, Effects of constant electric fields on the buoyant stability of reaction fronts. Phys. Rev. E **71**, 1–11 (2005)
26. W. Zhang, Nanoscale iron particles for environmental remediation: An overview. J. Nanopart. Res. **5**, 323–332 (2003)

Prediction of Thermal Deformation for a Ball Screw System Under Composite Operating Conditions

A. S. Yang, S. Z. Chai, H. H. Hsu, T. C. Kuo, W. T. Wu, W. H. Hsieh and Y. C. Hwang

Abstract The position error of a feed drive system is mostly caused by thermal deformation of a ball screw shaft. A high-speed ball screw system can generate massive heat with greater thermal expansion produced, and consequently have a negative effect on the positioning accuracy. In this study, we applied the computational approach using the finite element method (FEM) to simulate the thermal expansion process for estimating the deformation of the ball screw system. In the numerical analysis, the deformation of the ball screw shaft and nut was modeled via a linear elasticity approach along with the assumption that the material was elastic, homogeneous, and isotropic. To emulate the reciprocating movements of the nut at the speeds of 20, 40 and 60 m/min corresponding to the screw shaft, we also employed a three-dimensional unsteady heat conduction equation with the heat generation from the main sources including the ball screw shaft, nut and bearings as the heat transfer model to solve the temperature distributions for determining the temperature rises and axial thermal deformations in a ball screw

A. S. Yang · S. Z. Chai · H. H. Hsu
Department of Energy and Refrigerating Air-Conditioning Engineering, National Taipei University of Technology, Taipei 106, Taiwan
e-mail: asyang@ntut.edu.tw

T. C. Kuo · W. H. Hsieh (✉)
Department of Mechanical Engineering, and Advanced Institute of Manufacturing with High-tech Innovations, National Chung-Cheng University, Chiayi 621, Taiwan
e-mail: imewhh@ccu.edu.tw

W. T. Wu
Department of Biomechatronics Engineering Nation Pingtung University of Science and Technology, Pingtung 912, Taiwan
e-mail: azbennywu@gmail.com

Y. C. Hwang
HIWIN Technologies Corp, Taichung 408, Taiwan
e-mail: lawrence@mail.hiwin.com.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_2,

shaft under composite operating conditions. The simulated results demonstrated that the countermeasures must be taken to thermally compensate great deterioration of the positioning accuracy due to vast heat production at high rotating speeds of shaft for a ball screw system.

Keywords Ball screws · FEM · Heat transfer model · Machine tool · Positioning accuracy · Thermal deformation

1 Introduction

The performance of a ball screw feed drive system in terms of speed, positioning accuracy and machine efficiency plays a very important role in product quality and yield in manufacturing industries primarily including machine tools, semi-conductors, optoelectronics, and so on. Considering a high-speed precision ball screw system, the occurrence of contact surfaces (such as the interfaces between the ball and nut grooves, the ball and screw grooves, and the bearing and shaft) produces contact friction at these junctions. The friction of the nut and ball bearings entails a sudden and violent heating of balls, and in turn results in the temperature rises of the ball screw, leading to mechanical micro- deformations and an overheating of the coolant. Such a temperature heating of ball screw could also cause significant thermal deformations deteriorating the ball screw system accuracy in mechatronics tools or instruments [1].

The development of fabrication technology for a variety of applications necessitates high-precision apparatuses for achieving remarkably delicate goods with high output [2]. As indicated by Bryan [3], the thermal induced error in precision parts has still been the key setback in the industry. Substantial efforts were done on the machine tools, thermal behavior and thermal error compensation on the spindles, bearings and ball screws, respectively. Ramesh et al. [4] and Chen [5] carried out the air-cutting experiments to reproduce the loads under actual-cutting situations. To adjust the thermal conditions of the machine tool, Li et al. [6] conducted the tests of varying spindle speed for controlling the loads. The performance of a twin-spindle rotary center was experimentally evaluated by Lo et al. [7] for particular operating settings. Afterward, Xu et al. [8] incorporated the contact resistance effect into a thermal model for simulation of machine tool bearings. Koda et al. [9] produced an automatic ball screw thermal error compensation system for enhancement of position accuracy.

The frictional process from a high-speed ball screw system essentially released tremendous amounts of heat and results in the continuing temperature increase and thermal expansion, leading to deterioration of the positioning accuracy. In this investigation, we considered the heat generation from two bearings and the nut as the thermal loads with the prescribed heat flux values imposed on the inner surfaces of grooves between the bearings and nut of the ball screw system.

The convection boundary conditions were also treated for solid surfaces exposed to the ambient air. A FEM-based thermal model was developed to resolve the temperature rise distribution and in turn to predict the thermal deformation of the ball screw. In addition, simulations were conducted to appraise the influence of composite operating conditions in terms of different speeds (1,000, 2,000 and 3,000 rpm) as well as moving spans (500 and 900 mm) of the nut on the temperature increases and thermal deformations of a ball screw shaft.

2 Description of Ball Screw System

Figure 1 presents a schematic diagram of a ball screw feed drive system, encompassing a ball screw and driving unit. A continuous advance and return movement of the ball screw takes place in the range of 900 mm. It has 20-mm lead, 41.4-mm ball center diameter (BCD) and 1,715-mm total length. The outer and inner diameters of the screw shaft are 40 and 12.7 mm, respectively. Table 1 presents the parameters of main components for the ball screw drive system, which really contains the ball screw shaft, ball screw nut and bearings.

Figure 2 illustrates the moving velocity of the screw nut. This investigation considers the reciprocating movements of the nut at a maximum speed of 40 m/min pertaining to the screw shaft with a time period of 3.43 s and acceleration/deceleration of ± 2.1 m/s^2 as the baseline study case.

3 Computational Analysis

The physical model in this study investigates the thermal expansion process in a ball screw system. Essentially, heat is generated mainly from the friction between the ball and nut grooves as well as the ball and screw grooves. In view of the fact that a string of balls filled the grooves between the screw and nut are rotating very fast, the heat has been distributed evenly over the inner surface of raceways. The nut and two bearings are modeled as the fixed thermal loads imposed on the ball screw shaft. The thermal resistance resulting from the lubrication oil film between the balls and raceways is assumed to be ignored here attributable to a very thin layer of oil film, and the effect of heat conduction by means of the lubricant and thermal deterioration is negligible. Numerical calculations were performed by the FEM software ANSYS® to investigate the thermal behavior of a ball screw [10]. The theoretical formulation was based upon the time-dependent three-dimensional heat conduction equation for a ball screw system. The governing equations are stated as follows:

$$k\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}\right)=\rho c\frac{\partial T}{\partial t}. \tag{1}$$

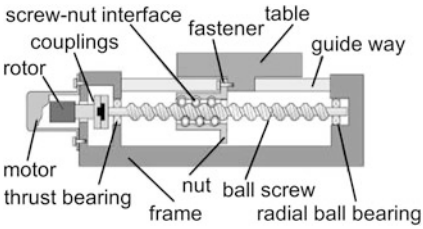


Fig. 1 Schematic of a ball screw feed drive system

Table 1 Main component parameters of the ball screw feed drive system

Ball screw shaft		Ball screw nut	
Total Length (mm)	1,715	Type	FDC
Thread Length (mm)	1,295	Length (mm)	143.4
LEAD (mm)	20	Diameter (mm)	70
BCD (mm)	41.4		
Outer diameter(mm)	40		
Inner diameter (mm)	12.7	*Bearing*	
Line number	2	Type	TAC
Contact type	4 points	OD (mm)	30
Ball diameter (mm)	6.35	ID (mm)	12.7

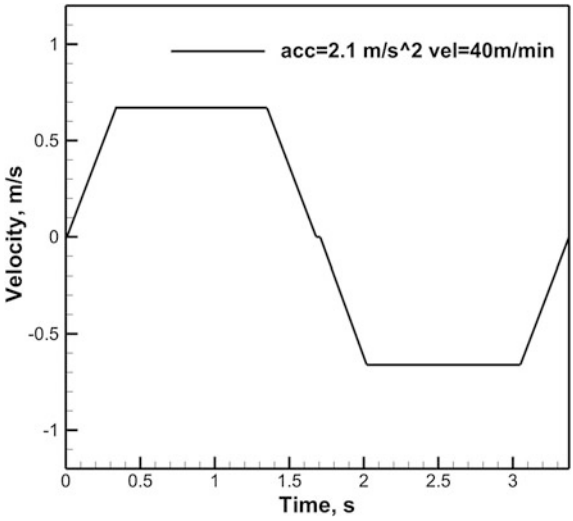


Fig. 2 Moving velocity of the screw nut with respect to the screw shaft

The symbols ρ, c, k and T mean the density, specific heat, thermal conductivity, and temperature of the ball screw shaft and nut, respectively. Here the temperature T is a function of the spatial coordinates (x, y, z) and time. The ρ, c, k values for computations are 7,750 kg/m^3, 480 J/kg-^{o}C and 15.1 W/m-^{o}C. Figure 3 exhibits the heat generation by the nut and bearing for a period of 3.43 s. The friction effect

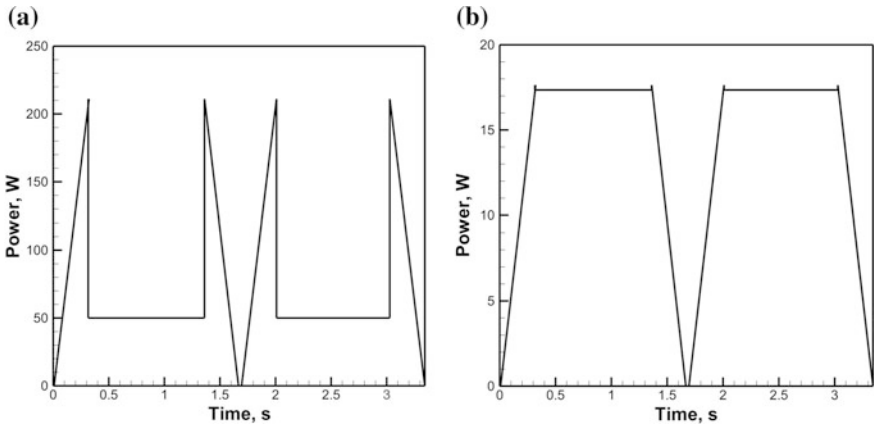


Fig. 3 Heat generation by **a** nut and **b** bearing for a period of 3.43 s

between the balls and raceways of the nut and bearings is the most important cause for temperature increase.

Given that the load of the nut contains two parts: the preload and dynamic load, $\dot{G}_{nut}$, the heat generation by the nut (in W), can be described as [11, 12]:

$$\dot{G}_{nut} = 0.12\pi f_0 v_0 nM. \tag{2}$$

Here f_0 is a factor determined by the nut type and lubrication method; v_0 is the kinematic viscosity of the lubricant (in m^2/s); n is the screw rotating speed (in rpm); M is the total frictional torque of the nut (in N-mm). In this research, $\dot{G}_{bearing}$ is the heat generated by a bearing (in W), defined as below [13].

$$\dot{G}_{bearing} = 1.047 \times 10^{-4} nM. \tag{3}$$

The variables n is the rotating speed of a bearing and M is the total frictional torque of bearings, including the frictional torque due to the applied load and the frictional torque due to lubricant viscosity.

The convective heat transfer coefficient h (in $W/m^2\text{-}^oC$) is computed [14] by

$$h = Nuk_{fluid}/d. \tag{4}$$

Here Nusselt number $Nu = 0.133Re^{2/3}Pr^{1/3}$, while the variables Re and Pr represent Reynolds number and the Prandtl number. The sign k_{fluid} is the thermal conductivity of the surrounding air and d is the outer or inner diameter of the screw shaft (mm). More detailed information can be found in Ref. [15].

In this study, the ball screw is modeled using a linear elasticity approach and assumed as the elastic, homogeneous, and isotropic material. The governing equations for the ball screw deformation are as follows:

$$\rho \frac{\partial^2 v_x}{\partial t^2} = \frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} + \rho g_x. \tag{5}$$

$$\rho \frac{\partial^2 v_y}{\partial t^2} = \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{zy}}{\partial z} + \rho g_y. \tag{6}$$

$$\rho \frac{\partial^2 v_z}{\partial t^2} = \frac{\partial \sigma_z}{\partial z} + \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \rho g_z. \tag{7}$$

The variables $\vec{v}$, $\vec{\sigma}$ and $\vec{\varepsilon}$ symbolize the displacement, stress and strain vectors (tensors).

$$\vec{\sigma} = D(\vec{\varepsilon} - \varepsilon^{th}) \tag{8}$$

The symbol ε^{th} is the thermal strain. Given the assumption of a linear elastic response, the stress-strain relationship is given by $\vec{\sigma} = D\vec{\varepsilon}$, where D has the form

$$D = \begin{pmatrix} \lambda + 2G & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2G & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2G & 0 & 0 & 0 \\ 0 & 0 & 0 & G & 0 & 0 \\ 0 & 0 & 0 & 0 & G & 0 \\ 0 & 0 & 0 & 0 & 0 & G \end{pmatrix} \tag{9}$$

$$\lambda = \frac{vE}{(1+v)(1-2v)}, \quad G = \frac{E}{2(1+v)} \tag{10}$$

$$\varepsilon^{th} = \Delta T(\ \alpha, \quad \alpha, \quad \alpha, \quad 0, \quad 0, \quad 0,)^T \tag{11}$$

D is the elasticity matrix consisting of the material properties, whereas the properties of E, v and α are the Young's modulus, the Poisson's ratio and the coefficient of thermal expansion (CTE). In this investigation, the values of Young's modulus, Poisson's ratio and CTE of the ball screw are set to be 1.93×10^{11} Pa, 0.31 and 1.16×10^{-5} m/m-°C. The term $\Delta T = T - T_{ref}$, T_{ref}, is the initial temperature of 27 °C. A finite element method was used to solve the ball screw model in accordance with the principal of virtual work. For each element, displacements were defined at the nodes and the associated displacements within the elements were subsequently obtained by means of interpolation of the nodal values by the shape functions. The strain-displacement and stress-strain equations for structure were solved with the Gaussian elimination method for sparse matrices [10].

4 Experimental Measurements

Experimental measurements were conducted to determine the time-dependent distributions of temperature, temperature rise and thermal deformation for validation of the thermal model by FEM and evaluate the performance of the cooling system. Figure 4 illustrates a schematic diagram of the experimental set up, containing the ball screw, driving unit, LNC controller, thermal/laser detection system for measuring temperature/deformation and linked data acquisition system. A continual forward and backward movement of ball screw ensued in a 900-mm range, having 20 mm lead, 41.4 mm BCD and 1,715 mm total length. The data processing module consisted of four thermal couples as well as a high-precision Renishaw XL-80 laser system with the test data recorded for every 600 s.

5 Results and Discussion

Simulations were attained by the FEM software ANSYS® to predict the thermal and deformation characteristics of a ball screw. In the analysis, we modeled a string of balls as a coil-like circular band fully filled the interior surface of raceways between the screw and nut. Figure 5 displays the numerical grids for system simulations. The mesh setup had three unstructured sections, such as the bearings, shaft and coil-like circular band. Finer grids were paced in the areas near the grooves and the solid surfaces. The average cell length was approximately 0.0015 m with the least spacing of 0.0006 m to resolve the steep variations of thermal properties near heat sources. The predictions of transient temperature distributions on the surface of the ball screw with different grids and time steps suggested that grid independence could be attained using a mesh setup of 1,367,867 grids with a time step of 5 s.

In simulations, the reciprocating movements of the nut was at a speed of 40 m/min respecting to the screw shaft with a 3.43-s period. The analysis was based on the rotational speed of 3,000 rpm and the initial temperature of 27 °C. During the continual operations, Fig. 6 illustrates the temperature and thermal deformation distributions along the axial distance of hollow ball screw shafts at t = 3,600 s. The frictional heat produces the temperature rise with high temperatures occurred in the core areas of the ball screws for hollow ball screws.

In the structure analysis, the left side of the ball screw was treated as the free end with the right side maintained as the fixed-end condition. The thermal loads from the temperature distribution predictions were then input to solve the thermal deformation. The predicted results indicated relatively larger local thermal deformations near the high temperature regions occurred at the center of the shaft in general. The thermal deformation distributions showed a great reduction in thermal expansion for the hollow shaft at the left end of the screw.

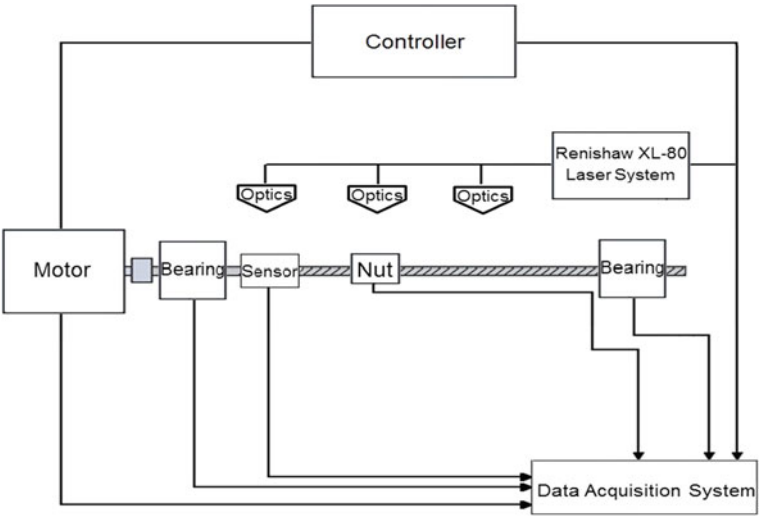


Fig. 4 Schematic diagram of the experimental set up

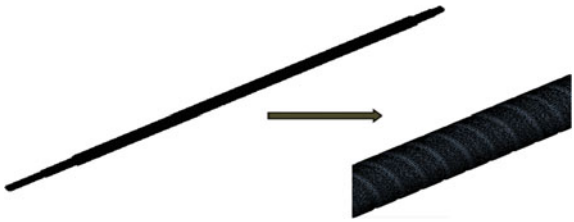


Fig. 5 Numerical grids of ball screw shaft

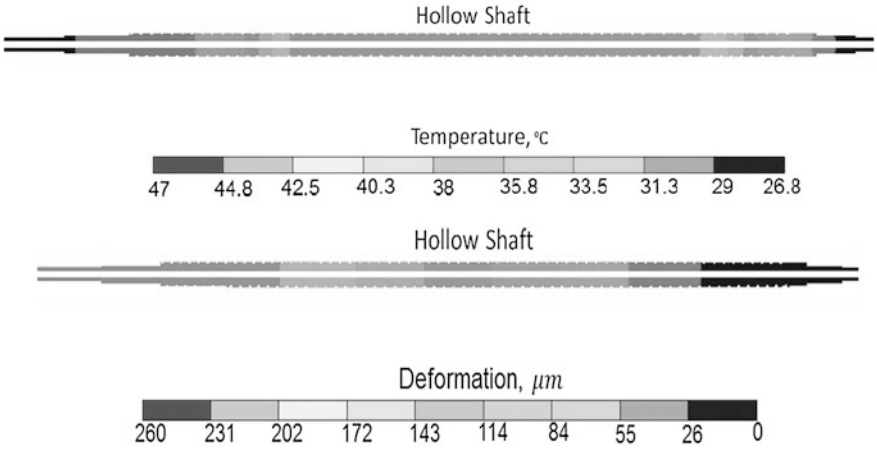


Fig. 6 Temperature and thermal deformation distribution of ball screw shafts at t = 3,600 s

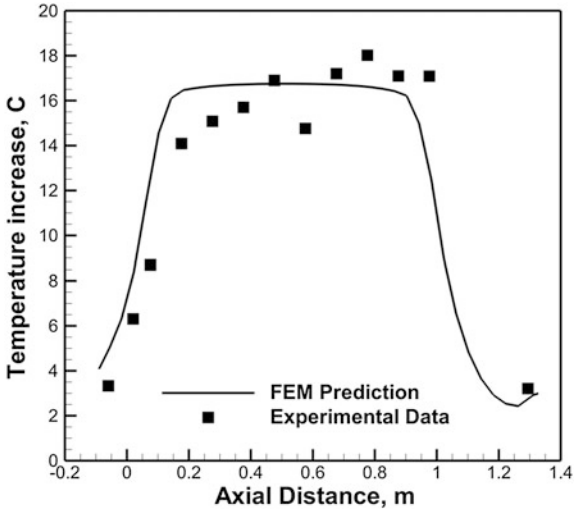


Fig. 7 Comparison of prediction with measured temperature rise data along the axial distance of ball screw at t = 3,600 s

Figure 7 illustrates a comparison of the prediction with the measured temperature rise data along the axial axis of hollow ball screw at t = 3,600 s. The calculated temperature rise distribution was compared with the measurements of different positions. Overall, the results clearly indicated that the FEM model relatively over-predicted the temperature increase (i.e. 17.7 °C in the central area of the shaft) with the discrepancy within approximately 18.7 % in the axial distance of 0.15 to 0.4 m, showing that the simulation software may reasonably simulate the thermal expansion phenomenon.

As far as the transient developments of temperature rises are concerned, Fig. 8 illustrates a comparison of time history of the predictions with measured temperature rises at three locations for the hollow ball screw. At the pre-specified axial distances of 0.05, 0.68 and 1.62 m, all three temperature rise profiles grew quickly in the early stage and tended to level off toward their steady-state values of 2.73, 17.2 and 2.66 °C for the test points of 1, 2 and 3, respectively, at t = 3,600 s. Due to massive heat generated in continuing operations, both the predicted measured results of the test point 2 show a higher temperature rise with the difference under 1.93 %, revealing the accuracy of the FEM simulations.

In order to verify the thermal and structural models by FEM, Fig. 9 illustrates a comparison of the prediction with measured thermal deformation data along the axial distance of hollow ball screw at t = 3,600 s. The thermal deformation along the axial direction was measured using a laser interferometer, for comparison with the calculated results. It can be clearly seen that the deformations from the FEM predictions and the experimental results were fairly close, deviating below 9.1 % for the axial distance from 0.2 to 0.9 m. Nevertheless, as compared to the test data, the over-predicted deformation was noted with a large error appeared at the axial distance of 0.1 m owing to the over-estimation of temperature rise in this associated area.

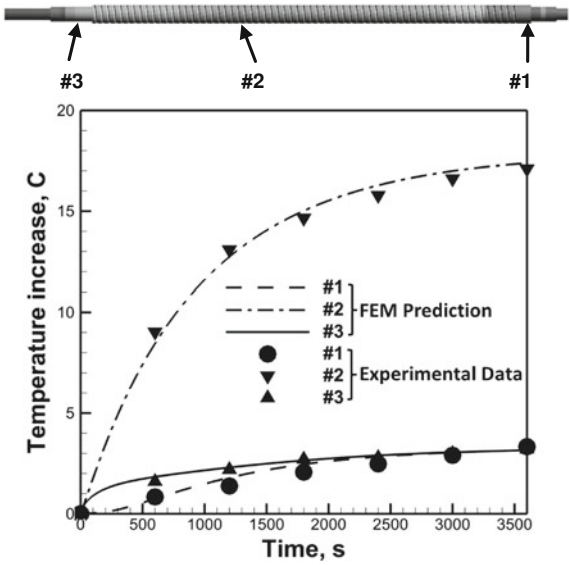


Fig. 8 Comparison of the time history of the predictions with measured temperature rises at three locations for ball screw

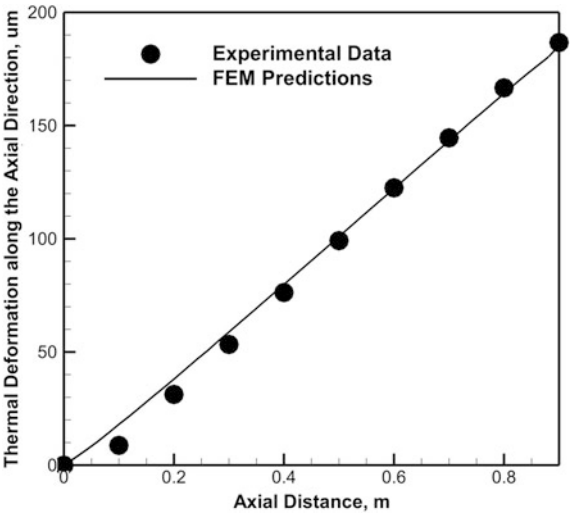


Fig. 9 Comparison of the prediction with measured thermal deformation data along the axial distance of ball screw at t = 3,600 s

The comparisons of predictions with the experimental data indicated that the FEM method could reasonably predict the thermal expansion process for determining the deformation of a ball screw system. Thus, calculations were extended to consider the influences of different spindle speeds and traveling distance of the nut on the temperature rises and thermal deformations of a ball screw shaft. Table 2 illustrates the numerical test details in terms of the rotating speed and

Table 2 Numerical test details in terms of the rotating speed and moving extent of the ball screw under the composite operating conditions

Time (s)	Speed (rpm)	Distance (mm)
300	2,000	900
300	2,000	500
300	3,000	900
300	1,000	500

moving extent of the ball screw under the composite operating conditions. In simulations, the shaft starts the operation at a speed of 2,000 rpm with a 900-mm moving distance of the nut for 300 s and next with a moving distance of 500 mm for 300 s, then changes to the speed of 3,000 rpm with a 900-mm moving distance for 300 s, and decreases the speed to 1,000 rpm with a 500-mm moving distance in the last 300 s. The axial spans, defined as the moving distance of the nut, are 900 and 500 mm from the center to the left and right ends of the ball screw, whereas the operating time and ambient temperature are 1,200 s and 27.7 °C, respectively.

In this analysis, the FEM method and experimental measurements were conducted simultaneously to predict the temperature increase and thermal deformation distributions for composite working states. Three monitoring points (corresponding to #1, #2 and #3) were selected at the axial distance of 0.045, 0.871 and 1.45 m to probe the temperature variations. Figure 10 shows a comparison of the predictions with the measured temperature data along the axial distance of ball screw at t = 1,200 s. The results revealed that the predicted temperatures relatively higher than those of the experimental values with the greatest discrepancy of around 1 °C under the composite operating conditions. Steep temperature gradients were viewed at two areas (with the axial distances of 0.3–0.5 and 0.8–1.2 m, respectively) of the screw shaft because of the frictional heat generation associated with the moving spans of the nut as well as the axial thermal conduction transfer to both ends of the shaft.

Under the composite operating conditions, Fig. 11 illustrates a comparison of the time histories of the predictions with the measured temperature of the ball screw at 1,200 s. At the pre-specified axial distances of 0.045, 0.871 and 1.45 m, all three temperature profiles increased in the first 600 s with the point #2 showing the steepest rate owing to a reduction of moving span from 900 to 500 mm during the period of 300 to 600 s. In response to an increase of the speed from 2,000 to 3,000 rpm with a 500-mm moving distance for 300 s, the higher heat production causes relatively faster temperature increases at both ends of the shaft during the period of 600 to 900 s. When the speed shifts to 1,000 rpm in the last 300 s, a lower heat source in a shorter central area (with a reduction of traveling span from 900 to 500 mm) can result in temperature drops at the points #1, 3 and a flatter temperature rise at the point #2 toward their measured values of 28.8, 41.2 and 32.5 °C, respectively, at t = 1,200 s. Overall, the FEM method reasonably predicted the temperature profiles with the maximum difference under 4.64 %.

Figure 12 shows a comparison of the prediction with measured positioning accuracy data from a laser interferometer along the axial distance of the ball screw

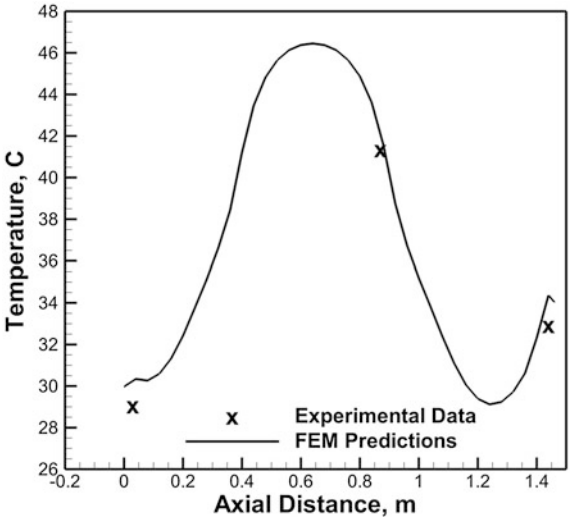


Fig. 10 Comparison of prediction with measured temperature data along the axial distance of ball screw at t = 1,200 s (under the composite operating conditions)

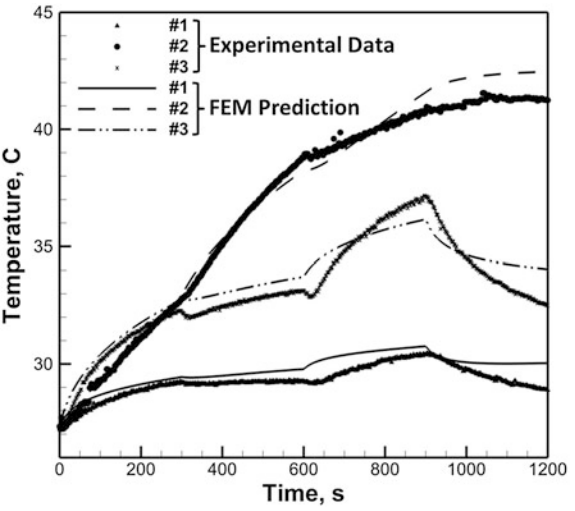


Fig. 11 Comparison of the time history of the predictions with measured temperature variation comparison at 1,200 s (under the composite operating conditions)

at t = 1,200 s. The FEM predictions were in reasonable agreement with the experimental results, and the highest discrepancy was 6.5 % at the axial distance of 0.8 m. Essentially, the FEM method slightly under-predicted the thermal deformation, as compared to the measurements. In addition, it can be obviously observed that the thermal deformations increase with the axial distance from both the FEM predictions and measured results, leading to the deterioration of the positioning accuracy up to 152 μm. The simulated results suggested that the significant decline of the positioning accuracy as a result of vast heat production at high rotating speeds of shaft must be thermally compensated for a ball screw system in operation.

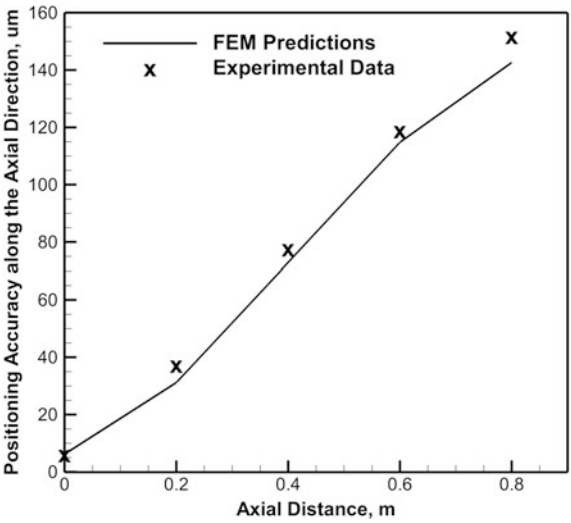


Fig. 12 Comparison of the prediction with measured positioning accuracy data along the axial distance of the ball screw at t = 1,200 s (under the composite operating conditions)

6 Conclusions

The position accuracy of a feed drive system was primarily influenced by the thermal deformation of a ball screw. A high-speed ball screw system can generate vast heat after long-term operations, with greater thermal expansion formed, and thereby negatively impact the positioning accuracy of the feed drive mechanism. In this research, the computational approach was applied using the FEM to simulate the thermal expansion development for solving the deformation of a ball screw. In simulations, we implemented the multi-zone heat loads to treat the heat generation sources from the frictions between the nut, bearings and the ball screw shaft to emulate reciprocating movements of the nut at a top speed of 40 m/min relative to the shaft in a time period of 3.43 s. We also employed a three-dimensional unsteady heat conduction equation to determine the steady and time-dependent temperature distributions, as well as the temperature increases for calculating the thermal deformations of the screw shaft. Simulations were extended to consider the composite operating conditions involving various spindle speeds and moving spans of the nut on the temperature rises and thermal deformations of a ball screw shaft. Both the FEM-based simulations and measurements found that the thermal deformations increased with the axial distance. The associated deformations can be up to 152 μm at 0.8 m in composite operating situations, and in turn depreciated the positioning accuracy. The computational and experimental results also indicated that the significant deterioration of the positioning accuracy due to massive heat production at high speeds of a shaft must be thermally compensated for a ball screw system in operations.

Acknowledgment This study represents part of the results under the financial support of Ministry of Economic Affairs (MOEA) and HIWIN Technologies Corp., Taiwan, ROC (Contract No. 100-EC-17-A-05-S1-189).

References

1. R. Ramesh, M.A. Mannan, A.N. Po, Error compensation in machine tools—a review. Part II: thermal error. Int. J. Mach. Tools Manuf. **40**, 1257–1284 (2000)
2. W.S. Yun, S.K. Kim, D.W. Cho, Thermal error analysis for a CNC lathe feed drive system. Int. J. Mach. Tools Manuf. **39**, 1087–1101 (1999)
3. J. Bryan, International status of thermal error research. Ann. CIRP. **39**(2), 645–656 (1990)
4. R. Ramesh, M.A. Mannan, A.N. Po, Thermal error measurement and modeling in machine tools. Part I. Influence of varying operation conditions. Int. J. Mach. Tools Manuf. **43**, 391–404 (2003)
5. J.S. Chen, A study of thermally induced machine tool errors in real cutting conditions. Int. J. Mach. Tools Manuf. **36**, 1401–1411 (1996)
6. S. Li, Y. Zhang, G. Zhang, A study of pre-compensation for thermal errors of NC machine tools. Int. J. Mach. Tools Manuf. **37**, 1715–1719 (1997)
7. C.H. Lo, J. Yuan, J. Ni, An application of real-time error compensation on a turning center. Int. J. Mach. Tools Manuf. **35**, 1669–1682 (1995)
8. M. Xu, S.Y. Jiang, Y. Cai, An improved thermal model for machine tool bearings. Int. J. Mach. Tools Manuf. **47**, 53–62 (2007)
9. S. Koda, T. Murata, K. Ueda, T. Sugita, Automatic compensation of thermal expansion of ball screw in machining centers. Trans. Jpn. Soc. Mech. Eng. Part C. **21**, 154–159 (1990)
10. ANSYS, 13 User Guide. ANSYS Inc. Canonsburg, PA, USA (2010)
11. A.S. Yang, S.Z. Cai, S.H. Hsieh, T.C. Kuo, C.C. Wang, W.T. Wu, W.H. Hsieh, Y.C. Hwang, in *Thermal deformation estimation for a hollow ball screw feed drive system*. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering, WCE 2013, 3–5 July, 2013, London, U.K., pp. 2047–2052
12. A. Verl, S. Frey, Correlation between feed velocity and preloading in ball screw drives. Ann. CIRP 59(2), 429–432 (2010)
13. T.A. Harris, *Rolling Bearing Analysis.* (Wiley & Sons, New York, 1991), pp. 540–560
14. H. Li, Y.C. Shin, Integrated dynamic thermo-mechanical modeling of high speed spindles, part I: model development. Trans. ASME, J. Manuf. Sci. Eng. **126**, 148–158 (2004)
15. Z.Z. Xu, X.J. Liu, H.K. Kim, J.H. Shin, S.K. Lyu, Thermal error forecast and performance evaluation for an air-cooling ball screw system. Int. J. Mach. Tools Manuf. **51**, 605–611 (2011)

Shear Strength and Fracture Surface Studies of Ball Grid Array (BGA) Flexible Surface-Mount Electronics Packaging Under Isothermal Ageing

Sabuj Mallik and Ahmed Z. El Mehdawi

Abstract Electronic systems are known to be affected by the environmental and mechanical conditions, such as humidity, temperature, thermal shocks and vibration. These adverse environmental operating conditions, with time, could degrade the mechanical efficiency of the system and might lead to catastrophic failures. The aim of this study is to investigate the mechanical integrity of lead-free ball grid array (BGA) solder joints subjected to isothermal ageing at 150 and 175 °C, for up to 1,000 h. Upon ageing at 150 °C the Sn-3.5Ag solder alloy initially age-softened for up to 200 h. This behaviour was linked to the coarsening of grains. When aged beyond 200 h the shear strength was found to increase up to 400 h. This age-hardening was correlated with precipitation of hard Ag_3Sn particles in Sn matrix. Further ageing resulted in gradual decrease in shear strength. This can be explained as the combined effect of precipitation coarsening and growth of intermetallic layer. Samples aged at 175 °C showed a similar behaviour with a reduced initial age-softening period and higher shear force values. Investigation of the fracture surfaces under a Scanning Electron Microscope (SEM) revealed that higher ageing temperature would expose the solder joints to brittle failures.

Keywords Ball grid array · Ductile and brittle fractures · Isothermal ageing · Lead free solder alloy · Shear strength · Solder joint

S. Mallik (✉) · A. Z. E. Mehdawi
Mechanical, Manufacturing and Design Engineering, Faculty of Engineering and Science, University of Greenwich, Central Avenue, Chatham, Kent, Greenwich ME4 4TB, UK
e-mail: S.Mallik@gre.ac.uk

A. Z. E. Mehdawi
e-mail: a.elmehdawi@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_3,

1 Introduction

Electronics manufacturing is one of the rapid changing industries in the world. Miniaturization of electronic products is at the heart of this change. Over the last two decades, the electronic manufacturing industries have experienced tremendous pressure to meet the requirements for miniaturized products, particularly, hand-held consumer products such as mobile phones, MP3 players etc. This is coupled with the requirement of electronic products with ever-higher performance. Functionality of these products has also evolved at the same pace through packing in more and more features.

To meet the demands many area-array packages are developed, such as Ball Grid Array (BGA), Chip-Scale Package (CSP) and Flip-Chip. Among these BGAs are now widely used as high performance miniature package and offer some distinct advantages over other surface mount packages. BGAs are mounted with the substrates at its bottom surface using solder balls. As there are no leads, BGAs have reduced co-planarity issues and are easier to handle [1]. Other important benefits of BGA include self-centering of solder balls during placement, higher ball pitch than the leaded flat packages, higher heat dissipation from the package to the substrate and also better electrical performance due to low induction leads.

In most cases the failures in electronics packaging originates at the solder joints between the electronic components and substrate. In a move towards more environment friendly electronic products, the lead-based solders are now being replaced with lead-free solders. It is therefore very important to study the new lead-free solder joints under harsh environmental conditions. The tiny solder joints at the bottom of BGA serve not only to provide electrical and mechanical connections but also to dissipate heat away from the chip. With further reductions in the size of solder joints, the reliability of solder joints has become more and more critical to the long-term performance of electronic products.

Several studies have been carried out to investigate the ageing behaviors of BGA solder joints. Ageing behavior of lead-free solder alloys differs from its Sn-Pb counterpart. Xiao et al. [2] reported that the ageing of Sn3.9Ag0.6Cu solder alloy at 180 °C initially leads to age-softening due to coarsening of grains. However, ageing at 180 °C beyond one day resulted in age hardening due to precipitation of hard Ag_3Sn particles. The same study also reported that the SnAgCu solder alloy offers better creep resistance than the Pb-Sn eutectic alloy. Koo and Jung [3] found that shear strength of BGA solder balls generally increased with increased test speed. Same study also demonstrated that the solder joint fractures at high shear speeds were mainly brittle fracture, with fractures occurring at the interface between the solder and pad. It was also concluded that the SnAgCu solder balls were more susceptible to interfacial brittle fractures than SnPb solder balls. In an investigation of ball shear strength of Sn-3.5Ag-0.75Cu solder balls on BGA pads with different surface finish by Lee et al. [4] it was demonstrated that solder ball shear strength

decreased with increasing ageing temperature and time. Fracture modes varied largely with pad metallization. For Cu pads, the fractures were mainly occurred at the interfacial intermetallic layer between the solder and the pad. Also for Cu pad metallization during ageing the fracture mode gradually changed from ductile fracture in the bulk solder to brittle interfacial fracture (at the intermetallic layers).

Previous research studies in this niche area have failed to establish a clear picture of the effect of isothermal ageing on the reliability of BGA solder joints. In fact, some of the conclusions from different sources are quite contradictory. This paper aims to clarify some of those issues through evaluating the mechanical reliably of BGA solder joints under isothermal aging.

2 Materials and Methods

Table 1 shows the materials used in the study. The solder balls used for the BGA attachment were made of lead-free Sn-3.5Ag solder alloy with 0.75 mm diameter. For the test vehicle flexible BGA substrate was used with 304 pads with a pad opening of 0.64 mm and pad metallization of electroplated Au/Ni over an underlying Cu pad (Fig. 1). Commercially available rosin-based, halide free and no-clean flux was used for surface mount attachment.

The BGA substrate was first covered with a thin layer of flux. The balls were then placed manually on to the substrate. For the ease of shear testing only limited number solder balls were placed as shown in Fig. 2. After placing the solder balls substrate test vehicles were then reflow soldered using a commercial scale six-zone convection reflow oven. The peak reflow temperature was maintained at around 230 °C. A total of eight flexible BGA substrate were reflow soldered. One of them was used as "as-soldered" sample and rest of them were placed in a climatic chamber for thermal ageing. Samples were aged at a constant temperature of 150 °C for up to 1,000 h, with samples taken out at 50, 100, 200, 400, 600, 800 and 1,000 h. A separate batch of samples were also isothermally aged at 175 °C for up to 500 h, with samples taken out at 50, 100, 200, 300, 400 and 500 h. Reliability of solder joints from the as-soldered and the aged samples 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/sec and 0.1 mm respectively for all the solder balls. For any particular ageing period the test sample size was 15, that is, 15 solder balls were sheared and the average was taken. The fractured solder balls were collected and fractured surfaces were investigated for brittle and ductile fractures under a scanning electron microscope (SEM).

Table 1 Test materials

Materials	Brief details
Solder balls	Sn-3.5Ag solder, 0.75 mm diameter
BGA substrate	Flexible substrate with 304 pads with a pad opening of 0.64 mm
Pad metallization	Electroplated Au/Ni over an underlying Cu pad
Flux	Rosin- based, halide free, no-clean

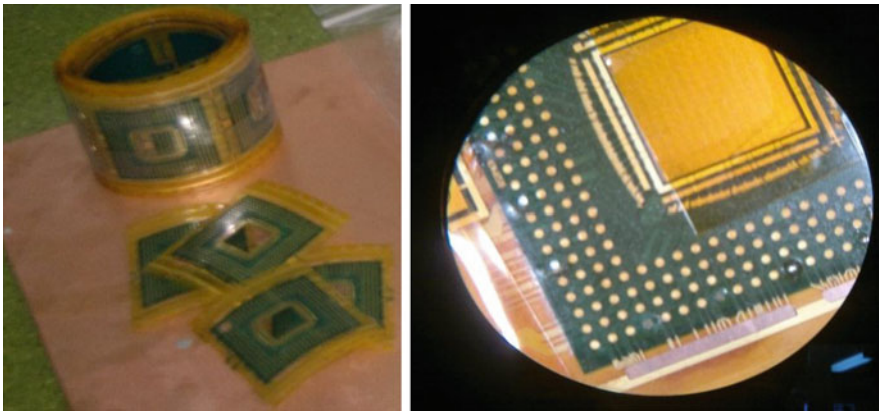


Fig. 1 Flexible BGA substrate (*left*) and magnified view of solder ball pads (*right*)

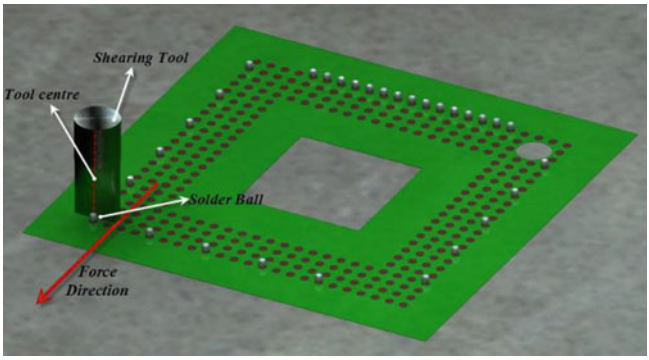


Fig. 2 Schematic of BGA substrate showing the location of mounted *solder balls* and positioning of *shear tool* during ball shear testing

3 Results and Discussion

The results are analyzed and discussed in three parts. First part looks at the effect of isothermal ageing on BGA solder joint shear strength. Second part outlines the effect of ageing temperature on BGA solder joint shear strength. The third and final part studies the solder ball fractured surfaces.

3.1 *Effect of Isothermal Ageing on the Shear Strengths of BGA Solder Joints*

Figure 3 presents the results from ball shear tests carried out on the as-soldered and isothermally (at 150 °C) aged BGA solder joints. The plot shows both the average shear force values and the variations in the collected data. The high degree of variation in shear force data could be due to uneven manual application of flux before the solder ball placement. Flux serves to facilitate solder wetting by cleaning surface oxides and preventing further oxidation during a soldering process. However, too-much or too-low quantity of flux could have an impact on solder joint integrity. In deed Painaik and Santos [5] found that the BGA solder ball shear strength affected by the flux quantity used in the BGA solder ball attachment process. Higher flux thickness and pad coverage were found to increase ball shear strength and vice versa.

A careful observation of Fig. 3 reveals three different patterns on how shear forces changed with increased ageing time. The shear force of Sn-3.5Ag solder joints was initially dropped till 200 h. This was then followed by an increase in shear force till 400 h. Ageing beyond 400 h resulted in gradual decrease in shear force. The initial decrease in shear force can be explained as due to the coarsening of grains [2, 3]. The coarsening of grains can be explained through a process called 'Ostwald Ripening' [6]. This natural spontaneous process occurs because larger grains are thermodynamically more stable than small grains. The whole process starts from the fact that molecules on the surface of grains are energetically less stable than the ones inside the grains. Due to lower surface to volume ratio large grains have a lower surface energy than the smaller grains. This creates a potential deference at the grain boundaries and as a result molecules from small grains diffuse through the grain boundaries and attach themselves to the larger grains. Therefore, the number of smaller grains continues to shrink, while the larger grains continue to grow. Without the application of adequate thermal or mechanical energy grain growth is very slow. The rate of grain growth increases with the application of thermal energy, because increased diffusion allows for more rapid movement of molecules. The reduced number of grain boundaries (due to grain coarsening) allows dislocations (crystal defects) to move easily through the boundaries. This results in the solder joints to deform at much lower shear loads. Xiao et al. [2] also reported softening of Sn3.9Ag0.6Cu solder alloy when aged at 180 °C. However the age-softening period was much shorter (24 h compared to 200 h) than what was observed in this study. This might be due to a lower ageing temperature of 150 °C. After 200 h of aging the solder joints shear strength was found to increase up to 400 h. This increase in shear strength could be due to precipitation hardening. In the same study on the Sn-Ag-Cu solder joints Xiao et al. [2] also observed the precipitation of hard Ag_3Sn particles after 1 day of ageing at 180 °C. Ageing beyond 400 h resulted in gradual drop in shear force. This behavior was quite expected and can be attributed due to combined effect of saturation of precipitants and further coarsening of grains and the growth of

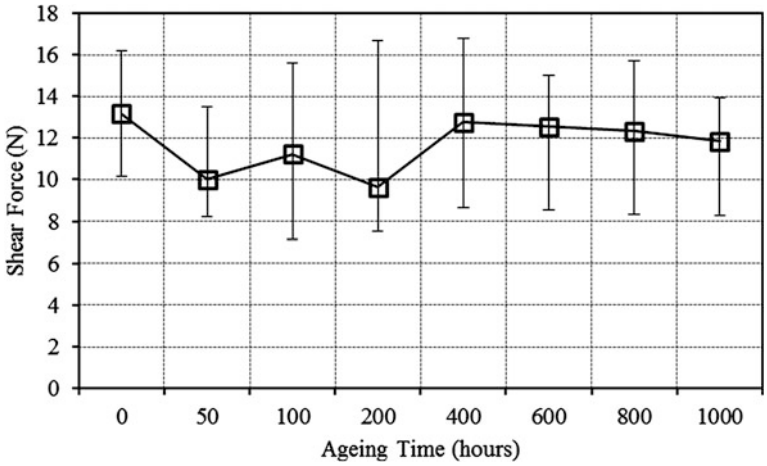


Fig. 3 Shear forces of Sn-3.5Ag BGA solder joints as a function of ageing time. Samples aged at 150 °C

intermetallic layers. In an ageing study on Sn-Ag-Cu solder joints Lee et al. [4] concluded that the Cu-Sn intermetallic layers are the major contributor of solder joint shear failures.

3.2 Effect of Ageing Temperature on Shear Strengths of BGA Solder Joints

In order to see the effect of ageing temperature, BGA solder joints were also separately aged at 175 °C, for a shorter period (up to 500 h). Although the plan was to pursue the isothermal ageing for 1,000 h, the authors could not continue the process after 500 h due to severe burning of the flexible substrate samples.

Figure 4 presents the ball shear test results for the as-soldered and isothermally aged samples for the two different ageing temperatures—150 and 175 °C. Figure 4 thus provides an opportunity for a comparative study of solder joint shear strengths, for different ageing temperatures. In general, the measured shear force profile for 175 °C ageing temperature followed a very similar pattern to the shear profile obtained for 150 °C ageing temperature. This means that the isothermal ageing behaviors at 175 °C can generally be explained in the same way as already done for 150 °C ageing temperature (Sect. 3.1). However, there are couple of discrepancies which require explanation.

Firstly, it was observed that the initial age hardening period (denoted by a decrease in shear force) were shorter for 175 °C compared to 150 °C. When aged at 175 °C, the shear force was initially decreased up to 100 h, whereas at 150 °C, the shear force was observed to decrease up to 200 h. As stated previously, Xiao

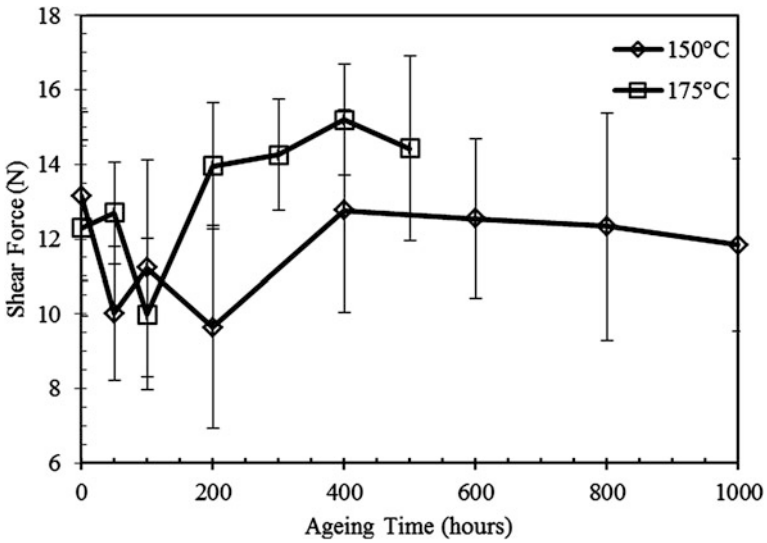


Fig. 4 Shear forces of Sn-3.5Ag BGA solder joints as a function of ageing time

et al. [2] also observed a shorter age-softening period at an ageing temperature of 180 °C. This implies that the initial grain coarsening period gets shorter with increased ageing temperature. In other words, at higher ageing temperature the effect of precipitation hardening kicks in earlier in the ageing process.

It was also observed that, for 200 h of ageing and beyond, the shear force values were higher when the solder joints aged at 175 °C. The higher shear forces could be due to higher rates of precipitation hardening at an elevated ageing temperature. This could also be a result of lower level of intermetallic formation, at higher ageing temperature. Nevertheless, this indicates that the solder joints become stronger when aged at higher temperature. However, this finding is not conclusive as this study only investigated two ageing temperatures. Further study with different isothermal ageing temperatures is required to come to a firm conclusion on this issue.

3.3 Study of Solder Ball Shear Fractures

The fracture behaviors of BGA 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 [7, 8]. Among these failures interfacial fractures are predominantly brittle and bulk solder fractures are tend to be ductile in nature.

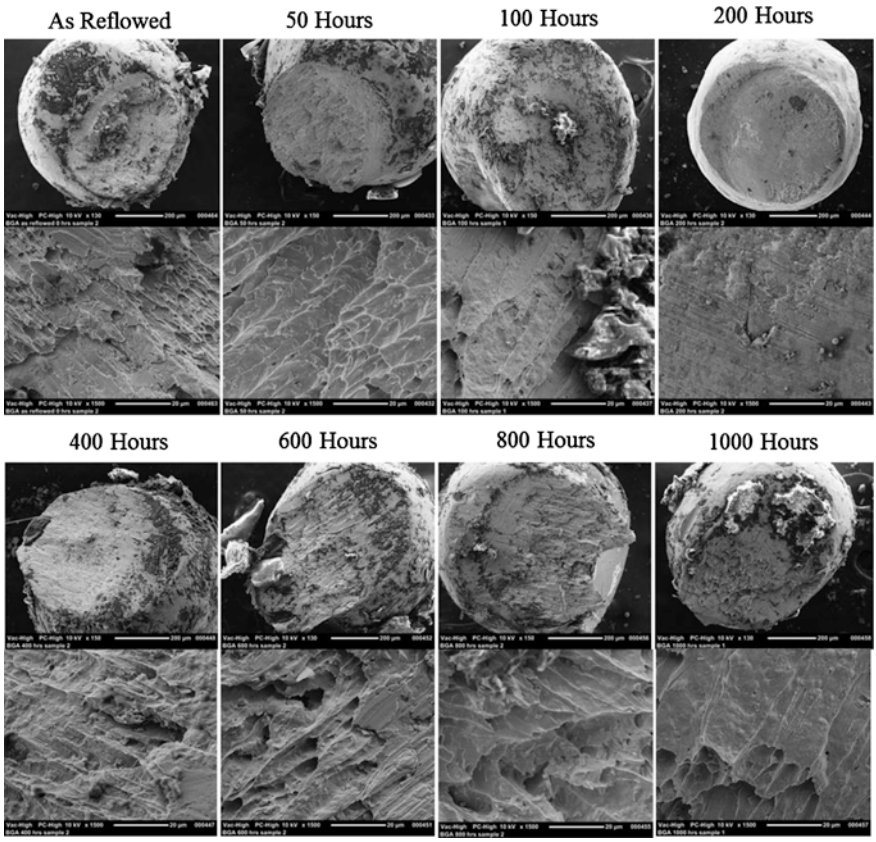


Fig. 5 Fracture surfaces of BGA joints on Au/Ni-Cu metallisation at different ageing times. Samples aged at 150 °C

However, solder ball failure through mixed fractures are also frequently observed by various researchers [3, 7].

Figure 5 presents the fractures surfaces from ball shear tests at different ageing time, for samples aged at 150 °C. From general understanding it was expected that the BGA solder joints will fail either at the interface between solder and substrate (at the intermetallic layers) or at a region of low strength (e.g. bulk solder). It was also expected that shear fracture mode will be ductile initially and then will show a transition towards brittle fracture with increased ageing time. However, the fracture surfaces didn't show any particular trend and were found to be generally ductile in nature. Although there was no particular trend in the shear fracture, coarsening of grains is evident from 400 to 1,000 h. This finding matches very well with the finding from shear force study where the shear force was found gradually decreasing after 400 h of ageing. The coarsening of grains on the fracture surface was also observed by Lee et al. [4] for shear fracture of BGA

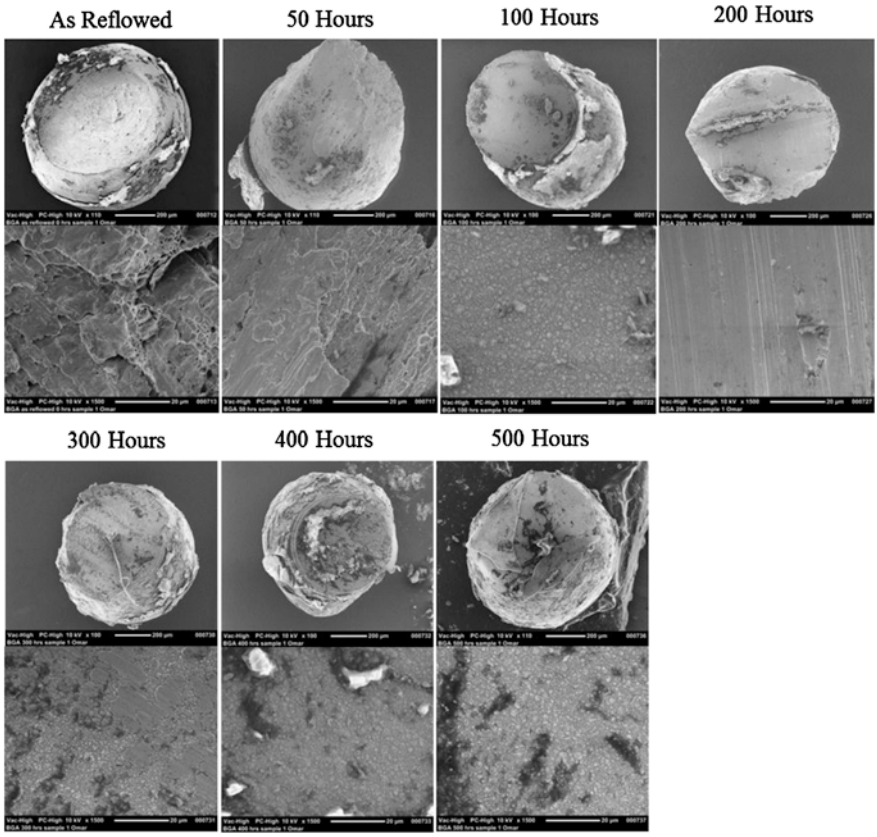


Fig. 6 Fracture surfaces of BGA joints on Au/Ni-Cu metallization at different ageing times. Samples aged at 175 °C

solder joint with Au/Ni-Cu pad under isothermal ageing. The ductile nature of the shear fractures and coarsening grain structure also indicate that fractures occurred in bulk solder irrespective of the ageing time. However, the author believes that a more detail elemental analysis of the fracture surfaces is required for an in-depth understanding of fracture locations.

Figure 6 presents the SEM images of the fracture surfaces from ball shear tests of samples aged at 175 °C. In the contrary to what was observed in Figs. 5, 6 shows a trend in the solder joint shear fracture behavior. Although the solder balls failed through mixed failures in most of the cases, as expected, the shear fracture mode was predominantly ductile initially (up to 50 h of ageing) and then showed a transition towards brittle fracture with increased ageing time. The increased proportion of brittle failures with increased ageing time indicates that although the initial ductile failure occurred in the bulk solder, the later brittle fractures did eventually initiated at the interface between solder and substrate. That is to say that the formation and growth of interfacial intermetallic layers was responsible for the

brittle fractures. This finding does contradict with the observed fracture behaviors for samples aged at 150 °C, but nonetheless gives an indication that higher ageing temperature could make the solder balls more vulnerable to brittle failures. Therefore, based on the findings from shear strength (reported in Sect. 3.2) and fracture studies it can be concluded that higher ageing temperature not only makes the solder joints stronger (in shear) but will also make them vulnerable to brittle failures.

4 Conclusion

In this study Sn-3.5Ag lead-free BGA solder joints were isothermally aged for up to 1,000 h, at two different ageing temperatures—150 and 175 °C. The solder joint shear strengths were evaluated and solder ball fracture surfaces were investigated. Upon ageing the solder joints at 150 °C, shear strength initially decreased as a result of grain coarsening. This was then followed by increase in shear strength, which was mainly due to precipitation hardening. A gradual decrease in shear strength was observed after 400 h of ageing. Overall, it can be concluded that the lead-free BGA solder joints are able to maintain the shear strength even after 1,000 h of isothermal ageing. BGA solder ball samples aged at 175 °C showed a similar shear force profile with reduced age-softening period and higher shear force values. Investigation of shear fractures, for samples aged at 150 °C, didn't show any particular trend. However, shear failures were mainly due to ductile fractures in bulk solders and therefore, the intermetallic interfacial layers were not responsible for shear failures. Observation of shear fractures of solder joints aged at 175 °C however, indicated that the higher ageing temperature would make the solder joint more vulnerable to brittle failures.

References

1. Intel, Ball grid array packaging: packaging databook, Chap. 14 [Online] (2000). Available: http://www.intel.co.uk/content/www/us/en/processors/packaging-chapter-14-databook.html
2. Q. Xiao, L. Nguyen, W.D. Armstrong, Aging and creep behavior of Sn3.9Ag0.6Cu solder alloy, in *Proceedings of the Electronic Components and Technology Conference*, 2004, pp. 1325–1332
3. J.M. Koo, S.B. Jung, 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)
4. C.B. Lee, S.B. Jung, Y.E. Shin, C.C. Shur, Effect of isothermal ageing on ball shear strength in BGA joints with Sn-3.5Ag-0.75Cu solder. Mater. Trans. **43**(8), 1858–1863 (2002)
5. M. Painaik, D.L. Santos, in *Effect of Flux Quantity on Sn-Pb and Pb-Free BGA Solder Shear Strength, SEMI Technology Symposium: International Electronics Manufacturing Technology (IEMT) Symposium*, 2002, pp. 229–237

6. C. Rauta, A. Dasgupta, C. Hillman, Solder phase coarsening, fundamentals, preparation, measurement and prediction [Online]. Available: http://www.dfrsolutions.com/wp-content/uploads/2012/06/Solder-Phase-Coarsening-Fundamentals-Preparation-Measurement-and-Prediction.pdf, 2009
7. K. Newman, BGA brittle fracture—alternative solder joint integrity test methods, in *Proceedings: 5th Electronic Components and Technology Conference*, 2005, pp. 1194–1201
8. S. Mallik, A.Z.E. Mehdawi, in *Evaluating the Mechanical Reliability of Ball Grid Array (BGA) Flexible Surface-mount Electronics Packaging under Isothermal Ageing*. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE 2013, London, 3–5 July 2013, pp. 1919–1922

Optimum Parameters for Machining Metal Matrix Composite

Brian Boswell, Mohammad Nazrul Islam and Alokesh Pramanik

Abstract The need for optimum machining parameters has always been of paramount importance for metal cutting. The economics of the process largely depends on selecting the best machining parameters. However, the additional challenge of being environmentally friendly in production while still being cost effective is now imperative. Machining conditions are not always conducive in reducing the carbon footprint when cutting material with a low machinability rating. This is particularly pertinent when aerospace material such as Boron Carbide Particle Reinforced Aluminium Alloy (AMC220bc) is machined. This material falls under the category of a particulate reinforced Metal Matrix Composite (MMC), where the ceramic fibers disrupt the flow of electrons. The result is a decrease in thermal conductivity causing the tool interface temperature to increase, reducing tool life. This research will determine the optimum economic and sustainable machining parameters for this material.

Keywords Aerospace material · Carbon footprint · Machinability · Machining parameters · Metal matrix composite · Taguchi method

B. Boswell · M. N. Islam (✉) · A. Pramanik
Department of Mechanical Engineering, Curtin University, GPO Box U1987
Perth, WA 6845, Australia
e-mail: m.n.islam@curtin.edu.au

B. Boswell
e-mail: b.boswell@curtin.edu.au

A. Pramanik
e-mail: alokesh.pramanik@curtin.edu.au

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_4,

1 Introduction

Aircraft parts by necessity should be made from lightweight, durable and fatigue resistant materials. Commonly used aerospace materials [1] which have these qualities are Aluminium alloys, Titanium alloys and Stainless Steels—of which Titanium alloys and Stainless Steels unfortunately have a low machinability rating, with Aluminium alloys suffering from galling and smearing [2]. Machining notoriously creates waste compelling companies to reduce their impact on the environment and put in place appropriate waste disposal measures. This in turn is necessitating Life Cycle Analysis (LCA) to be part of all manufacturing and aerospace design. Embrace sustainable manufacturing philosophy enables companies to reduce their carbon footprint, and improve their profitability. This requires that the best machining practices are used in an effort to reduce the total amount of greenhouse gas produced during cutting. The total waste produced by machining consists of metal chips, tool tips and coolant if used. In addition to the obvious waste produced during metal cutting is the amount of greenhouse gas produced from the electrical power used by the machine tool [3]. The technique used for assessing the environmental aspect and potential impact associated with machining is performed in accordance with the Environmental Management Life Cycle Assessment Principles and Framework ISO 14040 standard [4].

The challenging task of selecting the most environmental method of producing parts can be made simpler by using SimaPro software, which allows a number of scenarios to be evaluated. This analysis of the machining process identified the best reduction of greenhouse gasses, and how it was achieved. In practice, many cutting parameters need to be considered, such as cutting force, feed rate, depth of cut, tool path, cutting power, surface finish and tool life [2]. Dry machining is obviously the most ecological form of metal cutting as there are no environmental issues for coolant use or disposal to consider. For this reason the machining tests were all carried out by dry cutting [5]. Machining conditions and parameters are seen to be vital in order to obtain high quality products with the lowest environmental impact, at the lowest cost. The challenge that the manufacturing industry faces is how to find the optimum combination of cutting conditions in order to sustainably produce parts at a reduced cost to manufacture. To help achieve this goal the use of the Taguchi Method was used to establish the optimum cutting parameters to machine AMC220bc material. This method of statistical control allows the effect of many different machining parameters to be robustly tested on their machining performance. A three level L_{27} orthogonal array was selected where 0, 1 and 2 represent the different levels of the three control parameters, cutting speed, feed rate and depth of cut.

Analysis of the machining tests provided the deviation, and nominal values of the three quality measurements were used to determine the optimum machining parameters (length error, width error and surface roughness). Further analysis implemented the use of signal-to-noise ratios to differentiate the mean value of the experimental and nominal data of these quality measurements. A viable measure of

detectability of a flaw is its signal-to-noise ratio (S/N). Signal-to-noise ratio measures how the signal from the defect compares to other background noise [6]. The signal-to-noise ratio classifies quality into three distinct categories and the noise ratio differs with each category. The three different formulas are given below [7];

$$\frac{S}{N} = -10 \log \frac{1}{n}\left(\sum_{i=n}^{n} y_i^2\right) \quad [\text{dB}] \tag{1}$$

$$\frac{S}{N} = 10 \log \frac{1}{n}\left(\frac{\bar{y}}{s_{\bar{y}}^2}\right) \quad [\text{dB}] \tag{2}$$

$$\frac{S}{N} = -10 \log \frac{1}{n}\left(\sum_{i=n}^{n} \frac{1}{y_i^2}\right) \quad [\text{dB}] \tag{3}$$

The results from these formulas suggest that the greater the magnitude of the signal-to-noise ratio, the better the result will be because it yields the best quality with least variance [8]. The signal-to-noise ratio for each of the quality measurements; surface roughness, length error and width error were calculated and the mean signal-to-noise ratio for each parameter was found and tabulated. The results were graphed to illustrate the relationship that exists between S/N ratio, and the input parameters at different levels. The gradient of the graph represented the strength of the relationship for each of the machining parameters.

To help analyse the contribution of each variable and their interactions in terms of quality the Pareto ANOVA is implemented. The Pareto ANOVA was completed for each of the quality measures length error, width error and surface roughness. The Pareto ANOVA identified which control parameter affected the quality of the machined workpiece. By using the Pareto principle only 20 % of the total machining configuration is now needed to generate 80 % of the benefit of completing all machining test configurations [9]. This method separates the total variation of the S/N ratios. Each of the measured quality characteristics length error, width error and surface roughness, has its own S/N values for each of the 27 different tests. In order to obtain accurate result the S/N values are derived from an average value of 3 readings for each of the quality measurements. This research is intended to show all manufactures the effectiveness of using the optimum machining parameters for minimising their carbon footprint, and is based on a revised and extended version of our previous work [10].

2 Machine Test and Set-up

Normally a Polycrystalline Diamond (PCD) tool tip is used to machine MMC material as they can operate at speed due to their hardness. However, uncoated carbide tool tips were used to reduce the time needed for tool tips to exhibit wear

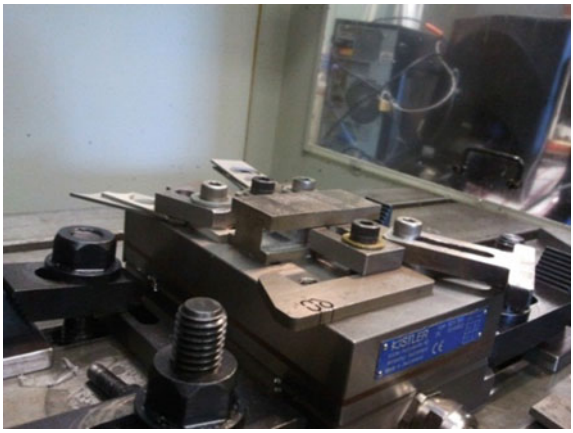


Fig. 1 Workpiece clamped onto dynamometer

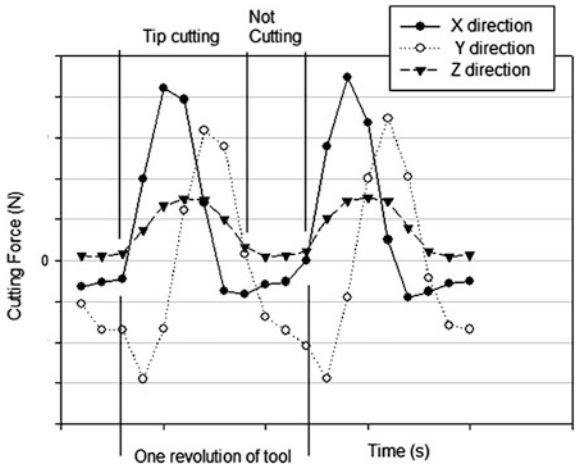


Fig. 2 Typical cutting forces for end milling

for analysis. Measurement of the cutting forces and power allows for analysis of the cutting operation and optimization of cutting parameters, as well as identifying wear of the tool tip. These important cutting forces were measured by a Kistler dynamometer which has a high natural frequency, and gives precise measurement. Figure 1 shows the workpiece securely clamped onto the dynamometer.

Dynoware28 software was used to provide high-performance real-time graphics of the cutting forces, and is used for evaluation of the forces. The end mill used a single tool tip to aid analysis of the cutting action shown in Fig. 2, which typically illustrating the intermittent engagement of the tool tip as it comes in contact with the material.

Real time machining power was measured by using a Yokogawa CW140 clamp on a power meter which was attached to the machines input power supply. The physical geometrical characteristics of the workpiece were precisely measured

Table 1 Control parameters and their levels

Control parameters	Units	Symbol	Levels		
			Level 0	Level 1	Level 2
Cutting speed	m/min	A	50	100	150
Feed rate	mm/rev	B	0.10	0.20	0.30
Depth of cut	mm	C	1.0	1.5	2.0

using a Discovery Model D-8 Coordinate Measuring Machine (CMM); the workpieces were split into 3 levels of 9 respectively. Mitutoyo Surftest SJ-201 portable stylus type surface roughness tester was used to measure the surface quality of the workpieces. The ideal roughness represents the best possible finish which can be obtained for a given tool geometry, and feed rate. This can only be achieved if inaccuracies such as chatter are completely eliminated. Natural roughness is greatly influenced by the occurrence of a built up edge. The larger the built up edge, the rougher the surface produced, factors tending to reduce chip-tool friction and to eliminate or reduce the built up edge would yield an improved finish.

For this research there were 27 different combinations of cutting speed, feed rate and depth of cut used, each match up with a trial level in the L_{27} orthogonal array. The values of the combinations of control parameters that correspond to the L_{27} orthogonal array can be found from Table 1. The best quality measurements will identify the optimal machining parameters for sustainable production. The Leadwell V30 CNC milling machine was used to machine the workpieces, allowing easy changes to the machining parameters for the different tests.

3 Results and Discussion

Variations in cutting power for input parameters; cutting speed, feed rate and depth of cut are shown in Figs. 3, 4, 5, 6, 7, 8. In these machining tests an unique combination of control parameters each have different combinations of different level values (0, 1, 2), and machining parameters A, B and C representing the cutting speed, feed rate and depth of cut respectively. For this research the 'smaller the better' category of the signal-to-noise ratio is chosen, which is shown as Eqs. 1–3. The results from this formula suggest that the greater the magnitude of the signal-to-noise ratio, the better the result will be because it yields the best quality with least variance.

Figure 9 shows that machining parameter with A (cutting speed) having the most significant effect on length error, followed by C (depth of cut) and then B (feed rate). The interaction between B $\times$ C also influences the machining process. The highest cutting speed, A2 was the best cutting speed to achieve a low length error. Since the interaction of B $\times$ C was also significant, it can be seen that the optimum combination for factors B and C in order to achieve a low length error

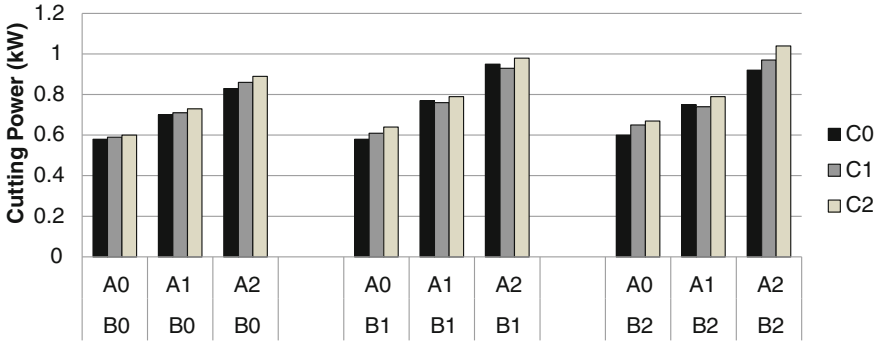


Fig. 3 Comparison of cutting power for different levels of cutting speed and feed rate versus different levels of depth of cut

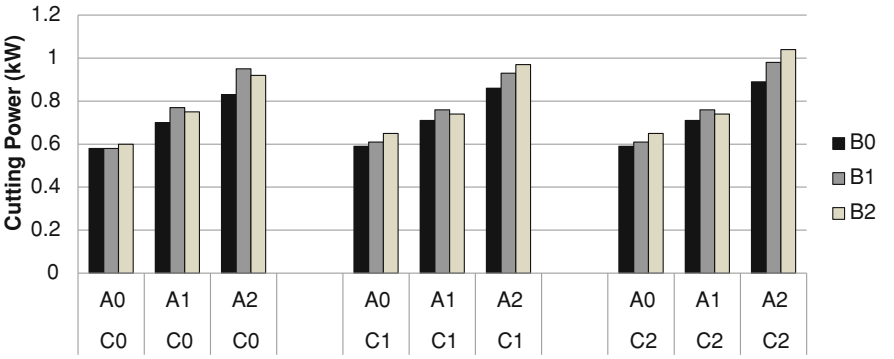


Fig. 4 Comparison of cutting power for different levels of cutting speed and depth of cut versus different levels of feed rate

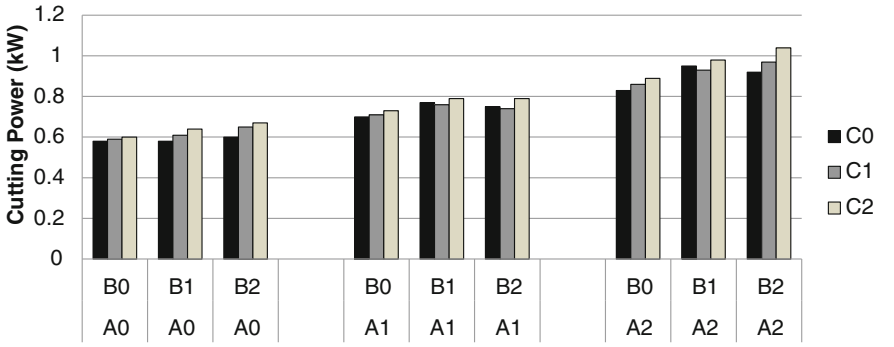


Fig. 5 Comparison of cutting power for different levels of feed rate and cutting speed versus different levels of depth of cut

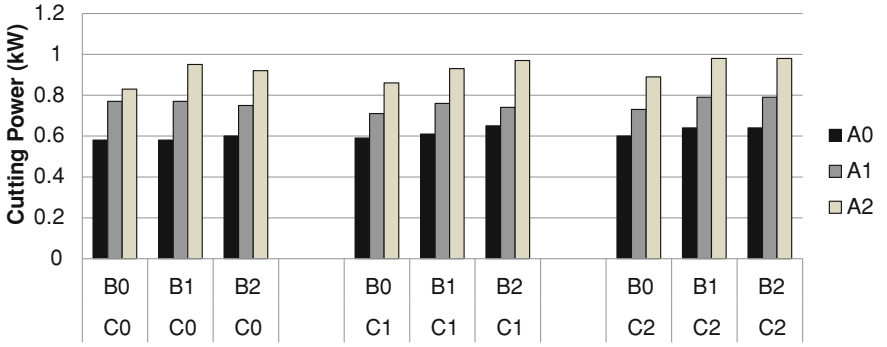


Fig. 6 Comparison of cutting power for different levels of feed rate and depth of cut versus different levels of cutting speed

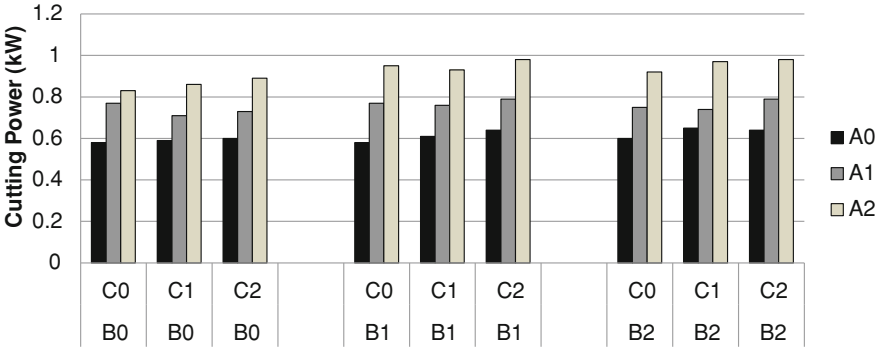


Fig. 7 Comparison of cutting power for different levels of depth of cut and feed rate versus different levels of cutting speed

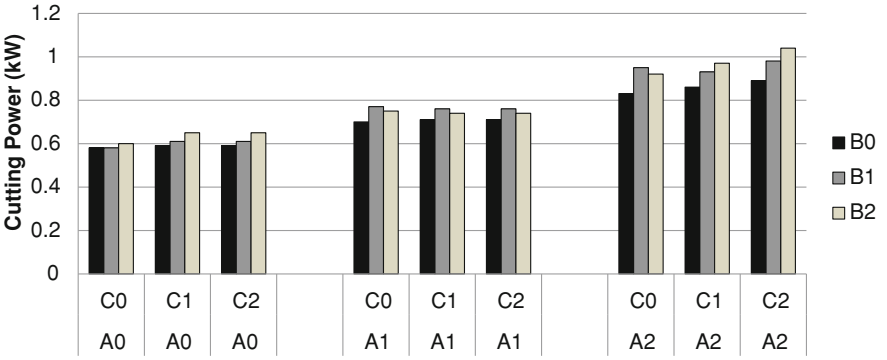


Fig. 8 Comparison of cutting power for different levels of depth of cut and cutting speed versus different levels of feed rate

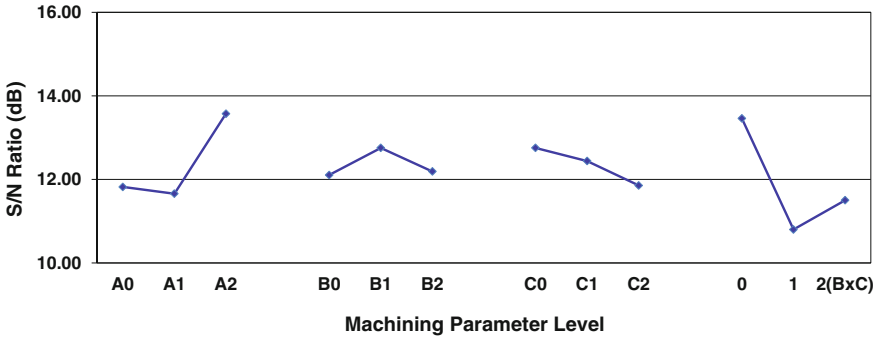


Fig. 9 Response graph for length error

was B1C0. Therefore, the combination to help achieve low length error is A2B1C0; i.e., the highest level of cutting speed, medium level of feed rate and low level of depth of cut.

The Pareto ANOVA for length error given in Table 2 confirms that the parameter that significantly affects the mean length error is cutting speed (with percentage contribution, P = 27.84 %). It is worth noting that the interactions B × C (P = 27.74 %) and B × C (P = 23.74 %) were more than the main effects for factors B (P = 3.07 %) and C (P = 5.13 %).

Figure 10 shows that in term of individual effects, machining parameter A (cutting speed) had the most significant effect on width error, followed by C (depth of cut) and then B (feed rate). This stays true with the individual parameter effects on length error. However, when considering all effects, i.e. individual and interaction effects, the interaction between B × C (feed rate and depth of cut) showed the greatest effect on width error. The medium level of cutting speed, A1 (100 m/min), was the best cutting speed to achieve a low width error. Since the interaction of B × C was also significant, it can be seen that the optimum combination for factors B and C in order to achieve a low width error was B1C2. Thus, the optimal combination to achieve the width error was A1B1C2; i.e., the medium level of cutting speed, medium level of feed rate and highest level of depth of cut.

The Pareto ANOVA for width error given in Table 3 illustrates that the most significant machining parameter affecting the width error was the interaction between the cutting speed and depth of cut (B × C) (P = 31.33 %), followed by cutting speed A (P = 14.49 %) and depth of cut C (P = 12.32 %). Also the total of all interaction effects is higher (P ≈ 66 %) than the total of all individual effects (P ≈ 34).

Figure 11 shows that parameter B (feed rate) had the most significant effect on surface roughness, followed by A (cutting speed) and then C (depth of cut). The interaction between B × C also played a role in this machining process. The medium level of cutting speed, A1 (100 m/min), was the best cutting speed to achieve a low surface roughness. Since the interaction of B × C was also significant, showing that the optimum combination for factors B and C in order to

Table 2 Pareto ANOVA for length error

Sum at factor level	Factor and interaction								
	A	B	A × B	A × B	C	A × C	A × C	B × C	B × C
0	106.37	108.94	110.34	107.33	114.78	111.53	106.91	103.16	121.12
1	104.91	114.78	112.68	110.13	111.96	112.07	111.02	121.71	107.96
2	122.15	109.70	110.40	115.96	106.69	109.82	115.49	108.55	104.34
Sum of squares of difference (S)	548.69	60.60	10.67	116.52	101.16	8.24	110.61	546.67	467.87
Contribution ratio (%)	27.84	3.07	0.54	5.91	5.13	0.42	5.61	27.74	23.74
	A	BxC	BxC	AxB	AxC	C	B	AxB	AxC
Cumulative contribution	27.84	55.57	79.31	85.22	90.83	95.97	99.04	99.58	100.00
Check on significant interaction			B × C two-way table						
Optimum combination of significant factor level			A2B1C0						

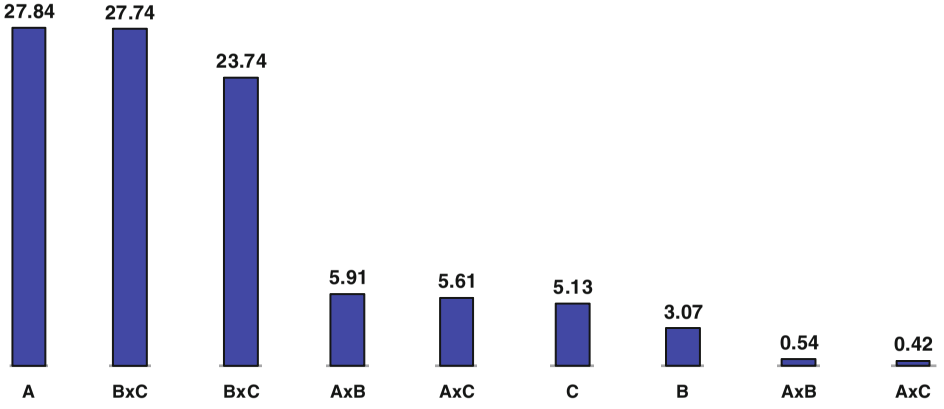


Fig. 10 Response graph for width error

achieve a low surface roughness was B2C0. Therefore, the optimal combination to achieve low surface roughness was A1B2C0; i.e., the medium level of cutting speed, highest level of feed rate and lowest level of depth of cut.

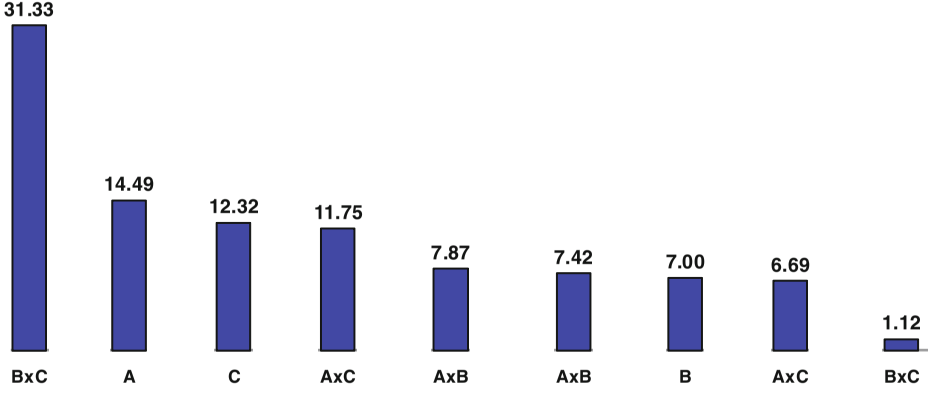
The Pareto ANOVA for surface roughness given in Table 4 confirms that the parameter that significantly affects the mean surface roughness is feed rate (with percentage contribution, P = 77.57 %). All other effects, both individual and interaction, had a minimal effect on surface roughness.

From the analysis of cutting power shown in Figs. 3–8 it can clearly be seen that the feed rate and depth of cut have a minimal independent effect on the cutting power. However, combining the cutting speed and feed rate, or cutting speed and depth of cut increases the required power, showing that the main machining parameter that affects the amount of power required is cutting speed. The mean resultant cutting force in Figs. 12 and 13 shows that generally a lower depth of cut in combination with a low cutting speed and feed rate generates lower resultant cutting forces. However, feed rate changes the cutting force significantly and its dependence is non-linear. Increasing the cutting speed slightly is found to reduce the cutting force. Cutting speeds at low range tend to form a built-up edge, and disappears at high cutting speeds; the dependence on cutting speed diminishes. Depth of cut also changes the cutting force significantly and the dependence is linear. Varying the depth of cut and the feed rate, yields a method of controlling cutting force [1]. Machining with a positive tool orthogonal rake angle will decrease the cutting force but at the same time increase the possibility of destruction of the tool. Dimensional error can be affected by cutting speed in various ways including increasing thermal distortion, altering tool wear, elastic deformation of the work piece and formation of a built-up edge (BUE).

The response graph from S/N ratio analysis for length error (Fig. 9) demonstrates that when cutting speed was increased from A0 to A1, the length error increased, but when cutting speed is further increased from A1 to A2, the length error was decreased significantly. The response graph from S/N ratio analysis for width error (Fig. 10) demonstrates that when cutting speed was increased from A0 to A1, the width error decreased, but when cutting speed is further increased from

Table 3 Pareto ANOVA for width error

Sum at factor level	Factor and interaction								
	A	B	A × B	A × B	C	A × C	A × C	B × C	B × C
0	121.28	127.97	136.69	121.18	133.66	123.68	137.32	134.02	113.14
1	148.22	141.96	117.24	143.16	114.32	120.50	135.22	125.45	156.32
2	121.00	120.57	136.56	126.15	142.51	146.32	117.96	131.03	121.04
Sum of squares of difference (S)	1466.94	708.38	751.51	796.61	1247.24	1189.34	677.32	113.51	3171.13
Contribution ratio (%)	14.49	7.00	7.42	7.87	12.32	11.75	6.69	1.12	31.33
	BxC	A	C	AxC	AxB	AxB	B	AxC	BxC
Cumulative contribution	31.33	45.82	58.14	69.89	77.76	85.16	92.19	98.88	100.00
Check on significant interaction	B × C two-way table								
Optimum combination of significant factor level	A1B1C2								



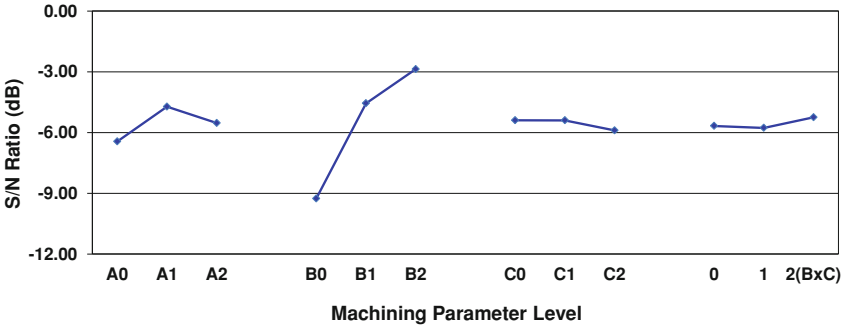


Fig. 11 Response graph for surface finish

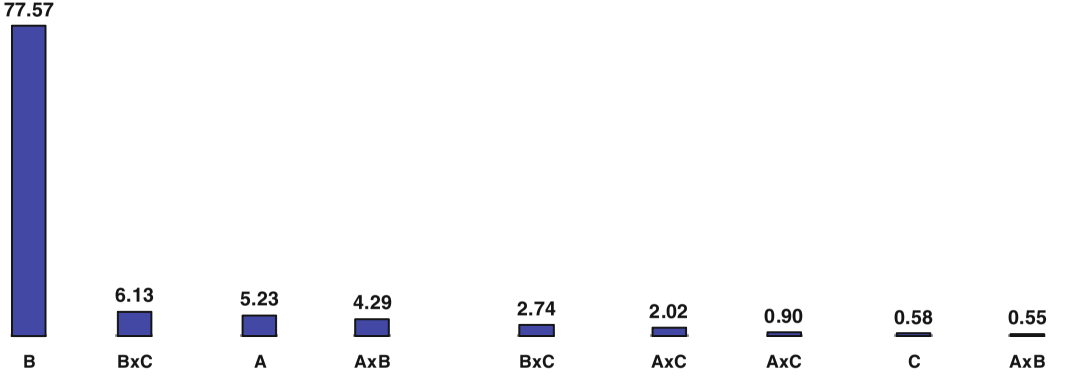
A1 to A2, the width error increased. Figures 9 and 10 show that a similar trend lies for the feed rate where, as the feed rate is increased from B0 to B1, the length and width errors decreases but when it is increased from B1 to B2 the length and width errors increase. Finally, when observing the depth of cut it can be seen that both the length and width errors increase from C0 to C1. When the depth of cut is increased from C1 to C2 however, the length error increases while the width error decreases.

The response graph from S/N ratio analysis for surface roughness shows that the cutting speed has varying effect on surface roughness (Fig. 11). When the cutting speed was changed from level 0 to level 1, the quality of the surface improved; whereas when speed was changed from level 1 to level 2, the quality of the surface deteriorated. When the depth of cut increased from level 0 to level 1, the surface roughness stayed fairly constant, but when the depth of cut was increased from level 1 to level 2, the surface roughness increased. These effects of cutting speed and depth of cut stay true to the results obtained by Rafai and Islam [8]. This may have been caused by the plastic deformation of the machined surface from the material softening, especially at high temperature due to dry machining [11]. Traditionally as the feed rate changes the cutting force alters in a nonlinear manner, whereas increasing the cutting speed only slightly reduces the cutting force, and for a low range of cutting speed there is a tendency to form a built-up-edge which disappears at higher speeds.

This research showed that as the feed rate increased, the surface roughness improved each time (Fig. 14) which is differs from traditional wisdom. Normally, the feed rate has a significant effect on surface finish and cutting force. However, here it is thought that the properties of the MMC are contributing to the improved surface finish at higher metal removal rates. The influence of the machining parameters on the surface roughness of MMCs is known to be complex compared to that of traditional steels and aluminium alloys. As it is believed that the size and amount of reinforcement phase of MMC is known to influence surface roughness of the cutting process [12].

Table 4 Pareto ANOVA for surface roughness

Sum at factor level	Factor and interaction								
	A	B	A × B	A × B	C	A × C	A × C	B × C	B × C
0	−57.94	−83.36	−57.96	−51.05	−48.54	−50.23	−53.72	−59.44	−56.52
1	−42.46	−41.01	−44.58	−51.91	−48.58	−45.16	−48.67	−47.44	−47.12
2	−49.76	−25.79	−47.63	−47.21	−53.04	−54.77	−47.78	−43.29	−46.53
Sum of squares of difference (S)	359.87	5339.06	295.00	37.61	40.18	138.74	61.75	421.94	188.45
Contribution ratio (%)	5.23	77.57	4.29	0.55	0.58	2.02	0.90	6.13	2.74
	B	BxC	A	AxB	BxC	AxC	AxC	C	AxB
Cumulative contribution	77.57	83.70	88.93	93.22	95.96	97.97	98.87	99.45	100.00
Check on significant interaction			B × C two-way table						
Optimum combination of significant factor level			A1B2C0						



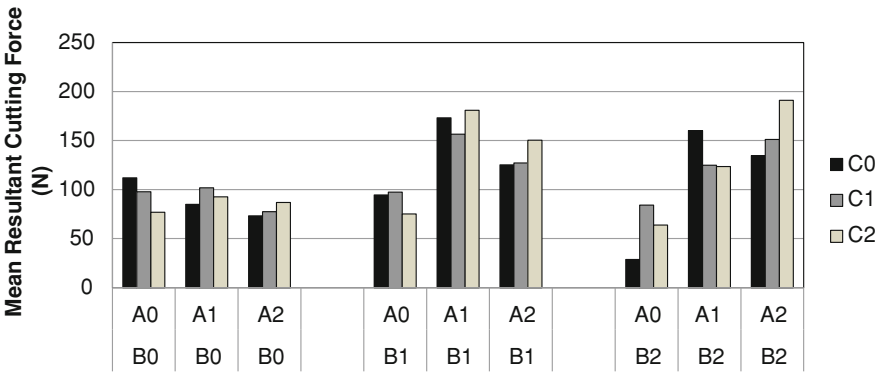


Fig. 12 Comparison of mean resultant cutting force for different levels of cutting speed and feed rate versus different levels of depth of cut

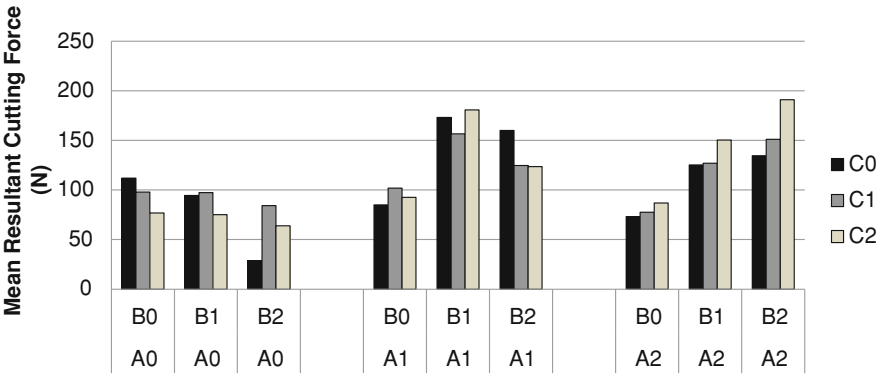


Fig. 13 Comparison of Cutting mean resultant cutting force for different levels of feed rate and cutting speed versus different levels of depth of cut

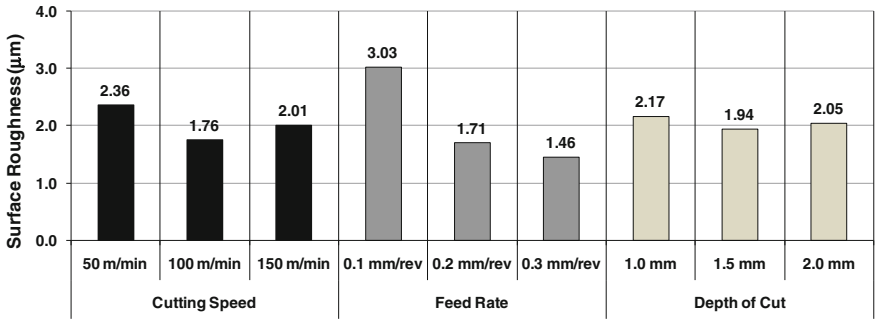


Fig. 14 Average variation of surface roughness for input parameters

4 Conclusion and Future Work

The investigation presented above demonstrate that end milling metal matrix composite comprised of aluminium alloy reinforced with 20 vol % of boron carbide particles in the approximate size range of $1 \sim 4$ μm is sustainable when the correct machining parameters are used. The cutting parameters, such as cutting speed, feed rate and depth of cut, have shown to have significant influences on the, dimensional errors surface roughness and economic aspect. However, the one machining parameter that seems to affect all three measures is the feed rate; a fast cut minimises dimensional error and produces a better surface finish and give a better material removal rate. Show that the optimum options for this material to be one of high cutting speed with high feed rate, to give sustainable and economic machining. The fact that the surface finish improves with an increase of feed rate is of interest, as this is the opposite found for traditional material. The up-cut milling is also deemed to have played an important element in this finding of improved surface finish. Further examination is necessary to help explain why the feed rate had such a major effect on the surface finish.

References

1. S. Kalpakjian, S.R. Schmid, *Manufacturing engineering and technology* (Prentice Hall, Singapore, 2010)
2. T.J. Drozda, C. Wick, J.T. Benedict, R.F. Veilleux, R. Bakerjian, P.J. Mitchell, *Tool and manufacturing engineers handbook : a reference book for manufacturing engineers, managers, and technicians*, 4th edn. (SME Publications, Michigan, 1983)
3. T. Gutowski, C. Murphy, D. Allen, D. Bauer, B. Bras, T. Piwonka, P. Sheng, J. Sutherland, D. Thurston, E. Wolff, Environmentally benign manufacturing: Observations from Japan, Europe and United States. J. Cleaner Prod **13**(1), 1–17 (2005)
4. A.N.I. 14040, *Environmental Management—Life Cycle Assessment—Principles and Framework* (Standards Australia and Standards New Zealand, The Crescent, Homebush, 1998)
5. P.S. Sreejith, B.K.A. Ngoi, Dry machining: Machining of the future. J. Mater Process. Technol. **101**(1), 287–291 (2000)
6. R. Roy, *A Primer on the Taguchi Method,* (Society of Manufacturing Engineers, USA 1990)
7. S. Tanaydin, Robust design and analysis for quality engineering. Technometrics **40**(4), 348 (1998)
8. N.H. Rafai, M.N. Islam, An investigation into dimensional accuracy and surface finish achievable in dry turning. Mach. Sci. Technol. **13**(4), 571–589 (2009)
9. D. Haughey, Pareto Analysis Step by Step [Online]. Available: http://www.projectsmart.co.uk/pareto-analysis-step-by-step.html (2010)
10. B. Boswell, M.N. Islam, A. *Pramanik, Sustainable machining of aerospace material*. In: Proceedings of the World Congress on Engineering 2013 (WCE 2013), Lecture Notes in Engineering and Computer Science. 3–5 July, 2013, (pp. 1869–1876). London (2013)
11. R. Azouzi, M. Guillot, On-line prediction of surface finish and dimensional deviation in turning using neural network based sensor fusion. Int. J. Mach. Tools Manuf. **37**(9), 1201–1217 (1997)
12. U.A. Dabade, H.A. Sonawane, S.S. Joshi, Cutting forces and surface roughness in machining Al/SiCp composites of varying composition. Mach. Sci. Technol. **14**(2), 258–279 (2010)

Base Isolation Testing Via a Versatile Machine Characterized by Robust Tracking

Salvatore Strano and Mario Terzo

Abstract A non-linear robust control of a multi-purpose earthquake simulator has been designed and experimentally tested. The test rig is characterized by a double functionality based on two configurations of the hydraulic actuation system. Due to the several operating conditions, the system is affected by structured and unstructured uncertainties that require a robust approach for the control of the position. Starting from a non-linear dynamic model, a sliding control is developed taking into account the incomplete knowledge of the system. The experimental results highlight the goodness of the proposed control in terms of stability and tracking error.

Keywords Earthquake simulator · Hydraulic actuator · Robust control · Seismic isolator · Shaking table · Vibration control

1 Introduction

This paper concerns a new hydraulically actuated multi-purpose earthquake simulator finalized to execute both shaking table test and seismic isolator characterization. The versatile earthquake simulator is essentially constituted by a hydraulically actuated shaking table and a suitable reaction structure.

Aside from the non-linear nature of the dynamics [1], the hydraulic systems also have a large extent of model uncertainties. The uncertainties can be classified

S. Strano · M. Terzo (✉)
Dipartimento di Ingegneria Industriale, Università degli Studi di Napoli Federico II,
via Claudio 21, 80125 Naples, Italy
e-mail: m.terzo@unina.it

S. Strano
e-mail: salvatore.strano@unina.it

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_5,

into two categories: structured (parametric uncertainties) and unstructured. These model uncertainties can lead to an unstable behaviour of the controlled system or a very degraded performance.

The earthquake simulator is characterized by two different hydraulic schemes that can be selected in dependence of the test that has to be executed. This determines an induced structured uncertainty. Moreover, the seismic isolator and the test structure can largely influence the controlled system performances due to very large restoring force or neglected dynamics (e.g. structural modes) that cause an induced unstructured uncertainty.

In order to meet this requirement, the basic idea is the choice of the non-linear hydraulic actuator model, with friction and dead band, as nominal one and characterized by external disturbances caused by seismic isolator or test structures.

2 The Earthquake Simulator

The earthquake simulator consists of movable and fixed parts made in structural steel. Particularly, it is constituted by:

- fixed base;
- hydraulically actuated sliding table with dimensions 1.8 m $\times$ 1.6 m;
- hydraulic actuator.

The table motion is constrained to a single horizontal axis by means of recirculating ball-bearing linear guides.

The hydraulic power unit consists of a variable displacement pump powered by a 75 kW AC electric motor and able to generate a maximum pressure of 210 bar and a maximum flow rate equal to 313 l/min. A pressure relief valve is located downstream of the pump.

The hydraulic circuit consists of a four way-three position proportional valve and a hydraulic cylinder. The cylinder is constituted by two equal parts separated by a diaphragm and contains two pistons which rods are connected to the base; so, the actuator is characterized by a mobile barrel and fixed pistons. The maximum horizontal force is 190 kN, the maximum speed 2.2 m/s and the maximum stroke 0.4 m ($\pm$0.2 m). Three way valves allow to select different configurations of the test rig activating different thrust area (active surface) and control volume. The selection of the maximum thrust configuration and the installation of suitable reaction structures allow the bench to be employed as seismic isolator test rig; conversely, the removal of the reaction structures, together with the selection of the maximum speed configuration, allows the earthquake simulator to be used as shaking table.

3 Nominal Model Derivation

The modelling refers to the testing machine in which no isolator or test structure is installed: the hydraulic cylinder has to move the sliding table only.

The modelling procedure is based on the following hypothesis: (a) fluid properties not depending on the temperature; (b) equal piston areas; (c) equal chamber volume for each side in the case of barrel in the centred position; (d) negligible internal and external fluid leakages.

The test rig is modelled as a single DOF system subjected to both actuation and friction force and can be considered equivalent to a double-ended hydraulic actuator, driven by a four-way spool valve, and with trust area and control volume depending on the selected configuration [2, 3].

The dynamics of the sliding table can be described by:

$$m\ddot{y} + \sigma\dot{y} + F_c \mathrm{sgn}(\dot{y}) + \mu N \mathrm{sgn}(\dot{y}) = A_p P_L \tag{1}$$

where

y	is the table displacement;
m	is the movable mass;
$P_L = P_A - P_B$	is the load pressure;
P_A and P_B	the pressures in the two chambers;
σ	is the viscous friction coefficient;
F_c	is the Coulomb friction force in the hydraulic actuator;
μ	is the Coulombian friction coefficient of the linear guides;
N	is the vertical load on the linear guides.

The actuator dynamics can be written as [1]:

$$\frac{V_0}{2\beta}\dot{P}_L = -A_p\dot{y} + Q_L \tag{2}$$

where

V_0	is the volume of each chamber for the centred position of the barrel;
β	the effective bulk modulus;
A_p	is the ram area;
$Q_L = (Q_A + Q_B)/2$	is the load flow;
Q_A and Q_B	are, respectively, the supplied flow rate and the return flow rate of the proportional valve.

An overlapped four-way valve is considered: this kind of valve is typically characterized by the lands of the spool greater than the annular parts of the valve body. Consequently, the flow rate is zero (dead band) when the spool is in the neighbourhood of its central position. Moreover, since the adopted valve is characterized by a high response, it is assumed that the control applied to the spool

valve is directly proportional to the spool position. Under the assumption of a tank pressure equal to zero, the load flow depends on the supply pressure, the load pressure and the input voltage in accordance with the following:

$$Q_L = DZ(u)\sqrt{P_s - \tilde{v}_e P_L} \tag{3}$$

where u is the input voltage, DZ(u) is the dead band function and $\tilde{v}_e$ defined as

$$\tilde{v}_e = \begin{cases} 1 & if \quad u > 0 \\ 0 & if \quad u = 0 \\ -1 & if \quad u < 0 \end{cases} \tag{4}$$

Without loss of generality, the slope of the static map beyond the dead band region can be assumed equal and the analytical expression of DZ(u) is:

$$DZ(u) = \begin{cases} K_Q(u - u^+) & if \quad u > u^+ \\ 0 & if \quad u^- \le u \le u^+ \\ K_Q(u - u^-) & if \quad u < u^- \end{cases} \tag{5}$$

where u^+ and u^- are the limits of the dead band and K_Q the slope.

Defined the state vector as $x = [\dot{y}\ y\ P_L]^T$, the system (sliding table + hydraulic actuator) is given by the following third order model non-linear in the state:

$$\begin{cases} m\ddot{y} = -(\mu N + F_C)\mathrm{sgn}(\dot{y}) - \sigma\dot{y} + A_p P_L \\ \dfrac{V_0}{2\beta}\dot{P}_L = -A_p\dot{y} + DZ(u)\sqrt{P_s - \tilde{v}_e P_L}. \end{cases} \tag{6}$$

4 Sliding Mode Control

The nominal model (6) has been adopted for the sliding control design.

The discontinuous nonlinearities in the friction force can be smoothly approximated taking into account that:

$$\mathrm{sgn}(\dot{y}) = \frac{2}{\pi}\arctan(\gamma\dot{y}) \tag{7}$$

where γ is the approximation parameter.

Differentiating the first equation of (6) and taking into account that the dead band can be expressed as

$$DZ(u) = K_Q u + S(u) \tag{8}$$

with

$$S(u) = \begin{cases} -K_Q u^+ & if \quad u > u^+ \\ -K_Q u & if \quad u^- \le u \le u^+ \\ -K_Q u^- & if \quad u < u^- \end{cases} \tag{9}$$

the following single expression is obtained for the nominal plant:

$$\dddot{y} = -\left(\frac{2\gamma}{\pi}\frac{\mu N + F_C}{m(1+\gamma^2\dot{y}^2)} + \frac{\sigma}{m}\right)\ddot{y} - \frac{2A_P^2\beta}{mV_0}\dot{y} + \frac{2A_P\beta K_Q}{mV_0}\sqrt{P_S - \tilde{v}_e P_L}u + \frac{2A_P\beta}{mV_0}\sqrt{P_S - \tilde{v}_e P_L}S(u) \quad (10)$$

At this step the plant model can be synthetically expressed as:

$$\dddot{y} = -\alpha_1\ddot{y} - \alpha_2\dot{y} + \alpha_3 u + \frac{\alpha_3}{K_Q}S(u) \quad (11)$$

where

$$\alpha_1 = \frac{2\gamma}{\pi}\frac{\mu N + F_C}{m(1+\gamma^2\dot{y}^2)} + \frac{\sigma}{m} \quad (12)$$

$$\alpha_2 = \frac{2A_P^2\beta}{mV_0} \quad (13)$$

$$\alpha_3 = \frac{2A_P\beta K_Q}{mV_0}\sqrt{P_S - \tilde{v}_e P_L}. \quad (14)$$

So the single-input dynamic system (11) is not exactly known but affected by uncertainties.

Taking into account the realistic hydraulic system together with the practical seismic isolator and test structures, the following realistic assumption is made.

Assumption. The modelling uncertainties (intrinsic and induced) in (11) are all bounded.

Given the table target displacement y_T, the objective is to design a bounded control input u so that the current table displacement y tracks as closely as possible the desired motion in spite of various model uncertainties, including parametric uncertainties and neglected dynamics due to both physical changing in the plant configuration and induced disturbances (i.e. seismic isolator or test structure).

The sliding mode control design procedure starts from the definition of a suitable sliding surface. With the intention of keeping stability conditions and improving closed loop system performance, the following sliding surface is defined [4]:

$$s = \left(\frac{d}{dt} + \lambda\right)^2 e = \ddot{e} + \lambda^2 e + 2\lambda\dot{e} \quad (15)$$

being λ a strictly positive constant and $e = y - y_T$ the tracking error.

The dynamics in sliding mode can be written as:

$$\dot{s} = \dddot{e} + \lambda^2\dot{e} + 2\lambda\ddot{e} = \dddot{y} - \dddot{y}_T + \lambda^2\dot{e} + 2\lambda\ddot{e} = 0 \quad (16)$$

Once the sliding surface is reached, the motion should continue on this surface with the application of the equivalent control law u_{eq} that is determined solving formally Eq. (16) referred to the nominal system

$$u_{eq} = \frac{\dddot{y}_T + \hat{\alpha}_1 \ddot{y} + \hat{\alpha}_2 \dot{y} - \lambda^2 \dot{e} - 2\lambda \ddot{e}}{\hat{\alpha}_3} - \frac{S(u_{ND})}{\hat{K}_Q} \tag{17}$$

where the superscript $\hat{\bullet}$ refers to the nominal parameter and $S(u_{ND})$ is the S function (9) evaluated for the equivalent control action with no dead band u_{ND} given by:

$$u_{ND} = \frac{\dddot{y}_T + \hat{\alpha}_1 \ddot{y} + \hat{\alpha}_2 \dot{y} - \lambda^2 \dot{e} - 2\lambda \ddot{e}}{\hat{\alpha}_3}. \tag{18}$$

In reality, system uncertainties make the state trajectories to oscillate in the neighbourhood of the ideal sliding mode and consequently and additional robust term u_r has to be considered for the control action in order to ensure the attractivity of the sliding surface:

$$u_r = -k(\mathbf{x})\mathrm{sgn}(s). \tag{19}$$

So, the robust control action is characterized by a term discontinuous across the sliding surface.

Taking into account a Lyapunov based design approach, the following Lyapunov function is selected:

$$V(s) = \frac{1}{2} s^2 \tag{20}$$

which is a measure of the squared distance to the sliding surface and generates the following sliding condition:

$$s\dot{s} < 0. \tag{21}$$

In order to guarantee that the system trajectories reach the sliding surface in a finite time, the sliding condition is modified to:

$$s\dot{s} < -\eta |s| \tag{22}$$

in which η is a strictly positive constant.

The sliding condition constraints the system subjected to the following control action

$$u = u_{eq} - k(\mathbf{x})\mathrm{sgn}(s) \tag{23}$$

to point towards the sliding surface in spite of uncertainties.

The robust control gain $k(\mathbf{x})$ can be derived taking into account the sliding condition (22):

$$k(\mathbf{x}) = |\alpha - 1|\left|\dddot{y}_T - \lambda^2\dot{e} - 2\lambda\ddot{e} + \hat{\alpha}_1\ddot{y} + \hat{\alpha}_2\dot{y}\right| + \alpha(F + \eta + \rho\alpha_{3_{MAX}}) \tag{24}$$

where F is the additive error bound given as $|(\hat{\alpha}_1 - \alpha_1)\ddot{y} + (\hat{\alpha}_2 - \alpha_2)\dot{y}| \leq F$, α the multiplicative error bound given as $\sqrt{\alpha_{3_{MIN}}^{-1}\alpha_{3_{MAX}}}$ and ρ the dead band error bound $\left(\left|S(u)/K_Q\right| \leq \rho\right)$. The subscript MIN, MAX indicate the bounds of the uncertainty ranges.

In order to counteract the chattering phenomenon caused by the discontinuous nature of the robust control action (19), a boundary layer is introduced around the sliding surface and the robust control action can be modified to:

$$u_r = -k(\mathbf{x})sat\left(\frac{s}{\phi}\right) \tag{25}$$

where ϕ represents the width of the boundary layer and sat is the saturation function defined as:

$$\begin{cases} sat\left(\frac{s}{\phi}\right) = \frac{s}{\phi} & if \ \left|\frac{s}{\phi}\right| \leq 1 \\ sat\left(\frac{s}{\phi}\right) = \mathrm{sgn}\left(\frac{s}{\phi}\right) & otherwise \end{cases}. \tag{26}$$

5 Simulation and Experimental Results

A variability is considered for all the parameters of the hydraulic actuator, including the geometric parameters and the supply pressure. The nominal values are set as:

$$\hat{m} = 540 \text{ kg}, \ \hat{\sigma} = 22,500 \ \frac{\text{Ns}}{\text{m}}, \ \hat{F}_C = 950 \text{ N}, \ \hat{\mu} = 0.03,$$
$$\hat{N} = 67,000 \text{ N}, \ \hat{A}_p = 0.0055 \text{ m}^2, \ \hat{V}_0 = 0.0035 \text{ m}^3, \ \hat{\beta} = 0.75e9 \text{ Pa},$$
$$\hat{K}_Q = 2.1082e-7 \ \frac{\text{m}^3}{\text{sVPa}^{\frac{1}{2}}}, \ \hat{u}^+ = 0.65 \text{ V}, \ \hat{u}^- = -0.65 \text{ V},$$
$$\hat{P}_S = 110e5 \text{ Pa}.$$

These nominal values contribute to determine the $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ values.

To demonstrate the robustness of the proposed controller, the following large bounds of the uncertain parameters are considered. They refer to the same measurement units previously adopted.

$$
\begin{aligned}
[m_{MIN}, m_{MAX}] &= [440, 640],\ [\sigma_{MIN}, \sigma_{MAX}] = [20,000, 25,000],\\
[F_{CMIN}, F_{CMAX}] &= [900, 1,000],\ [\mu_{MIN}, \mu_{MAX}] = [0.01, 0.05],\\
[N_{MIN}, N_{MAX}] &= [4, 316, 130,000],\ [A_{PMIN}, A_{PMAX}] = [0.003, 0.009],\\
[V_{0MIN}, V_{0MAX}] &= [0.003, 0.004],\ [\beta_{MIN}, \beta_{MAX}] = [5e8, 1e9],\\
[K_{QMIN}, K_{QMAX}] &= [1.5811e-7, 2.6352e-7],\ \left[u^{+}_{MIN}, u^{+}_{MAX}\right] = [0.3, 1],\\
\left[u^{-}_{MIN}, u^{-}_{MAX}\right] &= [-1, -0.3],\ [P_{sMIN}, P_{sMAX}] = [20e5, 200e5].
\end{aligned}
$$

Taking into account the above ranges, the bounds of α_1, α_2, α_3 are determined.

In the following, simulation and experimental results will be described for both the test rig configurations.

Simulations have been carried out in order to evaluate the preliminary results in terms of stability and robustness. To this aim, a seismic isolator and a test structure have been modelled to simulate the external disturbances. A noisy signal (white noise with upper and lower magnitude bounds ± 0.0001) has been taken into account in order to employ a realistic feedback affected by measurement noise and to further test the robustness performance of the proposed controller. The width of the boundary layer has been selected as $\phi = 1$.

As regards the seismic isolator configuration, a vibration absorber has been modelled taking into account both the viscous and the elastic forces. Consequently, the isolator equipped system consists of:

$$
\begin{cases}
m\ddot{y} = -F_f + \sigma_{is}\dot{y} + k_{is}y + A_p P_L \\
\frac{V_0}{2\beta}\dot{P}_L = -A_p\dot{y} + DZ(u)\sqrt{P_s - \tilde{v}_e P_L}
\end{cases}
\tag{27}
$$

being σ_{is} the damping coefficient and k_{is} the elastic one, assumed respectively equal to 1e5 Ns/m and to 1e6 N/m. A vertical load on the specimen of 1.25×105 N and a supply pressure Ps of 100 bar have been imposed. Moreover, the maximum values for both the A_p and V_0 parameters and the nominal values for all the other ones have been adopted.

Figure 1 represents the effective displacement (with and without the isolator) for a test characterized by a sinusoidal target of amplitude 0.1 m and a frequency of 0.5 Hz.

The result allows to appreciate the stability and the robustness properties in presence of induced uncertainties. The comparison between the table displacement with and without the specimen highlights the disturbance rejection of the controlled system.

Figure 2 illustrates the control input obtained with and without the isolator under test.

As regards the simulations concerning the shaking table configuration, a single DOF vibrating structure has been supposed in coupling with the hydraulically actuated sliding table in order to test the system.

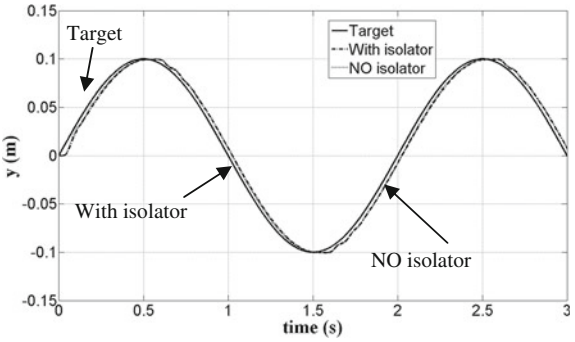


Fig. 1 Table displacement in the seismic isolator configuration (simulation results)

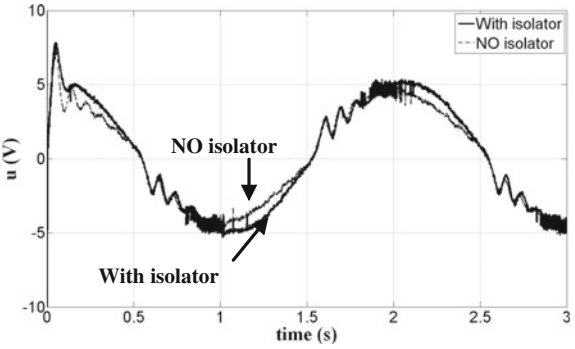


Fig. 2 Control action in the seismic isolator configuration (simulation results)

The equations of the system to be controlled are given by:

$$\begin{cases} m\ddot{y} = -F_f + A_p P_L + \sigma_s(\dot{y} - \dot{y}_s) + k_s(y - y_s) \\ \frac{V_0}{2\beta}\dot{P}_L = -A_p\dot{y} + DZ(u)\sqrt{P_s - \tilde{v}_e P_L} \\ m_s \ddot{y}_s = -\sigma_s(\dot{y}_s - \dot{y}) - k_s(y_s - y) \end{cases} \tag{28}$$

where the third Eq. (28) is derived taking into account the dynamic equilibrium of the forces acting on the m_s mass, which relative displacement to ground is y_s. σ_s and k_s are the damping and the elastic coefficients of the modelled structure and have been assumed equal to 2.4e3 Ns/m and 11e3 N/m respectively. The vibrating mass has been selected as $m_s = 200$ kg, giving a natural frequency of about 1.2 Hz. A supply pressure Ps of 30 bar has been employed, the minimum values for both the A_p and V_0 parameters and the nominal values for all the other ones have been considered.

Figure 3 illustrates the result, in terms of table displacement, obtained for a target signal given by a sinusoidal law of amplitude 0.01 m and a frequency of 1.2 Hz. This kind of test has been selected in order to emphasize the perturbation to the system, carrying the coupled vibrating structure around the resonance.

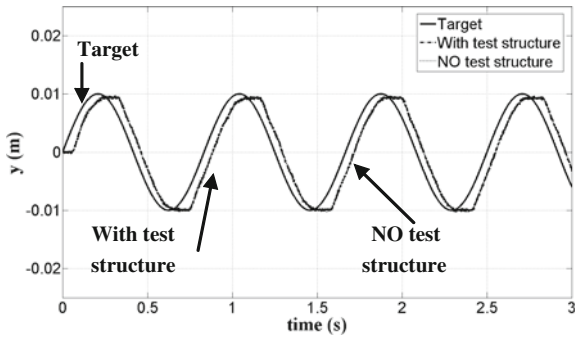


Fig. 3 Table displacement in the shaking table configuration (simulation results)

The controlled system exhibits stability and insensitiveness properties respect to the unstructured uncertainty represented by the additional dynamics of the vibrating structure.

The control action (Fig. 4) is slightly influenced by the test structure and is fully contained in its bounds (±10 V).

The designed control has been implemented in a DS1103 controller board equipped with a 16-bit A/D and D/A converter. A magnetostrictive position sensor is adopted to provide the table displacement and two strain gauge pressure sensors are located along the supply and the return pipeline of the proportional valve. In this way, the necessary feedback are given to the controller. The width of the boundary layer has been selected as $\phi = 5$.

With reference to the seismic isolator configuration [2, 3], a common elastomeric isolation bearing (Fig. 5) has been adopted to test the sliding control.

The isolator has been vertically loaded (1.25 × 105 N) by means of a hydraulic jack and subjected to the horizontal actuation force. A supply pressure Ps of 100 bar has been imposed and a target displacement of amplitude 0.1 m and a frequency of 0.5 Hz has been adopted.

The result in terms of table displacement is showed in Fig. 6. The performance of the controlled system is fully insensitive respect to the external disturbance due to the seismic isolator. Indeed, the table displacement in presence of the isolator and the same one with no isolator are practically superimposed. Moreover, the performance is fully appreciable for both tracking error and stability.

Figure 7 illustrates the control action with and without the seismic isolator. It has to be highlighted that the signal is not affected by the undesired chattering phenomenon.

The sliding mode control has been then experimentally tested on the seismic test rig in the shaking table configuration. To this aim, a suspended structure has been fixed on the sliding table (Fig. 8).

The structure consists of a rigid cabin (200 kg) equipped with suspensions, and is characterized by a 1.2 Hz resonant mode. As carried out in the simulation environment, a sinusoidal target displacement has been assigned (amplitude 0.01 m and frequency 1.2 Hz) together with a supply pressure Ps of 30 bar. The selection

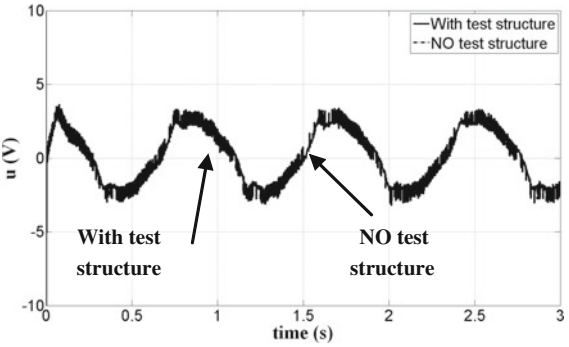


Fig. 4 Control action in the shaking table configuration (simulation results)

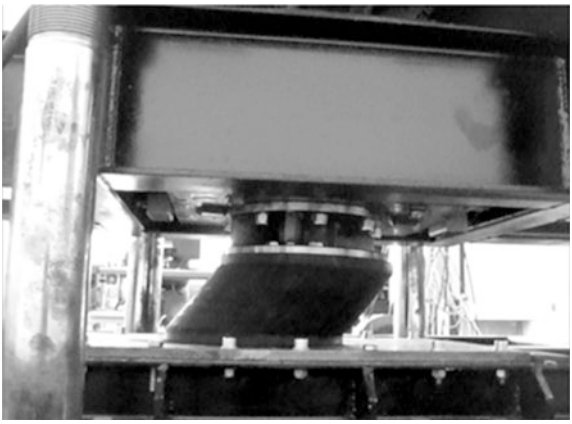


Fig. 5 Detail of the seismic isolator

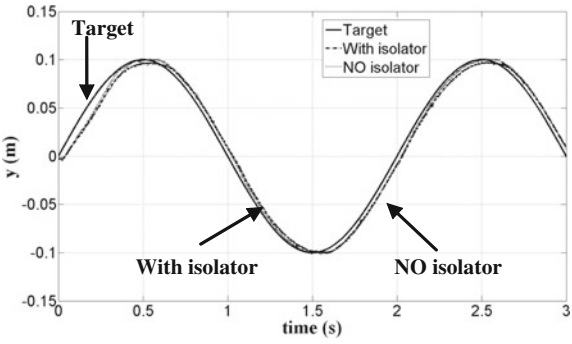


Fig. 6 Table displacement in the seismic isolator configuration (experimental results)

of the target frequency allows to evaluate the goodness of the proposed approach under the condition in which the unmodeled dynamics of the suspended structure highly perturbs the controlled system.

The controlled system is stable and shows (Fig. 9) a substantial robust performance, as can be observed focusing on the results obtained with and without

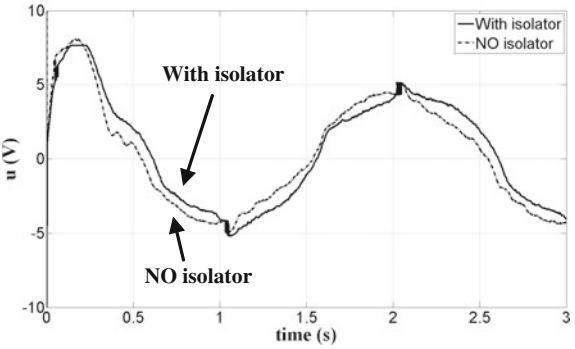


Fig. 7 Control action in the seismic isolator configuration (experimental results)

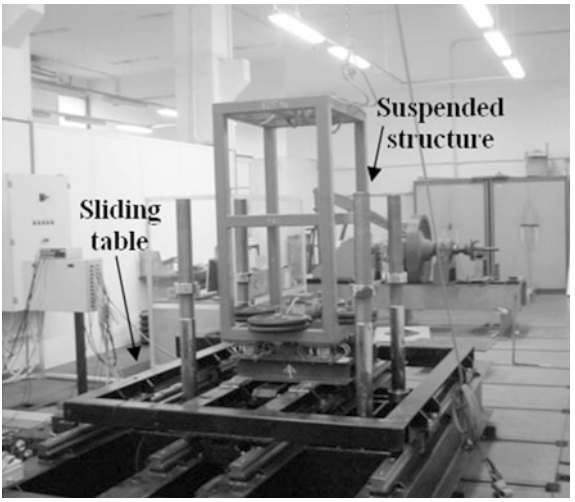


Fig. 8 The seismic test rig in the shaking table configuration

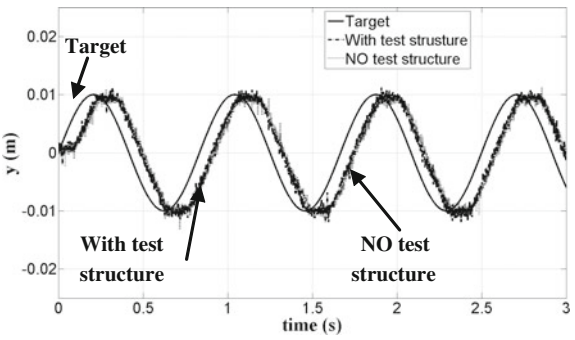


Fig. 9 Table displacement in the shaking table configuration (experimental results)

the test structure. In this test, the test structure is characterized by an acceleration increased of one order of magnitude respect to the table acceleration, and the excited additional dynamics doesn't cause effects on the controlled displacement.

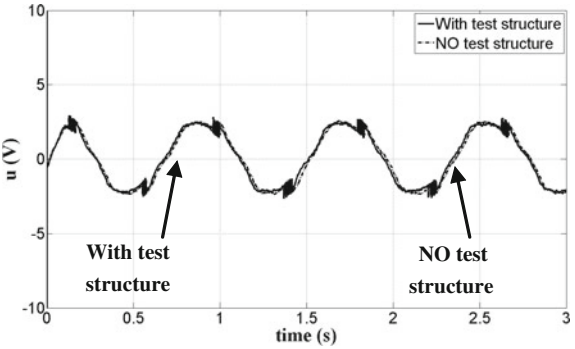


Fig. 10 Control action in the shaking table configuration (experimental results)

The target displacement amplitude is guaranteed, confirming the attitude to be employed as shaking table.

A contained influence of the test structure can be seen in the control action (Fig. 10) that is not contaminated by chattering.

6 Conclusion

A robust control has been designed for a multi-purpose earthquake simulator. A sliding mode approach has been followed starting from a third order non-linear dynamic model in presence of dead-band. The experimental results highlight the effectiveness in terms of stability, tracking error and robustness of the controlled earthquake simulator for both the configurations.

References

1. H.E. Merritt, *Hydraulic Control Systems* (Wiley, New York, 1967)
2. S. Strano, M. Terzo, A non-linear robust control of a multi-purpose earthquake simulator, *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013*, London, vol. 3, 3–5 July 2013, pp. 1687–1692
3. S. Strano, M. Terzo, A multi-purpose seismic test rig control via a sliding mode approach. Struct. Control Health Monit. (in press). doi: 10.1002/stc.1641
4. J.J.E. Slotine, W. Li, *Applied Nonlinear Control* (Prentice Hall, NJ, 1991)

Active Vibration Isolation Via Nonlinear Velocity Time-Delayed Feedback

Xue Gao and Qian Chen

Abstract This paper combines cubic nonlinearity and time delay to improve the performance of vibration isolation. By the multi-scale perturbation method, the average autonomous equations are first found to analyse local stability. Then with the purpose of obtaining the desirable vibration isolation performance, stability conditions are obtained to find appropriate the feedback parameters including gain and time delay. Last, the influence of the feedback parameters on vibration transmissibility is assessed. Results show that the strategy developed in this paper is practicable and feedback parameters are significant factors to alter dynamics behaviours, and more importantly, to improve the isolation effectiveness for the bilinear isolation system.

Keywords Active control · Cubic velocity feedback · Local stability · Piecewise bilinear · Stability boundary · Time delay · Vibration isolation

1 Introduction

Vibration isolation systems can be divided into three groups: passive, active and semi-active according to external energy requirement. The use of passive isolators is the most common method of controlling undesired vibrations in various engineering sectors such as aerospace engineering, transportation systems, marine engineering, civil engineering etc. [1–3]. The linear viscous damping is often

X. Gao · Q. Chen (✉)
State Key Laboratory of Mechanics and Control of Mechanical Structures, Nanjing University of Aeronautics and Astronautics, No. 29 Yudao Street, Nanjing 210016, China
e-mail: q.chen@nuaa.edu.cn

X. Gao
e-mail: xgao.detec@nuaa.edu.cn

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_6,

introduced to reduce vibration amplitude at resonance for such a vibration isolation device. Unfortunately, the transmissibility increases with the damping in the frequency region where isolation is required. This is a dilemma that the passive vibration isolation technique faces [4]. But it could be solved by the active or semi-active control method such as direct linear velocity feedback strategy, which is recognized as a simple and robust method. The feedback controller generates an additional force which is proportional to velocity of the equipment, and thus it is sometimes referred to a skyhook damping for it reacts off the structure at required frequencies [5]. The outstanding virtue of on–off skyhook damping is that the resonance peak is reduced without affecting the vibration transmission at higher frequencies [6]. On the other hand, conventional skyhook unfortunately introduces a sharp increase (jump or jerk often called in papers) in damping force, which, in turn, causes a jump in sprung-mass acceleration [7–9]. In order to figure out the dilemma of the design of passive linear damped vibration isolation and eliminate the acceleration jump induced by the sudden change of damping force, the present paper proposes an active controller of cubic velocity time-delayed feedback. By comparison with a passive device, the active controller is a practical approach to provide an exact cubic damping force as demanded. Besides, the feedback gain of the control strategy is fixed, not displacement- or velocity-dependent and consequently dynamics jerk induced by the sudden change of damping force could be avoided.

But in real active control system, one of open problems is the complicated system dynamics induced by the unavoidable time delay in controllers and actuators, especially in various analogue filters. The downsides of the time delay on the stability and performance of a dynamics system has drawn a great deal of attention from researchers in structural dynamics engineering [10–14]. However, if designed properly, the time delay existed in controller could suppress bifurcations and improve vibration control [15, 16]. This paper is to explore the dynamics of a bilinear vibration control system with cubic velocity time delay feedback and propose a proper design methodology for the controller.

2 Multi-scale Analysis

Figure 1 shows a single degree-of-freedom vibration system with an active vibration controller. The passive isolator including nonlinear stiffness and linear damping are abstracted from the Solid And Liquid Mixture (SALiM) vibration isolator [17–19]. And the passive isolator has been studied in [18].

The equation of motion of the controlled vibration isolation system with the time-delayed cubic velocity feedback can be represented by

$$\underbrace{M\ddot{X}(t) + C_1\dot{X}(t) + F_k[X(t)]}_{Passive\ model} + \underbrace{C_2\dot{X}^3(t-\delta)}_{Active\ control} = F\cos\omega t \tag{1}$$

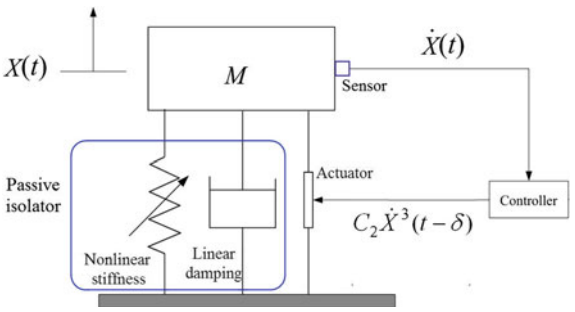


Fig. 1 The schematic of an active vibration isolation system

where M, C_1, $F\cos\omega t$ are mass, linear viscous damping coefficient and excitation force respectively. C_2 and δ denote the feedback gain and the designed time delay in the controller. The restoring force can be described by the piecewise linear function with respect to the displacement, as

$$F_{\mathrm{k}}(X) = \begin{cases} K_1 X & (X \geq a_{\mathrm{c}}) \\ K_2 X + (K_1 - K_2)a_{\mathrm{c}} & (X < a_{\mathrm{c}}) \end{cases} \tag{2}$$

where a_{c} is the coordinate value of discontinuity point on the displacement axis, and K_1 and K_2 are stiffness coefficients. Using the following dimensionless system parameters

$$\omega_0 = \sqrt{\frac{K_1}{M}},\ \Omega = \frac{\omega}{\omega_0},\ T = \omega_0 t,\ \tau = \omega_0 \delta,\ X = x a_{\mathrm{c}},$$
$$\xi_1 = \frac{C_1}{2M\omega_0}, \xi_2 = \frac{C_2(a_{\mathrm{c}}\omega_0)^3}{K_1 a_{\mathrm{c}}}, f = \frac{F}{K_1 a_{\mathrm{c}}},$$

one can easily obtain

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \omega_0 a_{\mathrm{c}} \frac{\mathrm{d}x(T)}{\mathrm{d}T},\ \frac{\mathrm{d}^2X(t)}{\mathrm{d}t^2} = \omega_0^2 a_{\mathrm{c}} \frac{\mathrm{d}^2x(T)}{\mathrm{d}T^2},\ \frac{\mathrm{d}X(t-\delta)}{\mathrm{d}t} = \omega_0 a_{\mathrm{c}} \frac{\mathrm{d}x(T-\tau)}{\mathrm{d}T}.$$

Substituting those transformations into (2) yields the dimensionless equation of motion as

$$\ddot{x}(T) + x(T) + 2\xi_1 \dot{x}(T) + \xi_2 \dot{x}^3(T-\tau) + \varepsilon g[x(T)] = f\cos\Omega T \tag{3}$$

where dot denotes differentiation with respect to T, and the nonlinearity factor is defined as

$$\varepsilon \overset{\mathrm{def}}{=} 1 - \frac{K_2}{K_1}, \tag{4}$$

and due to the fact that a_c is negative,

$$g(x) = \begin{cases} 0 & (x \le 1) \\ -(x-1) & (x > 1) \end{cases} \tag{5}$$

To analyse the primary resonance of the controlled system by using the multi-scale perturbation method, one confines the study to the case of small damping, weak nonlinearity, weak feedback and low level excitation [14]. i.e. $\xi_1 = \varepsilon\hat{\xi}_1,\ \xi_2 = \varepsilon\hat{\xi}_2,\ f = \varepsilon\hat{f},\ \Omega^2 = 1 + \varepsilon\sigma,\ \sigma = O(1)$.

Equation (3) can be rewritten as

$$\ddot{x}(T) + x(T) = \varepsilon\left\{-g[x(T)] - 2\hat{\xi}_1\dot{x}(T) - \hat{\xi}_2\dot{x}^3(T-\tau) + \hat{f}\cos\Omega T\right\} \tag{6}$$

For simplicity, first-order approximate with two time scales is introduced herein

$$x(T) = x_0(T_0, T_1) + \varepsilon x_1(T_0, T_1) + O(\varepsilon^2),\ T_r = \varepsilon^r T,\ r = 0,\ 1. \tag{7}$$

Using following differential operators and substituting them into (6),

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}T} = \frac{\partial}{\partial T_0} + \varepsilon\frac{\partial}{\partial T_1} + O(\varepsilon^2) \equiv D_0 + \varepsilon D_1 + O(\varepsilon^2) \\ \frac{\mathrm{d}^2}{\mathrm{d}T^2} = D_0^2 + 2\varepsilon D_0 D_1 + O(\varepsilon^2) \end{cases} \tag{8}$$

then equating the same power of ε produces

$$\varepsilon^0: \quad D_0^2 x_0(T_0, T_1) + \Omega^2 x_0(T_0, T_1) = 0 \tag{9}$$

$$\begin{aligned} \varepsilon^1: \quad & D_0^2 x_1(T_0, T_1) + \Omega^2 x_1(T_0, T_1) = -2D_0 D_1 x_0(T_0, T_1) - g[x_0(T_0, T_1)] - 2\hat{\xi}_1\dot{x}_0(T_0, T_1) \\ & -\hat{\xi}_2\dot{x}_0^3(T_0 - \tau,\ T_1) + \hat{f}\cos\Omega\tau + \sigma x_0(T_0, T_1) \end{aligned} \tag{10}$$

The solution of (9) is

$$x_0(T_0, T_1) = a(T_1)\cos[\Omega T_0 + \varphi(T_1)] \tag{11}$$

Substituting (11) into (10) yields

$$\begin{aligned} D_0^2 x_1(T_0, T_1) + \Omega^2 x_1(T_0, T_1) = & -2(-\Omega D_1 a\sin\phi - \Omega a D_1\varphi\cos\phi) - g[x_0(T_0, T_1)] \\ & + (\hat{f}\cos\phi\cos\varphi + \hat{f}\sin\phi\sin\varphi) \\ & + 2\hat{\xi}_1\Omega a\sin\phi - \hat{\xi}_2(-a\Omega)^3\sin^3(\phi - \Omega\tau) + \sigma a\cos\phi \end{aligned} \tag{12}$$

where $\phi = \Omega T_0 + \varphi(T_1)$. In order to eliminating secular term in (12), the coefficients of basic harmonic terms ($\sin\phi$ and $\cos\phi$) must be zero. Thus,

$$\sin\phi: \quad 2\Omega D_1 a + \hat{f}\sin\varphi + 2\hat{\xi}_1\Omega a + \tfrac{3}{4}\hat{\xi}_2 a^3\Omega^3\cos(\Omega\tau) + A_1 = 0 \tag{13}$$

$$\cos\phi: \quad 2\Omega a D_1\varphi + \hat{f}\cos\varphi + \sigma a - \tfrac{3}{4}\hat{\xi}_2 a^3\Omega^3\sin(\Omega\tau) + A_2 = 0 \tag{14}$$

where A_1 and A_2 are the Fourier coefficients of basic harmonic terms of $-g[x_0(T_0, T_1)]$.

Then $D_1 a$ and $D_1\varphi$ are obtained from (13) and (14).

$$D_1 a = -\frac{1}{2\Omega}\left[\hat{f}\sin\varphi + 2\hat{\xi}_1\Omega a + \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\cos(\Omega\tau)\right] \tag{15}$$

$$a D_1\varphi = -\frac{1}{2\Omega}\left[\hat{f}\cos\varphi + \sigma a - \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\sin(\Omega\tau) + A_2\right]. \tag{16}$$

3 Local Stability Analysis

The following discussion is divided into three cases depending on whether the system has nonlinearity or external excitation.

The statement begins with the simplest situation.

Case 1 $\varepsilon = 0$ and $f = 0$

The average equations becomes

$$D_1 a = -\frac{1}{2\Omega}\left[2\hat{\xi}_1\Omega a + \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\cos(\Omega\tau)\right] \tag{17}$$

$$a D_1\varphi = -\frac{1}{2\Omega}\left[\sigma a - \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\sin(\Omega\tau)\right] \tag{18}$$

Obviously, φ does not exist in Eq. (17) and thus Eq. (18) can be dropped. One can compute the fixed point by letting $D_1 a = 0$. A simple calculation gives

$$a = 0 \text{ and } a = \sqrt{-\frac{8\xi_1}{3\xi_2\Omega^2\cos(\Omega\tau)}} \tag{19}$$

To be precise, $a = 0$ corresponds to a fixed point and the value of $a > 0$ corresponds to periodic orbit (limit circle). The nature of their stability can be determined by finding the eigenvalue of linearization of Eq. (17).

1. $a = 0$. The fixed point is asymptotically due to the negative eigenvalue $-2\hat{\xi}_1 < 0$.
2. $a = \sqrt{-\frac{8\xi_1}{3\xi_2\Omega^2\cos(\Omega\tau)}}$. The existence condition of the periodic orbit is that $\cos(\Omega\tau) < 0$ because the given feedback gain $\xi_2 > 0$ is positive. In particular, in the situation $\Omega = 1$, $\cos(\tau) < 0$. Then it is easy to obtain the critical time

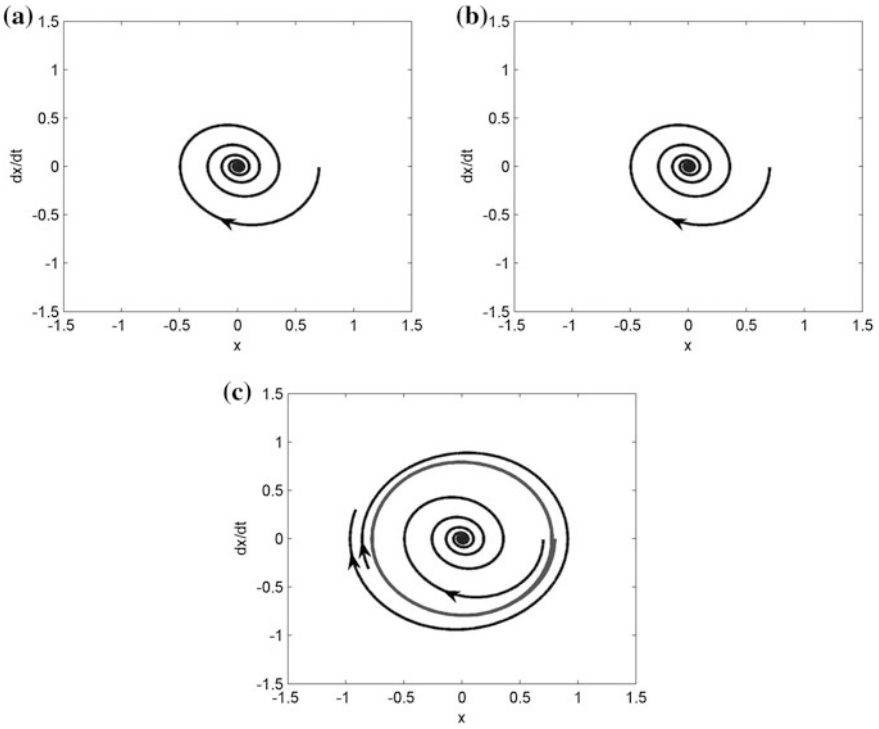


Fig. 2 Stability of fixed point and limit circle: **a** $\cos(\tau)<0$, **b** $\cos(\tau)=0$, **c** $\cos(\tau)>0$

delay $\tau = \frac{\pi}{2} + k\pi$, ($k$ = 0, 1, 2 …). Hence, it is found that the periodic emerges discontinuously and periodically. We next examine the stability of the limit circle. A simple calculation produces $-\hat{\xi}_1 - \frac{9}{8}\hat{\xi}_2 a^2\Omega^2\cos(\Omega\tau) = 2\hat{\xi}_1 > 0$, which results in an unstable orbit, as shown in Fig. 2. Readers should not the difference between the situation and Hopf bifurcation in which the limit circle's radius enlarges from zero as the bifurcation parameter varies. But in this case, the limit amplitude shrinks from infinity. From a physical view, it can be seen that the emergence of the periodic orbit indicates the energy dissipation of damping is compensated by the time-delayed feedback.

Case 2 $\varepsilon \neq 0$ and $f = 0$

In this case, can radically new dynamical behaviour occur? The following discussion is motivated by the question. The average equations are rewritten as

$$D_1 a = -\frac{1}{2\Omega}\left[2\hat{\xi}_1\Omega a + \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\cos(\Omega\tau)\right] \tag{20}$$

$$aD_1\varphi = -\frac{1}{2\Omega}\left[\sigma a - \frac{3}{4}\hat{\xi}_2 a^3\Omega^3 \sin(\Omega\tau) + A_2\right] \tag{21}$$

It is found that the nonlinearity merely appear in the phase equation. Hence, one can guess that there are still exist a fixed point and a periodic orbit. Depending on whether the radius of limit circle is over the break point, there are two possibilities: $a < 1$ and $a \geq 1$. In either case, no new dynamical behaviour emerges.

Case 3 $\varepsilon \neq 0$ and $f \neq 0$

The situation becomes more complicated. For a steady-state response, conditions of $D_1 a = 0$ and $D_1\varphi = 0$ have to be met. Therefore,

$$\hat{f}\cos\varphi = -\left[\sigma a - \frac{3}{4}\hat{\xi}_2 a^3\Omega^3 \sin(\Omega\tau) + A_2\right] \tag{22}$$

$$\hat{f}\sin\varphi = -\left[2\hat{\xi}_1\Omega a + \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\cos(\Omega\tau)\right] \tag{23}$$

To analyse the stability of steady state primary resonance, linearizing Eqs. (15) and (16) with respect to φ and a and combining (22) and (23) yield

$$D_1\Delta a = -\left[\hat{\xi}_1 + \frac{9}{8}\hat{\xi}_2 a^2\Omega^2\cos(\Omega\tau)\right]\Delta a + \frac{1}{2\Omega}\left[\sigma a - \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\sin(\Omega\tau) + A_2\right]\Delta\varphi \tag{24}$$

$$\begin{aligned} D_1\Delta\varphi = {} & \frac{1}{a}\left[-\frac{1}{2\Omega}\sigma - \frac{1}{2\Omega}A_2' + \frac{9}{8}\hat{\xi}_2 a^2\Omega^2\sin(\Omega\tau)\right]\Delta a \\ & - \frac{1}{2\Omega a}\left[2\hat{\xi}_1\Omega a + \frac{3}{4}\hat{\xi}_2 a^3\Omega^3\cos(\Omega\tau)\right]\Delta\varphi \end{aligned} \tag{25}$$

where $A_2 = \frac{1}{\pi}(a\varphi_0 - \sin\varphi_0)$ and $A_2' = \frac{1}{\pi}\left(\varphi_0 + a\frac{\mathrm{d}\varphi_0}{\mathrm{d}a} - \frac{\mathrm{d}\varphi_0}{\mathrm{d}a}\cos\varphi_0\right)$.

From the Routh–Hurwitz criterion, the steady-state vibration is asymptotically stable if and only if the following two inequalities hold simultaneously

$$\Sigma_1 \overset{\text{def}}{=} 2\hat{\xi}_1 + \frac{3}{2}\hat{\xi}_2 a^2\Omega^2\cos(\Omega\tau) > 0 \tag{26}$$

and

$$\Sigma_2 \overset{\text{def}}{=} S_1 S_4 - S_2 S_3 > 0 \tag{27}$$

As a matter of fact, if only condition (26) stands up but inequality (27) does not hold, the response is unstable due to the occurrence of saddle-node bifurcation. Actually, the stability boundary $\Sigma_1 = 0$ indicates the critical condition that the sign of real parts of both roots of characteristic equation changes. The Hopf

bifurcation may occur at the boundary. If Hopf bifurcation does not happen, $\Sigma_1 < 0$ indicates that the dynamic response of vibration system will diverge whether (27) holds or not, as shown in Fig. 4b.

4 Stable Feedback Parameter Combinations

Naturally, the task of this section is to find suitable feedback parameters which can give stable frequency responses. Figure 3 shows that the influences of time delay τ on the vibration amplitudes for given systems. The solid line determined by (20) depicts the variation of amplitude versus the time delay, and the broken line indicates stability boundary Σ_1. Compared with controlled system with time delay ($\tau \neq 0$), the controlled system without time delay ($\tau = 0$) can produce lower displacement amplitude, which is consistent with the conclusion of qualitative analysis using the equivalent damping. On the other hand, from the comparison between two figures, it is clear that the stronger feedback gain ξ_2 brings down the vibration amplitude, but shrinks the stable region. In the case of $\xi_2 = 0.1$ (Fig. 3b), some parts of the responses are rounded up in unstable regions. Figure 4 shows the time responses of the systems with different time delays. When the time delay varies cross the boundary, what does the steady-state periodic orbit. There are two possibilities: either the amplitude of the periodic solution goes to infinity or it becomes a more complicated bounded motion. In our study, the former happens in Fig. 4b.

For a vibration isolation system, the maximum displacement amplitude usually should be suppressed under a target level. Since the maximum displacement depends on the feedback parameters, how to choose gain ξ_2 and feedback time delay τ will be expatiated by the following example. Assume that for the given system with parameters ($f = 0.1786$; $\varepsilon = 0.6$, $\Omega = 1$) the dimensionless amplitude limit a_d equals 1.2.

Firstly, it is straightforward to determine the division line in parameter plane (ξ_2,τ) in Fig. 5 by using the frequency response (20). In this figure, the dash-dotted, broken and solid lines represent the division of $a_d = 1.2$ corresponding to dimensionless linear damping coefficients $\xi_1 = 0.03$, 0.05 and 0.07. The whole plane is divided into two parts by the division line, and on the upper plane, the displacement amplitude governed by parameter pair (ξ_2, τ) is less than the specified limit value. Therefore, the upper plane represents the feasible parameter combinations. However, not all parameter pairs located in the upper part can satisfy the stability conditions. Hence, it is necessary to identify the corresponding stability boundaries which could exclude those feedback parameters falling in the unstable region. In the figure, the solid lines respectively with triangle, square and circle are three stability boundaries corresponding to linear damping coefficient $\xi_1 = 0.03$, 0.05 and 0.07.

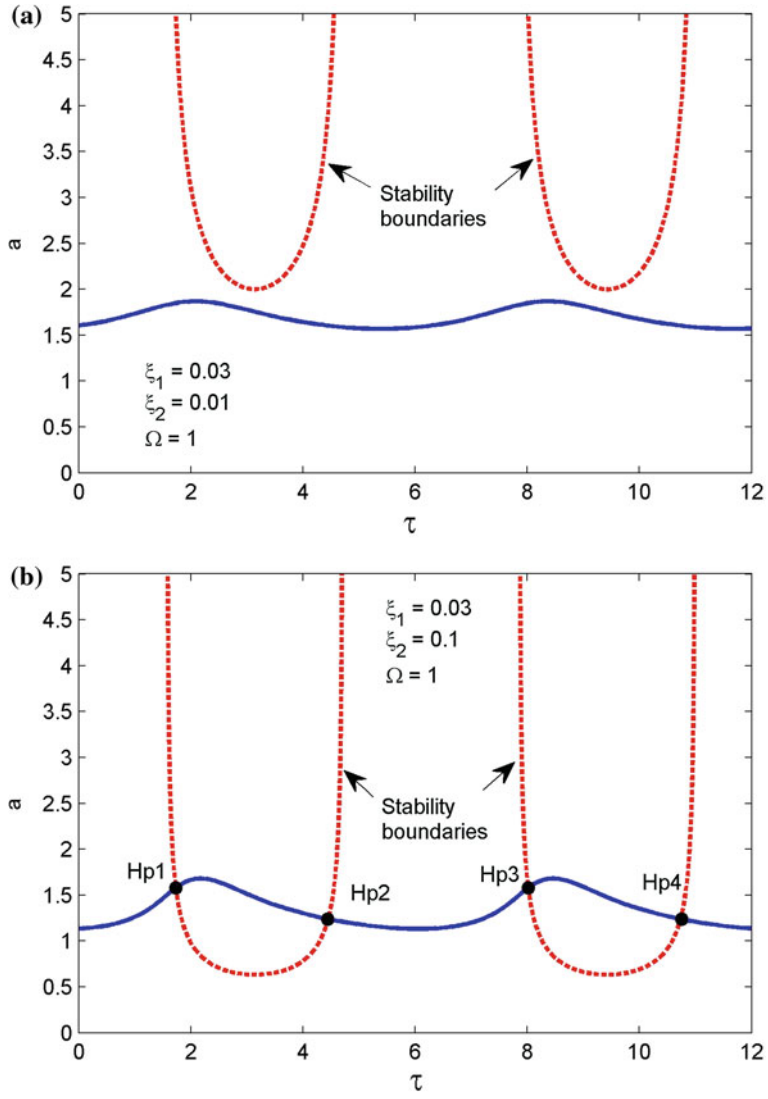


Fig. 3 Effect of the time delay on vibration amplitude for the system with different feedback gains. (Hp1, Hp2, Hp3 and Hp4 indicate the *points* where the bifurcation may occur, and the *solid lines* between Hp1–Hp2 and Hp3–Hp4 are unstable response branches)

5 Some Simple Illustrations on Transmissibility

The following discussion focuses on influence of ξ_1, ξ_2 and τ on vibration isolation performance which is evaluated by the force transmissibility. The Runge-Kutta algorithm is employed to estimate force transmissibility. The results are given in

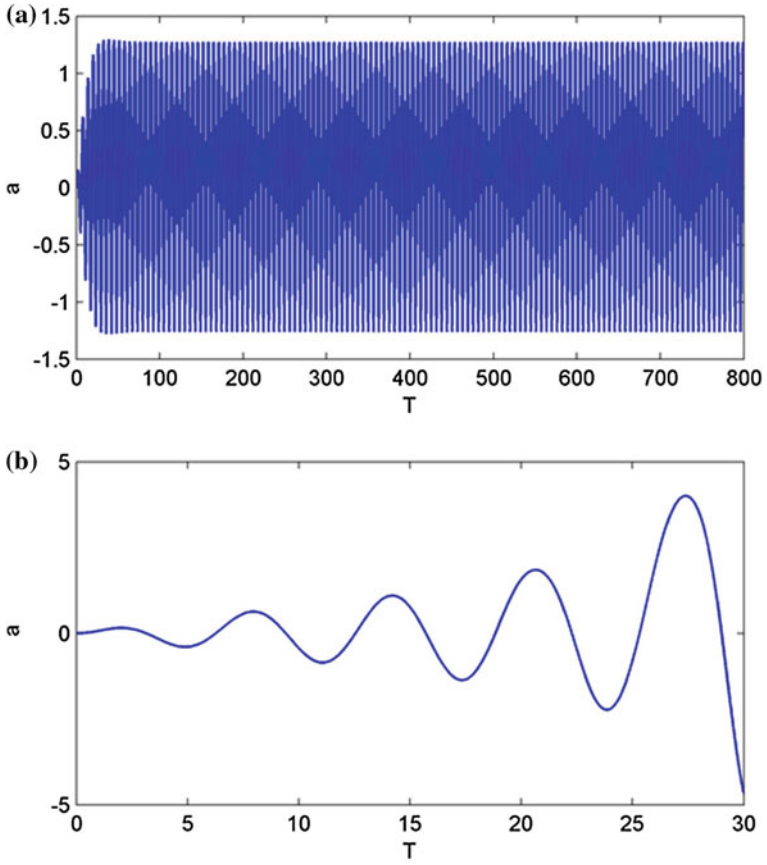


Fig. 4 Numerical responses of the system with different time delays. **a** $\tau = 1.0$. **b** $\tau = 3.0$

Figs. 6, 7, 8 and 9. Figure 6 shows the variation of transmissibility as the linear damping coefficient increases. As one might expect, the increase of linear damping reduces the transmissibility peak and consequently suppress the vibration in resonance region. However, the dilemma occurs that the increase of ξ_1 is detrimental for vibration isolation in frequency band where isolation is required. For controlled system without time delay, Fig. 7 compares the effect of feedback gain on the force transmissibility between three cases of $\xi_2 = 0$, 0.8 and 4.0. It is manifest that the increase of feedback gain can not only reduce transmissibility and suppress vibration in resonance region but keep them unchanged over higher frequency range. Therefore, cubic velocity feedback breaks through the barrier existing in the passive vibration isolation system with linear damping.

When the time delay is considered, the situation becomes a little complicated, as shown in Figs. 8 and 9. As be discussed in Sect. 3.2, the time delay can also affect the equivalent stiffness, and consequently the resonance peak shifts towards

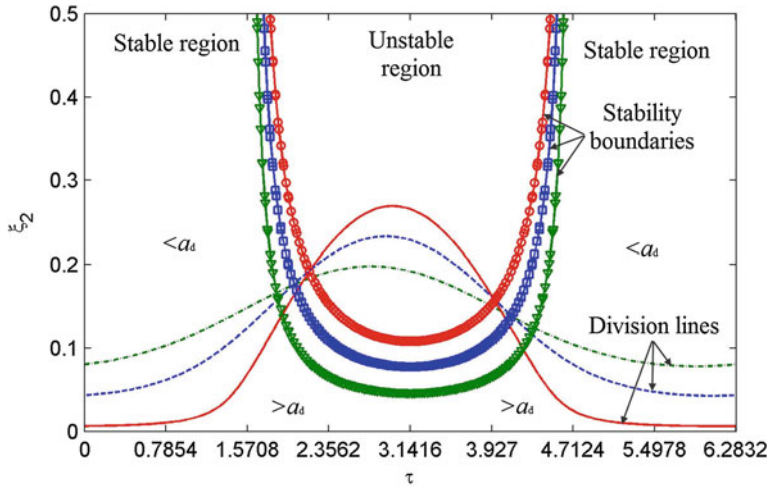


Fig. 5 Design illustration of feedback gain and time delay. (*Dash-dotted*, broken and *solid lines* correspond to the case of $\xi 1 = 0.03$, 0.05 and 0.07 respectively. *Solid lines* with *triangle*, *square* and *circle* are corresponding stability boundaries determined by (26))

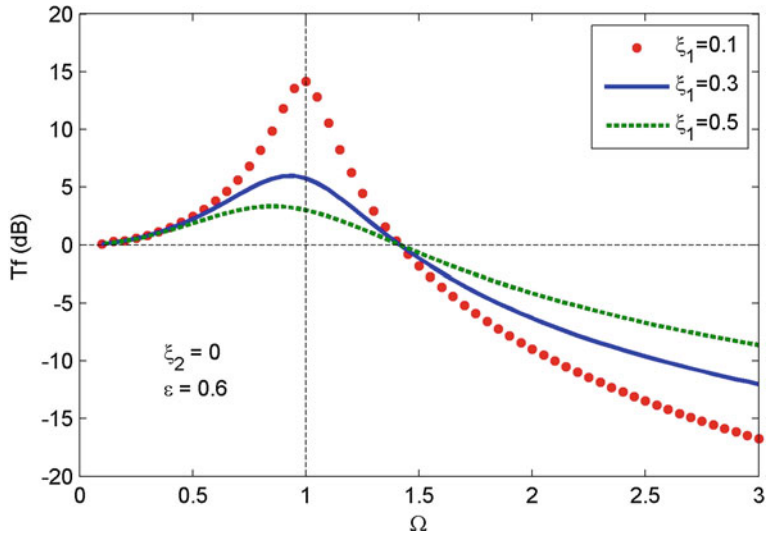


Fig. 6 Effects of linear damping on vibration transmissibility in case of $a < 1$

to higher frequency as the time delay increases. Meanwhile, force transmissibility peak rises. Furthermore, jump phenomenon will appear in the case of $\tau = 3\pi/8$. In fact, there are two potential cases in terms of system's resonance response.

The first is that the maximum frequency response is below the critical displacement point, i.e. $a < 1$. In this case, the increase of time delay produces

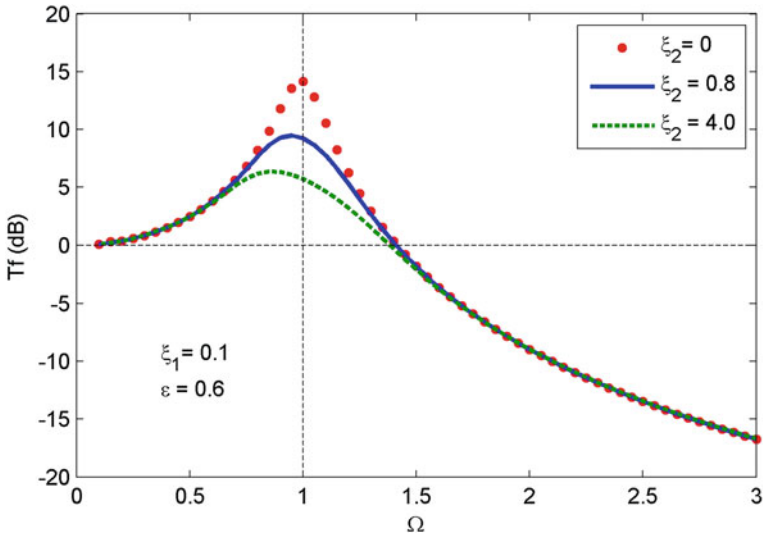


Fig. 7 Effects of feedback gain on vibration transmissibility in case of $a < 1$

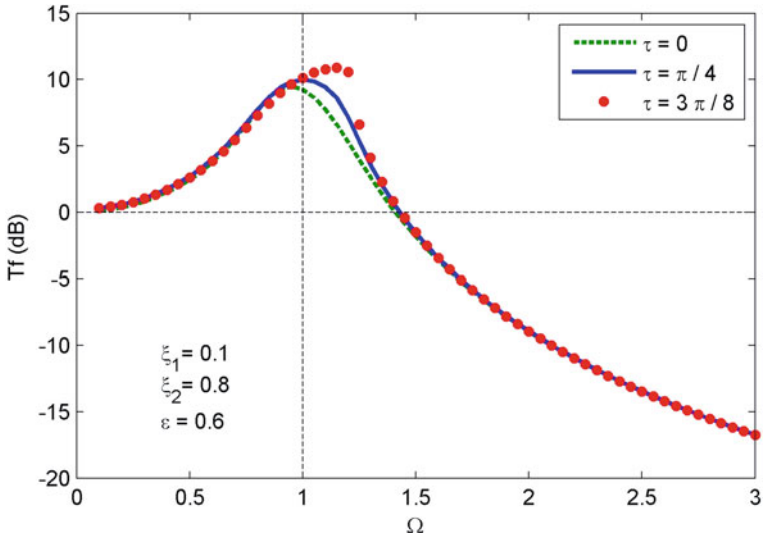


Fig. 8 Effect of time delay on vibration transmissibility for the controlled systems in case of $a < 1$

adverse influences on the force transmissibility, as shown in Fig. 8 and thus the control scheme with very small time delay is preferable to control vibration. The second situation is that the discontinuous point of the stiffness can be crossed, i.e. $a \geq 1$, and hence the softening stiffness property is exhibited. In this case, the time

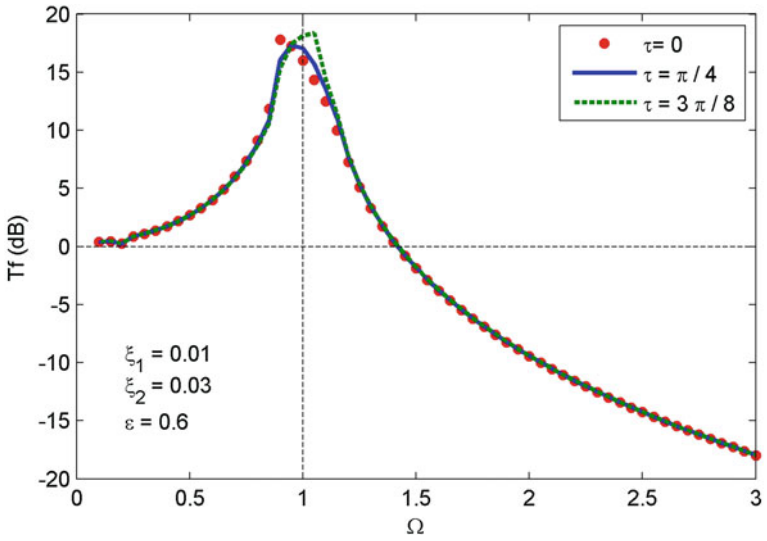


Fig. 9 Effect of time delay on vibration transmissibility for the controlled systems in case of $a \geq 1$

delay can exhibit its favorable effect on the vibration transmissibility. In Fig. 9, the dot line shows the emergence of jump phenomenon induced by the piecewise linear stiffness. As can be seen interestingly, the jump disappears as $\tau = \pi/4$.

6 Conclusions

This paper combines cubic nonlinearity with time-delayed feedback to achieve the improvement of the vibration isolation for a bilinear system. The average equations of the controlled system have been found analytically by utilising the multi-scale method and subsequently local dynamical behaviours were analysed. To control the resonance level under a specified value, the feedback parameters were determined by the frequency response together with stability boundaries which must be utlised to exclude the unstable parameter combinations. Although this paper allows for a series of satisfying parameters, the optimum feedback parameters for vibration isolation are still unknown and will be sought in future.

Lastly, the force transmissibility of the controlled system was studied, and it is concluded that the gain can not only reduce the whole force transmissibility level and greatly suppress vibration in the resonance region, but also can keep the transmissibility unchanged over higher frequency range where vibration isolation is required, and the larger feedback gain is beneficial to the vibration isolation. Besides, the jerk induced by sudden change of damping force in conventional skyhook cannot occur due to the smoothly imposed control force.

References

1. S.N. Mahmoodi, M. Ahmadian, Modified acceleration feedback for active vibration control of aerospace structures. Smart Mater. Struct. **19**(6), 065015(10 pp.) (2010)
2. Q. Hu, G. Ma. Variable structure control and active vibration suppression of flexible spacecraft during attitude maneuver. Aerosp. Sci. Technol. **9**(4), 307–317 (2005)
3. H. Ma, G.Y. Tang, Y.D. Zhao, Feed forward and feedback optimal control for offshore structures subjected to irregular wave forces. J. Sound Vib. **33**(8–9), 1105–1117 (2006)
4. Z.Q. Lang, X.J. Jing, S.A. Billings et al., Theoretical study of the effects of nonlinear viscous damping on vibration isolation of sdof systems. J. Sound Vib. **323**(1–2), 352–365 (2009)
5. S.J. Elliott, M. Serrand, P. Gardonio, Feedback stability limits for active isolation system with reactive and inertial actuators. J. Vib. Acoust. **123**(2), 250–261 (2001)
6. S.R. Will, M.R. Kidner, B.S. Cazzolato et al., Theoretical design parameters for a quasi-zero stiffness magnetic spring for vibration isolation. J. Sound Vib. **326**(1–2), 88–103 (2009)
7. Y. Liu, T.P. Waters, M.J. Brennan, A comparison of semi-active damping control strategies for vibration isolation of harmonic disturbances. J. Sound Vib. **280**(1–2), 21–39 (2005)
8. M. Ahmadian, X. Song, S.C. Southward, No-Jerk skyhook control methods for semiactive suspensions. J. Vib. Acoust. **126**(4), 580–584 (2004)
9. Y.M. Wang, Research on characteristics of on-off control for semi-active suspensions of vehicle. Chin. J. Mech. Eng. **38**(6), 148–151 (2002)
10. M.S. Ali, Z.K. Hou, M.N. Noori, Stability and performance of feedback control systems with time delays. Comput. Struct. **66**(2–3), 241–248 (1998)
11. A. Maccari, Vibration control for the primary resonance of the van der Pol oscillator by a time delay state feedback. Int. J. Non-Linear Mech. **38**(1), 123–131 (2003)
12. A. Maccari, The response of a parametrically excited Van der Pol oscillator to a time delay state feedback. Nonlinear Dyn. **26**(2), 105–119 (2001)
13. N. Eslaminasab, M.F. Golnaraghi, *The Effect of Time Delay of the Semi-Active Dampers of On–Off Control Schemes*. ASME 2007 International Mechanical Engineering Congress and Exposition, Seattle, USA
14. H. Hu, E.H. Dowell, L.N. Virgin, Resonances of a harmonically forced duffing oscillator with time delay state feedback. Nonlinear Dyn. **15**(4), 311–327 (1998)
15. Y. Zhao, J. Xu, Mechanism analysis of delayed nonlinear vibration absorber, Chin. J. Theor. Appl. Mech. **40**(1), 98–106 (2008)
16. N. Nayfeh, W. Baumann, Nonlinear analysis of time-delay position feedback control of container cranes. Nonlinear Dyn. **53**(1), 75–88 (2008)
17. X. Gao, Q. Chen, H.D. Teng, Modelling and dynamics properties of a novel solid and liquid mixture vibration isolator. J. Sound Vib. **331**(16), 3695–3709 (2012)
18. X. Gao, Q. Chen, Frequency response and jump avoidance of a vibration isolator with solid and liquid mixture. J. Vib. Shock **32**(12), 150–153 (2013)
19. X. Gao, Q. Chen, Active vibration control for a bilinear system with nonlinear velocity time-delayed feedback, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013*, London, UK, 3–5 July 2013, pp. 2037–2042

Project of Mechanical VVA Systems for Motorcycle Engines

Carmelina Abagnale, Mariano Migliaccio and Ottavio Pennacchia

Abstract This paper presents some results of the project of innovative mechanical variable valve actuation (hereafter VVA) systems, developed for high performance motorcycle engines, at University of Napoli Federico II, Department of Industrial Engineering–Section Mechanics and Energy (hereafter DiME). In addition to a first simple (and limited) system used just as a model for the previous analysis, the work has evolved through three basic steps leading to three types of VVA systems, all mechanical systems (as defined in literature and described later). The study has been conducted implementing a numerical procedure specifically designed to determine cam profile and kinematic and dynamic characteristics of the whole system, starting from some data (as described in the paper). The model has been validated against the conventional timing system using kinematic simulations. Results of the numerical procedure verify the validity of the VVA systems and particularly a better performance of the last one, in spite of its higher complexity.

Keywords Engine valves · Internal combustion engine timing · Valve timing · Valve timing variation · Variable valve actuation · VVA systems

C. Abagnale (✉) · M. Migliaccio · O. Pennacchia
Department of Industrial Engineering, University of Napoli Federico II, Via Claudio 21, 80125 Napoli, Italy
e-mail: c.abagnale@unina.it

M. Migliaccio
e-mail: mariano.migliaccio@unina.it

O. Pennacchia
e-mail: opennac@unina.it

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_7,

1 Introduction

The first step of the present work is represented by the study of different VVA mechanical systems in use to achieve the proposed objectives. The main strategies currently used in automotive field are: timing variation, duration variation, maximum lift variation, combined but not independent variation of timing, duration and lift (reference of the same authors for details [1–5]. The combined variation of the above parameters could enable several advantages in terms of performance, emissions and consumption. Even if complex, a mechanical system capable to implement this strategy is feasible (an example is the BMW Valvetronic). Generally this solution does not enable to reach independent variation of the three parameters (timing, duration and lift).

DiME is involved in study and manufacturing of new mechanical VVA (Variable Valve Actuation) systems to satisfy demands of weight and size for application on modern motorcycle engines designed by MotoMorini. This research aims to the design of a new mechanical VVA system for application on a single-cylinder motorcycle engine, to reach high performance, low specific consumption and low emissions [6–18].

2 A New Mechanical VVA System

The first proposed scheme (shown in Fig. 1 and studied just for its simplicity) is defined as a "3 elements-sliding system", because of its working: it enables the valve lift variation thanks to a sliding element (this first system has been designed to be applied on intake valve) and because it is a mechanical VVA system that consists of three elements: cam, main rocker arm with fixed fulcrum and secondary rocker arm with mobile fulcrum. This system enables valve lift variation through a simple sliding of one of the three elements (the secondary rocker arm). In this system (Fig. 1), fulcrum C of the auxiliary arm can slide from point A (maximum valve lift) to point B (minimum valve lift) along the segment AB.

The studied system presents a peculiarity: when valve is closed, fulcrum C sliding direction is parallel to the upper surface of the main rocker arm. This feature has two important consequences:

- Also the upper surface s of the secondary arm is plane and parallel to the sliding direction. The surface s must have a specific shape because it is necessary to keep contact between the working surfaces, during the fulcrum C sliding: between upper surface of the secondary arm and cam; between surface of the secondary arm and upper surface of the main arm. As shown in Fig. 1, the surface s is defined by forcing contact between surface s and base circumference of cam, when fulcrum moves, at closed valve. This contact is necessary to avoid unexpected valve lift and to ensure a closed valve configuration, when fulcrum is moving.

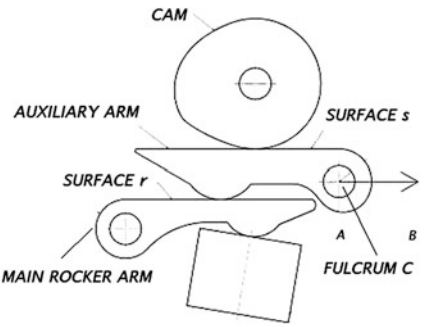


Fig. 1 Scheme of a preliminary system

- When fulcrum moves, valve lift law changes, but timing does not change: the start point of intake valve opening is the same (and also the start point of intake valve closing, since this type of distribution system characterized by constant angular duration of law).

The proposed VVA system 1 has been studied to be actuated by a DC engine: we evaluated to use a DC engine with an average power of 8 W and a maximum power of 45 W to reach a complete actuation. Performance obtained by this first VVA system in terms of valve lift law consists just in a variation of lift, when fulcrum of the auxiliary arm moves (with a maximum displacement of 20 mm).

3 Mathematical Model

The study began with the implementation of a numerical procedure (implemented in a program written by Mathcad) specifically designed to determine (in closed loop) cam profile and kinematic and dynamic characteristics of the whole system, starting from the following input data: rocker arm geometry, relative positions and inertial data of elements, spring stiffness and preloading, camshaft speed and valve lift law. The model was validated against the conventional timing system, using kinematic simulations, as described in this paper.

The input data for mathematical procedure are: geometric data of the system and valve lift law for maximum lift. Valve lift law is built starting from the maximum point: first we impose a cam inclination at constant speed (first by a polynomial acceleration, then by a polynomial valve lift law—in seventh grade. A maximum lift of 10 mm is imposed; maximum and minimum acceleration values are contained within the standard range recommended in the literature.

Starting from the valve lift law (Fig. 2) in terms of displacement (mm), speed (mm/rad) and acceleration (mm/rad^2), the cam profile is obtained in closed loop using the mathematical procedure: cam profile regularity is ensured by a continuous profile with continuous tangent and radius of curvature (Fig. 3).

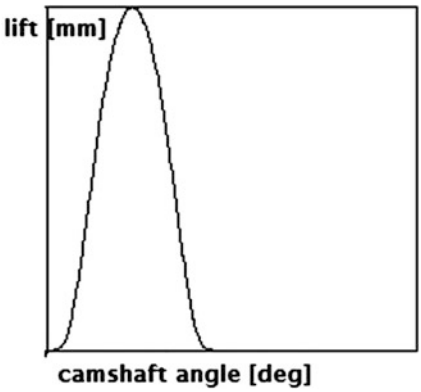


Fig. 2 Valve lift

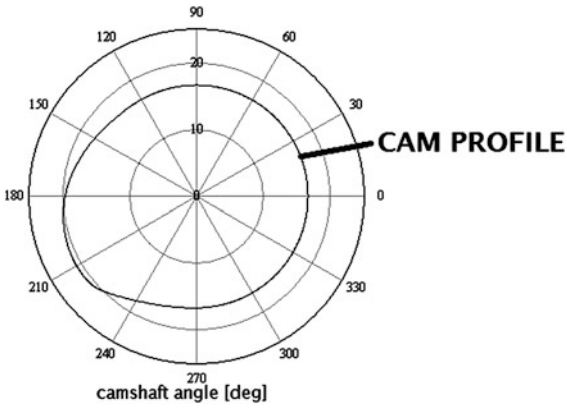


Fig. 3 Cam profile

Starting from input data, by the numerical procedure, it is possible to obtain contact forces between valve lifter and main rocker-arm (N0); between the two rocker-arms (N1); and between cam and auxiliary rocker-arm (N2) at a crankshaft speed of 8,500 RPM (hereafter we consider always this speed for study). The system is simplified and dynamic effects and vibrations due to components elasticity are not analyzed. All the equations have been solved considering non-deformable elements.

Starting from geometry and contact forces, it is possible to determine Hertzian pressure in contact areas, as function of camshaft angular position: Hertzian pressure is perfectly acceptable and compatible with the use of a good stainless steel.

The reactions of main and auxiliary arm pivots were calculated [3–6]. All the sliding speeds of the system were evaluated to estimate the power dissipated by the mechanism.

Starting from sliding speed among elements and from contact forces, it is possible to estimate global and instantaneous friction power absorbed by the

mechanism for the actuation of each valve. The absorbed power is referred to a friction coefficient equal to one: the diagrams are omitted for brevity. The results show that with a medium friction coefficient value of 0.04 there is a peak of friction power equal to 1.2 kW (=30 kW × 0.04) and a global power necessary to actuate each valve equal to 0.2 kW (=5 kW × 0.04). The instantaneous total power required for actuation of each valve can be evaluated as sum of the power required to move inertial masses and to deform springs and the power dissipated by friction, supposing an average coefficient of friction equal to 0.04. Since in ideal case of no friction total power (of each valve) is equal to zero (as the system is conservative), total power absorbed in real case coincides with power dissipated by friction (approximately 0.2 kW in this case). The energetic dissipation of a conventional actuation system is linked to the number of surfaces in contact and the relative speed. In this case we estimate a dissipation increase of about 10 %, largely recoverable by the advantages of variable actuation. The main cause of dissipation is the contact between cam and component, also in conventional systems.

A kinematic simulation of the system, to validate the mathematical procedure, was performed by Catia (Fig. 4): we used the first geometry and cam profile obtained with our model. Cam profile was exported as ASCI files by points (with a 1,000 points/rev resolution) and imported into Catia, then the profile was rebuilt by a "spline" function (to ensure profile regularity with its derivative). Simulation results are quite satisfying as shown in diagram of Fig. 5, which show comparison between kinematic parameters in the model (from which cam profile derived) and parameters obtained from simulation (based on cam profile derived from the model).

4 Results of the Preliminary System

The preliminary study of geometric, cinematic and dynamic features of the "3 elements—sliding VVA system", performed by means of the developed algorithm, conducted to results already reported in papers of the same authors [1–5] (reference for details). These results confirmed the potentiality of the "3 elements VVA system" and allowed to proceed to the design of the same system to be applied to the engine into account, providing for a maximum speed of 8,500 RPM. However, as already reported in [1–5], fluid dynamics analysis of the performance of the proposed system ("3 elements VVA system"), performed both at full load and at partial load of the considered engine, revealed limited possibilities to reduce consumption. In fact the system, although able to achieve especially at partial loads an increase of combustion speed (thanks to increased turbulence in the combustion chamber), it still requires a pumping work almost unaffected, that allows just slight advantages in terms of consumption. Analysis revealed, as it was to be expected, that the next step to a more effective mechanism must be able to vary, over lift, duration and timing.

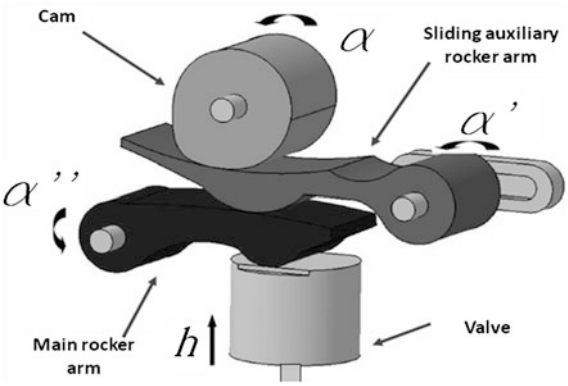


Fig. 4 3D Model of the preliminary VVA system

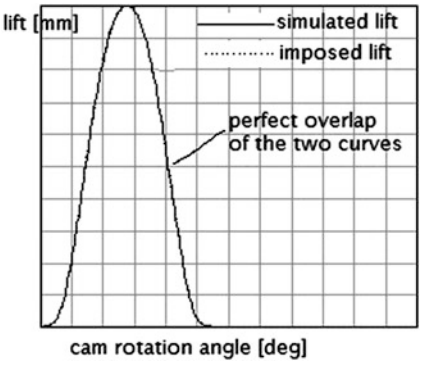


Fig. 5 Simulated versus imposed valve lift

5 3 and 4 Elements VVA Systems

In spite of its limited potential, the preliminary study has represented the essential starting point to address the study to new and more performance VVA systems. They are always operated by a mechanical cam, which can vary not only the maximum lift, but also the duration of valve law, in order to overcome the previous limitations.

In a 3 elements-VVA system, driven by a camshaft, the movement duration of a generic element in contact with the cam (our auxiliary rocker arm) is set by the cam profile itself and it is independent of its fulcrum position. For this reason, the only way to reduce the duration of valve lift is to achieve a free motion of the auxiliary rocker arm to the main one. A possible solution is to adopt a specific profile for the main rocker arm: a circular part concentric with the auxiliary rocker arm fulcrum (in condition of closed valve); an appropriate second part to produce the valve movement and to vary the starting point of the auxiliary rocker arm stroke. In the system A of Fig. 6 it is possible to achieve a simultaneous reduction of valve lift and duration by imposing the main rocker arm a circular motion

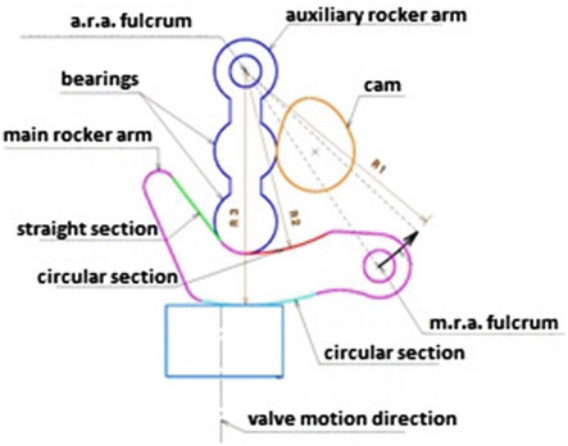


Fig. 6 System A: 3 elements VVA

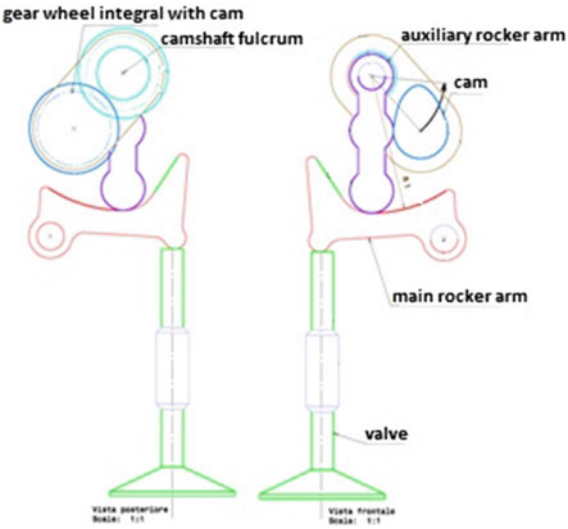


Fig. 7 System B: 3 elements VVA

around the fulcrum of the auxiliary rocker arm. The system B of Fig. 7 operates in a similar way of system 1: the difference is that it is not the main rocker arm fulcrum to move, but the camshaft support can rotate around a fixed fulcrum. Another system capable of variable lift and duration can be achieved by adopting a mechanism consisting of 4 elements in series, such as shown in Fig. 8 (system C). From the functional point of view, this system shows the advantages of both previous systems, but it also presents some problems due to increased complexity, size and losses. The system C, consisting of cam, main rocker arm, auxiliary rocker arm 1, auxiliary rocker arm 2, offers more potentialities, in terms of valve lift, duration and timing variation. The effect to achieve a simultaneous variation

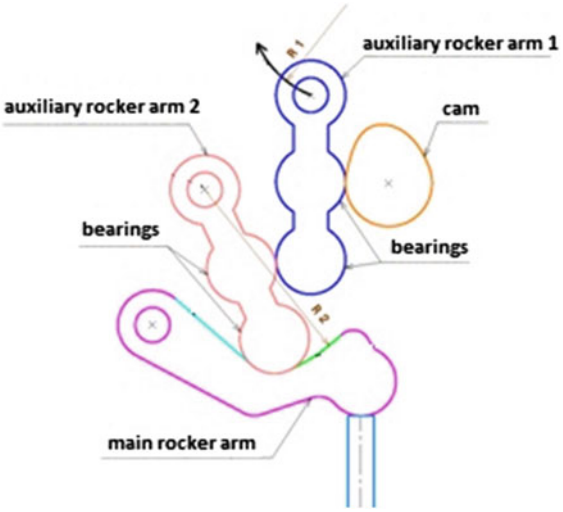


Fig. 8 System C: 4 elements VVA

of valve lift and duration is obtained by imposing to the auxiliary rocker arm fulcrum a movement on a trajectory belonging to an arc of a circle.

For all systems mentioned above, the first problem is the choice of a junction curve between the circular section and the straight one of the main rocker arm. To ensure continuity and regularity of the function of the angular velocity of the main rocker arm, the connection must necessarily have a radius of curvature greater than the local radius of the auxiliary rocker arm, at every point.

After geometric/mathematical analysis (the discussion is omitted), the result is that to ensure the continuity of angular acceleration law of the main rocker arm (and so the continuity of linear acceleration of the valve), it is necessary that the upper surface of the main rocker arm is made up of a class C2 curve (the curvature changes continuously along the profile). The continuity of the acceleration is necessary to ensure the continuity of inertial forces and consequently of contact forces, and the reduction of vibration phenomena. The easiest way to get a connection with the above properties is to use a cubic spline curve type.

6 The Problem of Accelerations

A common problem in the studied three systems is represented by the increase of maximum and minimum accelerations in condition of partial valve lift. A method to control acceleration values is to adopt a cam profile to ensure constant angular velocity to the auxiliary rocker arm along a sufficient portion of contact zone, followed by deceleration to reach the maximum valve lift. In this way, fixed maximum lift law, you can get the required lift law thanks to an appropriate top surface of the main rocker arm and of the cam. In fact, the contact area will always

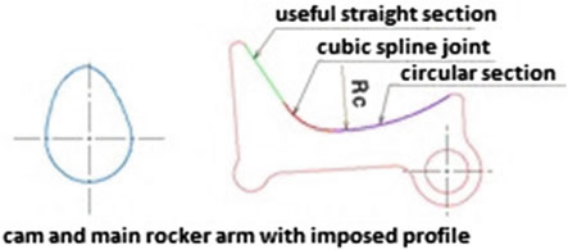


Fig. 9 System B with straight profile of the main rocker arm

be at constant angular velocity or during deceleration of the auxiliary rocker arm: the maximum acceleration of the valve will be (in the worst hypothesis) equal to the maximum acceleration at maximum lift. The objectives were fully achieved, in fact not only the maximum value of acceleration is contained in configuration of maximum lift, but also the minimum value of the acceleration decreases when the maximum lift decreases.

A common problem of the three studied systems is represented by the increase (in absolute value) of maximum and minimum accelerations of the valve in condition of partial lift.

The problem will be described in the following just for the system B (of the previous Fig. 7): the treatment is similar for all the systems.

There is an increase of maximum and minimum acceleration, referring to the system B with a straight profile of the main rocker arm (Fig. 9), where the same colors of lift and acceleration diagrams correspond to the same actuations of the system.

In this case there is an excessive acceleration value when the valve reverses its motion and closes (zone with negative acceleration) and the inertial forces (in the direction opposite to the acceleration but proportional to it) disagree with the reaction of the valve spring. This condition can cause a detachment of the valve tappet (flicker phenomenon).

The described problem depends on the not direct controllability of maximum and minimum accelerations (when the lift law is varying): in the particular case, these accelerations overly grow (in absolute value) when the maximum lift is decreasing.

A method to control acceleration values is to adopt a cam profile to ensure constant angular velocity to the auxiliary rocker arm along a sufficient portion of contact zone, followed by deceleration to reach the maximum valve lift (Fig. 10). In this way, fixed maximum lift law, you can get the required lift law thanks to an appropriate top surface of the main rocker arm and of the cam. In fact, the contact area will always be at constant angular velocity or during deceleration of the auxiliary rocker arm: the maximum acceleration of the valve will be (in the worst hypothesis) equal to the maximum acceleration at maximum lift.

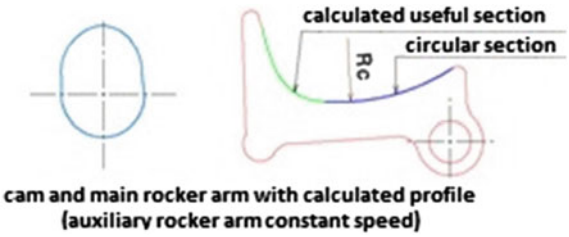


Fig. 10 System B with "constant speed" profile of the main rocker arm

However, the system presents the same problem on the minimum accelerations which are too high in absolute value. The problem can be explained in this way: when the maximum lift decreases, the compression of the valve spring (and then the total elastic force that opposes the phenomenon of detachment) decrease: in this condition, it would be desirable that the minimum value of the acceleration was reduced.

To overcome this problem, different solutions may be used, including:

1. to adopt an uniformly decreasing angular speed law of the auxiliary rocker arm (at constant deceleration after a first acceleration ramp in the neutral zone, until reaching the maximum oscillation), so that when the angle of the implementation of auxiliary rocker arm 1 gradually increases, the contact between the rocker arm in the useful zone occurs at increasingly reduced speed.
2. to make the deceleration of the auxiliary rocker 2 more gradual (by reducing the gradient by increasing the angular duration on the camshaft).

The first solution immediately has given very encouraging results (Fig. 11) and it has been chosen as the right way for the kinematic study of the three systems in object. The objectives have been fully achieved: in fact, not only the maximum value of the acceleration is contained in maximum value imposed by the configuration required for maximum lift, but also the minimum value of the acceleration decreases with decreasing of maximum lift.

7 4 Elements VVA System Design and Performance

In order to evaluate system C performance in terms of variability of the valve lift law, we used the kinematics module of software Catia. The simulations were performed up to a maximum of 11° of rotation of a rod (an element built in Catia to simulate rotation of the rocker arm in a simple way) corresponding to the maximum valve lift. A scheme of the followed procedure is shown in Fig. 12.

The results (shown in Fig. 13 just in terms of lift) are very encouraging: our mechanical VVA system enables the simultaneous variation of valve lift and duration. However, the proposed configuration has just been used to validate the

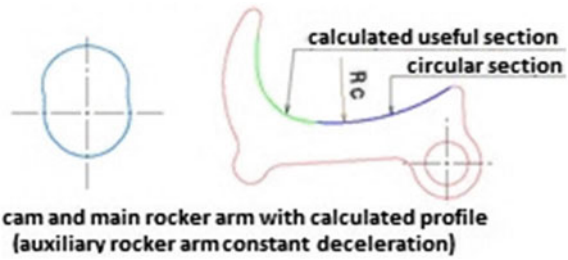


Fig. 11 System B with "constant acceleration" profile of the main rocker arm: lift and acceleration

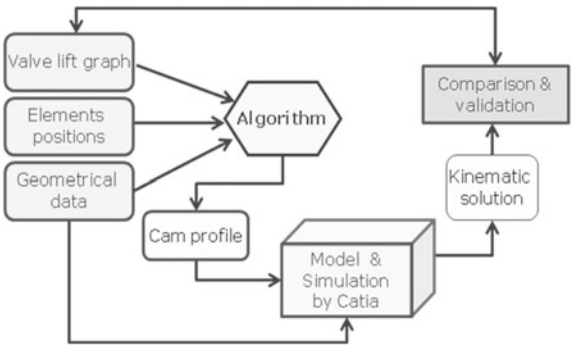


Fig. 12 The algorithm

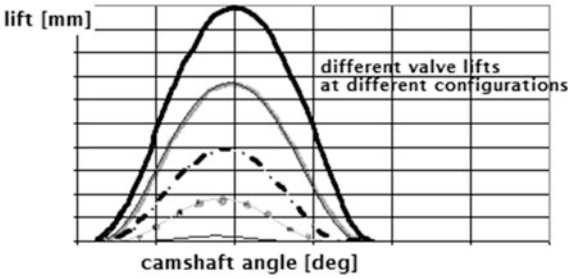


Fig. 13 System C: performance in terms of valve lift variation

analytical model, which will be used in future studies aimed to reach more adaptable configurations of 4 elements-systems able to give better performance on engines.

8 Test Bench

The results have led to design and construction of a prototype of the 4 elements-VVA system (system C). The rig consists of the engine Morini M638, modified to install VVA system, support plates and electrical actuator, as shown in Fig. 14.

In order to test the VVA system and evaluate its performance, a test bench has been designed and realized, with oscillating case motor controlled by an inverter.

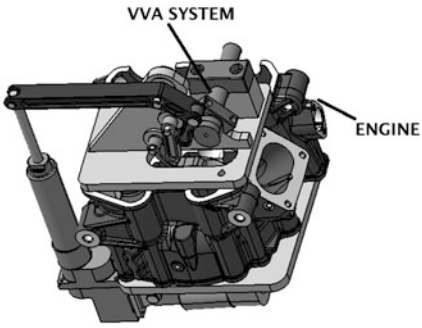


Fig. 14 Assembly system on engine

The test bench will be equipped with: load cell, encoder, position sensor on the actuator piston, high frequency position sensor for valve position, data acquisition system. With the described equipment the aim will be to evaluate the VVA system performance (reliability, torque, average and instantaneous power absorbed by the system, etc.) and to check the correspondence of the obtainable valve laws with the project ones.

9 Conclusion and Future Work

After this study, we concluded that the preliminary system presents advantages in terms of design and simplicity of manufacturing, but certainly it presents disadvantages in terms of use at high rotational speed and concerning impossible complete closing valve. DIME research group is performing simulations to verify and optimize VVA system object of study: results represent a basis for development of the project. A disadvantage of this new system is: if this VVA is not used in combination with a VVT, there are no significant advantages in terms of consumption. The advantages refer to an increase of turbulence and a consequent increase of combustion speed.

The future developments are about the opportunity to combine this VVA system with other systems capable to change duration and timing. In this way it would be possible to use different turbulences generated by the different lift laws under the best conditions of timing. This solution would lead to a greater efficiency of the system and to a significant reduction in terms of consumption.

Currently we are proceeding to build a prototype to test to verify the correspondence between numerical and experimental activities.

References

1. C. Abagnale, S. Caruso, A. Iorio, M. Migliaccio, O. Pennacchia, in *A New Mechanical Variable Valve Actuation System for Motorcycle Engines*. ICE 2009—9th International Conference on Engines and Vehicles, Capri (NA), 13–17 Sept 2009
2. C. Abagnale, PhD thesis, Sviluppo di un sistema di attuazione variabile VVA elettroidraulico per motori pluricilindrici, Università degli Studi di Napoli Federico II, 2009
3. C. Abagnale, A. Gimelli, M. Migliaccio, O. Pennacchia, Distribuzione variabile su motori alternativi a c.i.: VVA meccanici a 3 e a 4 elementi, 65° Congresso ATI, Cagliari, settembre 2010
4. C. Abagnale, M. Migliaccio, O. Pennacchia, Design of a new mechanical variable valve actuation system for motorcycle engines ESDA2012-82317, in *Proceedings of the ASME 2012 11th Biennial Conference on Engineering Systems Design and Analysis*, Nantes, France, 2–4 July 2012
5. C. Abagnale, M. Migliaccio, O. Pennacchia, Mechanical variable valve actuation systems for motorcycle engines, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Enegineering 2013*, WCE 2013, London, UK, 3–5 July, 2013, pp. 1809–1814
6. C. Abagnale, M. Migliaccio, O. Pennacchia, Brevetto italiano n. CE2010A000002 Sistemi di distribuzione variabile di tipo meccanico a 3 ed a 4 elementi attivi, industrial patent, 2010
7. C. Gray, *A Review of Variable Engine Valve Timing*, SAE paper 880386
8. M. Grohn, K. Wolf, *Variable Valve Timing in the new Mercedes–Benz Four-Valve Engines*, SAE 891990, 1989, doi:10.4271/891990
9. T.M. Lancefield, R.J. Gayler, A. Chattopadhay, *The Practical Application and Effects of a Variable Event Valve Timing System*. SAE paper 930825, SAE International Congress and Exposition, Detroit, MI, USA, marzo 1993
10. J.-C. Lee, C.W. Lee, J.A. Nitkiewicz, *The Application of a Lost Motion VVT System to a DOHC SI Engine*. SAE paper 950816, SAE International Congress and Exposition, Detroit, MI, USA, 1995
11. R.A. Stein, K.M. Galietti, T.G. Leone, *Dual Equal VCT—a Variable Camshaft Timing Strategy for Improved Fuel Economy and Emissions*. SAE paper 950975, 1995, doi:10.4271/950975
12. T.G. Leone, E.J. Christenson, R.A. Stein, *Comparison of Variable Camshaft Timing Strategies at Part Load*. SAE paper 960584, 1996, doi:10.4271/960584
13. Y. Moriya, A. Watanabe, H. Uda, H. Kawamura, M. Yoshioka, M. Adachi, *A Newly Developed Intelligent Variable Valve Timing System—Continuously Controlled Cam Phasing as Applied to a New 3 Liter Inline 6 Engine*. SAE paper 960579, SAE International Congress and Exposition, Detroit, MI, USA, 1996
14. K. Fukuo, T. Iwata, Y. Sakamoto, Y. Imai, K. Nakahara, K.A. Lantz, *Honda 3.0 Liter, New V6 Engine*. SAE paper 970916, SAE International Congress and Exposition, Detroit, MI, USA, 24–27 Feb 1997
15. Y. Urata, H. Umiyama, K. Shimizu, Y. Fujiyoshi, H. Sono, K. Fukuo, *A Study of Vehicle Equipped with Non-Throttling SI Engine with Early Intake Valve Closing Mechanism*. SAE paper 930820, 1993, doi:10.4271/930820
16. M. Hitomi, J. Sasaki, K. Hatamura, Y. Yano, *Mechanism of Improving Fuel Efficiency by Miller Cycle and Its Future Prospect*. SAE Paper 950974, 1995
17. Y. Wang, L. Lin, S. Zeng, J. Huang, A.P. Roskilly, Y. He, X. Huang, S. Li, Application of the Miller cycle to reduce NOx emissions from petrol engines. Appl. Energy **85**, 463–474 (2008)
18. B. Ludwig, in *Less CO2 Thanks to the BMW 4-Cyl. Valvetronic Engine*. ATA International Conference on Spark Ignition Engine: the CO2 Challenge, Paper 02A5011, Venezia, Italy, Nov 2002

Performance Evaluation of the Valveless Micropump with Piezoelectric Actuator

Chiang-Ho Cheng

Abstract To meet the rising need in biological and medical applications, the innovative micro-electro-mechanical systems (MEMS) technologies have realized an important progress of the micropump as one of the essential fluid handling devices to deliver and control precise amounts of fluids flowing along a specific direction. This paper aims to present the design, fabrication and test of a novel piezoelectrically actuated valveless micropump. The micropump consists of a piezoelectric actuator, a vibration plate, a stainless steel chamber plate with membrane and integrated diffuser/nozzle bulge-piece design, an acrylic plate as the top cover to form the channel with the channel plate and two glass tubes for delivery liquid. The chamber and the vibration plate were made of the stainless steel manufactured using the lithography and etching process based on MEMS fabrication technology. The experimental results demonstrate that the flow rate of micropump accurately controlled by regulating the operating frequency and voltage. The flow rate of 1.59 ml/min and back pressure of 8.82 kPa are obtained when the micropump is driven with alternating sine-wave voltage of 240 V_{pp} at 400 Hz. The micropump proposed in this study provides a valuable contribution to the ongoing development of microfluidic systems.

Keywords Actuator · Diffuser · Micropump · Nozzle · Piezoelectric · Valveless

1 Introduction

Microfluidic devices, such as micropumps, play a key role in micro-electro-mechanical systems (MEMS), particularly in the fields of biological, chemical, medical, and electronics cooling [1–3]. Micropumps exploit the MEMS technology

C.-H. Cheng (✉)
Department of Mechanical and Automation Engineering, Da-Yeh University, 168, University Rd., Dacun, 51591, Changhua, Taiwan
e-mail: chcheng@mail.dyu.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_8,

to provide the advantages including low cost, small size, low power consumption, reduction in the amount of reagents needed and small dead volume [4–6]. Various kinds of micropumping techniques have thus been developed. Nguyen et al. [7], Laser and Santiago [8], Iverson and Garimella [9] and Nabavi [10] have made the detailed reviews covering the fabrications, pumping mechanisms, actuations, valves, and operation characteristics of micropumps.

The actuation forms can be divided into two categories: mechanical and non-mechanical actuation in general. Since there are no moving elements, the structure of non-mechanical micropumps is simpler than that of mechanical micropumps. But the performance of non-mechanical micropumps are sensitive to the properties of working liquids, as discussed in the studies related to electrohydrodynamic (EHD) [11], magnetohydrodynamic (MHD) [12], electroosmotic [13] and electrochemical micropump [14]. Mechanical micropumps are relatively less sensitive to the liquid properties as compared to those non-mechanical micropumps; consequently, they may have much wider applications. The actuation mechanisms of mechanical micropumps include electrostatic [15], piezoelectric [16], electromagnetic [17], thermal pneumatic [18], bimetallic [19], shape memory alloy [20] and phase change [21] types. Due to the advantages of high stiffness, high frequency and fast response, piezoelectric actuation is very suitable to actuate micropumps especially.

In designing the vibrating displacement micropumps, a pumping chamber connected to the inlet and outlet microvalves is needed for flow rectification. Microvalves can be classified into check valve [16] and valveless [22–24] types. In check valve pumps, mechanical membranes or flaps are used with the concerned issues of wear, fatigue, and valve blocking in this type, resulting in limitation of its applications. The valveless micropumps, first introduced by Stemme and Stemme [22], implement diffuser and nozzle elements to function as a passive check valve. In addition, the peristaltic pumps or impedance pumps [4–6, 25, 26] and the Tesla-type pumps [27, 28] do not need passive check valves. The peristaltic pump consists of three chambers linked sequential. By creating peristaltic motion in these chambers, fluids can be pumped in a desired direction. Flow rectification can be also accomplished in Tesla microvalves by inducing larger pressure losses in the reverse direction compared to those in the forward direction assuming the same flow-rates. The above pump concepts have the major problems of requiring the complex design and fabrication processes. Thus, the valveless nozzle/diffuser micropumps are of particular interest for various microfluidic applications because of their simple configuration and low fabrication cost.

In order to characterize and optimize the performance of the valveless nozzle/diffuser micropumps, previous numerical and experimental studies have presented that the geometric design of the nozzle/diffuser elements can significantly affect the performance of valveless micropump. In this investigation, we proposed a high performance piezoelectric valveless micropump adopting an integrated nozzle/diffuser bulge-piece design. The micropump consisted of a stainless-steel structured chamber to strengthen its long-term reliability, low-cost production, and maximized liquid compatibility. A piezoelectric disc was also utilized to push

liquid stream under actuation. In simulating the inherently complex flow phenomena of pumping flowfield, the commercial computational fluid dynamics (CFD) software ESI-CFD ACE+® was used for numerical calculations [29].

2 Design

Figure 1 shows a novel valveless micropump proposed in this study and its cross-section view of the structure are shown in Fig. 2. The working principle of the proposed micropump is similar to that of the most contemporary valveless nozzle/diffuser micropumps. The flow is rectified owing to different pressure drops across the nozzle and the diffuser in changed flow directions. The pump cycle is divided into the supply and pump modes. During the supply model, the inlet and outlet regions operate as a diffuser and a nozzle respectively for the liquid flowing into the pump chamber. As a result, there is more fluid entering the chamber from the inlet side than that from the outlet side. Alternatively, the inlet region works as a nozzle and the outlet works as a diffuser, causing more fluid being expelled to the outlet side during the pump model. In this manner, net flow rate is generated from the inlet to the outlet. Figure 3 shows the schematic diagrams of operational principle in the supply and pump mode. In Fig. 4a, the dimensions of the etched pumping chamber were 8 mm in diameter and 70 μm in depth, respectively. Functioning as the flow-rectifying elements, the inlet width, length, height and divergence angle of the diffuser/nozzle were 800 μm, 3.1 mm, 70 μm and 10°, respectively. A 6 mm diameter and 70 μm high bulge-piece shown in Fig. 4b was right on the back side of the pumping chamber, as depicted in Fig. 4b. In this study, there were three bulge-piece diameters of 2, 4 and 6 mm at the same height tested to measure the delivered volumetric flow rates and pressures for evaluating the pumping performance in operations.

3 Fabrication

3.1 Piezoelectric Actuator

The piezoelectric disc with 200 μm thick was prepared by the commercial available piezoelectric powder (ARIOSE Company, B6 type) through the dry powder pressing technique, as illustrated in Fig. 5. The sintering process was performed in a tube furnace under a quiescent air atmosphere at a heating rate of 90 °C/min to the peak temperatures of 1,300 °C for maintaining a duration of 3 h, which followed by a 90 °C/min cooling rate to the room temperature. The poling electrodes were patterned using a screen-printing technique with silver paste. For poling the piezoelectric, the poling electric field was 2.5 V/μm under the temperature of 100 °C in 10 min.

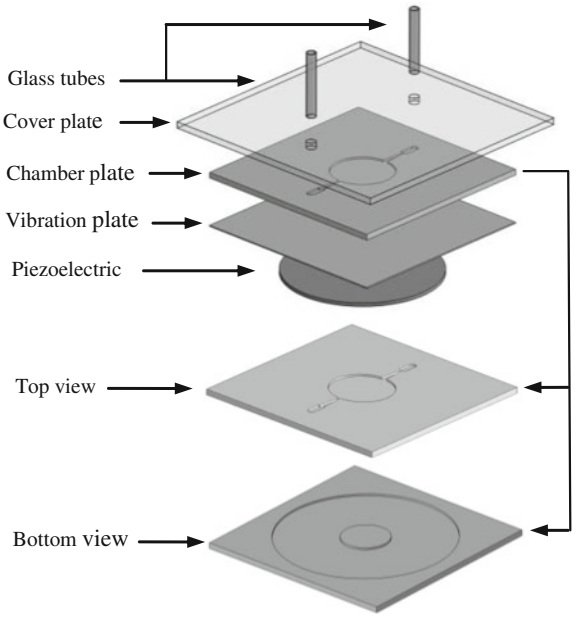


Fig. 1 Schematic of the novel valveless micropump

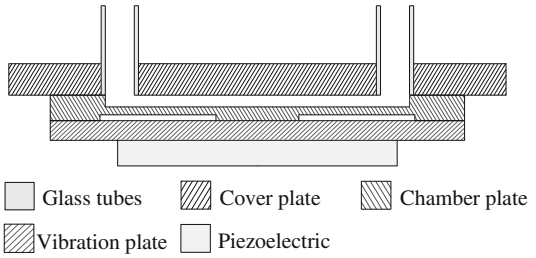


Fig. 2 The cross-section view of micropump

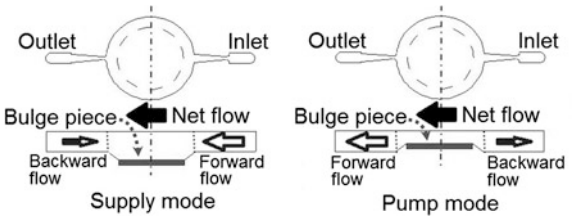


Fig. 3 The diagrams of operational principle in supply and pump mode

3.2 Chamber and Vibration Plate

The chamber and vibration plate were made of the stainless steel manufactured using the MEMS-based lithography and etching process. The chamber plate mainly included a pumping chamber (on the front side) and a bulge-piece diaphragm (on

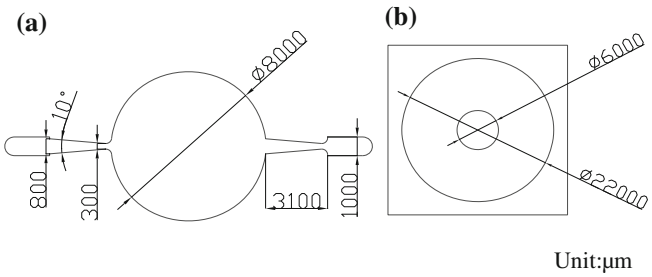


Fig. 4 The dimensions of **a** pumping chamber, and **b** bulge-piece of the micropump

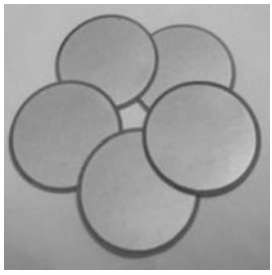


Fig. 5 The fabricated piezoelectric disc

the back side) made of a stainless-steel substrate (in 25 × 25 mm^2) left after wet etching processes. An etchant having 46-g ferric chloride ($FeCl_3$), 33-g hydrogen peroxide (HCl) and 54-g de-ionized (DI) water was used to obtain smooth uniform flow channels on stainless-steel substrates. At the start, the AZ 9260 photoresist was coated on the stainless-steel substrate by spin coater with both spreading step and thinning step. The photoresist on the substrate was baked on a hot plate or in an oven, and then exposed by a standard UV mask aligner (Karl Suess MA-6). The UV exposure process was performed under the hard contact mode with an intensity of 6 mW/cm^2 at a wavelength of 365 nm. The exposed photoresist was then developed in an immersion process via AZ400 K diluted developer. Finally, the samples that were wet etched were immersed in the etchant at 53–58 °C. Figure 6 presents a simple overview of the major steps performed in the fabrication procedure (not to scale). Figure 7a, b illustrate the schematic diagram and pictures of a vibration and a chamber plate, respectively. The chamber plate mainly included a pumping chamber (on the front side) and a 4-mm bulge-piece diaphragm (on the back side). Figure 8 is the SEM picture of the chamber plate (close view of diffuser channel).

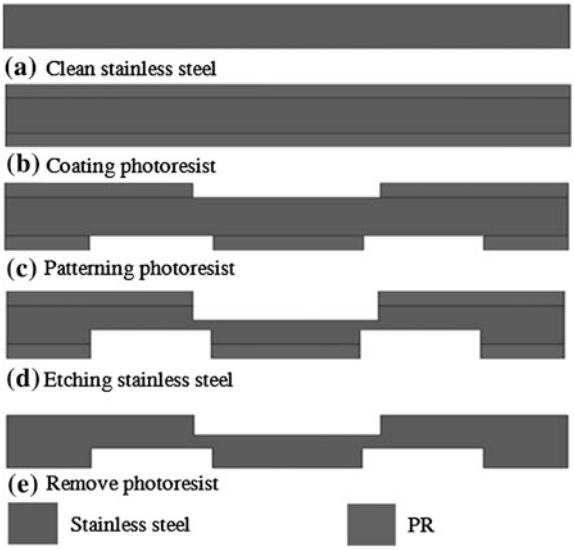


Fig. 6 Fabrication process of an etching stainless-steel micropump (not to scale)

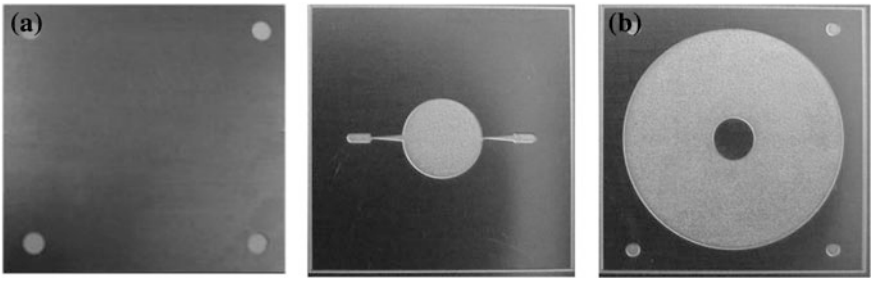


Fig. 7 schematic diagram and pictures of **a** a vibration, and **b** a chamber plate

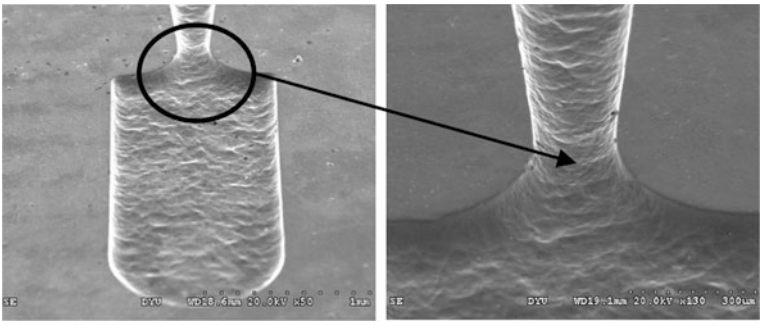


Fig. 8 SEM picture of the chamber plate (close view of diffuser channel)

3.3 Assembly

At first, applying epoxy adhesive (CIBA-GEIGY, AV 119) on the attached surfaces by screen printing, two components (chamber and vibration plate) with aligned marks were assembled by a CCD aligning system. The adhesive was cured in the oven kept at 120 °C for 2 h. Then, the piezoelectric actuator was attached by another epoxy adhesive (3 M, DP-460) cured at a lower temperature of 60 °C for 2 h to avoid depolarization. Finally, the inlet and outlet tubes were connected to the micropump inlet and outlet ports. The instant adhesive (LOCTITE, 403) was also employed to suppress permeating of epoxy adhesive into the glass tube and acrylic gap. Figure 9 shows the photo of the assembled micropump device.

4 Experimental Apparatus

The actuated displacements of the samples were measured by the 2-dimensional scanning laser vibrometer (Polytec MSV300), as displayed in Fig. 10. Each obtained data is the average value from the three measured samples with the same conditions. Figure 11 exhibits the schematic diagram of the experimental apparatus for flow rate measurements of the valveless micropump. Two silicone tubes were connected to the inlet and outlet of a micropump during the experiments. The volumetric flow rates were measured via reading the moving distance of the DI water column in the silicone tube per unit time. In the meantime, and the volumetric flow rates were determined from the mass change of the outlet reservoir [by a precision electronic balance (Precisa XS 365 M)] divided by the water density. The measurements were also conducted under various back pressures varied by lifting up the height of the downstream tube. In the tests, the piezoelectric disc was operated at the driving sinusoidal voltage of 160 V_{pp} and frequency ranging within 100–550 Hz from an electrical signal controller (Agilent 33120A function generator and Trek 50 Amplifier) with the electrical signals verified by an oscilloscope (Agilent 54622A). As the micropump was filled with DI water, gas bubble trapping was carefully eliminated to circumvent the inaccuracy in flow-rate measurements.

5 Theoretical Analysis

Simulations were conducted using the ESI-CFD ACE+® computer software to examine the internal flow field inside a micropump. The theoretical model was based on the transient, three-dimensional continuity and Navier-Stokes equations for incompressible laminar flows with a negligible temperature variation over the computational domain. The governing equations are stated as below.

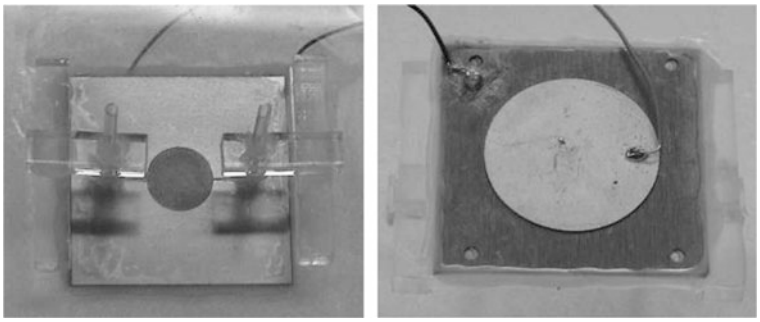


Fig. 9 The assembled valveless micropump

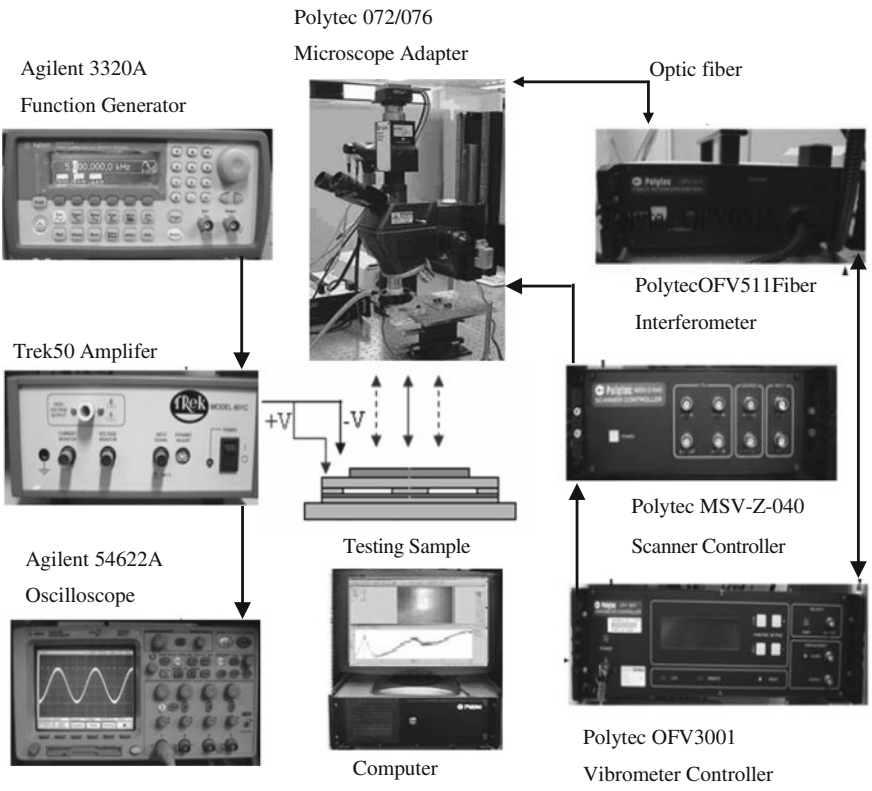


Fig. 10 Configuration of instruments for measuring the actuated displacement of the piezoelectric actuator

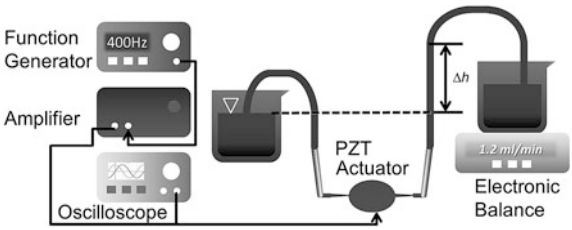


Fig. 11 Schematic diagram of the experimental setup for flow rate measurements of the micropump

$$\nabla \cdot \vec{V} = 0. \tag{1}$$

$$\frac{\partial}{\partial t}(\rho \vec{V}) + \rho(\vec{V} \cdot \nabla)\vec{V} = -\nabla p + \rho \vec{g} + \mu \nabla^2 \vec{V}. \tag{2}$$

The symbol $\vec{V}$ is the velocity vector; whereas, p, ρ and μ represent the pressure, density, and dynamic viscosity of the fluid. The term $\rho\vec{g}$ denotes the gravitational force. In this study, the ambient pressure outside the micropump was 1 atm. The no-slip condition and a zero normal pressure gradient were imposed on the solid walls of the micropump. Using the pure water as the working fluid in the experiments, the resultant values of the density and viscosity were 997 kg/m^3 and 8.55×10^{-4} N-s/m^2, respectively. The motion of the vibrating bulge-piece diaphragm was treated as the moving boundary under an applied driving voltage signal by prescribing the axial displacement on the diaphragm surface. The corresponding displacement was determined through the finite-element computer software ANSYS® to set the properties of the piezoelectric and stainless-steel materials as Young's moduli of 9.1×10^{10} and 2.13×10^{11} N/m^2, Poisson's ratios of 0.33 and 0.3, and densities of 7,900 and 7,780 kg/m^3. When considering a 6-mm-diameter and 70-μm thick bulge-piece diaphragm, a fixed boundary condition of the diaphragm was employed in the ANSYS® transient investigation for a case in which air and water were filled from two sides. The predictions showed that the displacement responded as a sinusoidal waveform having the same frequency of 400 Hz. The computed displacements were verified with the measured data from a Polytec scanning vibrometer (Polytec-MSV300®). The ANSYS® simulations and measurements indicated that the peak-to-peak amplitude of the bulge-piece diameters of 2, 4 and 6 mm at the voltage and frequency of 160 V_{pp} and 400 Hz were 10.8, 10.2 and 9.8 μm (5.4/5.1/4.9 μm outward and 5.4/5.1/4.9 μm inward), respectively. The simulations were carried out using the user-defined-function (UDF) module in the ESI-CFD ACE+® software for prescribing the moving boundary of the bulge-piece diaphragm displacement and inlet/outlet boundary conditions. In calculations, the zero gauge pressure was specified at the inlet with the measured back pressure (P_b) values ranging from 0 to 5.3 kPa set at the outlet, and the volumetric flow rates were computed by averaging the cyclical volumetric flow rates over one period.

6 Results and Discussion

To simulate the flow behavior of the valveless micropump, Fig. 12 showed the numerical grids of a full-size valveless micropump model. The mesh system included three major structured portions: the pumping chamber, nozzle/diffuser, and inlet and outlet tubes. In constructing the model, finer grids were disposed in the regions near the nozzle and diffuser throats as well as the moving and fixed wall boundaries. The average cell length in the chamber was about 45.3 μm with the smallest spacing of 1.2 μm for resolving the steep variations of flow properties. Calculations were also done on the total grids of 166,672, 191,673 and 216,677 points at CFL = 0.5 and 0.25. During transient calculations, the normalized residual errors of the flow variables (u, v, w and p) converged to 10^{-5} with the mass conservation check within 0.5 % for each time step. The calculated centerline velocity profiles across the pumping chamber at different grids and CFL values indicated that satisfactory grid independence could be achieved using a mesh setup of 191,673 grids with CFL = 0.5. A complete simulation for developing a pumping flow field (with the cycle-to-cycle variation of the delivered volumetric flow rates under 0.2 %) generally requires around 140 h of central processing unit (CPU) time on an Intel-Core2 Duo E6750-2.66 GHz personal computer.

To validate the present theoretical model, numerical calculations were conducted using the ESI-CFD ACE+® software by comparing the predictions with experimental data. The performance of a piezoelectric valveless micropump is essentially dependent on the frequency and voltage of excitation signals for the same geometric configuration. Figure 13 illustrates the measured and computed volumetric flow rates of the bulge-piece diameters of 2, 4 and 6 mm for (a) varied frequencies from 100 to 550 Hz at a zero back pressure, and (b) different back pressures at the frequency of 400 Hz with the sinusoidal voltage of 160 V_{pp} applied. A precision electronic balance was also used to compute the volumetric flow rates from the mass variations of the outlet reservoir. Both the predictions and measurements indicate that the maximum volumetric flow rate was 1.2 ml/min at the driving frequency of 400 Hz. A decay of volumetric flow rate was observed when the input frequency was shifted away from the optimal frequency. The maximum difference between the predictions and experiment results was below 9.1 %, showing that the CFD code can simulate the pumping process of piezoelectric valveless micropumps with a reasonable accuracy. Moreover, the maximum volumetric flow rates were substantially increased by 50 and 75.1 % when the bulge-piece diameters were enlarged from 2 to 4 mm and 4 to 6 mm, respectively.

In practice, a piezoelectric valveless micropump was operated to deliver a specified level of flow rate at a desired pressure head. Figure 13b illustrates the volumetric flow rates with respect to the back pressures for the micropumps having different bulge-piece diaphragms. In the experimental process, the volumetric flow rates were measured for different back pressures varied by changing the height of

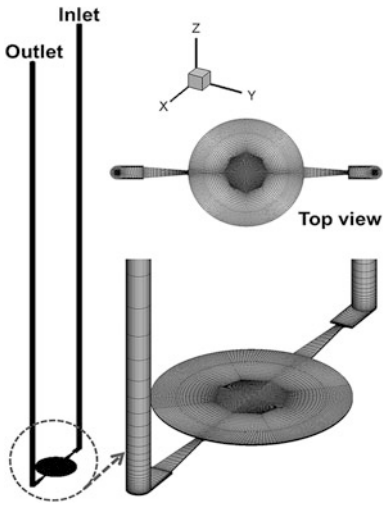


Fig. 12 Numerical grids of the valveless micropump: stereogram and *top view*

an outlet tube. The maximum pumping back pressure at a zero volumetric flow rate was determined by directly measuring the fluid height difference between the inlet and outlet tubes. Both the measured and simulated results showed that the volumetric flow rate decreased nearly linearly with the back pressure at the frequency of 400 Hz with the sinusoidal voltage of 160 V_{pp} applied. Enlargement of the bulge-piece diameter can essentially augment either the maximum volumetric flow rate or the maximum back pressure. The maximum back pressures were substantially increased by 23.3 and 32.9 % when the bulge-piece diameters were enlarged from 2 to 4 mm and 4 to 6 mm, respectively.

Figure 14 illustrates the volumetric flow rates with respect to the back pressures for the micropumps having different voltages at 6 mm bulge-piece diaphragms. Their pump characteristics are similar phenomenon, but their values have some different. Figure 14 shows the volumetric flow rate gradually rises as voltages increases in a good designed pump. The resultant maximum volumetric flow rate and back pressure were 1.59 ml/min and 8.82 kPa for a 6-mm bulge-piece diameter at the driving voltage and frequency of 240 V_{pp} and 400 Hz.

The position accuracy of a feed drive system was primarily influenced by the thermal deformation of a ball screw. A high-speed ball screw system can generate vast heat after long-term operations, with greater thermal expansion formed, and thereby negatively impact the positioning accuracy of the feed drive mechanism. In this research, the computational approach was applied using the FEM to simulate the thermal expansion development for solving the deformation of a ball screw. In simulations, we implemented the multi-zone heat loads to treat the heat generation sources from the frictions between the nut, bearings and the ball screw shaft to emulate reciprocating movements of the nut at a top speed of 40 m/min relative to the shaft in a time period of 3.43 s. We also employed a three-dimensional unsteady heat conduction equation to determine the steady and time-

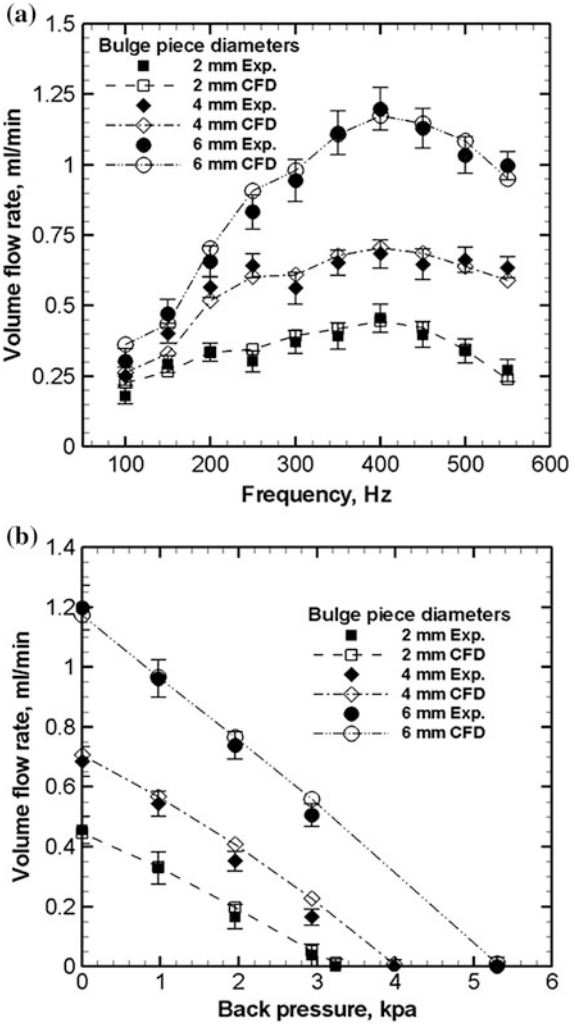


Fig. 13 Measured and computed volumetric flow rates of the bulge-piece diameters of 2, 4 and 6 mm for **a** varied frequencies from 100 to 550 Hz at zero back pressure, and **b** different back pressures at the frequency of 400 Hz with the sinusoidal voltage of 160 V_{pp} applied

dependent temperature distributions, as well as the temperature increases for calculating the thermal deformations of the screw shaft. Simulations were extended to consider the composite operating conditions involving various spindle speeds and moving spans of the nut on the temperature rises and thermal deformations of a ball screw shaft. Both the FEM-based simulations and measurements found that the thermal deformations increased with the axial distance. The associated deformations can be up to 152 μm at 0.8 m in composite operating situations, and in turn depreciated the positioning accuracy. The computational and experimental results also indicated that the significant deterioration of the positioning accuracy due to massive heat production at high speeds of a shaft must be thermally compensated for a ball screw system in operations.

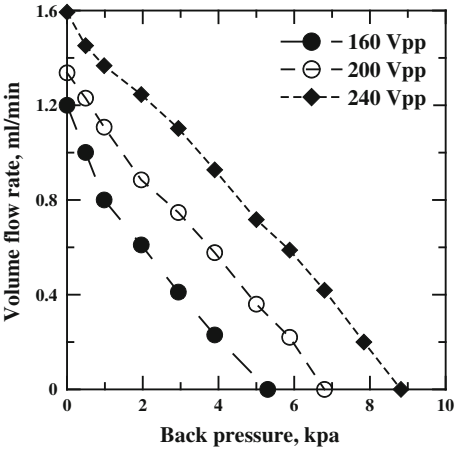


Fig. 14 Measured and computed volumetric flow rates of the bulge-piece diameters of and 6 mm for different back pressures at the frequency of 400 Hz with the sinusoidal voltage of 160, 200, 240 V_{pp} applied

7 Conclusion

The proposed piezoelectric valveless micropump with an integrated diffuser/nozzle bulge-piece design was developed, fabricated and tested to demonstrate an effective pumping performance. The micropump consisted of a stainless-steel structured chamber to strengthen its long-term reliability, low-cost production, and maximized liquid compatibility. The resultant maximum volumetric flow rate and back pressure were 1.59 ml/min and 8.82 kPa for a 6-mm bulge-piece diameter at the driving voltage and frequency of 240 V_{pp} and 400 Hz. When the back pressure was set at 3.9 kPa, the pump could still deliver a volumetric flow rate of 0.92 ml/min. The flow pattern inside the pumping chamber was examined via numerical calculations and experimental observations. We have investigated the time-recurring flow behavior in the chamber and its relationship to the pumping performance. It was found that the micropump with the 6-mm bulge-piece diameter produced a higher suction velocity, and led to a larger upstream and downstream vortex pair during the supply phase, as compared to the outcomes for the bulge-piece diameters of 2 and 4 mm.

Acknowledgment This paper represents part of the results obtained under the support of the National Science Council, Taiwan, ROC (Contract No. NSC101-2221-E-212-002-).

References

1. C. Yamahata, C. Vandevyver, F. Lacharme, P. Izewska, H. Vogel, R. Freitag, M.A.M. Gijs, Pumping of mammalian cells with a nozzle-diffuser micropump. Lab Chip **5**, 1083–1088 (2005)
2. S.S. Wang, X.Y. Huang, C. Yang, Valveless micropump with acoustically featured pumping chamber. Microfluid. Nanofluid. **8**, 549–555 (2010)

3. H.K. Ma, H.C. Su, J.Y. Wu, Study of an innovative one-sided actuating piezoelectric valveless micropump with a secondary chamber. Sens. Actuators, A **171**, 297–305 (2011)
4. L.S. Jang, W.H. Kan, Peristaltic piezoelectric micropump system for biomedical applications. Biomed. Microdevices **9**, 619–626 (2007)
5. C.Y. Lee, H.T. Chang, C.Y. Wen, A MEMS-based valveless impedance pump utilizing electromagnetic actuation. J. Micromech. Microeng. **18**, 035044 (2008)
6. Y.J. Yang, H.H. Liao, Development and characterization of thermopneumatic peristaltic micropumps. J. Micromech. Microeng. **19**, 025003 (2009)
7. N.T. Nguyen, X.Y. Huang, T.K. Chuan, MEMS-micropumps. J. Fluids. Eng. Trans. ASME **124**, 384–392 (2002)
8. D.J. Laser, J.G. Santiago, A review of micropumps. J. Micromech. Microeng. **14**, R35–R64 (2004)
9. B.D. Iverson, S.V. Garimella, Recent advances in microscale pumping technologies: a review and evaluation. Microfluid. Nanofluid. **5**, 145–174 (2008)
10. M. Nabavi, Steady and unsteady flow analysis in microdiffusers. Microfluid. Nanofluid. **7**, 599–619 (2009)
11. G. Fuhr, T. Schnelle, B. Wagner, Travelling wave-driven microfabricated electro-hydrodynamic pumps for liquids. J. Micromech. Microeng. **4**, 217–226 (1994)
12. J.C.T. Eijkel, C. Dalton, C.J. Hayden, J.P.H. Burt, A. Manz, A circular AC magnetohydrodynamic micropump for chromatographic applications. Sens. Actuators, A **92**, 215–221 (2003)
13. L. Chen, H. Wang, J. Ma, C. Wang, Y. Guan, Fabrication and characterization of a multi-stage electroosmotic pump for liquid delivery. Sens. Actuators, B **104**, 117–123 (2005)
14. J. Xie, Y.N. Miao, J. Shih, Q. He, J. Liu, Y.C. Tai, T.D. Lee, An electrochemical pumping system for on-chip gradient generation. Anal. Chem. **76**, 3756–3763 (2004)
15. M.M. Teymoori, E. Abbaspour-Sani, Design and simulation of a novel electrostatic peristaltic micromachined pump for drug delivery applications. Sens. Actuators, A **117**, 222–229 (2005)
16. G.H. Feng, F.S. Kim, Micropump based on PZT unimorph and one-way parylene valves. J. Micromech. Microeng. **14**, 429–435 (2004)
17. P. Dario, N. Croce, M.C. Carrozza, G. Varallo, A fluid handling system for a chemical microanalyzer. J. Micromech. Microeng. **6**, 95–98 (1996)
18. W.K. Schomburg, J. Vollmer, B. Bustgens, J. Fahrenberg, H. Hein, W. Menz, Microfluidic components in LIGA technique. J. Micromech. Microeng. **4**, 186–191 (1994)
19. Y. Yang, Z. Zhou, X. Ye, X. Jiang, in *Bimetallic Thermally Actuated Micropump*, vol. 59. American Society of Mechanical Engineers, Dynamic Systems and Control Division (Publication) DSC, (1996), pp. 351–354
20. E. Makino, T. Mitsuya, T. Shibata, Fabrication of TiNi shape memory micropump. Sens. Actuators, A **88**, 256–262 (2001)
21. W.Y. Sim, H.J. Yoon, O.C. Jeong, S.S. Yang, A phase change type of micropump with aluminum flap valves. J. Micromech. Microeng. **13**, 286–294 (2003)
22. E. Stemme, G. Stemme, A valve-less diffuser/nozzle based fluid pump. Sens. Actuators, A **39**, 159–167 (1993)
23. C.H. Cheng, C.K. Chen, in *WCE 2013*: *Characteristic Studies of the Piezoelectrically Actuated Valveless Micropump*. Proceedings of the World Congress on Engineering 2013. Lecture Notes in Engineering and Computer Science (London, 3–5 July 2013), pp. 1785–1790
24. B. Fan, G. Song, F. Hussain, Simulation of a piezoelectrically actuated valveless micropump. Smart Mater. Struct. **14**, 400–405 (2005)
25. L.S. Jang, Y.C. Yu, Peristaltic micropump system with piezoelectric actuators. Microsyst. Technol. **14**, 241–248 (2008)
26. J. gawa, I. Kanno, H. Kotera, T. Suzuki, Development of liquid pumping devices using vibrating microchannel walls. Sens. Actuators, A **152**, 211–218 (2009)

27. F.K. Forster, R.L. Bardell, M.A. Afromowitz, N.R. Sharma, A. Blanchard, Design, fabrication and testing of a fixed-valve micropump. IMECE FED **234**, 39–44 (1995)
28. C. Morris, F. Forster, Low-order modeling of resonance for fixed-valve micropumps based on first principles. J. Microelectromech. Syst. **12**, 325–334 (2003)
29. Y.Y. Tsui, S.L. Lu, Evaluation of the performance of a valveless micropump by CFD and lumped-system analyses. Sens. Actuators, A **148**, 138–148 (2008)

Role of Offset Factor in Offset-Halves Bearing

Amit Chauhan

Abstract The present paper aims at presenting the influence of offset factor on the thermal characteristics of offset-halves journal bearing profiles. To study the effect of offset factor, an attempt has been made to the solve the Reynolds's equation and energy equation using finite difference method on the performance characteristics such as oil-film temperatures, thermal pressures, load carrying capacity, power loss, and Sommerfeld's number of offset-halves journal bearing. To carry the numerical solution of Energy equation, the author has used Parabolic Temperature Profile Approximation technique in order to achieve faster computation. It has been observed that at constant radial clearance, a decrease in oil-film temperature, thermal pressure, load carrying capacity and an increase in Sommerfeld's number obtained with an increase in offset factor. During the study, it has been concluded that the offset factor significantly affects the performance parameters of the bearing under study and may be kept equal to around 0.4 to get the desired thermal characteristics.

Keywords Load carrying capacity · Offset-halves bearing · Offset factor · Oil-film temperature · Sommerfeld number · Thermal pressure

Nomenclature

e	Eccentricity, m
O_B	Bearing centre
O_j	Journal centre
O_L	Lower-lobe centre
O_U	Upper-lobe centre
P	Film pressure, Pa
R	Journal radius, mm

A. Chauhan (✉)
Department of Mechanical Engineering, UIET, Panjab University, Chandigarh 160014, India
e-mail: drchauhan98@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_9,

r	Bush radius, mm
T	Lubricating film temperature, °C
U	Relative velocity between journal and bearing surface, m/s
u, w	Velocity components in X- and Z-directions, m/s
u_L	Velocity of lower bounding surface, m/s
u_U	Velocity of upper bounding surface, m/s
θ	Angle measured from horizontal split axis in direction of rotation
ω	Angular velocity of shaft, rad/s
φ	Attitude angle
ε	Eccentricity ratio

1 Introduction

Hydrodynamic journal bearings are extensively used in high speed running machinery. These journal bearings operate under hydrodynamic lubrication regime in which a thick film of lubricant separates the bounding surfaces. Under the normal operating conditions, hydrodynamic journal bearings usually experiences a considerable variation in oil-film temperature due to viscous heat dissipation. This rise in oil-film temperature significantly affects the bearing performance since the lubricant viscosity is a strong function of temperature and, consequently, can lead to failure of the bearing. Computation of oil-film temperature is of great importance to predict bearing performance parameters. The non-conventional journal bearings, like lobed bearings and tilting pad bearings have a common feature that these bearings operate with more than one active oil film which accounts for superior stiffness and damping characteristics of these bearings as compared to the conventional circular bearings. The importance of thermal characteristics of such bearings has been explained in the number of studies reported in Refs. [1, 2, 3, 4]. The offset-halves journal bearing is commonly used as a lobed bearing in which two lobes are obtained by orthogonally displacing the two halves of a cylindrical bearing (Fig. 1) and frequently used in gear boxes connecting turbine and generator in power generation industries. In the design of offset-halves halves bearing, offset factor play an important role and is defined as the ratio of the minimum clearance to the radial clearance. The offset-halves journal bearing under study may also be called as horizontally displaced journal bearing since the two halves are horizontally shifted.

Offset-halves journal bearings show stiffness and damping properties which permit light loads at high rotational speeds [1]. The importance of thermal effects in hydrodynamic journal bearings has been long recognized, but very limited study about thermal effects in lobed bearing especially offset-halves bearing has been reported in literature. Chauhan et al. [2] has carried out a comparative study for rise in oil-film temperatures, thermal pressures and load capacity for three different commercially available grade oils.

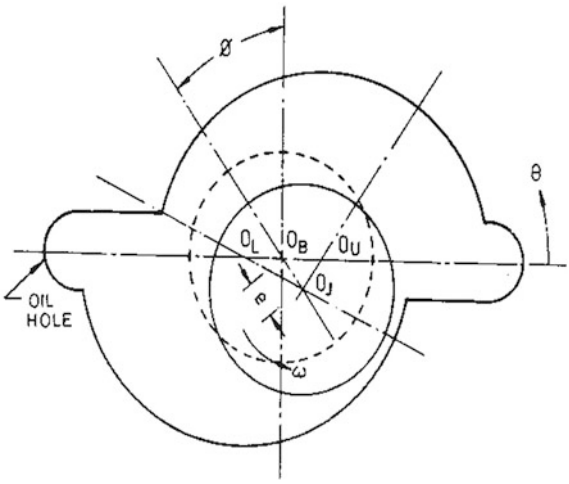


Fig. 1 Schematic diagram of offset-halves journal bearing

The authors have reported that with increase in speed, oil-film temperature, thermal pressure and load carrying capacity rises for all grade oils under study. Also, Chauhan et al. [3] have analyzed thermal performance of elliptical and offset-halves bearings by solving energy equation while assuming parabolic temperature profile approximation across the fluid film. Authors have been found that offset-halves journal bearing runs cooler with minimum power loss and good load capacity. Booker and Govindachar [5] have compared the performance of different bearing configurations namely offset-halves, lemon-bore, three-lobe and four-lobe bearing at the same load capacity and speed. During the comparison, the authors have considered the steady state and stability characteristics.

An attempt was made by Suganami and Sezri [6] to formulate a thermohydrodynamic model of film lubrication which is valid in both laminar and super-laminar flow regimes. The authors stated that energy equation retains heat conduction in direction of sliding motion, and is applicable even at large eccentricities. Boncompain et al. [7] have presented a general thermohydrodynamic theory. The authors have solved generalized Reynolds equation, energy equation in film and heat transfer equation in bush and shaft simultaneously. Read and Flack [8] have developed a test apparatus on which an offset-halves journal bearing of 70 mm diameter journal was tested at five vertical loads and two rotational speeds. Indulekha et al. [9] have solved three dimensional momentum and continuity equations, and three dimensional energy equations to obtain pressure, velocity and temperature field in the fluid of a hydrodynamic circular journal bearing using finite element method. Authors have computed attitude angle, end leakage and power loss, for a wide range of eccentricity ratios. Hussain et al. [10] have predicted temperature distribution in non-circular journal bearings (namely two-lobe, elliptical and orthogonally displaced). The work is based on a two-dimensional treatment following Mc Callion's approach (an approach in which Reynolds and energy equations in oil-film are decoupled by neglecting all pressure terms in

energy equation). Sehgal et al. [11] have presented a comparative theoretical analysis of three types of hydrodynamic journal bearing configurations namely: circular, axial groove, and offset-halves. It has been observed by the authors that the offset-halves bearing runs cooler than an equivalent circular bearing with axial grooves. A computer-aided design of hydrodynamic journal bearing is provided considering thermal effects by Singh and Majumdar [12]. In this design, Reynolds equation has been solved simultaneously along with energy equation and heat conduction equations in bush and shaft to obtain steady-state solution and a data bank is generated that consists of load, friction factor and flow rate for different L/D and eccentricity ratios. Sharma and Pandey [13] have carried out a thermohydrodynamic lubrication analysis of infinitely wide slider bearing assuming parabolic and Legendre polynomial temperature profile across film thickness. Authors have observed that temperature profile approximation across the film thickness by Legendre Polynomial yields more accurate results in comparison to Parabolic Temperature Profile approximation (PTPA).

In present paper, the thermal studies of offset-halves have been presented using thermohydrodynamic approach. Variation in Offset factor has been carried out while evaluating oil-film temperatures, thermal pressures, load carrying capacity, power loss, and Sommerfeld's number in the fluid film of an offset-halves journal bearing for the Mak Multigrade oil using PTPA. Mak Multigrade oil is recommended for use in heavy duty commercial vehicles, light commercial vehicles and multi-utility vehicles fitted with high speed naturally aspirated or turbo charged diesel engines operating at low speed and high torque conditions [2, 14].

2 Governing Equations

Reynolds equation:

For steady-state and incompressible flow, Reynolds equation is [10]:

$$\frac{\partial}{\partial x}\left(\frac{h^3}{\mu}\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial z}\left(\frac{h^3}{\mu}\frac{\partial p}{\partial z}\right)=6U\frac{\partial h}{\partial x} \tag{1}$$

This equation is then set into finite differences by using central difference technique. The final form is reproduced here.

$$P(i,j)_{iso}=A1P(i+1,j)_{iso}+A2P(i-1,j)_{iso}+A3P(i,j+1)_{iso}+A4P(i,j-1)_{iso}-A5$$
$$P(i,j)_{th}=E11P(i+1,j)_{th}+E22P(i-1,j)_{th}+E33P(i,j+1)_{th}+E44P(i,j-1)_{th}-E55$$

The equations above give isothermal pressures and thermal pressures respectively. The coefficients appearing in these equations have been given here as: where,

$$A11 = \left[\frac{h^3}{d\theta^2} + \frac{3h^2}{2d\theta}\frac{\partial h}{\partial \theta}\right]; \quad A22 = \left[\frac{h^3}{d\theta^2} - \frac{3h^2}{2d\theta}\frac{\partial h}{\partial \theta}\right]$$

$$A33 = \left[\frac{r^2h^3}{dz^2} + \frac{3r^2h^2}{2dz}\frac{\partial h}{\partial z}\right]; \quad A44 = \left[\frac{r^2h^3}{dz^2} - \frac{3r^2h^2}{2dz}\frac{\partial h}{\partial z}\right];$$

$$A55 = \left[6Ur\mu\frac{\partial h}{\partial \theta}\right]; \quad A = \left[\frac{2h^3}{d\theta^2} + \frac{2r^2h^2}{dz^2}\right];$$

$$A1 = A11/A; \quad A2 = A22/A; \quad A3 = A33/A; \quad A4 = A44/A; \quad A5 = A55/A$$

$$A6 = \frac{h^3}{\mu}; \quad A7 = \frac{\partial A6}{\partial \theta}; \quad A8 = \frac{\partial A6}{\partial z}; \quad A9 = \frac{2A6}{d\theta^2} + \frac{2r^2A6}{dz^2}; \quad F = \frac{A7}{2d\theta};$$

$$G = \frac{A6}{d\theta^2}; \quad B = \frac{\partial h}{\partial \theta}; \quad H = \frac{r^2A8}{2dz}; \quad H1 = \frac{r^2A6}{dz^2};$$

$$H2 = 6UrB; \quad E11 = \frac{F+G}{A9}; \quad E22 = \frac{-F+G}{A9}; \quad E33 = \frac{H+H1}{A9};$$

$$E44 = \frac{-H+H1}{A9}; \quad E55 = \frac{-H2}{A9}$$

The same forms have also been presented by the author in [2–4, 14]. The variation of viscosity with temperature and pressure has been simulated using following relation:

$$\mu = \mu_{ref} e^{\alpha P - \gamma(T - T_0)} \tag{2}$$

Energy Equation:

The energy equation for steady-state and incompressible flow is given as [13]:

$$\rho C_P \left(u\frac{\partial T}{\partial x} + w\frac{\partial T}{\partial z}\right) = \frac{\partial}{\partial y}\left(K\frac{\partial T}{\partial y}\right) + \mu\left[\left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2\right] \tag{3}$$

The term on left hand side in above equation represents energy transfer due to convection, whereas first term on right hand side represents energy transfer due to conduction and second term on right hand side represents energy transfer due to dissipation. The variation of temperature across film thickness in Eq. (3) is approximated by parabolic temperature profile for faster computation of temperatures [13]. The temperature profile across film thickness is represented by a second order polynomial as:

$$T = a_1 + a_2 y + a_3 y^2 \tag{4}$$

In order to evaluate constants appearing in Eq. (4), following boundary conditions are used:

$$\text{At } y = 0, T = T_L,$$

$$T_m = \frac{1}{h}\int_0^h T dy$$

$$\text{At } y = h, T = T_U,$$

Thus, temperature profile expression (written in Eq. 4) takes the following form:

$$T = T_L - (4T_L + 2T_U - 6T_m)\left(\frac{y}{h}\right) + (3T_L + 3T_U - 6T_m)\left(\frac{y}{h}\right)^2 \tag{5}$$

where, T_L, T_U, and T_m represent temperatures of lower bounding surface, upper bounding surface, and mean temperature across film respectively.

Substitution of '*u*', '*w*', and '*T*' expressions (Eqs. 4–6) in to energy Eq. (3) and subsequently integrating energy equation across film thickness from limit '0' to 'h' yields following form of energy equation.

$$\begin{aligned} &6T_L + 6T_U - 12T_m - \frac{\rho C_P h^4}{120K\mu}\frac{\partial P}{\partial x}\begin{pmatrix}\frac{\partial T_L}{\partial x} + \frac{\partial T_U}{\partial x}\\ -12\frac{\partial T_m}{\partial x}\end{pmatrix} - \frac{\rho C_P h^4}{120K\mu}\frac{\partial P}{\partial z}\\ &\left(\frac{\partial T_L}{\partial z} + \frac{\partial T_U}{\partial z} - 12\frac{\partial T_m}{\partial z}\right) - \frac{\rho C_P h^2(u_L + u_U)}{2K}\frac{\partial T_m}{\partial x} - \frac{\rho C_P h^2(u_U - u_L)}{12K}\\ &\left(\frac{\partial T_U}{\partial x} - \frac{\partial T_L}{\partial x}\right) + \frac{h^4}{12K\mu}\left[\left(\frac{\partial P}{\partial x}\right)^2 + \left(\frac{\partial P}{\partial z}\right)^2\right] + \frac{\mu(u_U - u_L)^2}{K} = 0 \end{aligned} \tag{6}$$

The film thickness (h) equations for offset-halves journal bearing are given as [11]:

$$h = c_m\left[\left(\frac{1+\delta}{2\delta}\right) + \left(\frac{1-\delta}{2\delta}\right)\cos\theta - \varepsilon\sin(\phi - \theta)\right] \quad (0 < \theta < 180) \tag{7}$$

$$h = c_m\left[\left(\frac{1+\delta}{2\delta}\right) - \left(\frac{1-\delta}{2\delta}\right)\cos\theta - \varepsilon\sin(\phi - \theta)\right] \quad (180 < \theta < 360) \tag{8}$$

3 Computational Procedure

Numerical solution of Reynolds's and energy equations has been obtained for offset-halves journal bearing through finite difference approach. The temperature of upper and lower bounding surfaces have been assumed constant throughout and have been set equal to oil inlet temperature for first iteration. A suitable initial value of attitude angle is assumed. During solution of Reynolds's equation over relaxation with relaxation factor of 1.7 has been taken for error convergence

whereas under relaxation factor of 0.7 has been taken in numerical solution of energy equation for error convergence. The load carrying capacity is obtained by applying Simpson's rule to pressure distribution. In computation, wherever reverse flow arises in domain, upwind differencing has been resorted. The boundary conditions used in the solution of governing equations are same as reported in Chauhan et al. [2, 4] which are

$$\begin{aligned} &P = 0 \text{ at } x = 0 \quad \text{and} \quad x = l \\ &u = u_L \text{ at } y = 0 \quad \text{and} \quad 0 \le x \le l \\ &u = 0 \text{ at } y = h \quad \text{and} \quad 0 \le x \le l \\ &T = T_0 \text{ at } x = 0 \quad \text{and} \quad 0 \le x \le h \\ &T = T_L \text{ at } y = 0 \quad \text{and} \quad 0 \le x \le l \\ &T = T_U \text{ at } y = h \quad \text{and} \quad 0 \le x \le l \end{aligned}$$

$$T(0,y) = T_0;\ T(x,0) = T_0\ ;\ k_{oil}\left(\frac{\partial T}{\partial y}\right)_{\substack{upper\\ bounding\\ surface}} = k_s\left(\frac{\partial T_s}{\partial y_s}\right)_{y_s=0};$$

$$-k_s\left(\frac{\partial T_s}{\partial y_s}\right)_{y_s=t} = h_c(T_s(x_s,t) - T_a);\ -k_s\left(\frac{\partial T_s}{\partial x_s}\right)_{x_s=0} = h_c(T_s(0,y_s) - T_a);$$

$$-k_s\left(\frac{\partial T_s}{\partial x_s}\right)_{x_s=l} = h_c(T_s(l,y_s) - T_a);\ k_s\left(\frac{\partial T_s}{\partial z_s}\right)_{z_s=0} = h_c(T_s(x_s,y_s,0) - T_a);$$

$$-k_s\left(\frac{\partial T_s}{\partial z_s}\right)_{z_s=b} = h_c(T_s(x_s,y_s,b) - T_a)$$

where K_s denotes thermal conductivity of bearing, h_c denotes convection heat transfer coefficient of bush,l denotes length of the bearing, s denotes bearing surface, t denotes thickness of bearing, b denotes width of bearing, and T_a ambient temperature. The solution of governing equations has been achieved by satisfying the convergence criterion given below:

For pressure:

$$\frac{\left|\left(\sum P_{i,j}\right)_{n-1} - \left(\sum P_{i,j}\right)_n\right|}{\left|\left(\sum P_{i,j}\right)_n\right|} \le 0.0001 \tag{9}$$

For temperature:

$$\frac{\left|\left(\sum T_{i,j}\right)_{n-1} - \left(\sum T_{i,j}\right)_n\right|}{\left|\left(\sum T_{i,j}\right)_n\right|} \le 0.0001 \tag{10}$$

where, n represents number of iterations.

Table 1 Input parameters [14]

Outer diameter of bearing, OD	85 mm
Inner diameter of bearing, ID	65 mm
Length, l	65 mm
Radial clearance, C	500 μm
Minimum clearance, C_m	200 μm
Oil inlet temperature, T_0	33 °C
Ambient temperature, Ta	30 °C
Viscosity, μ	0.200 Pas
Density of oil, ρ	885 Kg/m^3
Barus viscosity-pressure index, α	2.3e-8
Temperature viscosity—coefficient, γ	0.034

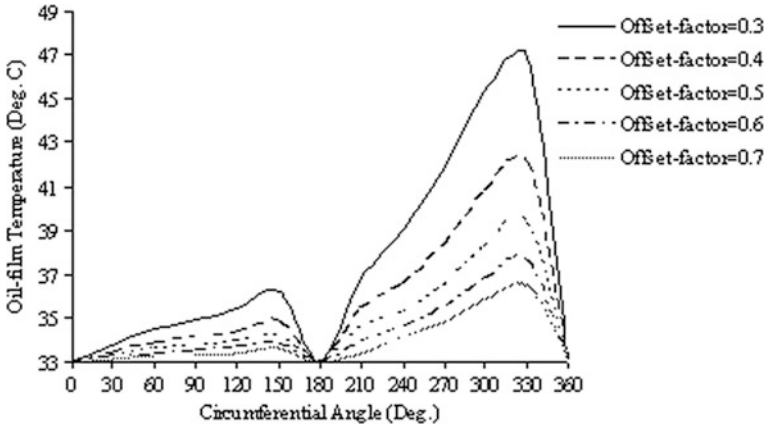


Fig. 2 Variation of oil-film temperature with circumferential angle for Mak Multigrade oil with different offset factor at speed = 5,000 rpm for offset-halves profile bearings

4 Results and Discussion

Input parameters and properties of oil used in computer simulations are given in Table 1. From computer simulation, various performance characteristics have been obtained and are discussed as [14]:

1. The variation of oil-film temperature and thermal pressure with offset factor (which is defined as the ratio of minimum clearance to the radial clearance) has been presented in Figs. 2 and 3 respectively. From Figs. 2 and 3, it has been observed that oil-film temperatures and thermal pressures decrease with increase in offset factor while keeping radial clearance constant which can be well explained with help of Fig. 1. However, oil-film temperatures and thermal pressures have been obtained same with variation in offset factor while keeping minimum clearance constant. Further, values of oil-film temperature and thermal pressures has been observed nearly same when offset factor is 0.4 for

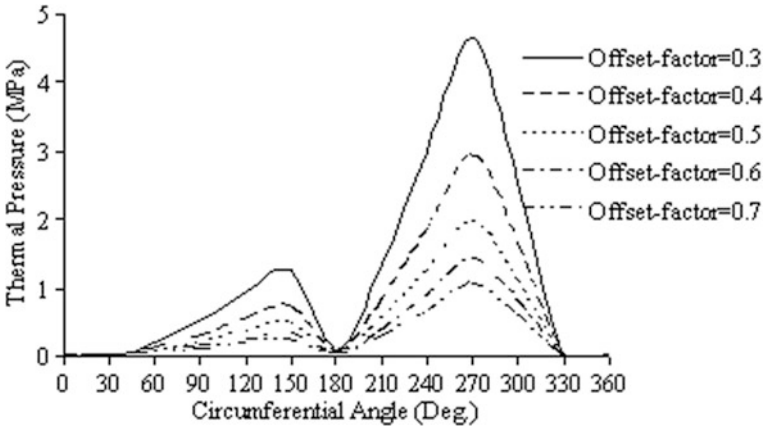


Fig. 3 Variation of thermal pressure with circumferential angle for mak multigrade oil with different offset factor at speed = 5,000 rpm for offset-halves profile bearings

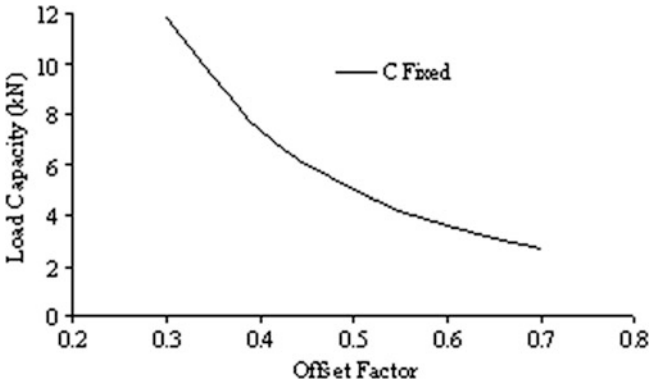


Fig. 4 Variation of load capacity with offset factor for mak multigrade oil at speed = 5,000 rpm for offset-halves profile bearings

both the cases. Hence, in further analysis the radial clearance has been kept constant only.

2. The load carrying capacity (Fig. 4) and power losses (Fig. 5) have been observed of decreasing nature with an increase in offset factor while keeping radial clearance constant. The Sommerfeld's number has been found of increasing nature with an increase in offset factor while keeping radial clearance or minimum clearance constant (Fig. 6). Further, values of load carrying capacity and power losses has been observed same when offset factor is 0.4 while keeping the radial clearance or minimum clearance constant and a slightly higher Sommerfeld's number has been observed when radial clearance is kept constant.

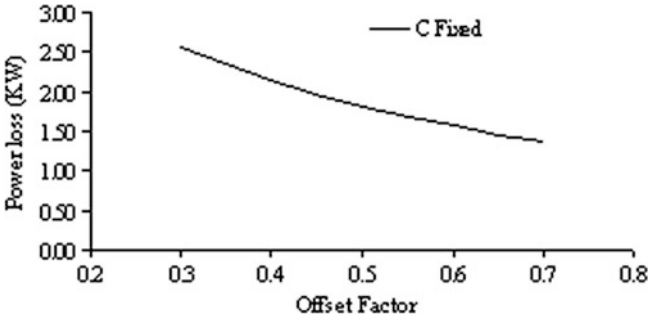


Fig. 5 Variation of power loss with offset factor for mak multigrade oil at speed = 5,000 rpm for offset-halves profile bearings

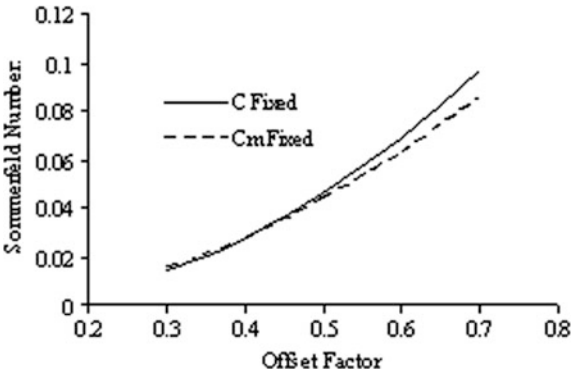


Fig. 6 Variation of sommerfeld's number with offset factor for mak multigrade oil at speed = 5,000 rpm for offset-halves profile bearings

5 Conclusion

The paper deals with the evaluation of different performance parameters of offset-halves journal bearing using various governing equations. The study on the effects of variation in offset factor on various performance characteristics of offset-halves journal bearing has been presented in the paper. From the results and discussion, at constant radial clearance, the oil-film temperatures and thermal pressures are found to decrease with increase in offset factor. However, oil-film temperatures and thermal pressures have been obtained same with variation in offset factor while keeping the minimum clearance constant. Keeping the radial clearance constant, the load carrying capacity and power losses have been observed of decreasing nature with an increase in offset factor. The Sommerfeld's number has been found of increasing nature with an increase in offset factor while keeping the radial clearance or minimum clearance constant. Finally, it has been concluded that during the design stage of offset-halves journal bearing, the value of offset

factor may be kept near to 0.4 to get desired characteristics. The paper presents different performance parameters like oil film temperature, thermal pressures, load carrying capacity, and power losses which will help the designer to design this type of non-circular journal bearings as well as analyze its performance. Presently the available design data/handbooks do not provide any direct analytical methods for the design and analysis of this non-circular journal bearing so the present methodology to a large extent will benefit the designer.

6 Scope for Future Work

The present work can further be extended to standardize the different equations for the offset-halves journal bearing and an attempt to develop the design charts should be made by conducting experimental work to find oil-film temperature and thermal pressure developed in the bearing under normal operating conditions.

References

1. A. Chauhan, R. Sehgal, An experimental investigation of the variation of oil temperatures in offset-halves journal bearing profile using different oils, Indian J. Tribol. **3**(2), 27–41 (2008)
2. A. Chauhan, R. Sehgal, R. K. Sharma, A study of thermal effects in offset-halves journal bearing profile using different grade oils, Lubr. Sci. **23**, 233–248 (2011)
3. A. Chauhan, R. Sehgal, R. K. Sharma, Investigations on the thermal effects in non-circular journal bearings, Tribol. Int. **44**, pp. 1765–1773 (2011)
4. A. Chauhan, R. Sehgal. Thermal studies of non-circular journal bearing profiles: Offset-halves and elliptical, Intech Publishers, (2012) pp. 3–24
5. J. F, Booker, S. Govindachar, Stability of offset journal bearing systems, in *Proceedings of IMechE, C283/84*, (1984) pp. 269–275
6. T. Suganami, A. Z. Sezri, A thermohydrodynamic analysis of journal bearings, J. Lubr. Technol. **101**, 21–27 (1979)
7. R. Boncompain, M. Fillon, J. Frene, Analysis of thermal effects in hydrodynamic bearings, J. Tribol. **108**, 219–224 (1986)
8. L. J. Read, R. D. Flack, Temperature, pressure and film thickness measurements for an offset half bearing, Wear; **117**(2), 197–210 (1987)
9. T. P. Indulekha, M. L. Joy, K. Prabhakaran Nair, Fluid flow and thermal analysis of a circular journal bearing, Warme-und stoffubertragung, **29**, 367–371 (1994)
10. A. Hussain, K. N. Mistry, S. K. Biswas, K. Athre, Thermal analysis of non-circular bearing, Trans ASME, **118**(1), pp 246–254 (1996)
11. R. Sehgal, K. N. S. Swamy, K. Athre, S. Biswas, A comparative study of the thermal behaviour of circular and non-circular journal bearings, Lubr Sci, **12**(4), pp. 329–344 (2000)
12. D. S. Singh, B. C. Majumdar, Computer-aided design of hydrodynamic journal bearings considering thermal effects, in *Proceedings of IMechE, Part J: Journal of Engineering. Tribology*, vol 219, 133–143 (2005)

13. R. K. Sharma, R. K. Pandey, Effects of the temperature profile approximations across the film thickness in thermohydrodynamic analysis of lubricating films, Indian J. Tribol. **2**(1), pp. 27–37 (2007)
14. A. Chauhan, Thermal characteristics of offset-halves bearing with offset factor, lecture notes in engineering and computer science: in *Proceedings of The World Congress on Engineering*, WCE 2013, London, UK, 03–05 July, 2013, pp. 1841–1846 (2013)

Relative Position Computation of Links in Planar Six-Bar Mechanisms with Joints Clearance and Complex Chain

Mohamad Younes and Alain Potiron

Abstract The presence of clearance in the mechanical joints leads to small position variation of the mechanism elements. The goal of this work is to model and analyze the equilibrium positions of elements in planar six-bar mechanisms with complex chain. To solve this subject, it is necessary to use a mathematical optimization code in order to obtain the optimal solution of the problem. To show the effectiveness of the proposed method, examples are presented and the numerical results obtained show that a good convergence was obtained in each case.

Keywords Complex chain · Joint clearance · Loaded link · Mechanical elements position · Optimization · Six-bar mechanism analysis

1 Introduction

The existence of clearance in the joints is necessary to allow the possibility of relative movements in the joints with satisfactory values of contact pressures. However, the presence of clearance induces errors in positioning the various components in the structure.

To minimize these errors, it is essential to take into account the presence of joint clearance for an accurate calculation of the elements' positions. This aims to give an optimal result of the mechanisms' studies.

M. Younes (✉)
Lebanese University, University Institute of Technology, Saida, Lebanon
e-mail: younesmhd@hotmail.com

A. Potiron
Ecole Nationale Supérieure d'Arts et Métiers, Laboratoire LAMPA Arts et Métiers Paris Tech Angers, 2 boulevard du Ronceray, BP 3525, 49035 Angers Cedex, France
e-mail: alain.potiron@sfr.fr

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_10,

Potiron et al. [1] proposed a new method of static analysis in order to determine the arrangement of the various components of planar mechanisms subjected to mechanical loadings. This study concerns the planar mechanism with closed chain and parallel joints. The study takes into account the presence of linkage clearance and allows for the computation of the small variations of the parts position compared to the large amplitude of the movements useful for the power transmission.

It appears that a rather small number of research tasks were carried out in this particular field. Funabashi et al. [2] tackled the problem by carrying out a dynamic, theoretical and experimental study of some simple mechanisms. In order to specify the influence of the clearance in the links on machine operations, they derived the equations of the movement of links including parts stiffnesses, viscous friction and Coulomb's friction in joints. The results are interesting for the specific models suggested but they don't lead to a general usable method suited for the study of complex mechanisms.

A model of mechanism with joint's clearance was defined by Giordano et al. [3] when researching the dimensional and geometrical tolerances associated with machine elements. The method is based on the definition of small rigid-body displacements and the use of closed loops equations for the associated kinematic chains.

To improve the quality of manufactured products and reduce their total cost, Gao et al. [4] and Chase et al. [5] have developed a method for the tolerance analysis of two and three-dimensional mechanical assemblies. This method is carried out by a direct linearization of a geometrical non-linear problem. It was implemented in a commercial C.A.D. code, in order to extract from the results, acceptable tolerances and the dimensions of the related parts. In the same topic, Chase and Parkinson [6] presented an outline on recent research in the analysis of mechanical tolerances, from which it is possible to have an idea of how to handle the study of the joints' clearance in mechanisms.

In the study of Erkaya and Uzmay [7] a dynamic response of mechanism having revolute joints with clearance is investigated. A four-bar mechanism having two joints with clearance is considered as a model of mechanism. A neural network was used to model several characteristics of joint clearance. Kinematic and dynamic analyses were achieved using continuous contact mode between journal and bearing. A genetic algorithm was also used to determine the appropriate values of design variables for reducing the additional vibration effect due primarily to the joint clearance.

Hsieh [8] has proposed a method allowing for the kinematic description of mechanisms containing prismatic, revolute, helical and cylindrical joints. Unfortunately, it cannot be directly applied to mechanical systems containing spherical pairs.

Younes and Potiron [9] presented a method for modeling and analyzing the position of planar six-bar mechanisms with clearance in the joints.

In this work, a method is proposed to analyze the six-bar mechanisms with complex chain. Given a geometrical position, resulting from the great amplitude of

movements in the mechanism, it will be possible to compute the equilibrium positions of the various parts in the six-bar mechanisms with complex chain by taking into account the joint clearance. The main idea is to define and minimize an objective function and to take into account the geometrical constraints imposed by the clearance on infinitely small displacements in the joints.

During these studies, we suppose that the joints in the mechanism are carried out with clearances, the solids are undeformable, the solids are geometrically perfect, i.e. the defects due to the tolerances of forms and of positions are ignored and the gravity force is neglected.

2 Six-Bar Mechanism with Complex Chain

The mechanism is constituted by six bars linked one with the other by a seven simple revolute joints L_i having a clearance joint J_i and an origin O_i ($i = 1,\ldots,7$). To show the influence of the presence of clearances in the mechanism, the scale of these clearances are much large compared with the mechanism dimensions as shown in Fig. 1.

Since the solid S_0 is connected to S_1, S_5 and S_3 and this latter is connected to S_0, S_2 and S_4, the chain mechanism is complex. The joints between the elements of six-bar mechanism are shown in the Fig. 2.

The relative positioning of parts can be reduced to the study of the relative positions of the references associated with each piece of mechanism.

Consider $R_1(O_1, X_1, Y_1, Z_1)$ the fixed reference connected to frame "S_0". The origin O_1 is theoretically the geometric center of the joint L_1 between the two solids "S_0" and "S_1".

The other references $R_i(O_i, X_i, Y_i, Z_i)$ ($i = 2,\ldots,7$) are movable. The point O_i is the geometric center of the joint L_i.

In our work, the abscissa axes X_i ($i = 1,\ldots,7$) are parallel. Also, Y_i axes are parallel.

3 Design Variables of Six-Bar Mechanism

Consider A_i ($i = 1,\ldots,7$) as the points which coincide initially with the origins O_i. If the mechanism is stressed by a mechanical load, the points A_i move into the empty space of clearances joints. In the two-dimensional study and in the fixed coordinate system (O_1, X_1, Y_1, Z_1), each solid of the mechanism has the possibility of two translation along the X_1 and Y_1 axes and rotation relative to the Z_1 axis.

In the local coordinate system (O_i, X_i, Y_i, Z_i) ($i = 1,\ldots,7$), each point A_i has three degrees of freedom. It has the possibility of two displacements u_i and v_i respectively along the X_i and Y_i axes and a rotation γ_i with respect to the Z_i axis. These parameters represent the relative movements of the solid with respect to

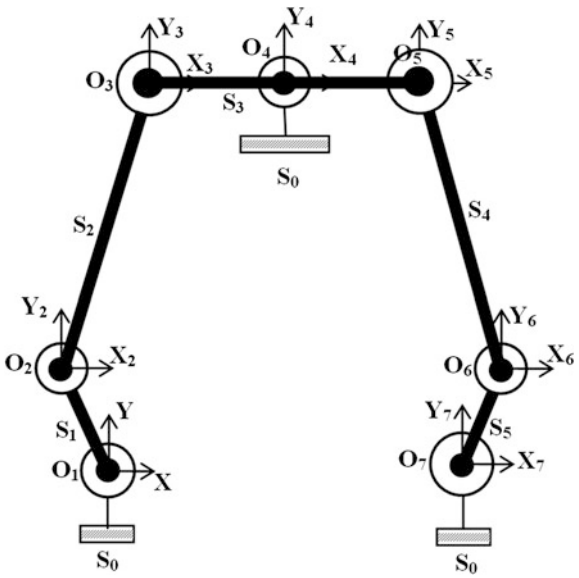


Fig. 1 Six-bar mechanism schematization

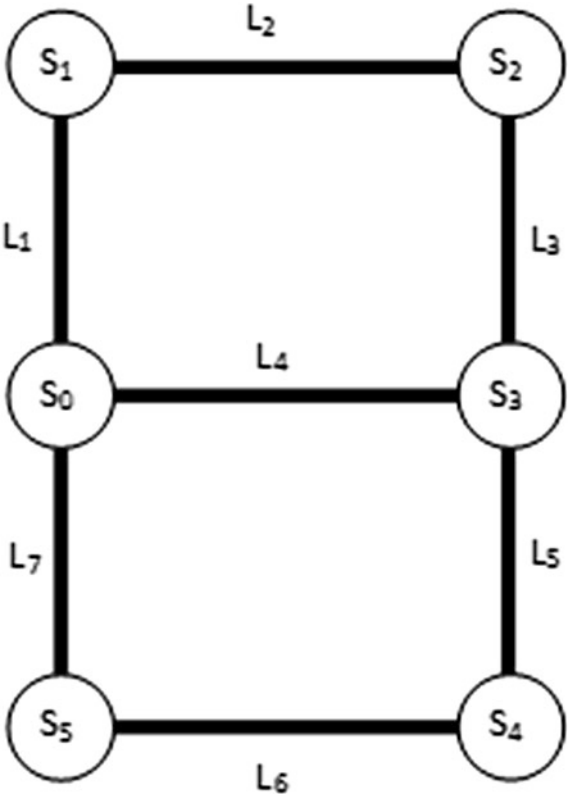


Fig. 2 Joints between the elements of six-bar mechanism

each other and they are the design variables of the problem. In the absence of great amplitude movements, the displacement and rotation of solid S_1 compared to S_0 in the point A_1 are defined in the motion vector as follows:

$$\mathbf{D}_{A_1}(S_1/S_0) = \begin{Bmatrix} u_1 \\ v_1 \\ \gamma_1 \end{Bmatrix} \tag{1}$$

Similarly, other parameters are contained in the following vectors:

$$\mathbf{D}_{A_2}(S_2/S_1) = \begin{Bmatrix} u_2 \\ v_2 \\ \gamma_2 \end{Bmatrix}, \tag{2a}$$

$$\mathbf{D}_{A_3}(S_3/S_2) = \begin{Bmatrix} u_3 \\ v_3 \\ \gamma_3 \end{Bmatrix}, \tag{2b}$$

$$\mathbf{D}_{A_4}(S_3/S_0) = \begin{Bmatrix} u_4 \\ v_4 \\ \gamma_4 \end{Bmatrix}, \tag{2c}$$

$$\mathbf{D}_{A_5}(S_3/S_4) = \begin{Bmatrix} u_5 \\ v_5 \\ \gamma_5 \end{Bmatrix}, \tag{2d}$$

$$\mathbf{D}_{A_6}(S_4/S_5) = \begin{Bmatrix} u_5 \\ v_5 \\ \gamma_5 \end{Bmatrix}, \tag{2e}$$

$$\mathbf{D}_{A_7}(S_5/S_0) = \begin{Bmatrix} u_7 \\ v_7 \\ \gamma_7 \end{Bmatrix} \tag{2f}$$

Therefore, the six-bar mechanism has 21 design variables: the components u_i, v_i and γ_i, of the vectors $\mathbf{D}_{A_1}(S_1/S_0)$, $\mathbf{D}_{A_2}(S_2/S_1)$, $\mathbf{D}_{A_3}(S_3/S_2)$, $\mathbf{D}_{A_4}(S_3/S_0)$, $\mathbf{D}_{A_5}(S_3/S_4)$, $\mathbf{D}_{A_6}(S_4/S_5)$ and $\mathbf{D}_{A_7}(S_5/S_0)$ which are the unknowns of the problem. The vector x contains these 21 design variables:

$$\mathbf{x} = \{ u_1 \quad v_1 \quad \gamma_1 \quad u_2 \quad v_2 \quad \gamma_2 \quad u_3 \quad v_3 \quad \gamma_3 \quad u_4 \quad v_4 \quad \gamma_4 \quad u_5 \quad v_5 \quad \gamma_5 \quad u_6 \quad v_6 \quad \gamma_6 \quad u_7 \quad v_7 \quad \gamma_7 \}^T \tag{3}$$

4 Method for Search the Equilibrium Position of Six-Bar Mechanism

4.1 Optimization Method

From a mathematical point of view, the optimization problem consists of minimizing the objective function Obj(**x**) subjected to constraints imposed by the problem. It follows that the problem can be defined as:

$$\text{Minimize Obj}(\mathbf{x}) \tag{4}$$

Subjected to the following optimization constraints:

$$g_i(\mathbf{x}) \leq 0 \quad i = 1, \ldots, m \tag{5}$$

$$h_j(\mathbf{x}) = 0 \quad j = 1, \ldots, n \tag{6}$$

$g_i(\mathbf{x})$ and $h_j(\mathbf{x})$ are respectively the constraints of inequality and equality equations of the problem.

The purpose of this study is the computation of the equilibrium positions of various elements in spatial mechanisms with parallel joints subjected to mechanical loadings. The values of the design variables being initialized, the initial values of the objective function and the constraints are calculated.

The resolution of this problem is considered here by using mathematical algorithms and iterative methods which require the calculation of the derivative, or the sensitivity, of the objective function and the constraints with respect to the design. This stage of calculation is integrated into the optimization process. The implemented optimization method is the sequential quadratic programming method [10].

The algorithm is based upon an iterative process in which, at each stage, the design parameters take new values, allowing for the convergence towards the optimal solution. In the case of non-convergence, new values are assigned to the design variables and a new iteration is carried out. This process is repeated until convergence is reached.

The design variables are limited by the geometry:

$$-\frac{J_i}{2} \leq u_i \leq \frac{J_i}{2}, \tag{7a}$$

$$-\frac{J_i}{2} \leq v_i \leq \frac{J_i}{2}, \tag{7b}$$

$$-\infty \leq \gamma_i \leq \infty, \quad i = 1, \ldots, 7. \tag{7c}$$

4.2 Objective Function

The objective function is the potential energy of the six-bar mechanism calculated by means of a kinematically admissible field.

The potential energy V(x) of a body, resulting from a kinematically admissible displacement-field is defined by Germain and Muller [11]:

$$V(C') = W(C') - \iiint f_k.U'_k \, dv - \iint T_k.U'_k \, ds \tag{8}$$

Then, the objective function is given by:

$$Obj(\mathbf{x}) = -\sum \begin{Bmatrix} F_{ix} \\ F_{iy} \\ C_{iz} \end{Bmatrix} \begin{Bmatrix} u_{B_i \in S_i/S_0} \\ v_{B_i \in S_i/S_0} \\ \gamma_{B_i \in S_i/S_0} \end{Bmatrix} \tag{9}$$

B_i is the application point of the mechanical load defined by F_{ix} and F_{iy} along the X_i and Y_i axes and by the torque C_{iz} with respect to Z_i axis. Components $u_{B_i \in S_i/S_0}$ $v_{B_i \in S_i/S_0}$ and $\gamma_{B_i \in S_i/S_0}$ are respectively the X and Y displacements and the rotation with respect to Z of B_i belonging S_i in the global reference.

4.3 Inequality Constraints

In the local reference (O_i, X_i, Y_i, Z_i), the point A_i can move in the inner surface of the circle with center O_i and radius $\frac{J_i}{2}$:

$$0 \leq u_i^2 + v_i^2 \leq \left(\frac{J_i}{2}\right)^2 \quad (i = 1, \ldots, 7) \tag{10}$$

Since the origins O_3, O_4 and O_5 belong to the same solid S_3, inequality constraints must be imposed. Indeed, the movements of A_3 and A_5 belonging to S_3 with respect to S_0 depends on the displacement of A_4 belonging to S_3 with respect to S_0. These are between $-\frac{J_4}{2}$ and $\frac{J_4}{2}$.

$$\begin{Bmatrix} -\frac{J_4}{2} \\ -\frac{J_4}{2} \end{Bmatrix} \leq \begin{Bmatrix} u_{O_3 \in (S_3/S_0)} \\ v_{O_3 \in (S_3/S_0)} \end{Bmatrix} \leq \begin{Bmatrix} \frac{J_4}{2} \\ \frac{J_4}{2} \end{Bmatrix} \tag{11}$$

$$\begin{Bmatrix} -\frac{J_4}{2} \\ -\frac{J_4}{2} \end{Bmatrix} \leq \begin{Bmatrix} u_{O_5 \in (S_3/S_0)} \\ v_{O_5 \in (S_3/S_0)} \end{Bmatrix} \leq \begin{Bmatrix} \frac{J_4}{2} \\ \frac{J_4}{2} \end{Bmatrix}. \tag{12}$$

4.4 Equality Constraints

Based on Fig. 2, relations between the different movements vectors, defined before, are:

$$\mathbf{D}_{A_1}(S_1/S_0) + \mathbf{D}_{A_2}(S_2/S_1) + \mathbf{D}_{A_3}(S_3/S_2) - \mathbf{D}_{A_4}(S_3/S_0) = \left\{\vec{0}\right\} \tag{13}$$

$$\mathbf{D}_{A_7}(S_5/S_0) + \mathbf{D}_{A_6}(S_4/S_5) + \mathbf{D}_{A_5}(S_3/S_4) - \mathbf{D}_{A_4}(S_3/S_0) = \left\{\vec{0}\right\} \tag{14}$$

These relationships should be reduced to the same point. The development gives six linear equations. In matrix form, we obtain:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ YO_1 - YO_4 & XO_4 - XO_1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ YO_2 - YO_4 & XO_4 - XO_2 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ YO_3 - YO_4 & XO_4 - XO_3 & 1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}^T \begin{Bmatrix} u_1 \\ v_1 \\ \gamma_1 \\ u_2 \\ v_2 \\ \gamma_2 \\ u_3 \\ v_3 \\ \gamma_3 \\ u_4 \\ v_4 \\ \gamma_4 \end{Bmatrix} = \left\{\vec{0}\right\} \tag{15}$$

and

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ YO_5 - YO_4 & XO_4 - XO_5 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ YO_6 - YO_4 & XO_4 - XO_6 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ YO_7 - YO_4 & XO_4 - XO_7 & 1 \end{bmatrix}^T \begin{Bmatrix} u_4 \\ v_4 \\ \gamma_4 \\ u_5 \\ v_5 \\ \gamma_5 \\ u_6 \\ v_6 \\ \gamma_6 \\ u_7 \\ v_7 \\ \gamma_7 \end{Bmatrix} = \left\{\vec{0}\right\}. \tag{16}$$

5 First Numerical Application

Consider the case where the geometry of the mechanism and the applied load are symmetrical with respect the plan $O_4Y_4Z_4$. The middle of two solids S_1 and S_2 are subjected to identical negative forces F_{1x} and F_{2x} following the opposite direction of the X axis. Two other forces F_{4x} and F_{5x} are applied in the middle of elements S_4 and S_5 having the same modules of F_{1x} and F_{2x} but in the opposite direction.

The initial coordinates of joint centers are:

$$O_1 = \begin{Bmatrix} -200\,\text{mm} \\ -400\,\text{mm} \end{Bmatrix}, \quad O_2 = \begin{Bmatrix} -300\,\text{mm} \\ -200\,\text{mm} \end{Bmatrix}, \quad O_3 = \begin{Bmatrix} -200\,\text{mm} \\ 0 \end{Bmatrix}, \quad O_4 = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix},$$
$$O_5 = \begin{Bmatrix} 200\,\text{mm} \\ 0 \end{Bmatrix}, \quad O_6 = \begin{Bmatrix} 300\,\text{mm} \\ -200\,\text{mm} \end{Bmatrix}, \quad O_7 = \begin{Bmatrix} 200\,\text{mm} \\ -400\,\text{mm} \end{Bmatrix}.$$

The clearances in the joints are identical:

$$J_1 = J_2 = J_3 = J_4 = J_5 = J_6 = J_7 = 0.2\,\text{mm}$$

The proposed optimization algorithm requires an iterative calculation for the convergence of the design variable **x** to the optimal solution. The final numerical values are placed beside each joint origin. In this case, the equilibrium position of the six-bar mechanism is (Fig. 3):

Since the studied mechanism has a symmetrical geometry and loading case with respect to the plane $O_4Y_4Z_4$, the displacements of A_1, A_2 and A_3 are respectively symmetrical with respect to the movement of A_7, A_6 and A_5. In addition, we find that the solid S_3 has no displacement along the X-axis or rotation relative to the Z axis.

Since $u_i^2 + v_i^2 = (J_i/2)^2$, all the points A_i lie on the circle representing the joints clearances.

6 Second Numerical Application

In this section, another form of six-bar mechanism will be processed. The two elements S_1 and S_5 are vertical while the other elements S_2, S_3 and S_4 are horizontal. The clearance joint of L_1 is equal to the sum of the clearances in the joints L_2 and L_3 ($J_1 = J_2 + J_3$). In the same way, the clearance J_7 is the sum of J_5 and J_6.

A horizontal force is applied to the middle of the bar S_1 in the negative direction of the X axis while the middle of bar S_5 is loaded by another force having the same modulus of the first but in opposite direction.

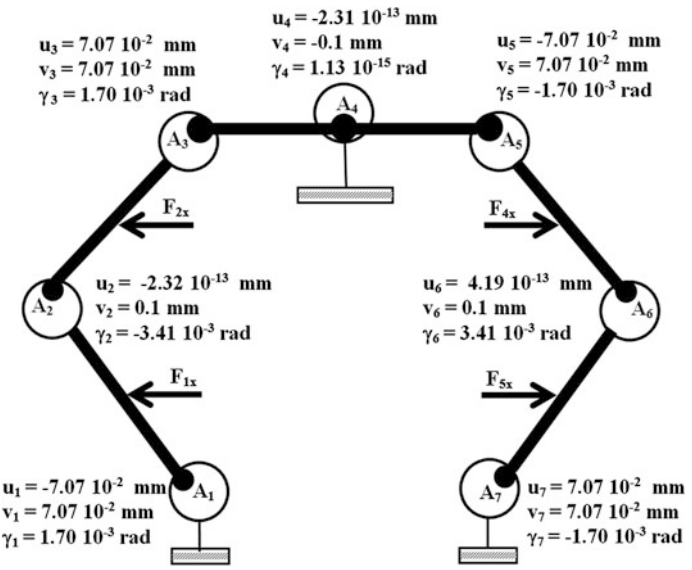


Fig. 3 Equilibrium position of six-bar mechanism after loading in the first numerical application

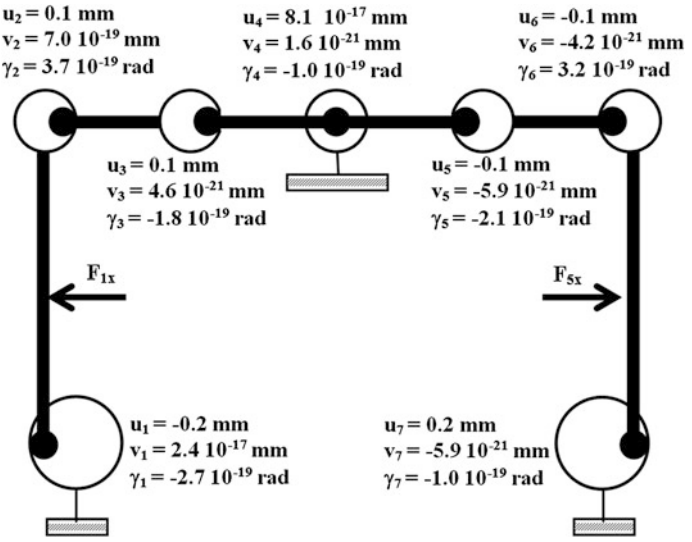


Fig. 4 Displacements of the links after loading in the second numerical application

The initial coordinates of the joint centers are:

$$O_1 = \begin{Bmatrix} -200\,\text{mm} \\ -400\,\text{mm} \end{Bmatrix}, \quad O_2 = \begin{Bmatrix} -200\,\text{mm} \\ 0 \end{Bmatrix}, \quad O_3 = \begin{Bmatrix} -100\,\text{mm} \\ 0 \end{Bmatrix}, \quad O_4 = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix},$$
$$O_5 = \begin{Bmatrix} 100\,\text{mm} \\ 0 \end{Bmatrix}, \quad O_6 = \begin{Bmatrix} 200\,\text{mm} \\ 0 \end{Bmatrix}, \quad O_7 = \begin{Bmatrix} 200\,\text{mm} \\ -400\,\text{mm} \end{Bmatrix}.$$

The clearances of the joints L_2, L_3, L_4, L_5 and L_6 are identical ($J_2 = J_3 = J_4 = J_5 = J_6 = 0.2$ mm) while others are: $J_1 = J_7 = 0.4$ mm (Fig. 4).

The results show that there is no movement of point A_4. For the elements S_1, S_2, S_3, S_5 and S_6, there is no rotation movement relative to the Z. This is normal because $J_1 = J_2 + J_3$ and $J_7 = J_5 + J_6$ but these elements move only along the X axis.

7 Conclusion

To provide accurate relative movement and to minimize geometrical errors in a mechanism, it is essential to control the clearance in joints between parts. The purpose of this study is to propose an analytical method for determining the static equilibrium positions of the various components of six-bar mechanisms with complex chain and subjected to mechanical loads. The study takes into account the presence of the joint clearance in the mechanism. The method is based on the minimization of potential energy, taking into account the constraints imposed by the geometry of the joints. The results show the effectiveness of the method.

Acknowledgment This work was carried out under the support of the research program of Lebanese University

References

1. A. Potiron, P. Dal Santo, M. Younes, Étude bidimensionnelle du positionnement relatif des éléments de mécanismes avec jeu dans les liaisons par une méthode d'optimisation. Mécanique Ind. **4**, 229–238 (2003)
2. H. Funabashi, K. Ogawa, M. Horie, A dynamic analysis of mechanisms with clearance. Bull JSME **21**(161), 1652–1659 (1978)
3. M. Giordano, collective, Modèle de détermination des tolérances géométriques, in *Conception de produits mécaniques*, chapter 13, ed. by M. Tollenaere (HERMES, Paris, 1998)
4. J. Gao, K.W. Chase, S.P. Magleby, General 3-D tolerance analysis of mechanical assemblies with small kinematic adjustments. IIE Trans. **30**, 367–377 (1998)
5. K.W. Chase, J. Gao, S.P. Magleby, General 2-D tolerance analysis of mechanical assemblies with small kinematic adjustments. J. Des. Manuf. **5**, 263–274 (1995)
6. K.W. Chase, A.R. Parkinson, A survey of research in the application of tolerance analysis to the design of mechanical assemblies. Res. Eng. Des. **3**, 23–37 (1991)

7. S. Erkaya, I. Uzmay, Investigation on effect of joint clearance on dynamics of four-bar mechanism. Nonlinear Dyn. **58**(1–2), 179–198 (2009)
8. J.-F. Hsieh, Numerical analysis of displacements in spatial mechanisms with spherical joints utilizing an extended D-H notation. Trans. Can. Soc. Mech. Eng. **34**(3–4), 417–431 (2010)
9. M. Younes, A. Potiron, in *Influence of Clearance Joints on the Elements Position of Planar Six-Bar Mechanisms with Complex Chain.* Lecture notes in Engineering and Computer Science: Proceeding of The World Congress on Engineering (WCE 2013), London, UK, 3–5 July 2013, pp. 1899–1903
10. G.N. Vanderplaats, *Numerical optimization techniques for engineering design with applications* (Mc Graw-Hill Book Company, New York, 1984)
11. P. Germain, P. Muller, in *Introduction à la mécanique des milieux continus* (Masson, Paris, 1980)

Flutter Analysis of an Aerofoil Using State-Space Unsteady Aerodynamic Modeling

Riccy Kurniawan

Abstract This paper deals with the problem of the aeroelastic stability of a typical airfoil section with two-degree of freedom induced by the unsteady aerodynamic loads. A method is presented to model the unsteady lift and pitching moment acting on a two-dimensional typical aerofoil section, operating under attached flow conditions in an incompressible flow. Starting from suitable generalisations and approximations to aerodynamic indicial functions, the unsteady loads due to an arbitrary forcing are represented in a state-space form. From the resulting equations of motion, the flutter speed is computed through stability analysis of a linear state-space system.

Keywords Aeroelasticity · Aerodynamics · Aerofoil · Dynamics · Flutter · Incompressible · Modeling · Stability · State-space · Unsteady

1 Introduction

Aeroelasticity is the complex interaction among aerodynamic, elastic and inertia forces acting on a structure. Structures subject to an air-flow are not entirely rigid, so they deflect under aerodynamic forces. Meanwhile, the aerodynamic forces depend on structural deformation, so a coupled problem arises, the equations of motion for both the structure and fluid must be solved simultaneously.

Flutter is the dynamic aeroelasticity phenomenon whereby the inertia forces can modify the behavior of a flexible system so that energy is extracted from the incoming flow. The flutter or critical speed V_F is defined as the lowest air speed at which a given structure would exhibit sustained, simple harmonic oscillations.

R. Kurniawan (✉)
Department of Mechanical Engineering, Atma Jaya Catholic University of Indonesia, Jenderal Sudirman 51, Jakarta 12930, Indonesia
e-mail: riccy.kurniawan@atmajaya.ac.id

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_11,

V_F represents the neutral stability boundary: oscillations are stable at speeds below it, but they become divergent above it.

The aeroelastic analysis and hence the flutter computation heavily relies on an accurate description of the unsteady aerodynamics. Unsteady flow arises for three reasons: the body is in unsteady motion (vibrating aerofoil), the incident flow contains unsteady disturbances (gust and turbulence), and the body wake in unsteady (von Karman vortex street, for instance). The main feature of wakes is that they contain shear layers. These shear layers are usually unstable causing them to roll up into concentrated vortices and tending to imprint dominant frequencies of unsteady motion into the flow.

Theodorsen [1] obtained closed-form solution to the problem of an unsteady aerodynamic load on an oscillating aerofoil. This approach assumed the harmonic oscillations in inviscid and incompressible flow subject to small disturbances. Wagner [2] obtained a solution for the so-called indicial lift on a thin-aerofoil undergoing a transient step change in angle of attack in an incompressible flow. The indicial lift response makes a useful starting point for the development of a general time domain unsteady aerodynamics theory. A practical way to tackle the indicial response method is through a state-space formulation in the time domain, as proposed, for instance by Leishman and Nguyen [3].

The main objective of this chapter is to investigate the flutter analysis of a typical aerofoil section with two-degree of freedom induced by the unsteady aerodynamic loads defined by the Leishman's state-space model. This paper will discuss extended the state-space unsteady aerodynamics modeling on flutter analysis in [4].

2 Aeroelastic Model Formulation

The mechanical model under investigation is a two-dimensional typical aerofoil section in a horizontal flow of undisturbed speed V, as shown in Fig. 1. Its motion is defined by two independent degrees of freedom, which are selected to be the vertical displacement (plunge), h, positive down, and the rotation (pitch), α. The structural behaviour is modeled by means of linear bending and torsional springs, which are attached at the elastic axis of the typical aerofoil section. The springs in the typical aerofoil section can be seen as the restoring forces that the rest of the structure applies on the section.

The equations of motion for the typical aerofoil section have been derived in many textbooks of aeroelasticity, and can be expressed in non-dimensional form as

$$r_\alpha^2 \ddot{\alpha} + \frac{x_\alpha}{b}\ddot{h} + \omega_\alpha^2 r_\alpha^2 \alpha = 2\frac{\kappa}{\pi}\left(\frac{V}{b}\right)^2 C_M(t) \tag{1}$$

$$x_\alpha \ddot{\alpha} + \frac{1}{b}\ddot{h} + \frac{\omega_h^2}{b} h = \frac{\kappa}{\pi}\left(\frac{V}{b}\right)^2 C_L(t) \tag{2}$$

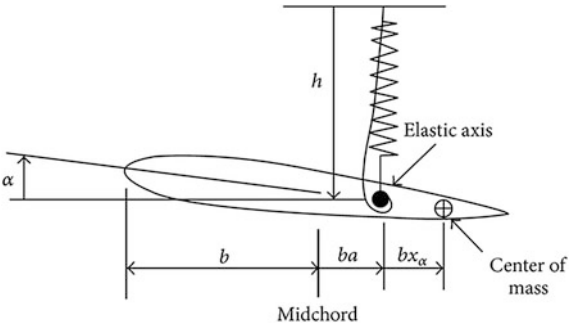


Fig. 1 A typical aerofoil section with two degrees of freedom

where $C_M(t)$ and $C_L(t)$ denote the coefficients of the aerodynamic forces corresponding to pitching moment and lift, respectively. For a general motion, where an aerofoil of chord $c = 2b$ is undergoing a combination of pitching and plunging motion in a flow of steady velocity V, Theodorsen [1] obtained the aerodynamic coefficients

$$C_M(t) = -\frac{\pi}{2V^2}\left[\left(\frac{1}{8}+a^2\right)\ddot{\alpha} - ab\ \ddot{h}\right] + \pi\left(a+\frac{1}{2}\right)C(k)\alpha_{qs} - \frac{\pi}{2V^2}\left[V\left(\frac{1}{2}-a\right)\dot{\alpha}\right] \tag{3}$$

$$C_L(t) = \frac{\pi b}{V^2}\left(V\dot{\alpha} + \ddot{h} - ba\ddot{\alpha}\right) + 2\pi C(k)\alpha_{qs} \tag{4}$$

The first term in (3) and (4) is the non-circulatory or apparent mass part, which results from the flow acceleration effect. The second group of terms is the circulatory components arising from the creation of circulation about the aerofoil. Theodorsen's function $C(k) = F(k) + iG(k)$ is a complex-valued transfer function which depends on the reduced frequency k, where

$$k = \frac{\omega b}{V} \tag{5}$$

Theodorsen's function $C(k)$ is expressed in terms of Hankel functions, H, with the reduced frequency k as the argument, where

$$C(k) = F(k) + iG(k) = \frac{H_1^{(2)}(k)}{H_1^{(2)}(k) + iH_0^{(2)}k} \tag{6}$$

The Hankel function is defined as $H_V^{(2)} = J_V - iY_V$ with J_V and Y_V being Bessel functions of the first and second kind, respectively. Implicitly recognizing that each Bessel function has an argument k, then the real or in-phase ($\Re$) part and imaginary or out-of-phase ($\Im$) can be written as

$$\Re C(k) = F = \frac{J_1(J_1 + Y_0) + Y_1(Y_1 - J_0)}{(J_1 + Y_0)^2 + (J_0 - Y_1)^2} \tag{7}$$

and

$$\Im C(k) = G = -\frac{Y_1 Y_0 + J_1 J_0}{(J_1 + Y_0)^2 + (J_0 - Y_1)^2} \tag{8}$$

The amplitude and phase of Theodorsen's function are given by

$$|C(k)| = \sqrt{F^2 + G^2} \quad \text{and} \quad \phi = \tan^{-1}\left(\frac{G}{F}\right) \tag{9}$$

α_{qs} represents a quasi-steady aerofoil angle of attack, i.e.

$$\alpha_{qs} = \frac{\dot{h}}{V} + \alpha + b\left(\frac{1}{2} - \alpha\right)\frac{\dot{\alpha}}{V} \tag{10}$$

Theodorsen's theory is formulated in the frequency domain, as a function of the parameter k, which is implicit in the solution since it depends on the frequency of oscillation. However, a theory formulated in the time domain is more generally applicable. For incompressible flow, Wagner [2] obtained a solution for the indicial lift on a thin aerofoil undergoing a step change in angle of attack.

The indicial response method is the response of the aerodynamic flowfield to a step change in a set of defined boundary conditions such as a step change in aerofoil angle of attack, in pitch rate about some axis, or in a control surface deflection (such as a tab of flap). If the indicial aerodynamic responses can be determined, then the unsteady aerodynamic loads due to arbitrary changes in angle of attack can be obtained through the superposition of indicial aerodynamic responses using the Duhamel's integral.

Assuming two-dimensional incompressible potential flow over a thin aerofoil, the circulatory terms in (3) and (4) can be written as

$$C(k)\alpha_{qs} = \alpha_{qs}(0)\varphi_w(s) + \int_0^s \frac{d\alpha_{qs}}{dt}\varphi_w(s-t)dt \tag{11}$$

where s is the non-dimensional time, given by

$$s = \frac{1}{b}\int_0^t V\,dt \tag{12}$$

which represents the relative distance traveled by the aerofoil through the flow in terms of semi-chords during a time interval t.

Wagner's function, which accounts for the influence of the shed wake, as does Theodorsen's function. In fact, both Wagner's and Theodorsen's function

represents a Fourier transform pair. Wagner's function is known exactly in terms of Bessel functions (see [2] for details), but for practical implementation it is useful to represent it approximately. One of the most useful expressions is an exponential of the form

$$\varphi_w(s) \approx 1 - A_1 e^{-b_1 s} - A_2 e^{-b_2 s} \tag{13}$$

One exponential approximation is given by Jones [5] as

$$\varphi_w(s) \approx 1.0 - 0.165 e^{-0.0455 s} - 0.335 e^{-0.3 s} \tag{14}$$

One of the most fundamental concepts associated with the description of any dynamic system, aerodynamic or otherwise, is the state of the system. The state describes the internal behavior of that system and simply the information required at a given instant in time to allow the determination of the outputs from the system given future inputs. The state-space formulation of an nth-order ODEs system with m inputs and p outputs is of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{15}$$

with the output equations

$$\dot{\mathbf{y}} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \tag{16}$$

where $\mathbf{x} = x_i$, $i = 1, 2,\ldots,n$ are the states of the system. The inputs are denoted by $\mathbf{u} = u_i$, $i = 1, 2,\ldots,m$ and the outputs are denoted by $\mathbf{y} = y_i$, $i = 1, 2,\ldots,p$. The state-space constant matrices **A**, **B**, **C**, and **D** can be found from the transfer function using a method described by Franklin et al. [6].

The state-space equations describing the behavior of a 2-D unsteady aerofoil can be obtained through application of Laplace transform methods to the indicial response. Consider the indicial response, φ, which is to be approximated by the exponential function given previously by Eq. (13). This function can be written in the time domain as

$$\varphi(t) = 1.0 - A_1 e^{-b_1\left(\frac{2V}{c}\right)t} - A_2 e^{-b_2\left(\frac{2V}{c}\right)t} \tag{17}$$

Initially, let $(0) = 0 = 1 - A_1 - A_2$. Then the corresponding impulse response, $h(t)$, is given by

$$h(t) = A_1 b_1 \left(\frac{2V}{c}\right) e^{-b_1\left(\frac{2V}{c}\right)t} + A_2 b_2 \left(\frac{2V}{c}\right) e^{-b_2\left(\frac{2V}{c}\right)t} \tag{18}$$

The Laplace transform of the impulse response is

$$L[h(t)] = \frac{A_1 b_1\left(\frac{2V}{c}\right)}{p + b_1\left(\frac{2V}{c}\right)} + \frac{A_2 b_2\left(\frac{2V}{c}\right)}{p + b_2\left(\frac{2V}{c}\right)} \tag{19}$$

which can be arranged to yield the transfer functions as the Padé approximant

$$L[h(t)] = \frac{(A_1 b_1 + A_2 b_2)(\frac{2V}{c})p + (A_1 + A_2)(b_1 b_2)(\frac{2V}{c})^2}{p^2 + (b_1 + b_2)(\frac{2V}{c})p + (b_1 b_2)(\frac{2V}{c})^2} \tag{20}$$

From this transfer function, the response to an input α_{qs} can be directly written in state-space form as

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -b_1 b_2 \left(\frac{V}{b}\right)^2 & -(b_1 + b_2)\frac{V}{b} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \alpha_{qs} \tag{21}$$

with the outputs

$$C(k)\alpha_{qs} = \begin{bmatrix} \frac{b_1 b_2}{2}\left(\frac{V}{b}\right)^2 & (A_1 b_1 + A_2 b_2)\left(\frac{V}{b}\right) \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \frac{1}{2}\alpha_{qs} \tag{22}$$

The main benefit of the state-space formulation is that the equations can be appended to the equations of motion directly, very useful in aeroservoelastic analysis. Furthermore, it permits the straightforward addition of more features to the model, such as gust response and compressibility.

The indicial approach and the state-space formulation lead to a dynamic matrix that governs the behaviour of the system and enables future prediction. The analysis of flutter in this case is straightforward and it can be performed in the frequency domain, since the eigenvalues of the dynamic matrix directly determine the stability of the system. If, for a given velocity, any of the eigenvalues has a zero real part, the system is neutrally stable, i.e., it defines the flutter onset.

3 Results and Discussion

In this section, the stability analysis of the state-space aeroelastic equation is presented. The results have been validated against published results. Theodorsen and Garrick [7] presented a graphical solution of the flutter speed of the two-dimensional aerofoil for the flexture-torsion case. In order to validate the present model, a flutter speed computation is performed with varying combinations of aeroelastic parameters, as used by Theodorsen and Garrick, as shown in Table 1.

Figure 2 shows the comparison of the flutter margin from Theodorsen and Garrick's work with the present computation. In the graph, non-dimensional flutter speed V_F* is presented as a function of the frequency ratio ω_h/ω_α. As can be seen, the present method provides a good agreement with the published figures only for low frequency ratios. In fact, as the ratio approaches unit value, the actual curve drifts to generally lower speeds.

Table 1 Aeroelastic parameters for the validation

Case	x_α	κ	a	r_α^2
a	0.2	1/3	−0.4	0.25
b	0.2	1/4	−0.2	0.25
c	0	1/5	−0.3	0.25
d	0.1	1/10	−0.4	0.25

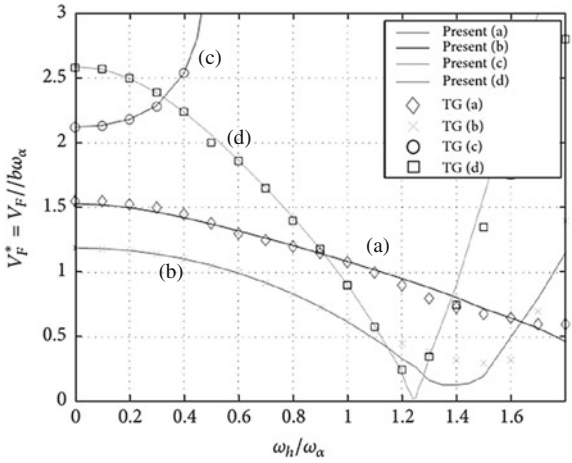


Fig. 2 Comparison of flutter boundaries from Theodorsen and Garrick [7] with present computations

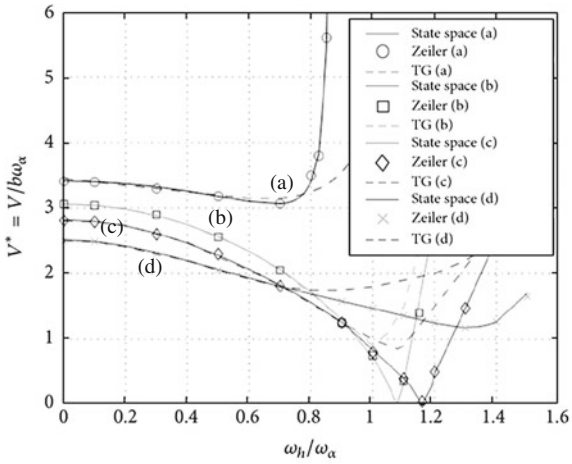


Fig. 3 Comparisons of flutter boundaries from Zeiler [8] and Theodorsen and Garrick [7] with present computations. The parameters used are $a = -0.3$, $\kappa = 0.05$, $r_\alpha^2 = 0.25$, $b = 0.3$. (*a*) $x_\alpha = 0$, (*b*) $x_\alpha = 0.05$, (*c*) $x_\alpha = 0.1$, and (*d*) $x_\alpha = 0.2$

This discrepancy is probably due to numerical inaccuracies in the curves presented in the original work. Zeiler [8] found a number of erroneous plots in the reports of Theodorsen and Garrick and provided a few corrected plots. In order to

verify the validity of Zeiler's statement, the numerical computation of the flutter speed is conducted using the aeroelastic parameters used by Zeiler.

Figure 3 shows some of the results obtained by Zeiler, compared to the figures obtained by Theodorsen and Garrick and those obtained using the present state-space method. As can be observed, the agreement with Zeiler is very good, whereas Theodorsen and Garrick's results deviate considerably. This confirms the validity of Zeiler's statement and provides evidence of the validity of the results obtained here.

4 Conclusion

A model to determine the flutter onset of a two-dimensional typical aerofoil section has been implemented and then validated. A traditional aerodynamic analysis, based on Theodorsen's theory and Leishman's state-space model was used. The validation was performed by solving Theodorsen and Garrick's problem for the flexture-torsion flutter of a two-dimensional typical aerofoil section. The stability curves obtained are in close agreement with the results reported by more recent solutions of the same problem, whereas the original figures from Theodorsen and Garrick are found to be biased, as was previously reported by Zeiler.

References

1. T. Theodorsen, in *General Theory of Aerodynamics Instability and the Mechanism of Flutter*. NACA Report number: 496 (1934)
2. H. Wagner, Über die Entstehung des dynamischen Auftriebes von Tragflügeln. Zietschrift für Angewandte Mathematik und Mechanik **5**(1), 17–35 (1925)
3. J.G. Leishman, K.Q. Nguyen, State-space representation of unsteady airfoil behavior. AIAA J. **28**(5), 863–844 (1989)
4. R. Kurniawan, in *Numerical Study of Flutter of a Two-Dimensional Aeroelastic System*. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering (WCE 2013), London, UK, 3–5 July 2013, pp. 1858–1860
5. R.T. Jones, in *The Unsteady Lift of a Wing of Finite Aspect Ratio*. NACA report number: 681 (1940)
6. G.F. Franklin, J.D. Powell, E.N. Abbas, *Feedback Control Dynamics System*, 3rd edn. (Addison Wesley Publishing Co., Reading, MA, 1994)
7. T. Theodorsen, I.E. Garrick, in *Mechanism of Flutter: A Theoretical and Experimental Investigation Of Flutter Problem*. NACA Report number: 685 (1938)
8. T.A. Zeiler, Results of Theodorsen and Garrick revisited. J. Aircr. **37**(5), 918–920 (2000)

WRS-BTU Seismic Isolator Performances

Renato Brancati, Giandomenico Di Massa, Stefano Pagano, Ernesto Rocca and Salvatore Strano

Abstract This paper describes an experimental investigation conducted on WRS-BTU seismic isolators that are constituted by a wire rope spring coupled with a ball transfer unit. The device can be considered rigid along the vertical direction while the horizontal stiffness can be independently chosen to shift the natural period away from the period range having the most of earthquake energy. Two kinds of tests were carried out: the first ones regard the isolator characterization; then, to evaluate the isolation efficiency, a small laboratory structure was equipped with four isolators and several tests on a shake-table were performed.

Keywords Ball transfer unit · Hysteresis · Nonlinear dynamics · Seismic isolators · Shake-table tests · Wire rope springs

R. Brancati · G. Di Massa · S. Pagano · E. Rocca · S. Strano (✉)
Dipartimento di Ingegneria Industriale, Università degli Studi di Napoli Federico II,
Via Claudio 21 80125 Naples, Italy
e-mail: salvatore.strano@unina.it

R. Brancati
e-mail: renato.brancati@unina.it

G. Di Massa
e-mail: giandomenico.dimassa@unina.it

S. Pagano
e-mail: pagano@unina.it

E. Rocca
e-mail: erocca@unina.it

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_12,

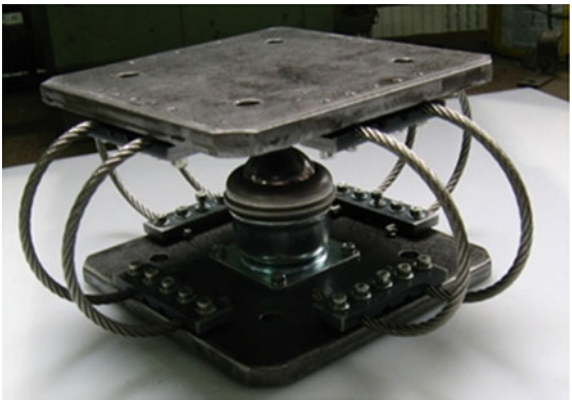


Fig. 1 WRS-BTU isolator

1 Introduction

A new type of seismic isolator was developed and realized at the Department of Industrial Engineering (DII) of the University of Naples Federico II; it is constituted (Fig. 1) by the coupling of a ball transfer unit (BTU) with a wire rope spring (WRS).

BTU is an omni-directional load-bearing spherical balls mounted inside a restraining fixture, having the task to bear the structure weight with a neglecting vertical deformation (Fig. 2); it allows the structure to translate along any horizontal direction with low friction.

The isolator restoring force is provided by a wire rope spring (WRS) constituted by several ropes connecting the two plates of the device.

WRSs are widely adopted in industrial field as they are particularly effective in vibration isolation and energy absorption. In fact they dissipate energy through friction forces acting among the wires of each rope due to their relative movement; the produced thermal energy is easily exchanged with the environment thanks to favourable heat exchange surface of the ropes.

Compared with the best known elastomeric seismic isolators, WRSs have a longer service life since they are not very sensitive to temperature changes and resists aggressive environments caused by the presence of ozone, oil, grease and salt spray.

The most common WRS are helical type (Fig. 3a) or circular type (Fig. 3b). These springs are sometimes called "cable isolators" even if the term *cable* generally refers to a flexible tension member (*rope*) which, in addition to a strength member, includes power and/or signal conductors within its structure.

The WRS-BTU isolator can be obtained by coupling the two main component, not necessarily integrated into a unique support; the two main component can be chosen in function of the vertical load acting on the BTU and of the WRS horizontal stiffness required to shift the natural period away from the period range having the most of earthquake energy; the choices are facilitated by the numerous

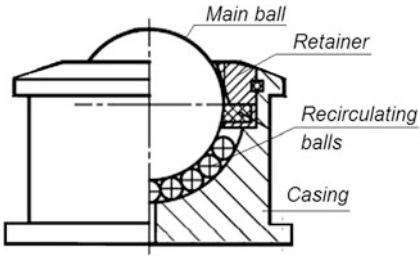


Fig. 2 Ball transfer unit (BTU) scheme

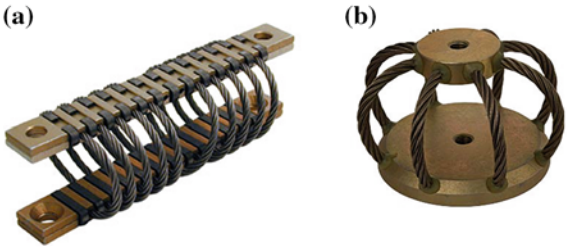


Fig. 3 Commercial wire rope springs: **a** helical type; **b** polygonal type

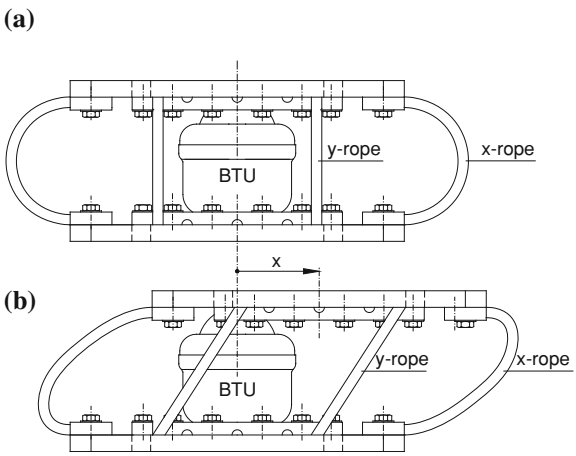


Fig. 4 Isolator configurations: **a** equilibrium configuration; **b** deformed configuration

technical catalogues available on the web and no additional costs are required for the devices characterization.

The WRS-BTU prototype [1, 2], developed at the DII, adopts a simple WRS characterized by the possibility to easily change the rope length or the ropes with others having different diameter or different rope configuration.

WRSs provide a restoring force due to the deflection of the ropes; if the upper plate is in the centred position the ropes, are stressed by bending moment due to the curvature imposed during isolator assembly (Fig. 4a).

With reference to Fig. 4b, it can be noted that an upper plate displacement along the x direction, induces:

- torque and bending in the *y*-ropes;
- a change of the curvature of the *x*-ropes that induces an increase or a decrease of the bending moment depending on the length of the rope.

While the additional deflection of *y*-ropes induces a positive restoring force, the change of the bending moment in *x*-ropes can even determine a set of forces that take away the plate from the central position if the ropes are long enough (negative stiffness). Shortening the ropes the bending stiffness becomes positive and all the ropes (*x* and *y*) contribute to re-centre the upper plate.

The BTU, is mounted in ball-up configuration, i.e. with the main ball in contact with the intrados of the upper plate so that dust or debris cannot settle on the rolling surface and cannot affect the regular rolling of the ball.

The isolator is characterized by a hysteretic non-linear behaviour; a specific hysteresis model, able to describe its behaviour, is reported in [3].

The present paper describes some tests conducted at the DII; the first set of test regards the isolators characterization and hence the force-displacement diagram obtained superimposing and a periodic relative horizontal displacement with a constant vertical load and the friction force exerted by the BTU.

To verify the insulation efficiency, a light structure was realized and was insulated by means of four WRS-BTU isolators. The experimental tests were conducted by fixing the structure on the moving platform of the shake-table developed at the DII; it is driven by a servo-hydraulic actuator able to simulate a wide range of simulated ground motions including the reproductions of recorded earthquakes time-histories.

2 BTU Friction

The isolator is constituted by 8 ropes, each having a length of 90 mm and a diameter of 5 mm.

To characterize the BTU friction coefficient, some tests were performed on a bi-axial press (Fig. 5). The apparatus consists of a 800 kN hydraulic press equipped with a slide that can translate on two linear duct trolleys and rails in a horizontal direction. A mechanical actuator moves the slide with a harmonic motion of assigned amplitude and frequency. The instrumentation enables the slide position and the force transmitted by the actuator to be detected.

The tests were conducted by fixing the isolator lower plate onto the horizontal slide and the upper one to a vertical slide onto which the vertical load is applied; the slide was driven to perform a horizontal harmonic movement with an amplitude of 20 mm (i.e. a stroke of 40 mm) and a frequency of 0.05 Hz.

To characterize the BTU rolling resistance, the isolator without wire ropes was fixed onto the bi-axial press and by superimposing the above defined harmonic motion, the horizontal force was detected; the contribution exerted by the BTU

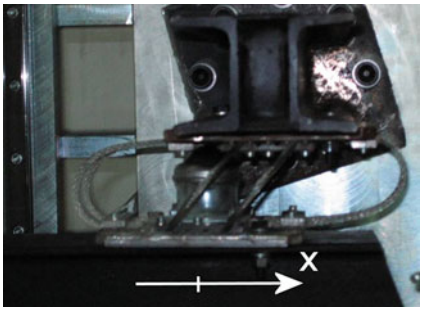


Fig. 5 Isolator on the bi-axial press for the BTU friction coefficient evaluation

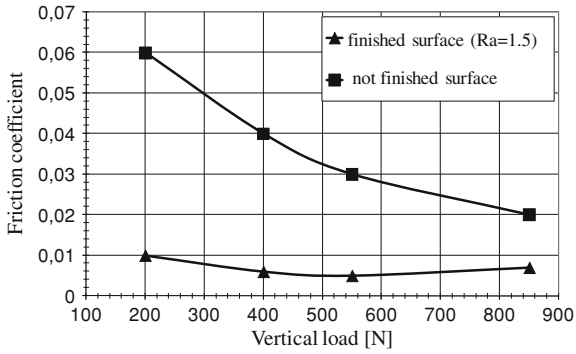


Fig. 6 Friction test results

was obtained subtracting the slide friction force previously determined detecting the force required to move the slide with different masses fixed on the slide.

The tests, conducted for two different levels of rolling surface mechanical finishing, are reported in Fig. 6; the upper curve regards the "rough" rolling surface (without surface finishing) whereas the lower one was obtained finishing the rolling surface (R_a roughness equal to about 1.5).

3 Shear Tests on the WRS-BTU Isolator

The shear tests were performed using the isolator testing machine (BPI) available at the DII laboratory (Fig. 7). The BPI mainly consists of a movable horizontal platform driven by a hydraulic actuator that allows to impose periodic shear deformations to the device under test, simultaneously loaded with a constant vertical compression [4].

The platform motion is constrained to a single horizontal axis by means of recirculating ball-bearing linear guides.

The removal of the reaction structures (Fig. 7) allows the testing machine to be used as a shake-table able to verify the seismic performances of isolated structures [5].

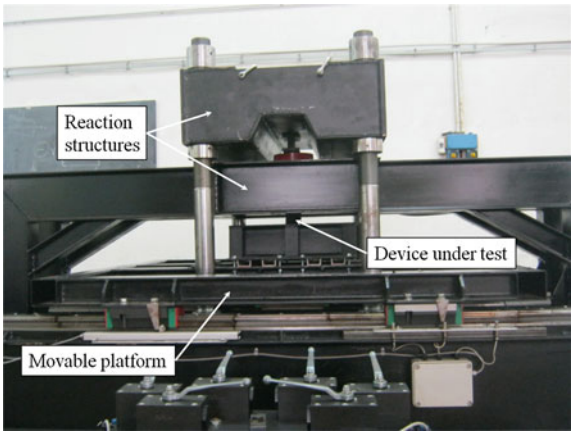


Fig. 7 Isolator testing machine (BPI)

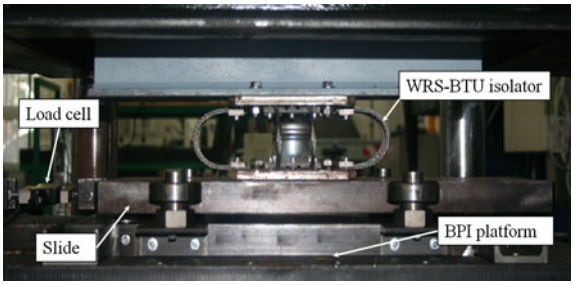


Fig. 8 Slide on the BPI platform

Since the BPI was designed to characterize more rigid isolators [6], the sensitivity of the load cell and the friction forces of the platform linear guides do not allow to make accurate investigation for the isolator described in the present paper. For this reason a suitable device was realized (Figs. 8, 9) constituted by a slide horizontally connected to the BPI platform by means of a load cell and vertically supported by four BTUs; the slide is laterally guided by two couples of rolling bearing.

When BPI platform moves, the isolator restoring force can be obtained by the load cell signals, subtracting the platform inertia force and the friction force exerted by the four supporting BTUs.

The inertia force is given by the measurement of the platform horizontal acceleration and the by its mass.

Figures 10, 11 and 12 show the results of some tests carried out by imposing sinusoidal shear deformation for different values of frequency (f) and amplitude (A). Moreover, the tests were conducted with and without the BTU contact in order to appreciate its effect on the restoring force.

Figure 10 shows two hysteresis cycles obtained without BTU contact for two different platform motion amplitude and with the same frequency.

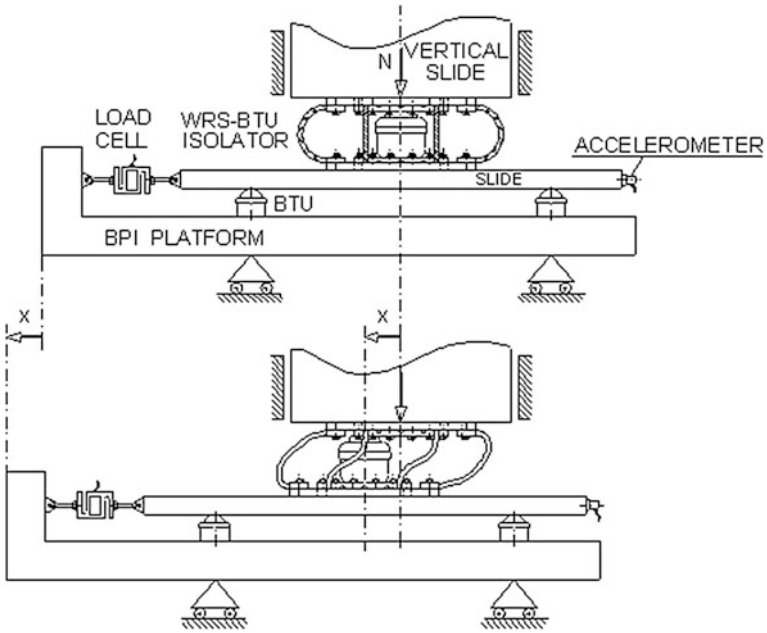


Fig. 9 Scheme of the system adopted to measure the isolator restoring force

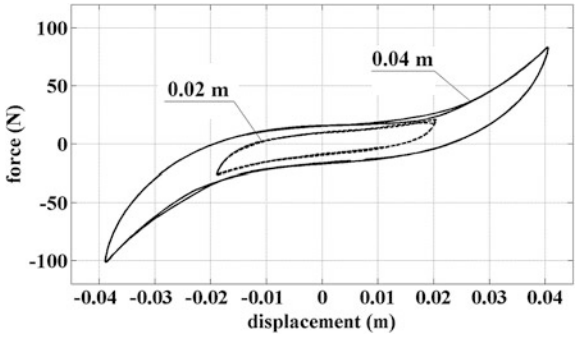


Fig. 10 Hysteresis cycles, $f = 0.5$ Hz, $A = 0.02$ m (*dashed line*), $A = 0.04$ m (*solid line*), without BTU contact

The results reported in Fig. 11 show the influence of the excitation frequency on the restoring force. Also in this case the cycles were obtained without the BTU contact.

To highlight the influence of the BTU contact, Fig. 12 compares two cycles obtained with and without the BTU contact; the two tests were performed keeping constant the vertical distance between the isolator plates and so with the same rope deformations. In the BTU contact case, a vertical load of 1,220 N was applied on the isolator (smaller than the maximum allowed, equal to 1,400 N).

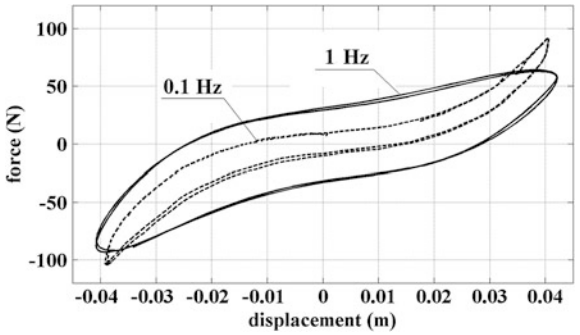


Fig. 11 Hysteresis cycles, $A = 0.04$ m, $f = 0.1$ Hz (*dashed line*), $f = 1$ Hz (*solid line*), without BTU contact

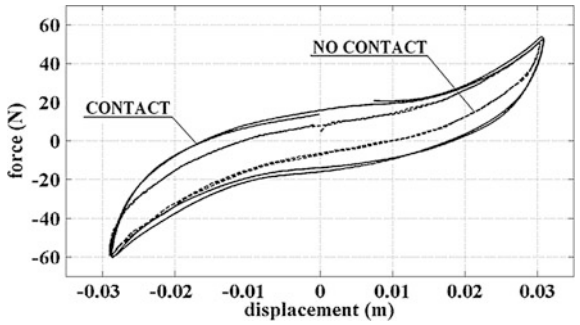


Fig. 12 Hysteresis cycles, $A = 0.03$ m, $f = 0.1$ Hz, without BTU contact (*dashed line*), with BTU contact (*solid line*)

The detected cycles show that the device has a hardening behaviour and the hysteresis cycle area increases with the forcing frequency.

To study the isolator dynamic properties, the nonlinear energy dissipation represented by the preview hysteresis cycles was approximated with the one corresponding to an equivalent linear spring-damper system [7].

In Table 1 the main results of the sinusoidal shear tests in terms of the equivalent shear stiffness k_{eq}, the equivalent damping coefficient σ_{eq} and the equivalent damping ratio ξ_{eq}, are summarized.

The equivalent shear stiffness was obtained considering the following relation:

$$k_{eq} = \frac{F_{\max} - F_{\min}}{d_{\max} - d_{\min}} \tag{1}$$

where $F_{\max}$ and $F_{\min}$ are the maximum and minimum values of horizontal force, $d_{\max}$ and $d_{\min}$ are the horizontal displacement corresponding to $F_{\max}$ and $F_{\min}$. The equivalent damping coefficient was determined by means of the equation:

$$\sigma_{eq} = \frac{2\Delta W}{\pi^2 f (d_{\max} - d_{\min})^2}, \tag{2}$$

Table 1 Results of the shear tests

Frequency (f) (Hz)	Amplitude (A) (m)	BTU contact	Equivalent shear stiffness (k_{eq}) (N m^{-1})	Equivalent damping coefficient (σ_{eq}) (N s m^{-1})	Equivalent damping ratio (ξ_{eq}) (–)
0.1	0.02	No	1,443	584	0.13
0.1	0.03	No	1,841	505	0.09
0.1	0.03	Yes	1,880	913	0.15
0.1	0.04	No	2,416	519	0.06
0.5	0.02	No	1,209	166	0.21
0.5	0.03	No	1,563	142	0.14
0.5	0.03	Yes	1,736	213	0.19
0.5	0.04	No	2,323	152	0.10
1	0.02	No	1,148	142	0.39
1	0.03	No	1,416	142	0.31
1	0.03	Yes	1,482	177	0.37
1	0.04	No	1,884	138	0.23

where ΔW is the area enclosed by one hysteresis cycle. The equivalent damping ratio is given by:

$$\xi_{eq} = \frac{2\Delta W}{\pi k_{eq}(d_{\max} - d_{\min})^2}. \tag{3}$$

The results reported in Table 1, concerning the BTU without contact, clearly show the increasing of k_{eq} with the increasing of the shear deformation amplitude A; moreover, k_{eq} decreases with the increasing of the forcing frequency f. The equivalent damping coefficient σ_{eq} decreases with the increasing of f. Another observation is that σ_{eq} changes with respect to A, while in the case $f = 1$ Hz, its value seems to be almost constant. The results also show that the parameter ξ_{eq} increases with the increasing of f and decreases with the increase of A.

In the case of BTU contact, it is possible to note only a slight variation of $k_{eq,}$ while the results clearly highlight the increasing of σ_{eq} and ξ_{eq}. The effect of the BTU contact on the isolator dissipated energy, represented by the hysteresis cycle, is more relevant for small frequency values.

4 Shake-Table Tests of a Structure with WRS-BTU Isolators

WRS-BTU isolators were adopted to insulate a laboratory cabinet (Figs. 13, 14) constituted of a rigid steel frame (0.60 × 0.70 × 1.20 m) on which may be fixed additional masses to modify the inertial properties.

In Fig. 13 it is possible to note that the shaking table tests were performed with the BPI used as a shake-table.

Fig. 13 Cabinet on the shake-table

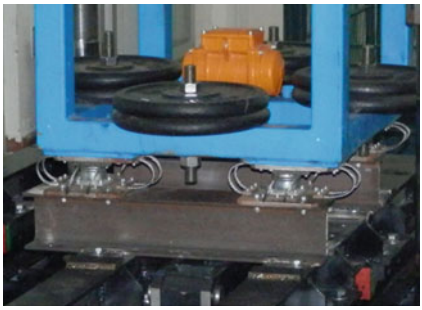


Fig. 14 Additional masses

The tests were conducted with the cabinet overall mass of 165 kg. The additional masses were arranged in the lower part of the cabinet to avoid the cabinet overturning.

The performance of the isolated cabinet was evaluated by means of the acceleration response of the isolated structure and its relationship with the ground acceleration.

Harmonic and sweep excitations were used to investigate the dynamic characteristics of the isolated cabinet; furthermore, an earthquake record was used to evaluate the seismic behaviour of the cabinet in presence of a realistic seismic excitation [8].

4.1 Harmonic Type Excitations

This kind of tests were conducted for two different amplitude values (0.005 and 0.010 m) and adopting a forcing frequency in the 0.5–5.0 Hz range; the accelerations of the platform and of the cabinet were detected and the two time histories were compared.

For each test, a frequency analysis of the two time histories (platform and cabinet accelerations) was performed and the amplitudes of the components, synchronous with that one imposed to the platform, were compared; their ratio,

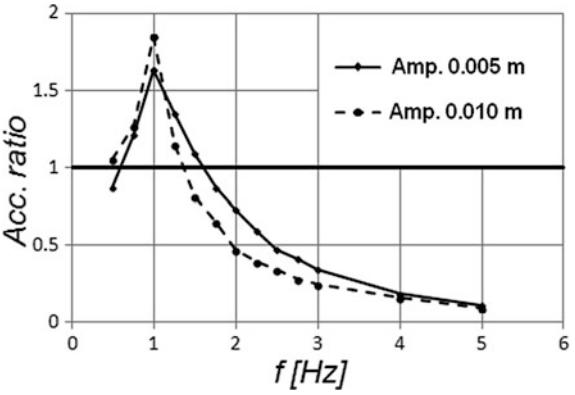


Fig. 15 Acceleration ratio versus forcing frequency

reported as function of the forcing frequency (Fig. 15), shows that the cabinet acceleration is amplified in the neighbourhood of 1 Hz and that the system is isolated for forcing frequencies greater than about 1.5 Hz.

An approximation of the isolated cabinet natural frequency can be easily obtained considering an equivalent linear mass-spring-damper system [9]; hence, the cabinet was modelled as a single degree of freedom system with four linear spring-damper devices.

Since the shear tests have shown that it is not possible to find constant values of stiffness and damping for the considered system, the isolator equivalent shear stiffness k_{eq}, adopted in the natural frequency calculation, was chosen equal to the mean of the 12 values reported in Table 1, so that:

$$f_n = \frac{1}{2\pi}\sqrt{\frac{4k_{eq}}{m}} = \frac{1}{2\pi}\sqrt{\frac{4 \cdot 1695}{165}} = 1.02 \text{ Hz} \tag{4}$$

where m is the cabinet mass.

The value obtained in (4) is in agreement with the experimental results reported in Fig. 15.

The dependence of the dynamic response of the isolated cabinet on the intensity of the ground motion highlights the nonlinear property of the WRS-BTU isolator.

It must be noted that the system under test is characterized by two vertical planes of symmetry and that the cabinet center of mass lies on the same vertical axis of the isolators stiffness center; therefore the three cabinet planar modes are uncoupled and the platform motion along the x (longitudinal) direction does not excite cabinet yaw rotation whose natural frequency can be estimated considering that the structure mass moment of inertia, with respect to its vertical symmetry axis, is equal to about 35 kgm^2 and that the yaw stiffness, provided by the isolators, is equal to:

$$k_\psi = 4\,k_{eq}b^2 = 1{,}186\,\text{Nm/rad} \tag{5}$$

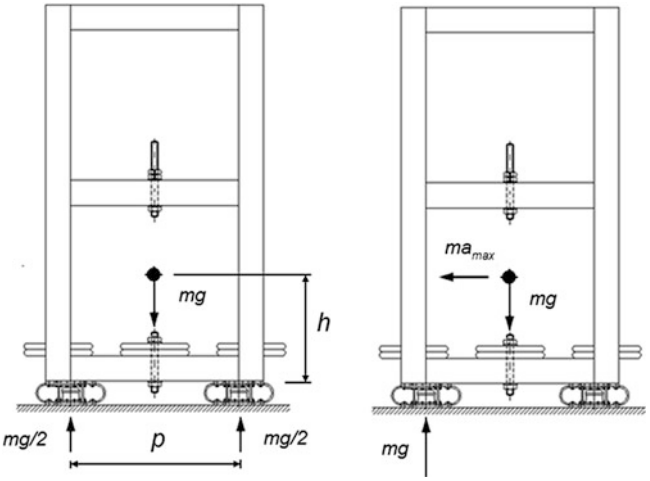


Fig. 16 BTU vertical reactions

being b the distance between each isolators and the vertical symmetry axis. The yaw natural frequency is therefore:

$$f_{\psi} = \frac{1}{2\pi}\sqrt{\frac{k_{\psi}}{I}} = 0.92\,\text{Hz} \tag{6}$$

Finally it can be notated that each isolator realizes an unilateral vertical constrain for the cabinet as it is unable to exert a downward reaction, until the ropes are not strained; for this reason the acceleration must not exceed the maximum value $a_{\max}$ for which a couple of devices exert null vertical reaction (Fig. 16). From a simple equilibrium condition it follows:

$$a_{\max} = \frac{g}{2}\frac{p}{h} \tag{7}$$

being p the distance between the BTUs along the motion direction and h the heigh of the center of mass.

The minimum cabinet p/h ratio is equal to 1.8 and therefore: $a_{\max} = 0.9g$, that is greater than the maximum seismic acceleration ($0.5g$ is a very high level seismic acceleration).

4.2 Sweep Responses

Driving the platform with a frequency sweep motion characterized by a constant amplitude equal to 5 mm and a frequency increasing with a rate of 0.01 Hz/s until 3 Hz both, platform and cabinet accelerations have been detected; reporting in the

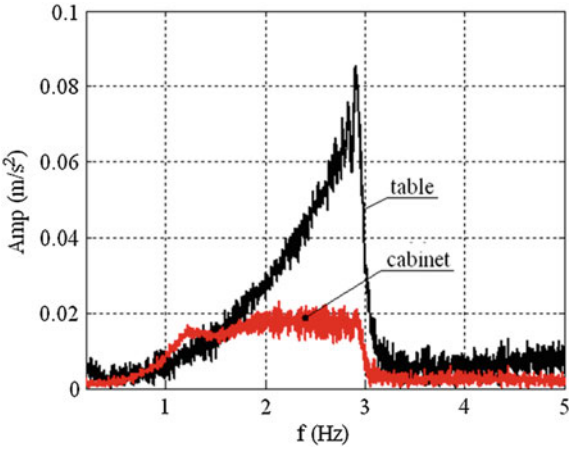


Fig. 17 Acceleration frequency spectrum for the sweep motion

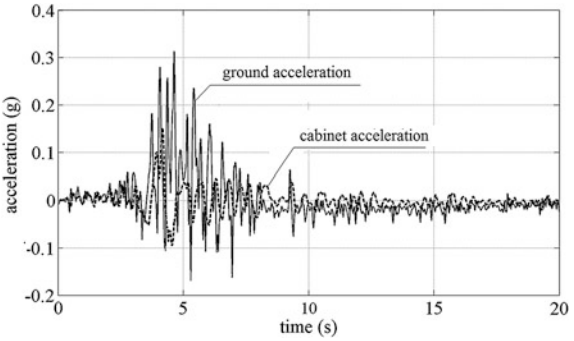


Fig. 18 Comparison between the ground and the cabinet acceleration

same diagram the corresponding frequency spectrum curves, it can be noted that while the acceleration platform amplitude grows with parabolic trend the cabinet acceleration achieves a peak in correspondence of a frequency value slightly greater than the horizontal natural frequency for which the cabinet acceleration exceeds the platform one; for excitation frequencies higher than 1.5 Hz the isolation system reduces the amount of acceleration transmitted to the cabinet.

Figure 17 shows that the cabinet acceleration is limited in all the tested frequency range.

4.3 Earthquake Responses

Figure 18 shows the cabinet and the platform acceleration for an earthquake ground excitation (Friuli, Italy 1976). In this case, the cabinet overall mass was chosen equal to 185 kg.

The isolation system reduces the acceleration transmitted to the cabinet by 52 % over the ground acceleration.

5 Conclusion

An experimental investigation of the performances of a WRS-BTU seismic isolator was presented. The proposed isolator is cheap and easy to realize by coupling a BTU with WRS springs; both the components are of normal industrial production and are already characterized by the manufacturers.

The tests on the WRS-BTU isolator were performed with a multi-purpose machine able to execute both shear tests on seismic isolators and shake-table tests on seismically isolated structures.

The results concerning the shear tests showed a nonlinear behaviour of the device. Moreover, shake-table experiments conducted on a laboratory cabinet, equipped with four WRS-BTU isolators, demonstrated their nonlinear dynamic characteristic and, at the same time, highlighted their good seismic isolation performance in terms of ground acceleration reduction.

References

1. R. Brancati, G. Di Massa, S. Pagano, E. Rocca, S. Strano, in *Experimental Investigation of the Performances of a WRS-BTU Seismic Isolator*. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013, London, UK, 3–5 July 2013, pp. 1646–1651
2. S. Pagano, S. Strano, Wire rope springs for passive vibration control of a light steel structure. WSEAS Trans. Appl. Theor. Mech. **8**(3), 212–222 (2013)
3. M. Kikuchi, ID Aiken, An analytical hysteresis model for elastomeric seismic isolation bearings. Earthq. Eng. Struct. Dyn. **26**(2), 215–231 (1997)
4. M. Cardone, S. Strano, Fluid-dynamic analysis of earthquake shaking table hydraulic circuit, in *ASME 11th Biennial Conference on Engineering Systems Design and Analysis (ESDA2012)*, vol 2 (2012), pp. 343–350
5. S. Strano, M. Terzo, A multi-purpose seismic test rig control via a sliding mode approach. Struct. Control Health Monit. doi:10.1002/stc.1641, in press
6. S. Pagano, M. Russo, S. Strano, M. Terzo, 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
7. C. Onorii, M. Spizzuoco, A. Calabrese, G. Serino, in *Applicability and Reliability of Innovative Low-Cost Rubber Isolators*. 14th ANIDIS Congress, Bari, Italy, 18–22 Sept 2011
8. A.H. Muhr, G. Bergamo, Shaking table tests on rolling-ball rubber-layer isolation system, in *14th European Conference on Earthquake Engineering*, vol 7, Ohrid, Republic of Macedonia, 30 Aug–3 Sept 2010, pp. 5703–5710
9. G. Di Massa, S. Pagano, S. Strano, F. Timpone, A comparison between linear and nonlinear modelling of a wire rope seismic isolator. Int. Rev. Model. Simul. **6**(4), 1307–1313 (2013)

A CFD Study of a pMDI Plume Spray

Ricardo F. Oliveira, Ana C. Ferreira, Senhorinha F. Teixeira, José C. Teixeira and Helena Cabral-Marques

Abstract Asthma is an inflammatory chronic disease characterized by airway obstructions disorders. The treatment is usually done by inhalation therapy, in which pressurized metered-dose inhalers (pMDIs) are preferred devices. The objective of this paper is to characterize and simulate a pMDI spray plume by introducing realistic factors through a computational fluid dynamics (CFD) study. Numerical simulations were performed with Fluent® software, by using a three-dimensional "testbox" for room environment representation. A salbutamol/HFA-134a formulation was used for characterization, whose properties taken as input for the CFD simulations. Spray droplets were considered to be composed by ethanol, salbutamol and HFA-134a. Propellant evaporation was taken into consideration, as well as, drag coefficient correction. Results showed an air temperature drop of 3.3 °C near the nozzle. Also, an increase in air velocity of 3.27 m/s was noticed. The CFD results seem to be in good agreement with Dunbar (1997) data on particle average velocity along the axial distance from the nozzle.

R. F. Oliveira (✉) · J. C. Teixeira
CT2M R&D Center, Department of Mechanical Engineering, University of Minho,
4800-058 Guimarães, Portugal
e-mail: ricardo.falcao.oliveira@gmail.com

J. C. Teixeira
e-mail: jt@dem.uminho.pt

A. C. Ferreira · S. F. Teixeira
CGIT R&D Center, Department of Production and Systems, University of Minho,
4800-058 Guimarães, Portugal
e-mail: acferreira@dps.uminho.pt

S. F. Teixeira
e-mail: st@dps.uminho.pt

H. Cabral-Marques
iMed.UL R&D Center, Faculty of Pharmacy, University of Lisbon,
1649-003 Lisbon, Portugal
e-mail: hcmarques@ff.ul.pt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_13,

Keywords Computational fluid dynamics · Discrete phase model · Drug particles · Lagrangian tracking · pMDI · Spray characterization

1 Introduction

The inhalation therapy is a cornerstone in the treatment of airway diseases. Asthma is a chronic inflammatory disorder associated with airway hyper responsiveness, which can be characterized by episodes of wheezing, breathing difficulties, chest tightness and coughing [1]. More than 300 million worldwide are affected by this disease which is responsible for the death of 220 thousand per year, growing at a rate of 50 % per decade [2]. Anti-inflammatory and bronchodilator drugs are used with the objective of reducing the inflammation of the pulmonary tissue, which causes the diameter reduction of the bronchus [3].

Pressurized metered-dose inhalers (pMDIs) are one of the major aerosol-generating devices used for aerosol delivery of bronchodilators in ambulatory patients [4]. Drug dose effectiveness in inhaled delivery is difficult to measure due to the fact that only a small fraction of the pMDI nominal dose reaches the lower respiratory tract. The pMDI is a small, cost-effective and very portable device containing between 100 and 400 doses. This device comprises a disposable canister with a pressurized mixture of propellants, surfactants, preservatives, flavoring agents and active drugs. This mixture is released from the canister through a metering valve [4].

The particle size of the aerosol produced by a pMDI depends on the pressure of the propellant mixture, ambient temperature, valve design, drug concentration and actuator orifices. In fact, there is a relationship between the actuator nozzle diameter and the particle size distribution, as well as, the ethanol concentration [5].

Moreover, the effectiveness of pMDIs is deeply associated with how the metering valve delivers, in an accurately and reproducibly manner, a measured volume and how it forms a propellant-tight seal for high pressure. According to Dhand [4], high-vapor-pressure propellants produce finer aerosol sprays, whereas increasing the drug concentration increases aerosol particle size. The actuator nozzle controls the atomization process in order to guarantee the formation of a spray plume. The canister, typically made of aluminum, holds a high internal pressure of 3–5 atm [6–8].

The application of computational fluid dynamics (CFD) in the design of aerosol drug delivery technologies has been proved to be a valuable tool when inhaler performance is investigated. The pMDI actuation is a complex phenomenon which involves turbulent flow, multiple phases, heat and mass transfer between the droplets and the environment. Several studies have been developed in order to model, numerically, pharmaceutical aerosols as a multi-phase flow, in which inhaled air is the continuous phase and the particles or droplets the discrete phase.

Dunbar et al. [9], performed a theoretical investigation of a pMDI spray by a CFD study consisting on the construction of actuator flow from the metered

chamber to the nozzle, which was based on a quasi-steady-state for flow analysis during a single actuation. The objective was to examine droplet formation and its trajectory during the inhaler actuation. The predicted results were validated against experimental data obtained using Phase Doppler Particle Anemometry (PDPA). Comparing the numerical results with the experimental data, it was observed that at a distance of 25 mm from the spray orifice, the droplet velocity and size distributions are in agreement, although such correlation does not hold further downstream [9].

Kleinstreuer et al. [10], experimentally validated a computational fluid-particle dynamics model developed to simulate the airflow, droplet spray transport and aerosol deposition in a pMDI considering several conditions, including different nozzle diameters and the use of a spacer. Also, the properties of both chlorofluorocarbon (CFC) and hydrofluoroalkane-134a (HFA) were investigated. The results indicated that the use of HFA, smaller valve orifices and the inclusion of spacers yields the best performance in terms of droplets deposition. Smyth et al. [11] also performed a spray pattern analysis for pMDIs, studying the influence of orifice size, particle size, and droplet motion correlations.

Recently, Ruzycki et al. [12] presented a comprehensive review in the use of CFD in inhaler design. The authors enlightened that the application of CFD modeling techniques for pMDIs, nebulizers and DPIs improves the aerosol transport and deposition understanding and, therefore, allows for an intuitive optimization of inhaler technologies whereas saving time and resources.

Previous studies in the simulation of the pMDI spray plume were made by the authors [13–16]. After an extensive review of the pMDI properties and characteristics, a CFD simulation was made but considering the particles to be solid (i.e. made of active pharmaceutical ingredient) [15].

This work aims to characterize and simulate a pMDI spray by means of a commercial CFD software (i.e. Fluent® v14.0 from ANSYS®). A pMDI salbutamol formulation was used for characterization. CFD simulations were performed in a three-dimensional "testbox". Spray droplets were injected and tracked accounting for propellant evaporation, aerodynamic size distribution, gravity, Brownian motion, drag coefficient corrections, turbulence and energy exchange. The input injection file was created using a Python language script.

2 Spray Characterization

2.1 Spray Dynamics

The spray dynamics can be effectively evaluated through images obtained by using a high speed digital video camera. This technique is able to record up to 10,000 frames per second, which is very suitable for understanding and capturing details of transient phenomena despite the difficulties related with the illumination required. Specifically for the delivery of aerosol drugs, the potential of such technique is

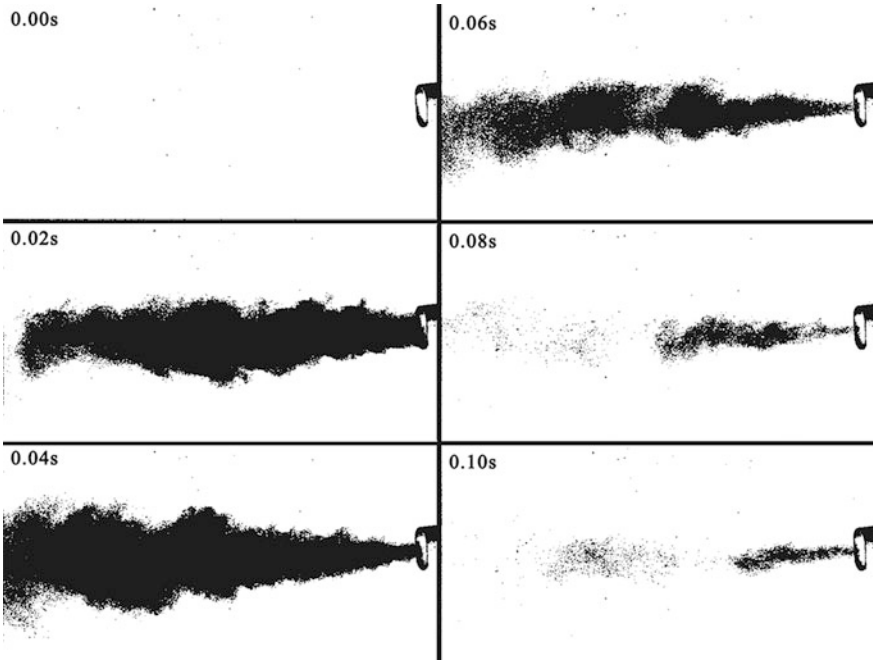


Fig. 1 High speed images of a puff taken from a salbutamol HFA-134a pMDI. These images were treated with greyscale and inverted colors after application of a threshold filter for easier visualization of the plume

suitable due to the very nature of the spray exiting from a high pressure canister. Nevertheless, the greatest advantage of this technique is its ability to capture the transient nature of the aerosol formation over the delivery time.

Using a high-speed camera (FASTCAM-APX RS 250KC), a puff event from the pMDI HFA-134a spray was recorded. Images were captured with an interval of 0.02 s (see Fig. 1). Those were taken at a rate of 6,000 frames per second, allowing to confirm the duration of the spray (0.1 s) and to calculate the spray angle (approximately 17°) by visual analysis.

2.2 Aerodynamic Size Distribution

Particle size distribution can be characterized either as Probability Density Function (PDF) or as a Cumulative Distribution Function (CDF). A particle size distribution is usually denoted by an independent variable, x, and two additional adaptable parameters [17].

The spray particle/droplets have been described through different mathematical distributions, being the Log-Normal, Rosin-Rammler and Nukiyama-Tanasawa the most cited. Amongst these distributions, it is well accepted that the pharmaceutical

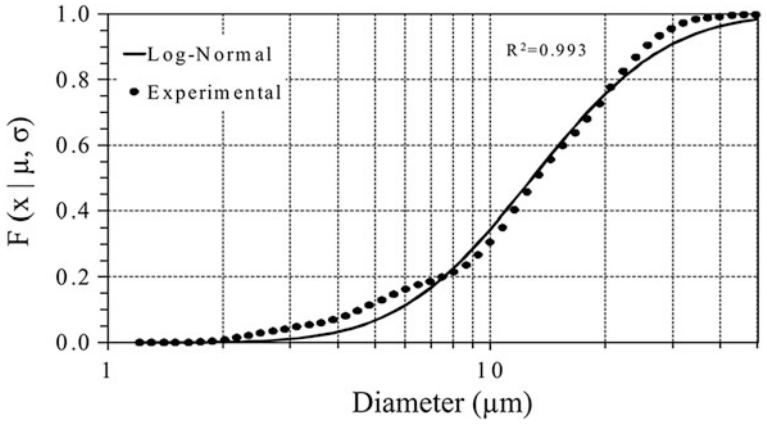


Fig. 2 Graphical representation of the pMDI HFA-134a salbutamol experimental data and their fitting for the log-normal CDF distribution. Measurements were obtained at 100 mm from the laser beam

aerosols can be reasonably represented by the Log-Normal distribution fitting the measured data to the CDF as shown by Eq. (1). The Log-Normal PDF (Eq. 2) was derived from the normal distribution [17].

$$F(x;\,\mu,\sigma) = \frac{1}{2}\,erfc\left[-\frac{\ln x - \mu}{\sigma\sqrt{2}}\right] \tag{1}$$

$$f(x;\,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\,e^{-(\ln x-\mu)^2/2\sigma^2} \tag{2}$$

where σ is the geometric standard deviation (which shall be $\neq 0$), μ represents the mean diameter and *erfc* is the complementary error function. Using the laser diffraction analysis technique (Malvern 2,600 particle sizer), a pMDI spray plume of HFA-134a formulation of salbutamol was measured. Data were fitted to the Log-Nornal CDF distribution model (1) using the least-squares method. Through the calculation of the Pearson coefficient of determination (i.e. R^2-squared value) and its maximization ($R^2 = 0.993$), the distribution parameters were obtained: $\mu = 2.55$ and $\sigma = 0.634$. Such values are in agreement with those usually reported in the literature. The experimental results and its Log-Normal CDF curve are shown in Fig. 2.

2.3 Axial Velocity

The velocity of the droplets decreases along the axial distance from the nozzle, due to the momentum exchange with the air. As reported by Dunbar [18], using PDPA measurements of a HFA-134a spray plume during an actuation, values were taken

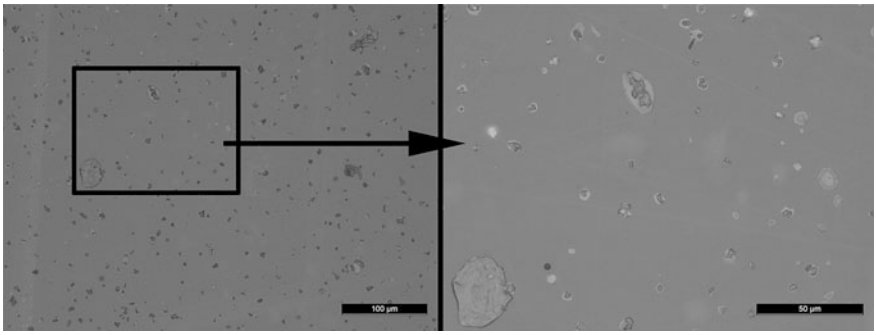


Fig. 3 pMDI spray particles and droplets observed through the optical microscope at two different magnifications

at different distances from the nozzle of the actuator. Consistent with their measurements, Dunbar concluded that a HFA propellant formulation produces a spray with higher velocities than a CFC formulation. This feature is due to the higher vapor pressure used in the HFA formulation. The plume behaves like a spray up to a distance of 75 mm from the nozzle and as an aerosol downstream that distance, where the droplet motion is being influenced by the gas [18].

2.4 Particle Shape and Size

Using an optical microscope (Leica DM2500 M) for the visual analysis of the particle shape and size, a set of images were taken. After a single puff being discharged against a glass plate, it was observed under the microscope at two different magnifications (see Fig. 3). It is possible to observe that the particles present a very irregular shape, although a limitation of the technique is the reduced depth of field. Also, it can be noticed that some particles present a solid craggy surface (i.e. salbutamol sulfate crystals) and others are encapsulated within a smooth spherical droplet of propellant that did not evaporate.

3 Spray Simulation

The pMDI formulation mainly consists of salbutamol, which is the most frequently prescribed short-acting β-agonist (SABA) [7, 19, 20].

This CFD study accounts for the temperature, velocity, turbulence, droplet tracking and evaporation of its propellant, as well as, its concentration in the air.

Table 1 Thermo-physical properties of the formulation

Properties	HFA-134a	Ethanol	Salbutamol
Density (kg/m^3)	1,311	790	1,230
Specific heat (J/kg K)	982	2,470	–
Latent heat (J/kg)	182,000	855,237	–
Boiling point (K)	247	351	–
Binary diffusivity (m^2/s)	9.709E−6	1.370E−5	–

3.1 Spray Injection Properties

The most important characteristics of the pMDI spray for the simulation are: spray cone angle (see Sect. 2.1); initial velocity (considered 100 m/s [8, 21]); aerodynamic size distribution; components present in the formulation; nozzle diameter (i.e. 0.25 mm [8, 10, 11, 21]); temperature (i.e. 215 K [22]); and mass flow rate. The aerodynamic size distribution parameters, discussed above in Sect. 2.2, used to configure the injection input file ranged from 1.22 to 49.5 μm, distributed along 80 intervals. From the knowledge of the drug dose delivered per puff labeled by the manufacturer (i.e. 100 μg) and the puff duration discussed in Sect. 2.1 (i.e. 0.1 s), a spray mass flow rate of 1.0E−6 kg/s was estimated.

For the creation of the input injection file, a Python language script was programmed. Into this file, the injections are placed, by a uniform random distribution, within the nozzle area. Properties such as diameter, temperature, and mass flow rate are attributed to each injection. After that assignment, the corresponding velocity components for each injection are calculated. They are calculated according to their distance to the center of the nozzle, so that their initial velocity vectors form a solid cone. The algorithm also calculates the corresponding mass flow rate value for each injection, as a function of its diameter assuming a Log-Normal distribution. It is ensured that the sum of all mass flow rates in the file equals the total mass flow rate defined initially. The total number of injections on the file was 16,200.

The formulation properties of the pMDI spray droplets were assumed to be composed by partial factions of HFA-134a (91.1 % w/w), ethanol (8.5 % w/w) and salbutamol (0.4 % w/w) [23]. The properties for each component are listed in Table 1.

The spray parameters used to configure the solver were obtained from various references, though some caution is required.

3.2 Geometry and Grid

The domain geometry was taken as a “testbox” consisting of a simple parallelepiped form with the dimensions of 0.2 × 0.2 × 0.3 m representing a fraction of a room environment. The pMDI actuator and canister were included in the middle of

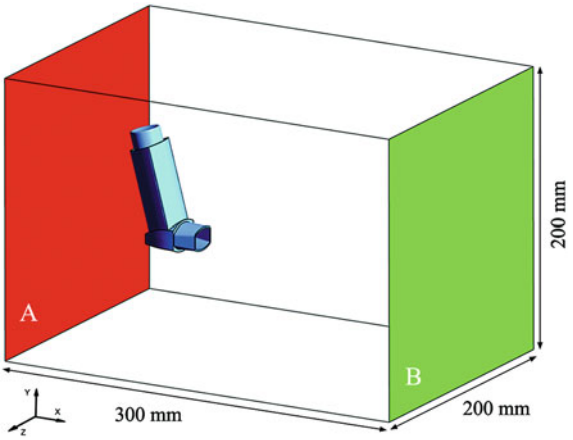


Fig. 4 A "testbox" representation where, the *red* plane (*A*) is the boundary condition 'velocity inlet' and the *green* plane (*B*) is the boundary 'outflow'

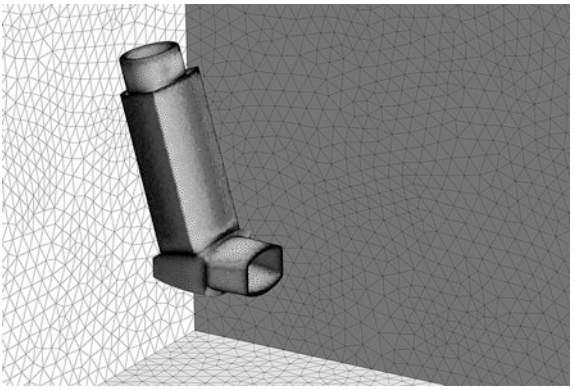


Fig. 5 3D domain grid representation, focusing the pMDI walls refinements for proximity and curvature

it. The spray injection point, the actuator's nozzle, is located in the origin point, see Fig. 4. The geometry was drawn using an external design program and then loaded into the ANSYS® platform.

The numerical grid, by discretization of the domain, was generated, consisting in tetrahedral and wedge elements, with sizes ranging from 0.1 to 20.0 mm. That resulted in a computational grid of 3,060,339 elements and 1,022,403 nodes. Several refinements, close to the wall zones of high proximity and curvature, were included (as shown in Fig. 5). The grid quality reports showed a good quality according to the Skewness parameter, with an average value of 0.21.

The boundary conditions were defined as: a 'Velocity Inlet' (see Fig. 4 "A"), forcing air to move uniformly inside the domain at 0.01 m/s and with a temperature of 293 K; and an 'Outflow' (see Fig. 4 "B"), enabling the freely motion of the air, as well as particles. For the remaining four external walls, a 'Symmetry' boundary condition was assumed. The pMDI actuator and canister boundaries were considered 'Wall', trapping all the particles that collide with them.

3.3 CFD Configuration

To account for the transient effects of a real pMDI spray plume, an unsteady simulation was made using a time step of 0.01 s for the flow field and 0.005 s for the particle tracking. The solution of the differential equations for mass and momentum was done in a sequential manner, using the SIMPLE algorithm [24–26]. The standard discretization scheme was used for the pressure and the second order upwind scheme for the energy, turbulence, momentum and air species concentration equations.

For the turbulence calculation, the SST k–ω model was used. This model is adequate for low-Re simulations and it has been used in the literature for this type of flow [10, 27–29].

Convergence was reached in the simulation by using a criterion value of 1.0E−5 for the continuity (pressure), velocity components, turbulence, species and a value of 1.0E−10 for the energy.

Droplets were considered as being multi component, as described above, where only the HFA is evaporating into the environment. This is initially simulated without any HFA gas, as well as, the air entering by the "velocity Inlet" boundary. As the HFA fraction is evaporating, it drastically reduces the diameter of the droplet, and changes its trajectory. On the other hand, the HFA concentration in the environment increases, making it harder for more particles to evaporate in areas of high concentration. The gravitational acceleration was assumed 9.81 m/s^2 along the -y axis direction. For the configuration of the Discrete Phase Model (DPM) droplet tracking model, the drag between both phases, Brownian motion for small particles and turbulence exchange were accounted. Through a User Defined Function (UDF), a customized drag law was included in the solver. This law was based in the work of Clift and his collaborators plus a correction for particles below 1 μm known as the Cunningham correction slip factor [28, 30].

The total number of particle streams injected during the simulation was approximately 323,200.

4 Results

4.1 Contour Fields

Figures 6, 7 and 8 show the contours for air velocity magnitude, air temperature and spray mass concentration, respectively, taken at a XY plane located at $z = 0$ and $t = 0.1$ s. The air velocity field (see Fig. 6) ranges from 0.0 to 3.27 m/s, being the lowest value found in almost all domain, because the ambient air was assumed stagnant. The maximum value is found at the nozzle exit, resulting from momentum exchange imposed by the high velocity spray particle injection.

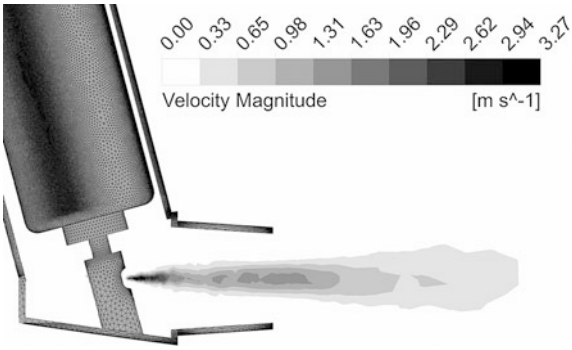


Fig. 6 Air velocity magnitude contours took at the XY plane (z = 0 and t = 0.1 s)

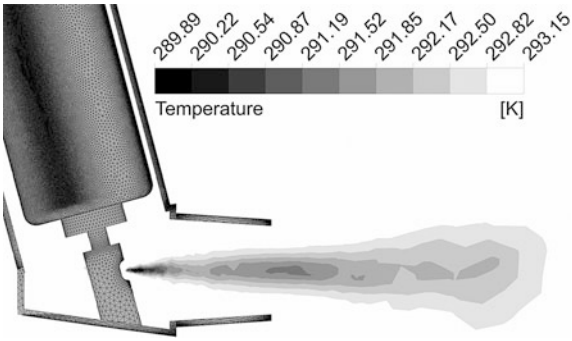


Fig. 7 Air temperature contours took at the XY plane (z = 0 and t = 0.1 s)

Observing the air temperature (see Fig. 7), it can be concluded that it ranges from 289.89 to 293.15 K, where the higher variation can be found at the spray plume formation zone. A sudden drop of 3.3 K occurs at the nozzle exit, due to the injection of droplets with an initial temperature of 215.15 K [22]. This temperature drop results from the energy exchange needed to evaporate the propellant.

Analyzing the HFA mass fraction present in the air (see Fig. 8), it can be perceived that its value is between 0 and 2.9E−3. The higher concentration zone, at the end of the injection period ($t = 0.1$ s), is located ahead of the nozzle, more specifically at the exit of the pMDI actuator mouthpiece. It has the shape of a spray plume. As expected, the droplets evaporate more in the periphery of the actuator zone, where HFA diffusion into the air is more effective.

4.2 Particle Trajectory

As shown in Fig. 9, the particle velocity magnitude in the spray plume ranges from approximately 100 m/s (as described in Sect. 3.1) to 0.01 m/s (the input air velocity as described in Sect. 3.2). The particles located downstream the actuator mouthpiece decelerate rapidly until they match the air velocity. Larger particles

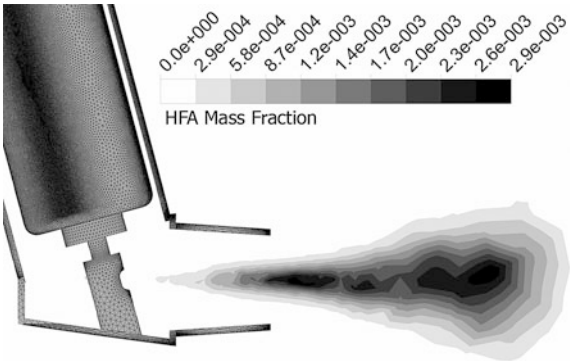


Fig. 8 HFA mass fraction contours took at the XY plane (z = 0 and t = 0.1 s)

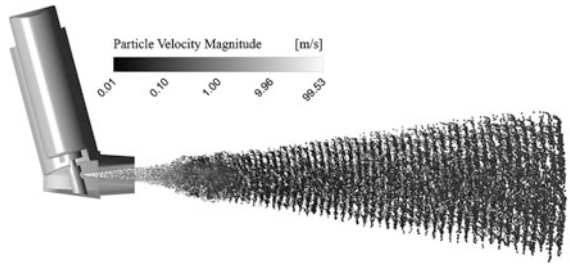


Fig. 9 Representation of the particle streams at the end of the injection (t = 0.11 s), with particles colored by its velocity magnitude. The particle streams are draw as spheres with proportional size scaled 20 times more than the real diameter

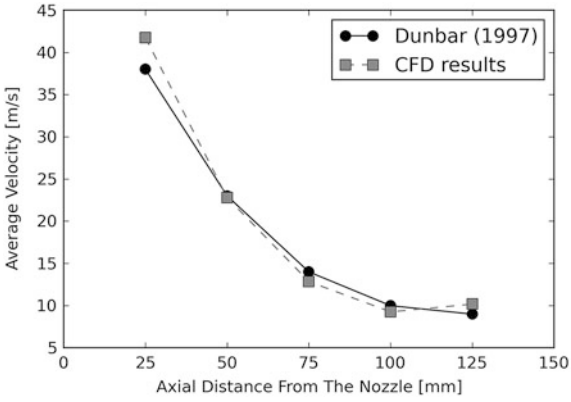


Fig. 10 Comparison of Dunbar [18] data with our CFD results, in terms of particle averaged velocity along the axial distance from the nozzle of a pMDI using a formulation only with HFA-134a

travel further into the still air than smaller ones. The influence of the gravitational acceleration is noticeable for higher particles when the slip velocity equals zero.

After post-processing of the CFD results, at different nozzle distances, results were obtained (see Figs. 10, 11).

In Fig. 10, the average velocity of the drug particles/droplets, at the spray plume centerline, as a function of the distance from the nozzle, is reported. The results are then compared against data reported by Dunbar [18] for a formulation

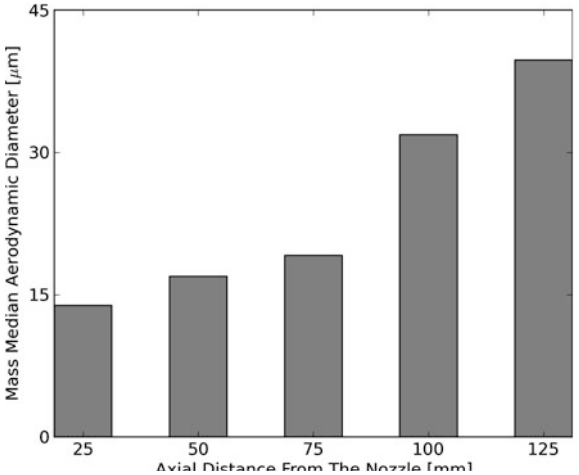


Fig. 11 Representation of the mass median aerodynamic diameter along the axial distance from the nozzle

containing only HFA 134a. A good agreement between the CFD and the experimental results was found.

Figure 11 reports the mass median aerodynamic diameter (MMAD), which is a common pharmaceutical measure to evaluate the size distribution of particles. This was calculated by considering the mass in the cumulative form and plotted against its corresponding diameter. Then, the diameter for which the mass fraction reached 50 % of the total was calculated. The results show that the value of MMAD increases with the distance from the nozzle, which can be indicative that particles of higher diameter travel further than the smaller ones.

5 Conclusions and Future Work

The study herein reported the characterization of a pMDI spray plume, of a salbutamol/HFA-134a formulation. The spray plume showed to be a transient jet, with effects that are dependent on the canister pressure. There is no constant delivery of the spray plume, but almost all the dose is aerosolized in the first 4/10 of the spray duration.

Microscopy showed that particles do not present a regular shape when in solid/dry state, although if they are involved in propellant they are almost spherical. The existence of a multi-component droplet confirmed the need for that approach in the simulation.

The aerodynamic size distribution of the pMDI sprays are usually accurately fitted by Log-Normal distribution. This was confirmed showing a good coefficient of determination for the experimental data obtained by laser diffraction analysis.

The spray characteristics were introduced in the CFD model and the spray droplets trajectory calculated in still air. Results showed that air temperature can

drop over 3.3 K and increase its velocity 3.27 m/s, in the proximities of the actuator's nozzle.

A good agreement between the CFD and the experimental data from Dunbar [18] was found, regarding the average velocity along the axial distance. Also, it was found that the MMAD value increases with the axial distance.

Spray characteristics used contributed to a correct configuration of the spray in the CFD software.

Acknowledgments The first author would like to express his acknowledgments for the support given by the Portuguese Foundation for Science and Technology (FCT) through the PhD Grant SFRH/BD/76458/2011. This work was financed by National Funds-Portuguese Foundation for Science and Technology, under Strategic Project PEst-C/EME/UI4077/2011 and PEst-OE/EME/UI0252/2011.

References

1. Global Initiative for Asthma: Global strategy for asthma management and prevention. GINA (2012)
2. M. Masoli, D. Fabian, S. Holt, R. Beasley, The global burden of asthma: executive summary of the GINA dissemination committee report. Allergy **59**, 469–478 (2004)
3. M.B. Dolovich, In my opinion—interview with the expert. Pediatr. Asthma Allergy Immunol. **17**, 292–300 (2004)
4. R. Dhand, Inhalation therapy with metered-dose inhalers and dry powder inhalers in mechanically ventilated patients. Respir. Care. **50**, 1331–1345 (2005)
5. S.W. Stein, Estimating the number of droplets and drug particles emitted from MDIs. AAPS PharmSciTech. **9**, 112–115 (2008)
6. S.P. Newman, in *Aerosols*, eds. by G.J. Laurent, S.D. Shapiro. Encyclopedia of Respiratory Medicine (Elsevier, Amsterdam, 2006), pp. 58–64
7. G. Crompton, A brief history of inhaled asthma therapy over the last fifty years. Prim. Care Respir. J. **15**, 326–331 (2006)
8. S.P. Newman, Principles of metered-dose inhaler design. Respir. Care. **50**, 1177–1190 (2005)
9. C.A. Dunbar, A.P. Watkins, J.F. Miller, An experimental investigation of the spray issued from a pMDI using laser diagnostic techniques. J. Aerosol Med. **10**, 351–368 (1997)
10. C. Kleinstreuer, H. Shi, Z. Zhang, Computational analyses of a pressurized metered dose inhaler and a new drug-aerosol targeting methodology. J. Aerosol Med. **20**, 294–309 (2007)
11. H. Smyth, A.J. Hickey, G. Brace, T. Barbour, J. Gallion, J. Grove, Spray pattern analysis for metered dose inhalers I: orifice size, particle size, and droplet motion correlations. Drug Dev. Ind. Pharm. **32**, 1033–1041 (2006)
12. C.A. Ruzycki, E. Javaheri, W.H. Finlay, The use of computational fluid dynamics in inhaler design. Expert Opin. Drug Deliv. **10**, 307–323 (2013)
13. R.F. Oliveira, S.F.C.F. Teixeira, L.F. Silva, J.C. Teixeira, H. Antunes, in *Study of a Pressurized Metered-dose Inhaler Spray Parameters in Fluent™*. WCE 2010—World Congress on Engineering 2010, London, UK (2010), pp. 1083–1087
14. R.F. Oliveira, S.F.C.F. Teixeira, L.F. Silva, J.C. Teixeira, H. Antunes, in *CFD Study of the Volumatic Spacer: A Realistic Approach*, eds. by J.C.F. Pereira, A. Sequeira, J.M.C. Pereira. V European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010. ECCOMAS, Lisbon, Portugal (2010)

15. R.F. Oliveira, S.F.C.F. Teixeira, J.C. Teixeira, L.F. Silva, H. Antunes, in *pMDI Sprays: Theory, Experiment and Numerical Simulation*, ed. by C. Liu. Advances in Modeling of Fluid Dynamics (Intech, Rijeka, Croatia, 2012), p. 300
16. R.F. Oliveira, A.C.M. Ferreira, S.F.C.F. Teixeira, J.C. Teixeira, H.M.C. Marques, in *pMDI Spray Plume Analysis: A CFD Study*, eds. by S.I. Ao, L. Gelman, D.W.L. Hukins, A. Hunter, A.M. Korsunsk. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE 2013 (Newswood Limited, London, UK, 2013), pp. 1883–1888
17. C.A. Dunbar, A.J. Hickey, Evaluation of probability density functions to approximate particle size distributions of representative pharmaceutical aerosols. J. Aerosol Sci. **31**, 813–831 (2000)
18. C.A. Dunbar, Atomization mechanisms of the pressurized metered dose inhaler. Part. Sci. Technol. **15**, 253–271 (1997)
19. G. Jepson, T. Butler, D. Gregory, K. Jones, Prescribing patterns for asthma by general practitioners in six European countries. Respir. Med. **94**, 578–583 (2000)
20. M.G. Zuidgeest, H.A. Smit, M. Bracke, A.H. Wijga, B. Brunekreef, M.O. Hoekstra, J. Gerritsen, M. Kerkhof, J.C. Jongste, H.G. Leufkens, Persistence of asthma medication use in preschool children. Respir. Med. **102**, 1446–1451 (2008)
21. A.R. Clark, MDIs: physics of aerosol formation. J. aerosol Med. **9**(Suppl. 1), S19–S26 (1996)
22. S.W. Stein, P.B. Myrdal, The relative influence of atomization and evaporation on metered dose inhaler drug delivery efficiency. Aerosol Sci. Technol. **40**, 335–347 (2006)
23. S.W. Stein, P. Sheth, P.B. Myrdal, A model for predicting size distributions delivered from pMDIs with suspended drug. Int. J. Pharm. **422**, 101–115 (2012)
24. J.H. Ferziger, M. Peric, *Computational Methods for Fluid Dynamics* (Springer, Berlin, 2003)
25. H.K. Versteeg, W. Malalasekera, *An Introduction to Computational Fluid Dynamics: The Finite Method* (Longman, Harlow, England, 1995)
26. ANSYS: *ANSYS Fluent Theory Guide* (ANSYS Inc, Canonsburg, PA, USA 2011)
27. P.W. Longest, G. Tian, R.L. Walenga, M. Hindle, Comparing MDI and DPI aerosol deposition using in vitro experiments and a new stochastic individual path (SIP) model of the conducting airways. Pharm. Res. **29**, 1670–1688 (2012)
28. P.W. Longest, J. Xi, Effectiveness of direct lagrangian tracking models for simulating nanoparticle deposition in the upper airways. Aerosol Sci. Technol. **41**, 380–397 (2007)
29. G. Tian, P.W. Longest, G. Su, R.L. Walenga, M. Hindle, Development of a stochastic individual path (SIP) model for predicting the tracheobronchial deposition of pharmaceutical aerosols: effects of transient inhalation and sampling the airways. J. Aerosol Sci. **42**, 781–799 (2011)
30. R. Clift, J.R. Grace, M.E. Weber, *Bubbles, Drops, and Particles* (Dover Publications, New York, 2005)

Harmonic Decomposition of Elastic Constant Tensor and Crystal Symmetry

Çiğdem Dinçkal

Abstract This paper presents a new outlook on harmonic decomposition method for elastic constant tensor. Harmonic decomposition method is developed in such a way that it is applied to anisotropic engineering materials exhibiting different crystal symmetry. The explicit results for each crystal symmetry types are presented. Numerical examples serve to illustrate and verify the developed method. This new representation of elastic constant tensor is compared with other theories such as orthogonal and non-orthogonal irreducible decompositions in literature. The results demonstrate that there are significant relationships between harmonic, non-orthogonal irreducible and orthogonal irreducible decomposition methods. While in harmonic and non-orthogonal irreducible decomposition methods, decomposition of total scalar part is not orthogonal. It is proposed that it is possible to make these parts orthogonal to each other.

Keywords Anisotropic engineering materials · Crystal symmetry · Elastic constant tensor · Harmonic decomposition method · Non-orthogonal irreducible decomposition method · Orthogonal irreducible decomposition method · Orthogonal

1 Introduction

Most of the elastic materials in engineering are anisotropic; metal crystals, fiber-reinforced composites, polycrystalline textured materials, biological tissues and rock structures. In order to gain a clear understanding the physical properties of

Ç. Dinçkal (✉)
Department of Civil Engineering, Çankaya University, Eskişehir yolu, 29.km 06810 Yenimahalle, Ankara, Turkey
e-mail: cdinckal@cankaya.edu.tr

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_14,

those materials, use of tensors by decomposing them is inevitable. Tensors are the most significant mathematical entities which describe direction dependent physical properties of solids and the tensor components characterizes physical properties which must be specified without reference to any coordinate system.

The well-known linear relation between the stress tensor σ_{ij} and the strain tensor ε_{kl} is the generalized Hooke's law

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl} \tag{1}$$

where σ_{ij}, ε_{kl} are second-order tensors and C_{ijkl} is of fourth order and C_{ijkl} is the elastic constant tensor (elastic coefficients). Each of the indices i, j, k and l takes all the values 1, 2 and 3. So the tensors are of dimension 3.

Elastic constant tensor must satisfy three important symmetry restrictions which are

$$C_{ijkl} = C_{jikl} \quad C_{ijkl} = C_{ijlk} \quad C_{ijkl} = C_{klij}. \tag{2}$$

These restrictions indicate the symmetry of the stress tensor, the symmetry of the strain tensor and the elastic strain energy. As a result, number of independent elastic constants is reduced from 81 to 21 [1]. In an anisotropic material, elastic constant tensor generally contains 21 independent (non-zero) coefficients. When the material has some kind of crystal symmetry, for instance orthorhombic or trigonal, the number of independent elastic constants is reduced if the coordinate axes coincide with symmetry axes for the material.

The indices are abbreviated according to the replacement rule given in Table 1.

In literature, the decomposition methods under the name of irreducible decomposition method had been studied extensively. To name few; [2, 3] and quoted by [4–9] carried out non-orthogonal irreducible decomposition for elastic constant tensor. Moreover, according to Sirotin [9], elastic constant tensor was decomposed with respect to general linear group and then orthogonal group O(3). He derived certain results for the irreducible tensors in their natural form. Besides Jerphagnon et al. [10] derived certain results for the irreducible tensors in their natural form. Andrews and Ghoul [11] followed the technique of Jerphagnon et al. [10] and gave the reduction of a fourth rank Cartesian tensor into irreducible parts under the three-dimensional rotation group. On the other hand Dinçkal [12] developed the work of Andrews and Ghoul [11] and realized orthogonal irreducible decomposition for construction materials such as tool steel and rock types by use of orthogonal decomposed parts.

Besides that there are also other works for harmonic decomposition of tensors. Firstly Backus [13] proposed a representation of elastic constant tensor in terms of harmonic tensors. These are based on an isomorphism between the space of homogeneous harmonic polynomials of degree q and the space of totally symmetric tensors of order q. Baerheim [14] followed Backus [13] and developed the method. Forte and Vianello [15] decomposed elastic constant tensor by use of harmonic decomposition method.

Table 1 Abbreviation of indices

Four index notation	11	22	33	23.32	13.31	12.21
Double index notation	1	2	3	4	5	6

This work is an extension of the study of Dinçkal [16]. It elaborates on harmonic decomposition method and new aspects are also presented. Main outline of the paper is summarized as; the method is described briefly, it is applied to various crystal symmetry types such as triclinic, monoclinic, trigonal, tetragonal, cubic and isotropic. Numerical examples are presented for each symmetry types. Developed method here, is compared with the other theories in literature. Finally, conclusions and future works pertinent to this work are stated.

One of the purposes of this work is to develop harmonic decomposition method and apply to various anisotropic materials from crystal symmetry types (excluding those presented in [16]) explicitly. For the first time in the literature, this decomposition process is applied to elastic constants of materials possessing anisotropic crystal symmetry types such as isotropy, cubic, tetragonal, trigonal, monoclinic and triclinic.

Another purpose is to figure out the relations between the results of the developed method and others (under the title of irreducible and harmonic decomposition methods) in the literature.

2 Harmonic Decomposition

The action of $SO(3)$ on a vector space is irreducible when there are no proper invariant subspaces. This infers that there is a decomposition of the space of elastic constant tensors (**E**la) into a direct sum of orthogonal subspaces on which the action of $SO(3)$ is irreducible. An important theorem of group representation theory is that every space on which the group of rotations acts irreducibly is isomorphic through an $SO(3)$-invariant map spaces of harmonic tensors (See for instance [5, 15]).

Moreover, there is an $SO(3)$-invariant isomorphism between **E**la and the direct sum R⊕ R⊕Dev⊕Dev⊕**Hrm**.

In harmonic decomposition, elastic constant tensor with fourth rank in three dimensions can be expressed as:

$$\mathrm{C} = \mathrm{S} + \mathrm{A} \tag{3}$$

S denotes the symmetric part of elastic constant tensor and it is defined as

$$\boldsymbol{S}_{ijkl} = \frac{1}{3}(C_{ijkl} + C_{iklj} + C_{iljk}) \tag{4}$$

$\boldsymbol{A}$ denotes asymmetric part and it is defined as

$$\boldsymbol{A}_{ijkl} = C_{ijkl} - \boldsymbol{S}_{ijkl} = \frac{2}{3}C_{ijkl} - \frac{1}{3}C_{iklj} - \frac{1}{3}C_{iljk} \tag{5}$$

The total symmetric part can be expressed in terms of H and H_{ij}.

$$\begin{aligned}\boldsymbol{S}_{ijkl} = H_{ijkl} + \left[\delta_{ij}H_{kl} + \delta_{kl}H_{ij} + \delta_{ik}H_{jl} + \delta_{jl}H_{ik} + \delta_{il}H_{jk} + \delta_{jk}H_{il}\right] \\ + H\left(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)\end{aligned} \tag{6}$$

The total asymmetric part can be rewritten in terms of h and h_{ij} as

$$\begin{aligned}\boldsymbol{A}_{ijkl} = \delta_{ij}h_{kl} + \delta_{kl}h_{ij} - \frac{1}{2}\delta_{ik}h_{jl} - \frac{1}{2}\delta_{jl}h_{ik} - \frac{1}{2}\delta_{il}h_{jk} - \frac{1}{2}\delta_{jk}h_{il} + h(\delta_{ij}\delta_{kl} - \frac{1}{2}\delta_{ik}h_{jl} \\ - \frac{1}{2}\delta_{il}h_{jk})\end{aligned} \tag{7}$$

By adding Eqs. 6 and 7 comes at harmonic decomposition of elastic constant tensor for anisotropic materials exhibiting triclinic symmetry. In this method, the notation of Baerheim [14] is used.

The details of harmonic decomposition for triclinic, monoclinic, trigonal, tetragonal, cubic and isotropic symmetry are presented respectively.

For triclinic symmetry:

The representation of elastic constant tensor in terms of harmonic tensors is

$$\begin{aligned}C_{ijkl} = H_{ijkl} + \left[\delta_{ij}H_{kl} + \delta_{kl}H_{ij} + \delta_{ik}H_{jl} + \delta_{jl}H_{ik} + \delta_{il}H_{jk} + \delta_{jk}H_{il}\right] + H\left[\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right] \\ + \delta_{ij}h_{kl} + \delta_{kl}h_{ij} - \frac{1}{2}\delta_{ik}h_{jl} - \frac{1}{2}\delta_{jl}h_{ik} - \frac{1}{2}\delta_{il}h_{jk} - \frac{1}{2}\delta_{jk}h_{il} + h\left(\delta_{ij}\delta_{kl} - \frac{1}{2}\delta_{ik}\delta_{jl} - \frac{1}{2}\delta_{il}\delta_{jk}\right)\end{aligned} \tag{8}$$

where $H = \frac{1}{45}\left(C_{ppqq} + 2C_{pqpq}\right)$, $h = \frac{1}{9}\left(C_{ppqq} - C_{pqpq}\right)$,

$$H_{ij} = \frac{1}{21}\left(C_{ijkk} - \frac{1}{3}C_{jjkk}\delta_{ij} + 2\left(C_{ikjk} - \frac{1}{3}C_{jkjk}\delta_{ij}\right)\right),$$

$$h_{ij} = \frac{2}{3}\left(C_{ijpp} - C_{ipjp}\right) - \frac{2}{9}\delta_{ij}\left(C_{rrpp} - C_{rprp}\right).$$

The total scalar (isotropic) part is obtained from Eq. 8 and it is expressed as

$$S = \frac{\delta_{ij}\delta_{kl}}{15}\left(2C_{ppqq} - C_{pqpq}\right) + \frac{\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)}{30}\left(3C_{pqpq} - C_{ppqq}\right) \tag{9}$$

where, first and second parts are respectively

$$S^{1}_{ijkl} = \frac{\delta_{ij}\delta_{kl}}{15}\left(2C_{ppqq} - C_{pqpq}\right) \tag{10}$$

$$S^2_{ijkl} = \frac{\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)}{30}\left(3C_{pqpq} - C_{ppqq}\right) \tag{11}$$

Moreover the total deviator part or second rank traceless tensor consists of summation of the linear combination of second order tensors (H_{ij} and h_{ij}) presented in Eq. 8.

$$\begin{aligned} D = {} & \frac{1}{7}\delta_{kl}\left(5C_{ijpp} - 4C_{ipjp}\right) + \frac{1}{7}\delta_{ij}\left(5C_{klpp} - 4C_{kplp}\right) + \frac{1}{7}\delta_{ik}\left(3C_{jplp} - 2C_{jlpp}\right) \\ & + \frac{1}{7}\delta_{il}\left(3C_{jpkp} - 2C_{jkpp}\right) + \frac{1}{7}\delta_{jl}\left(3C_{ipkp} - 2C_{ikpp}\right) + \frac{1}{7}\delta_{jk}\left(3C_{iplp} - 2C_{ilpp}\right) \\ & + \left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)\left(\frac{4}{21}C_{pprr} - \frac{2}{7}C_{rprp}\right) + \delta_{ij}\delta_{kl}\left(\frac{8}{21}C_{rprp} - \frac{10}{21}C_{pprr}\right). \end{aligned} \tag{12}$$

Similarly, from Eq. 8 harmonic part can be obtained as

$$\begin{aligned} H = {} & \frac{1}{3}\left(C_{ijkl} + C_{iklj} + C_{iljk}\right) - [\delta_{kl}\left(C_{ijmm} + 2C_{imjm}\right) + \delta_{ij}(C_{klmm} + 2C_{kmlm}) \\ & + \delta_{jl}(C_{ikmm} + 2C_{imkm}) + \delta_{ik}\left(C_{jlmm} + 2C_{jmlm}\right) + \delta_{jk}(C_{ilmm} + 2C_{imlm}) \\ & + \frac{1}{21}\delta_{il}\left(C_{jkmm} + C_{jmkm}\right)] + \frac{1}{105}\left(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) \end{aligned} \tag{13}$$

For monoclinic symmetry (classes: 2, m, 2/m):

In the case of monoclinic symmetry, there are five decomposed parts which are two scalars, two deviators and one harmonic part. By considering the symmetry condition of monoclinic symmetry, scalar parts in Eqs. 10 and 11 become

$$S^1_{ijkl} = \frac{\beta}{15}\delta_{ij}\delta_{kl}, \tag{14}$$

$$S^2_{ijkl} = \frac{\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)}{30}\alpha, \tag{15}$$

The deviators parts are found by rearranging the Eq. 12 and these parts are

$$\begin{aligned} D^1_{ijkl} = {} & \frac{1}{15}(2o\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + (2p + r)\left(\delta_{1i}\delta_{1k}\delta_{jl} + \delta_{1j}\delta_{1k}\delta_{il} + \delta_{1i}\delta_{1l}\delta_{jk} + \delta_{1j}\delta_{1l}\delta_{ik}\right) \\ & + (r - o)\left(\delta_{2i}\delta_{2k}\delta_{jl} + \delta_{2j}\delta_{2k}\delta_{il} + \delta_{2i}\delta_{2l}\delta_{jk} + \delta_{2j}\delta_{2l}\delta_{ik}\right)) + s(2\delta_{1i}\delta_{1j}\delta_{1k}\delta_{3l} \\ & + 2\delta_{1i}\delta_{1j}\delta_{3k}\delta_{1l} + 2\delta_{3i}\delta_{3j}\delta_{1k}\delta_{3l} + 2\delta_{3i}\delta_{3j}\delta_{3k}\delta_{1l} + 2\delta_{3i}\delta_{1j}\delta_{1k}\delta_{1l} + 2\delta_{1i}\delta_{3j}\delta_{1k}\delta_{1l} \\ & + 2\delta_{1i}\delta_{3j}\delta_{3k}\delta_{3l} + 2\delta_{3i}\delta_{1j}\delta_{3k}\delta_{3l} - \delta_{1i}\delta_{2j}\delta_{3k}\delta_{2l} - \delta_{1i}\delta_{2j}\delta_{2k}\delta_{3l} - \delta_{2i}\delta_{1j}\delta_{3k}\delta_{2l} - \delta_{2i}\delta_{1j}\delta_{2k}\delta_{3l}), \end{aligned} \tag{16}$$

$$D^2_{ijkl} = \frac{1}{105}(-4v(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + 10v\delta_{ij}\delta_{kl} - (4q + 2y)(\delta_{1i}\delta_{1k}\delta_{jl} + \delta_{1j}\delta_{1k}\delta_{il} + \delta_{1i}\delta_{1l}\delta_{jk} + \delta_{1j}\delta_{1l}\delta_{ik}) + (4v + 2q)(\delta_{2i}\delta_{2k}\delta_{jl} + \delta_{2j}\delta_{2k}\delta_{il} + \delta_{2i}\delta_{2l}\delta_{jk} + \delta_{2j}\delta_{2l}\delta_{ik})) + (5y - 5v)(\delta_{2k}\delta_{2l}\delta_{ij} + \delta_{2i}\delta_{2j}\delta_{kl}) + (5q - 5v)(\delta_{1k}\delta_{1l}\delta_{ij} + \delta_{1i}\delta_{1j}\delta_{kl}) + z(\delta_{1i}\delta_{1j}\delta_{1k}\delta_{3l} + \delta_{1j}\delta_{1k}\delta_{3k}\delta_{1l} + \delta_{3i}\delta_{1j}\delta_{1k}\delta_{1l} + \delta_{1i}\delta_{3j}\delta_{1k}\delta_{1l} + 5\delta_{2i}\delta_{2j}\delta_{1k}\delta_{3l} + 5\delta_{2i}\delta_{2j}\delta_{3k}\delta_{1l} + 5\delta_{1i}\delta_{3j}\delta_{2k}\delta_{2l} + 5\delta_{3i}\delta_{1j}\delta_{2k}\delta_{2l} + \delta_{3i}\delta_{3j}\delta_{1k}\delta_{3l} + \delta_{3i}\delta_{3j}\delta_{3k}\delta_{1l} + \delta_{1i}\delta_{3j}\delta_{3k}\delta_{3l} + \delta_{3i}\delta_{1j}\delta_{3k}\delta_{3l} - 2\delta_{2i}\delta_{3j}\delta_{1k}\delta_{2l} - 2\delta_{2i}\delta_{3j}\delta_{2k}\delta_{1l} - 2\delta_{3i}\delta_{2j}\delta_{1k}\delta_{2l} - 2\delta_{3i}\delta_{2j}\delta_{2k}\delta_{1l} - 2\delta_{1i}\delta_{2j}\delta_{2k}\delta_{3l} - 2\delta_{1i}\delta_{2j}\delta_{3k}\delta_{2l} - 2\delta_{2i}\delta_{1j}\delta_{2k}\delta_{3l} - 2\delta_{2i}\delta_{1j}\delta_{3k}\delta_{2l}), \tag{17}$$

Rearranging Eq. 13 gives harmonic part for monoclinic symmetry. This part is explicitly shown as

$$H_{ijkl} = \frac{1}{35}(-(a + b)(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + \delta_{ij}\delta_{kl}) + (3a + 2b)\delta_{ij}\delta_{kl} + (b - c)(\delta_{1i}\delta_{1k}\delta_{jl} + \delta_{1j}\delta_{1k}\delta_{il} + \delta_{1i}\delta_{1l}\delta_{jk} + \delta_{1j}\delta_{1l}\delta_{ik}) - (b - d)(\delta_{2i}\delta_{2k}\delta_{jl} + \delta_{2j}\delta_{2k}\delta_{il} + \delta_{2i}\delta_{2l}\delta_{jk} + \delta_{2j}\delta_{2l}\delta_{ik}) + (2e - 3a + 3c)(\delta_{2k}\delta_{2l}\delta_{ij} + \delta_{2i}\delta_{2j}\delta_{kl}) + (2f - 3a + 3d)(\delta_{1k}\delta_{1l}\delta_{ij} + \delta_{1i}\delta_{1j}\delta_{kl}) + 35(C_{12} + C_{33} + 2C_{66} - C_{23} - C_{13} - 2C_{44} - 2C_{55})(\delta_{1i}\delta_{1j}\delta_{2k}\delta_{2l} + \delta_{2i}\delta_{2j}\delta_{1k}\delta_{1l}) + 35(C_{11} + C_{33} - 2C_{13} - 4C_{55})\delta_{1i}\delta_{1j}\delta_{1k}\delta_{1l} + (35(C_{33} + C_{22} - 2C_{23} - 4C_{44})\delta_{2i}\delta_{2j}\delta_{2k}\delta_{2l}) + g(\delta_{1i}\delta_{1j}\delta_{1k}\delta_{3l} + \delta_{1i}\delta_{1j}\delta_{3k}\delta_{1l} + \delta_{1i}\delta_{3j}\delta_{1k}\delta_{1l} + \delta_{3i}\delta_{1j}\delta_{1k}\delta_{1l}) + h(\delta_{2i}\delta_{2j}\delta_{1k}\delta_{3l} + \delta_{2i}\delta_{2j}\delta_{3k}\delta_{1l} + \delta_{1i}\delta_{3j}\delta_{2k}\delta_{2l} + \delta_{3i}\delta_{1j}\delta_{2k}\delta_{2l} + \delta_{3i}\delta_{2j}\delta_{1k}\delta_{2l} + \delta_{3i}\delta_{2j}\delta_{2k}\delta_{1l} + \delta_{2i}\delta_{3j}\delta_{1k}\delta_{2l} + \delta_{2i}\delta_{3j}\delta_{2k}\delta_{1l} + \delta_{1i}\delta_{2j}\delta_{3k}\delta_{2l} + \delta_{1i}\delta_{2j}\delta_{2k}\delta_{3l} + \delta_{2i}\delta_{1j}\delta_{3k}\delta_{2l} + \delta_{2i}\delta_{1j}\delta_{2k}\delta_{3l}) + i(\delta_{3i}\delta_{3j}\delta_{3k}\delta_{1l} + \delta_{3i}\delta_{3j}\delta_{1k}\delta_{3l} + \delta_{3i}\delta_{1j}\delta_{3k}\delta_{3l} + \delta_{1i}\delta_{3j}\delta_{3k}\delta_{3l}) \tag{18}$$

where

$$\beta = C_{11} + C_{22} + C_{33} + 4C_{12} + 4C_{13} + 4C_{23} - 2C_{44} - 2C_{55} - 2C_{66},$$
$$\alpha = 2(C_{11} + C_{22} + C_{33}) + 6(C_{44} + C_{55} + C_{66}) - 2(C_{12} + C_{13} + C_{23}),$$
$$p = 2C_{11} - C_{22} - C_{33} - 2C_{44} + C_{55} + C_{66},$$
$$r = 2C_{22} - C_{11} - C_{33} + C_{44} - 2C_{55} + C_{66},$$
$$o = -p - r,$$
$$s = 3(C_{15} + C_{46} + C_{35}),$$
$$q = 2C_{11} - C_{22} - C_{33} + 5C_{12} + 5C_{13} - 10C_{23} + 8C_{44} - 4C_{55} - 4C_{66},$$
$$y = 2C_{22} - C_{11} - C_{33} + 5C_{12} - 10C_{13} + 5C_{23} - 4C_{44} + 8C_{55} - 4C_{66},$$
$$v = -q - y,$$
$$z = 3(C_{15} + C_{35} + 5C_{25} - 4C_{46}),$$
$$c = C_{11} - 4C_{22} - 4C_{33} - C_{12} - C_{13} + 9C_{23} + 18C_{44} - 2C_{55} - 2C_{66},$$
$$d = C_{22} - 4C_{11} - 4C_{33} - C_{12} + 9C_{13} - C_{23} - 2C_{44} + 18C_{55} - 2C_{66},$$
$$a = -c - d,$$
$$e = -b - c,$$

$$f = -b - d,$$
$$g = 5(4C_{15} - 3C_{35} - C_{25} - 2C_{46}),$$
$$h = 5(2C_{25} - C_{35} - C_{15} + 4C_{46}),$$
$$i = -g - h,$$

For trigonal symmetry (Classes: 3, $\bar{3}$):

For trigonal materials, there are five parts (two scalars, two deviators and one harmonic part). By taking into account the symmetry condition of trigonal symmetry, scalar parts are same as those found in Eqs. 14 and 15.

Deviator parts are obtained by arranging Eq. 12, these parts are illustrated as

$$D^1_{ijkl} = \frac{j}{30}\left(2\delta_{ik}\delta_{jl} + 2\delta_{il}\delta_{jk} - 3\delta_{3i}\delta_{3k}\delta_{jl} - 3\delta_{3j}\delta_{3k}\delta_{il} + \delta_{3i}\delta_{3l}\delta_{jk} + \delta_{3j}\delta_{3l}\delta_{ik}\right), \tag{19}$$

$$D^2_{ijkl} = \frac{k}{105}(-4\delta_{ik}\delta_{jl} - 4\delta_{il}\delta_{jk} + 10\delta_{ij}\delta_{kl} - 15\delta_{3i}\delta_{3j}\delta_{kl} - 15\delta_{3k}\delta_{3l}\delta_{ij} + 6\delta_{3i}\delta_{3k}\delta_{jl} + 6\delta_{3j}\delta_{3l}\delta_{ik} + 6\delta_{3j}\delta_{3k}\delta_{il} + 6\delta_{3i}\delta_{3l}\delta_{jk}), \tag{20}$$

According to symmetry conditions of trigonal materials, Eq. 13 becomes

$$\begin{aligned} H_{ijkl} = \frac{l}{35}&(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + \delta_{ij}\delta_{kl} - 5\delta_{3i}\delta_{3j}\delta_{kl} - 5\delta_{3k}\delta_{3l}\delta_{ij} - 5\delta_{3i}\delta_{3k}\delta_{jl} - 5\delta_{3j}\delta_{3k}\delta_{il} \\ &- 5\delta_{3i}\delta_{3l}\delta_{jk} - 5\delta_{3j}\delta_{3l}\delta_{ik} + 35\delta_{3i}\delta_{3j}\delta_{3k}\delta_{3l}) + C_{14}(\delta_{1i}\delta_{1j}\delta_{2k}\delta_{3l} + \delta_{1i}\delta_{1j}\delta_{3k}\delta_{2l} \\ &+ \delta_{3i}\delta_{2j}\delta_{1k}\delta_{1l} + \delta_{2i}\delta_{3j}\delta_{1k}\delta_{1l} + \delta_{3i}\delta_{1j}\delta_{2k}\delta_{1l} + \delta_{3i}\delta_{1j}\delta_{1k}\delta_{2l} + \delta_{1i}\delta_{3j}\delta_{2k}\delta_{1l} \\ &+ \delta_{1i}\delta_{3j}\delta_{1k}\delta_{2l}\delta_{1i}\delta_{2j}\delta_{1k}\delta_{3l} + \delta_{1i}\delta_{2j}\delta_{3k}\delta_{1l} + \delta_{2i}\delta_{1j}\delta_{1k}\delta_{3l} + \delta_{2i}\delta_{1j}\delta_{3k}\delta_{1l} \\ &- \delta_{2i}\delta_{2j}\delta_{2k}\delta_{3l} - \delta_{2i}\delta_{2j}\delta_{3k}\delta_{2l} - \delta_{3i}\delta_{2j}\delta_{2k}\delta_{2l} - \delta_{2i}\delta_{3j}\delta_{2k}\delta_{2l}) \end{aligned} \tag{21}$$

where $j = 3C_{11} - 2C_{33} - 2C_{44} - C_{12}$, $\quad k = 7C_{12} - C_{33} - C_{11} - 5C_{13} + 4C_{44}$
$l = C_{11} + C_{33} - 2C_{13} - 4C_{44}$.

For tetragonal symmetry (Classes: 4mmm, $\bar{4}2m$, 422, $4/mmm$):

In the case of tetragonal symmetry, elastic constant tensor can be represented in terms of five parts (two scalars, two deviators and one harmonic part). By considering the symmetry condition of tetragonal symmetry, scalar parts in Eqs. 10 and 11 take the following forms respectively:

$$S^1_{ijkl} = \frac{c}{15}\delta_{ij}\delta_{kl}, \tag{22}$$

$$S^2_{ijkl} = \frac{\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)}{30}d, \tag{23}$$

Similarly rearranging Eq. 12 gives the following deviator parts respectively

$$D^1_{ijkl} = \frac{e}{15}\left(2\delta_{ik}\delta_{jl} + 2\delta_{il}\delta_{jk} - 3\delta_{3i}\delta_{3k}\delta_{jl} - 3\delta_{3j}\delta_{3k}\delta_{il} + \delta_{3i}\delta_{3l}\delta_{jk} + \delta_{3j}\delta_{3l}\delta_{ik}\right), \tag{24}$$

$$\begin{aligned} D^2_{ijkl} = \frac{f}{105}(&-4\delta_{ik}\delta_{jl} - 4\delta_{il}\delta_{jk} + 10\delta_{ij}\delta_{kl} - 15\delta_{3i}\delta_{3j}\delta_{kl} - 15\delta_{3k}\delta_{3l}\delta_{ij} + 6\delta_{3i}\delta_{3k}\delta_{jl} \\ &+ 6\delta_{3j}\delta_{3l}\delta_{ik} + 6\delta_{3j}\delta_{3k}\delta_{il} + 6\delta_{3i}\delta_{3l}\delta_{jk}), \end{aligned} \tag{25}$$

With use of tetragonal symmetry conditions, Eq. 13 takes the following form

$$\begin{aligned} H_{ijkl} = \frac{1}{35}(&g\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + \delta_{ij}\delta_{kl}\right) - (g+h)\left(\delta_{3i}\delta_{3j}\delta_{kl} + \delta_{3k}\delta_{3l}\delta_{ij}\right) \\ &- (g+h)\left(\delta_{3i}\delta_{3k}\delta_{jl} + \delta_{3j}\delta_{3k}\delta_{il} + \delta_{3i}\delta_{3l}\delta_{jk} + \delta_{3j}\delta_{3l}\delta_{ik}\right) \\ &+ 35(C_{33} + C_{12} + 2C_{66} - 2C_{13} - 4C_{44})\delta_{3i}\delta_{3j}\delta_{3k}\delta_{3l} \\ &+ 35(C_{11} - C_{12} - 2C_{66})\left(\delta_{1i}\delta_{1j}\delta_{1k}\delta_{1l} + \delta_{2i}\delta_{2j}\delta_{2k}\delta_{2l}\right)\Big) \end{aligned} \tag{26}$$

where

$c = 2C_{11} + C_{33} + 4C_{12} + 8C_{13} - 4C_{44} - 2C_{66},$
$d = (2(2C_{11} + C_{33}) + 6(2C_{44} + C_{66}) - 2(C_{12} + 2C_{13})),$
$e = C_{11} - C_{33} - C_{44} + C_{66},$
$f = C_{11} - C_{33} + 5C_{12} - 5C_{13} + 4C_{44} - 4C_{66},$
$g = 9C_{12} + 18C_{66} - 8C_{11} - 2C_{13} - 4C_{44} + C_{33},$
$h = 4C_{33} + 3C_{11} - 8C_{13} + C_{12} - 16C_{44} + 2C_{66}.$

For cubic symmetry:

There are three decomposed parts which are two scalars and one harmonic part for cubic materials. According to the symmetry conditions of cubic materials, Eqs. 10 and 11 become

$$S^1_{ijkl} = \frac{1}{5}(C_{11} + 4C_{12} - 2C_{44}), \tag{27}$$

$$S^2_{ijkl} = \frac{1}{30}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)(6C_{11} + 18C_{44} - 6C_{12}), \tag{28}$$

Harmonic part of cubic symmetry is found similarly by arranging Eq. 13 and the explicit form of the harmonic part is

$$H_{ijkl} = \frac{c}{15}\left(5\delta_{ij}\delta_{kl}\delta_{jk} - \delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk} - \delta_{ij}\delta_{kl}\right) \tag{29}$$

where

$$c = C_{11} - C_{12} - 2C_{44}$$

For isotropic symmetry:

There are two independent components for isotropic elastic constant tensor. So it must have two decomposed parts. By considering the symmetry conditions in Eq. 2 and matrix structure of isotropic symmetry, Eqs. 10 and 11 are rearranged. As a result the explicit forms of scalar parts are

$$S^1_{ijkl} = C_{12}\delta_{ij}\delta_{kl} \tag{30}$$

$$S^2_{ijkl} = \frac{1}{2}(C_{11} - C_{12})\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right). \tag{31}$$

3 Numerical Examples

To support the analytic results of harmonic decomposition method, numerical examples are presented for different materials which exhibit triclinic, monoclinic, trigonal, tetragonal, cubic, isotropic symmetry. All units are in GPa.

Low Albite [17]:

As an example for triclinic symmetry, elastic coefficients of Low Albite have the following matrix structure:

$$\left[C_{ij}\right] = \begin{bmatrix} 69.1 & 34 & 30.8 & 5.1 & -2.4 & -0.9 \\ 34 & 183.5 & 5.5 & -3.9 & -7.7 & -5.8 \\ 30.8 & 5.5 & 179.5 & -8.7 & 7.1 & -9.8 \\ 5.1 & -3.9 & -8.7 & 24.9 & -2.4 & -7.2 \\ -2.4 & -7.7 & 7.1 & -2.4 & 26.8 & 0.5 \\ -0.9 & -5.8 & -9.8 & -7.2 & 0.5 & 33.5 \end{bmatrix} \tag{32}$$

By applying harmonic decomposition to elastic constant tensor in Eq. 32, elastic coefficients for Low Albite can be decomposed as

$$S_1 = \begin{bmatrix} 36.19 & 36.19 & 36.19 & 0 & 0 & 0 \\ 36.19 & 36.19 & 36.19 & 0 & 0 & 0 \\ 36.19 & 36.19 & 36.19 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{33}$$

$$S_2 = \begin{bmatrix} 82.32 & 0 & 0 & 0 & 0 & 0 \\ 0 & 82.32 & 0 & 0 & 0 & 0 \\ 0 & 0 & 82.32 & 0 & 0 & 0 \\ 0 & 0 & 0 & 41.16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 41.16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 41.16 \end{bmatrix} \tag{34}$$

$$D_1 = \begin{bmatrix} -57.147 & 0 & 0 & 0 & -1 & -3.64 \\ 0 & 32.853 & 0 & 1.48 & 0 & -3.64 \\ 0 & 0 & 24.293 & 1.48 & -1 & 0 \\ 0 & 1.48 & 1.48 & 14.287 & -1.82 & -0.5 \\ -1 & 0 & -1 & -1.82 & -8.213 & 0.74 \\ -3.64 & -3.64 & 0 & -0.5 & 0.74 & -6.073 \end{bmatrix} \tag{35}$$

$$D_2 = \begin{bmatrix} 0.042 & -0.433 & 0.538 & -1.357 & 0.737 & -0.197 \\ -0.433 & -0.215 & -0.105 & -0.271 & 3.686 & -0.197 \\ 0.538 & -0.105 & 0.173 & -0.271 & 0.737 & -0.986 \\ -1.357 & -0.271 & -0.271 & 0.0419 & 0.394 & -1.474 \\ 0.737 & 3.686 & 0.737 & 0.394 & -0.215 & 0.543 \\ -0.197 & -0.197 & -0.986 & -1.474 & 0.543 & 0.173 \end{bmatrix} \tag{36}$$

$$H = \begin{bmatrix} 7.692 & -1.76 & -5.931 & 6.457 & -2.137 & 2.937 \\ -1.76 & 32.349 & -30.589 & -5.109 & -11.386 & -1.963 \\ -5.931 & -30.589 & 36.52 & -9.909 & 7.363 & -8.814 \\ 6.457 & -5.109 & -9.909 & -30.589 & -0.974 & -5.226 \\ -2.137 & -11.386 & 7.363 & -0.974 & -5.932 & -0.783 \\ 2.937 & -1.963 & -8.814 & -5.226 & -0.783 & -1.76 \end{bmatrix} \tag{37}$$

where S_1, S_2 are scalars, D_1, D_2 are deviators and H is the harmonic part of elastic constant tensor for low Albite.

Gypsum [18]:

Gypsum is presented as an example of monoclinic symmetry and its elastic coefficients are

$$[C_{ij}] = \begin{bmatrix} 78.6 & 41 & 26.8 & 0 & -7 & 0 \\ 41 & 62.7 & 24.2 & 0 & 3.1 & 0 \\ 26.8 & 24.2 & 72.6 & 0 & -17.4 & 0 \\ 0 & 0 & 0 & 9.10 & 0 & -1.55 \\ -7 & 3.1 & -17.4 & 0 & 26.4 & 0 \\ 0 & 0 & 0 & -1.55 & 0 & 10.4 \end{bmatrix} \tag{38}$$

By applying harmonic decomposition, the matrix in Eq. 38 is represented in terms of following decomposed parts.

$$S_1 = \begin{bmatrix} 32.67 & 32.67 & 32.67 & 0 & 0 & 0 \\ 32.67 & 32.67 & 32.67 & 0 & 0 & 0 \\ 32.67 & 32.67 & 32.67 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{39}$$

$$S_2 = \begin{bmatrix} 34.61 & 0 & 0 & 0 & 0 & 0 \\ 0 & 34.61 & 0 & 0 & 0 & 0 \\ 0 & 0 & 34.61 & 0 & 0 & 0 \\ 0 & 0 & 0 & 17.31 & 0 & 0 \\ 0 & 0 & 0 & 0 & 17.31 & 0 \\ 0 & 0 & 0 & 0 & 0 & 17.31 \end{bmatrix} \tag{40}$$

$$D_1 = \begin{bmatrix} 10.8 & 0 & 0 & 0 & -10.38 & 0 \\ 0 & -15.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4.96 & 0 & -10.38 & 0 \\ 0 & 0 & 0 & -2.7 & 0 & -5.19 \\ -10.38 & 0 & -10.38 & 0 & 3.94 & 0 \\ 0 & 0 & 0 & -5.19 & 0 & -1.24 \end{bmatrix} \tag{41}$$

$$D_2 = \begin{bmatrix} 0.848 & 9.995 & -7.876 & 0 & -0.077 & 0 \\ 9.995 & 3.151 & -2.119 & 0 & -0.386 & 0 \\ -7.876 & -2.119 & -3.998 & 0 & -0.077 & 0 \\ 0 & 0 & 0 & 0.848 & 0 & 0.154 \\ -0.077 & -0.386 & -0.077 & 0 & 3.151 & 0 \\ 0 & 0 & 0 & 0.154 & 0 & -3.998 \end{bmatrix} \tag{42}$$

$$H = \begin{bmatrix} -0.334 & -1.669 & 2.003 & 0 & 3.457 & 0 \\ -1.669 & 8.023 & -6.354 & 0 & 3.486 & 0 \\ 2.003 & -6.354 & 4.3509 & 0 & -6.94 & 0 \\ 0 & 0 & 0 & -6.354 & 0 & 3.486 \\ 3.457 & 3.486 & -6.94 & 0 & 2.003 & 0 \\ 0 & 0 & 0 & 3.486 & 0 & -1.669 \end{bmatrix} \tag{43}$$

where S_1, S_2 are scalars, D_1, D_2 are deviators and H is the harmonic part of elastic constant tensor for Gypsum.

Haematite ($Fe_2 O_3$) [18]:

As an example of trigonal symmetry, Haematite is presented and its elastic coefficients are given

$$[C_{ij}] = \begin{bmatrix} 242 & 54.9 & 15.7 & -12.7 & 0 & 0 \\ 54.9 & 242 & 15.7 & 12.7 & 0 & 0 \\ 15.7 & 15.7 & 228 & 0 & 0 & 0 \\ -12.7 & 12.7 & 0 & 85.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 85.3 & -12.7 \\ 0 & 0 & 0 & 0 & -12.7 & 93.55 \end{bmatrix} \tag{44}$$

Applying harmonic decomposition to Eq. 44 yields

$$S_1 = \begin{bmatrix} 35.26 & 35.26 & 35.26 & 0 & 0 & 0 \\ 35.26 & 35.26 & 35.26 & 0 & 0 & 0 \\ 35.26 & 35.26 & 35.26 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{45}$$

$$S_2 = \begin{bmatrix} 189.09 & 0 & 0 & 0 & 0 & 0 \\ 0 & 189.09 & 0 & 0 & 0 & 0 \\ 0 & 0 & 189.09 & 0 & 0 & 0 \\ 0 & 0 & 0 & 94.54 & 0 & 0 \\ 0 & 0 & 0 & 0 & 94.54 & 0 \\ 0 & 0 & 0 & 0 & 0 & 94.54 \end{bmatrix} \tag{46}$$

$$D_1 = \begin{bmatrix} 5.93 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5.93 & 0 & 0 & 0 & 0 \\ 0 & 0 & -11.87 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1.48 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.48 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.97 \end{bmatrix} \tag{47}$$

$$D_2 = \begin{bmatrix} 3.37 & 16.857 & -8.43 & 0 & 0 & 0 \\ 16.857 & 3.37 & -8.43 & 0 & 0 & 0 \\ -8.43 & -8.43 & -6.74 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.37 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.37 & 0 \\ 0 & 0 & 0 & 0 & 0 & -6.74 \end{bmatrix} \tag{48}$$

$$H = \begin{bmatrix} 8.35 & 2.783 & -11.13 & -12.7 & 0 & 0 \\ 2.783 & 8.35 & -11.13 & 12.7 & 0 & 0 \\ -11.13 & -11.13 & 22.26 & 0 & 0 & 0 \\ -12.7 & 12.7 & 0 & -11.13 & 0 & 0 \\ 0 & 0 & 0 & 0 & -11.13 & -12.7 \\ 0 & 0 & 0 & 0 & -12.7 & 2.783 \end{bmatrix} \tag{49}$$

where S_1, S_2 are scalars, D_1, D_2 are deviators and H is the harmonic part of elastic constant tensor for Haematite.

Zircon ($ZrSiO_4$, metamict) [19]:

Zircon is given as an example of tetragonal symmetry, the elastic coefficients of Zircon are

$$[C_{ij}] = \begin{bmatrix} 284 & 73 & 119 & 0 & 0 & 0 \\ 73 & 284 & 119 & 0 & 0 & 0 \\ 119 & 119 & 309 & 0 & 0 & 0 \\ 0 & 0 & 0 & 77.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 77.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 47.7 \end{bmatrix} \tag{50}$$

By applying harmonic decomposition, the matrix in Eq. 50 is decomposed as

$$S_1 = \begin{bmatrix} 114.37 & 114.37 & 114.37 & 0 & 0 & 0 \\ 114.37 & 114.37 & 114.37 & 0 & 0 & 0 \\ 114.37 & 114.37 & 114.37 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{51}$$

$$S_2 = \begin{bmatrix} 156.55 & 0 & 0 & 0 & 0 & 0 \\ 0 & 156.55 & 0 & 0 & 0 & 0 \\ 0 & 0 & 156.55 & 0 & 0 & 0 \\ 0 & 0 & 0 & 78.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 78.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 78.3 \end{bmatrix} \tag{52}$$

$$D_1 = \begin{bmatrix} -14.6 & 0 & 0 & 0 & 0 & 0 \\ 0 & -14.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 29.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.65 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.65 & 0 \\ 0 & 0 & 0 & 0 & 0 & -7.31 \end{bmatrix} \tag{53}$$

$$D_2 = \begin{bmatrix} -2.59 & -12.9 & 6.47 & 0 & 0 & 0 \\ -12.9 & -2.59 & 6.47 & 0 & 0 & 0 \\ 6.47 & 6.47 & 5.17 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2.59 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2.59 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5.17 \end{bmatrix} \tag{54}$$

$$H = \begin{bmatrix} 30.28 & -28.44 & -1.84 & 0 & 0 & 0 \\ -28.44 & 30.28 & -1.84 & 0 & 0 & 0 \\ -1.84 & -1.84 & 3.68 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1.84 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.84 & 0 \\ 0 & 0 & 0 & 0 & 0 & -28.44 \end{bmatrix} \tag{55}$$

where S_1, S_2 are scalars, D_1, D_2 are deviators and H is the harmonic part of elastic constant tensor for Zircon.

Aluminium Antimonide (AlSb) [18]:

As an example of cubic symmetry, the elastic coefficients of AlSb are presented.

$$[C_{ij}] = \begin{bmatrix} 87.7 & 43.4 & 43.4 & 0 & 0 & 0 \\ 43.4 & 87.7 & 43.4 & 0 & 0 & 0 \\ 43.4 & 43.4 & 87.7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 40.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 40.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 40.8 \end{bmatrix} \tag{56}$$

By employing harmonic decomposition, the matrix in Eq. 56 is decomposed as

$$S_1 = \begin{bmatrix} 35.9 & 35.9 & 35.9 & 0 & 0 & 0 \\ 35.9 & 35.9 & 35.9 & 0 & 0 & 0 \\ 35.9 & 35.9 & 35.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{57}$$

$$S_2 = \begin{bmatrix} 66.68 & 0 & 0 & 0 & 0 & 0 \\ 0 & 66.68 & 0 & 0 & 0 & 0 \\ 0 & 0 & 66.68 & 0 & 0 & 0 \\ 0 & 0 & 0 & 33.34 & 0 & 0 \\ 0 & 0 & 0 & 0 & 33.34 & 0 \\ 0 & 0 & 0 & 0 & 0 & 33.34 \end{bmatrix} \tag{58}$$

$$H = \begin{bmatrix} -14.9 & 7.5 & 7.5 & 0 & 0 & 0 \\ 7.5 & -14.9 & 7.5 & 0 & 0 & 0 \\ 7.5 & 7.5 & -14.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 7.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7.5 \end{bmatrix} \tag{59}$$

where S_1 and S_2 are scalar and H is the harmonic part of elastic constant tensor for AlSb.

Reactor Pressure Vessel (RPV) Steel [20]:

Reactor pressure vessel (RPV) steel is presented as an example for the isotropic material. Especially textured and non-crystalline materials possess isotropic symmetry. There are two independent elastic constants which are C_{11}, C_{12}. For RPV steel, the elastic coefficients are

$$[C_{ij}] = \begin{bmatrix} 277.001 & 118.715 & 118.715 & 0 & 0 & 0 \\ 118.715 & 277.001 & 118.715 & 0 & 0 & 0 \\ 118.715 & 118.715 & 277.001 & 0 & 0 & 0 \\ 0 & 0 & 0 & 79.143 & 0 & 0 \\ 0 & 0 & 0 & 0 & 79.143 & 0 \\ 0 & 0 & 0 & 0 & 0 & 79.143 \end{bmatrix} \tag{60}$$

Applying harmonic decomposition to Eq. 60 gives

$$S_1 = \begin{bmatrix} 118.715 & 118.715 & 118.715 & 0 & 0 & 0 \\ 118.715 & 118.715 & 118.715 & 0 & 0 & 0 \\ 118.715 & 118.715 & 118.715 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{61}$$

$$S_2 = \begin{bmatrix} 158.286 & 0 & 0 & 0 & 0 & 0 \\ 0 & 158.286 & 0 & 0 & 0 & 0 \\ 0 & 0 & 158.286 & 0 & 0 & 0 \\ 0 & 0 & 0 & 79.143 & 0 & 0 \\ 0 & 0 & 0 & 0 & 79.143 & 0 \\ 0 & 0 & 0 & 0 & 0 & 79.143 \end{bmatrix} \tag{62}$$

where S_1 and S_2 are scalar parts for RPV steel.

4 Comparing the Decomposition Theories

For comparison purposes, critical relationships between the decomposition theories are found out.

Recall that there are many works done on non-orthogonal irreducible and harmonic decomposition method in the literature. All these methods have several characteristics in common. Elastic constant tensor is decomposed into five parts which are two scalars, two deviators and the nonor part by irreducible and harmonic methods. The only difference is the term 'harmonic' is used instead of 'nonor' in harmonic decomposition. Total scalar, total deviators and nonor part are identical in all methods. Contrary, the components of scalar and deviator parts in orthogonal irreducible decomposition are not equal to those in non-orthogonal

irreducible decomposition and harmonic decomposition methods. Hence it proves that there is not a unique decomposition for both scalar and deviator parts. In other words, total scalar and deviator parts can be decomposed into infinitely many independent components. This is the first relationship between these methods.

In orthogonal irreducible decomposition, components of total scalar parts are orthogonal to each other [12]. On the other hand, decomposed parts of total scalar part are not orthogonal to each other in harmonic decomposition method due to the expression for decomposition of elasticity coefficients given in Eqs. 10 and 11.

Since $\delta_{ij}\delta_{kl}$ is not orthogonal to $\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$ (which is represented as $2I_{ijkl}$). If $\delta_{ij}\delta_{kl}$ is replaced with $2I_{ijkl}$ by the hydrostatic and deviatoric operators.

$$I^h_{ijkl} = \frac{1}{3}\delta_{ij}\delta_{kl}, \quad I^d_{ijkl} = I_{ijkl} - \frac{1}{3}I^h_{ijkl}, \tag{63}$$

respectively, then the expressions in Eqs. 6 and 7 in the work of Dinçkal [12] are obtained in which decomposed parts of total scalar parts are orthogonal to each other. So the components of total scalar part in harmonic decomposition method take the form of $C^{(0;1)}_{ijkl}$ and $C^{(0;2)}_{ijkl}$ (See [12]). This case is a significant innovation for the orthogonal irreducible and non-orthogonal irreducible and harmonic decomposition methods for elastic constant tensor. It is the second relationship between decomposition theories.

5 Conclusion and Future Work

Consequently, comparison of the decomposition theories reveals the following major results:

1. Harmonic decomposition method can be made orthogonal only if it converts into orthogonal irreducible method by defining and using hydrostatic and deviatoric operators.
2. Components of scalar and deviator parts for both orthogonal irreducible harmonic and non-orthogonal irreducible decomposition methods are different. Therefore it proves that there is not a unique decomposition for both deviator and scalar parts. Those parts can be decomposed into infinitely many independent components, while total scalar, total deviator and nonor parts are identical for all methods.

Representation of elastic constant tensor in terms of harmonic tensors provides not only a deeper understanding of tensor structure but also simplify immensely the calculations of sums, products, inverses and inner products. Moreover, harmonic decomposition method has many applications in different subjects of physics and engineering (atomic and molecular physics and the physics of

condensed matter). It has more significant effects on many applications in different fields such as:

1. Examining material symmetry types in detail. It is possible to decide which type of symmetry a material has when the elastic constants are measured relative to an arbitrary coordinate system. A second rank symmetric tensor associated to the elastic constant tensor can be used to verify if the coordinate axes are the symmetry axes of the material and determine a symmetry coordinate system.
2. Determination of materials possessing same crystal symmetry type which are highly anisotropic or close to isotropy by use of norm concept.
3. Understanding the mechanical and elastic behavior of natural composites such as bone and wood types.

As future works, the following problems can be studied:

- pure shear and pure longitudinal wave propagation in different anisotropic materials,
- the relation between the decomposed parts and the angle of orientation of fibers,
- the relation between the decomposed parts and the material properties of fibers and matrix in fiber reinforced composites.

References

1. J.F. Nye, *Physical properties of crystals: their representation by tensors and matrices* (Clarendon Press, Oxford, 1957)
2. E.T. Onat, Effective properties of elastic materials that contain penny shaped voids. J. Elast. **22**, 1013–1021 (1984)
3. J.P. Boehler, A.A. Kirillov, E.T. Onat, On the polynomial invariants of the elasticity tensor. J. Elast. **34**, 97–110 (1994)
4. S.C. Cowin, Properties of the anisotropic elasticity tensor. Q. J. Mech. Appl. Mech. **42**, 249–267 (1989)
5. S. Forte, M. Vianello, Restricted invariants on the space of elasticity tensor. Math. Mech. Solids **11**, 48–82 (2006)
6. T.C.T. Ting, Q.C. He, Decomposition of elasticity tensors and tensors that are structurally invariant in three dimensions. Q. J. Mech. Appl. Mech. **59**, 323–341 (2006)
7. X. Zheng, Identification of symmetry planes in weakly anisotropic elastic media. Geophys. J. Int. **170**, 468–478 (2007)
8. B.D. Annin, N.I. Ostrosablin, Anisotropy of elastic properties of materials. J. Appl. Mech. Tech. Phys. **49**, 998–1014 (2008)
9. Y.I. Sirotin, Decomposition of material tensors into irreducible elements. Kristallografiya **19**(5), 909–915 (1974)
10. J. Jerphagnon, D. Chemla, R. Bonneville, The description of the physical properties of condensed matter using irreducible tensors. Adv. Phys. **8**, 633–671 (1970)
11. D.L. Andrews, W.H. Ghoul, Irreducible fourth-rank Cartesian tensors. Phys. Rev. A **25**, 2647–2657 (1982)

12. Ç. Dinçkal, A new decomposition method for elastic constant tensor to study the anisotropy of construction materials; tool steel and rock types. Elixir Int. J. Appl. Math. **39**, 5090–5092 (2011)
13. G. Backus, A geometrical picture of anisotropic elastic tensors. Rev. Geophys. Space Phys. **8**, 633–671 (1970)
14. R. Baerheim, Harmonic decomposition of the anisotropic elasticity tensor. Q. J. Mech. Appl. Mech. **46**, 391–418 (1993)
15. S. Forte, M. Vianello, Symmetry classes for elasticity tensors. J. Elast. **43**, 81–108 (1996)
16. Ç. Dinçkal, On the mechanical and elastic properties of anisotropic engineering materials based upon harmonic representations, in *Proceedings of the World Congress on Engineering 2013*, *WCE 2013*, 3–5 July 2013. Lecture Notes in Engineering and Computer Science, London, UK (2013), pp. 2177–2183
17. J.M. Brown, E.H. Abramson, R.J. Angel, Triclinic elastic constants for low albite. Phys. Chem. Miner. **33**(4), 256–265 (2006)
18. H. Landolt, R. Börnstein, M.M. Choy, K.H. Hellwege, A.M. Hellwege, Numerical Data and Functional Relationships in Science and Technology, New Series, Group III (Crystal and Solid State Physics), vol. 11 (Springer, Berlin, 1979)
19. H. Özkan, L. Cartz, in *AIP Conference 1973*, *AIP Conference Proceedings*, *1974*, vol. 17, p. 21
20. Y.-M. Cheong, H.-K. Jung, Y.-S. Joo, S.-S. Kim, Y.-S. Kim, Dynamic elastic constants of Weld HAZ of SA 508 CL.3 steel using resonant ultrasound spectroscopy. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **47**(3), 559–564 (2000)

DLC Coated Piston Skirts Behavior at Initial IC Engine Start Up

Zahid ur Rehman, S. Adnan Qasim and M. Afzaal Malik

Abstract The diamond-like carbon (DLC) based coating protects the lubricant-starved dry surface of the piston skirts against wear in a low-load and speed initial start-up of an internal combustion (IC) engine. Despite a relatively large radial clearance a physical contact between the skirts and the cylinder liner causes an elastic deformation of the DLC coated surface producing elastic stress and strains of noticeable amplitudes. The elastic stress accumulation may facilitate a subsequent de-lamination of the DLC coating. This study develops the numerical piston dynamics model by incorporating the secondary eccentric displacements of the piston skirts and their contact with the engine cylinder liner during the 720-degree crank rotation cycle. The contact zone in the boundary-value problem is discretized using the finite difference method. The elastic surface displacements, stresses and strains are determined by applying the theory of elasticity and solving the Navier's or Lame's equation numerically. The nature and extent of the displacements, stresses and strains produced at the interface of a fairly thin DLC coating with the substrate are analyzed. The results highlight the extent of the depth of the elastic surface displacements, the stresses and strains produced in the coating and the substrate materials. The simulation results show that the dry piston skirts establish a physical contact with the liner in the compression stroke that is maintained in the expansion and exhaust strokes. The stresses produced at the low

Z. ur Rehman (✉) · S. A. Qasim
Department of Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan
e-mail: zahid.rehman@yahoo.com

S. A. Qasim
e-mail: adnan_qasim@yahoo.com

M. A. Malik
Department of Mechanical and Aerospace Engineering, Air University, Islamabad, Pakistan
e-mail: drafzaalmalik@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_15,

engine start-up speed are significant, increase beneath the surface and accumulate at the interface between the coating and the substrate. The elastic displacements of a fairly thin DLC coating prevents the stress accumulation on the substrate and protects it from adhesive wear.

Keywords DLC coating · Dry contact · Elasticity · IC engine · Initial engine start-up · Piston skirt

Nomenclature

a	Vertical distance from the top of piston skirt to the piston-pin
b	Vertical distance from the top of piston skirt to the piston center of gravity
C_g	Horizontal distance between piston center of mass and piston pin
C_p	Distance of the piston-pin from the axis of piston
$\ddot{e}_b$, $\ddot{e}_t$	Acceleration term of piston skirts bottom and top eccentricities
E1, *E2*	Young's modulus of coating and piston
F	Normal force acting on piston skirts
F_{fc}	Friction force due to coulombs' friction
F_G	Combustion gas force acting on the top of piston
F_{IC}	Inertia force due to piston mass
F_{IP}	Inertia force due to piston pin mass
G, μ	Lam's constants
I_{pis}	Piston rotary inertia about its center of mass
l	Connecting rod length
L	Piston skirt length
m_{pin}	Mass of piston-pin
m_{pis}	Mass of piston
M_f	Moment about piston-pin due to friction force
ϕ	Connecting rod angle
r	Crank radius
R	Radius of piston
τ	Shear stress
U	Velocity of piston
υ_1, υ_2	Poisson's ratio of coating and piston
ω	Crankshaft speed
y	Distance to the direction of sliding
x	Normal to the direction of sliding in depth of coating-substrate
z	Normal to the direction of sliding in width of coating-substrate
ψ	Crank angle

1 Introduction

In the last few decades, serious environmental concerns emerged due to an excessive use of petroleum-based liquid lubricants in the commercial and automotive IC engine applications. The demands at the global level ask for the fuel economy by minimizing the energy losses inviting an extensive research in reducing the sliding and viscous friction of the reciprocating engine components. The DLC coatings offer a promise of reducing friction and accompanying losses significantly. In a normal IC engine operation the piston skirts act as the piston assembly brakes to retard the inertial thrust of the flywheel. In the process the possibility of an excessive heat generation and damage in the form of piston scuffing cannot be ruled out completely. The DLC based solid coating is expected to behave as an effective lubricant as a substitute of a petroleum-based engine oil. When an IC engine starts up a significant amount of energy dissipation in the form of heat occurs due to the rubbing of the skirts surface against the liner wall during the expansion stroke of the piston. The presence of a fairly thin surface layer of the DLC coating is anticipated to act as an effective solid lubricant. The piston skirts of an IC engine establish a physical contact with the liner in the absence of a liquid lubricant in a few initial cycles of a low-speed engine start up. It is followed by an appreciable amount of adhesive wear of the opposing surfaces even if the engine is initially lubricated. It happens because the hydrodynamic pressures do not develop in a few initial engine start-up cycles. The dry surfaces of the piston skirts and the cylinder liner contact each other during the cyclic axial piston motion and its eccentric transverse displacements despite a large radial clearance. Under the thermal loading conditions due to combustion a physical contact invites the elastic surface displacements, producing strains and stresses. The stresses and strains accumulate to ultimately produce the plastic material flow and wear of the surfaces [1, 2].

Tribologists have been working on reducing wear of the tribological surfaces by applying fairly thick hard laminates on the piston substrate. Many people have worked on modeling the contact mechanics and the forces associated with it to study the elastic and plastic deformation of the contacting bodies [3, 4]. The work was extended to the single-layered coated surfaces to study the action of the concentrated normal and tangential forces analytically [4]. Some workers extended the work on the coated surfaces to 3-D problems employing the theory of elasticity [5]. The bi-layered isotropic coated surfaces were also considered and analytical solutions were obtained for the stress and displacement fields [6]. In some of the cases the normal and frictional distributed loads were applied to study the properties of the stress fields responsible for surface cracks and de-lamination of the coating [7]. In that context DLC is identified as suitable for improving fuel economy because of its superior wear resistance and low friction properties. A DLC coating on the piston may offer advantages such as better scuffing resistance and wear protection. To study the behavior of the DLC coatings and their response to the applied loads some researchers developed specific analytical functions but later, various numerical schemes were adopted to seek the approximate solutions of the

elastic and elasto-plastic response of the DLC coated surfaces [8]. Apart from using the finite element formulations some researchers worked on the finite difference schemes for the single-layered DLC coatings. The mathematical modeling of the response of the DLC coated skirts surfaces encompasses certain logical assumptions like ignoring the thermal effects due to the dry contact initially, apart from the transient elastic effects. The applied loads are assumed to be uniformly distributed over the homogeneous and isotropic lamination of the coated skirts surface, free from the effects of the rough substrate material.

2 Mathematical Model

A single layered DLC coating applied on the piston surface is modeled by simultaneously solving the piston dynamics model and the solid lubricant coating model.

2.1 Piston's Dynamics (Equations of Motion)

The position, velocity and acceleration along the axis of the cylinder of the piston depend on the crank angle. For the constant low rotational speed of the crankshaft, the linear or primary motion of the piston is determined by a set of equations [8, 9]:

$$\begin{aligned} Y &= \left((l+r)^2 - C_P^2\right)^{0.5} - \left(l^2 - B^2\right)^{0.5} - r\cos\psi \\ U &= \dot{Y} = r\omega\sin\psi + r\omega B\cos\psi\left(l^2 - B^2\right)^{-0.5} \\ \ddot{Y} &= r\omega^2\cos\psi + (r\omega B\cos\psi)^2\left(l^2 - B^2\right)^{-1.5} + \left((r\omega\cos\psi)^2 - r\omega^2 B\sin\psi\right)\left(l^2 - B^2\right)^{-0.5} \end{aligned} \tag{1}$$

where $B = C_p + rsin\Psi$.

During its axial translation, the piston displaces eccentrically in the transverse direction. The second-order secondary piston displacements along the direction perpendicular to the axis of the liner are defined by the relationships, which are incorporated in the mathematical model. To calculate the eccentricities of the top and the bottom surfaces of the skirts, the forces and the moments are used in the form of the force and moment with the help of reciprocating inertial forces equation [9] (Fig. 1, Table 1):

$$\begin{bmatrix} m_{pin}\left(1-\frac{a}{L}\right) + m_{pis}\left(1-\frac{b}{L}\right) & m_{pin}\frac{a}{L} + mpis\frac{b}{L} \\ \frac{I_{pis}}{L} + m_{pis}(a-b)\left(1-\frac{b}{L}\right) & m_{pis}(a-b)\frac{b}{L} - \frac{I_{pis}}{L} \end{bmatrix} \begin{bmatrix} \ddot{e}_t \\ \ddot{e}_b \end{bmatrix} = \begin{bmatrix} F + F_s + F_f\tan\Phi \\ M + M_s + M_f \end{bmatrix} \tag{2}$$

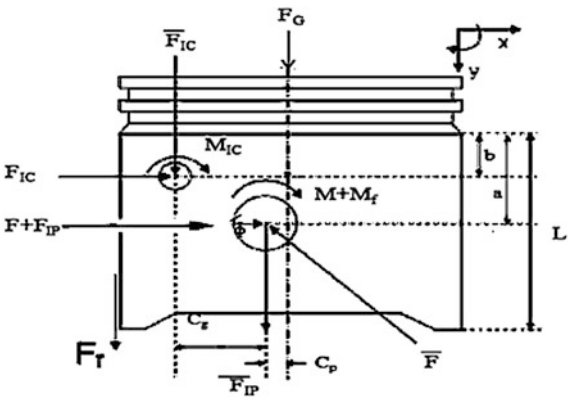


Fig. 1 Forces and moments on piston [8]

Table 1 Input parameters

Parameter	Value	Parameter	Value
m_{pis}	0.295 kg	E_1	350 GPa
m_{pin}	0.09 kg	E_2	200 GPa
L	0.133 m	v_1	0.28
$\theta = \theta_1 + \theta_2$	75°	v_2	0.3
R	0.0415 m	L	0.0338 m

where

$$F_s = tan\varnothing(F_G + \tilde{F}_{IP} + \tilde{F}_{IC})$$
$$M_s = F_G C_p + \tilde{F}_{IC} C_g$$
$$Ff = Ffh + Ffc \qquad F = Fh + Fc$$
$$Mf = Mfh + Mfc \qquad M = Mh + Mc$$

2.2 Contact Geometry Profile

The skirts get displaced eccentrically during the primary piston motion. The absence of a liquid lubricant invites a physical contact between the skirts and the liner during the 4-stroke cycle. Considering the radial clearance C, the equation representing the contact profile curve as a function of 720-degree cycle is [9, 10]:

$$h = C + e_t(t)\cos\theta + [e_b(t) - e_t(t)]\frac{y}{L}\cos\theta \tag{3}$$

where $\theta \approx \frac{x}{R}$.

2.3 Calculation of Elastic Displacements

The analytical closed-form solutions of the 3D problems are quite difficult to achieve due to the complexity of the elasticity field equations. Many solutions are developed for the reduced-order problems. The two dimensional elastic problem has been solved using plain strain theory. The desired equilibrium equations in terms of the surface displacements are the Navier's or Lame's equation. The scalar form of the system of equations in 2-D is [7, 11]:

$$\mu\nabla^2 w + (\lambda+\mu)\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial x}+\frac{\partial w}{\partial z}\right) = 0 \tag{4}$$

$$\mu\nabla^2 u + (\lambda+\mu)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}+\frac{\partial w}{\partial z}\right) = 0 \tag{5}$$

The Lame's system of equations can be represented in the vector form as [7, 11]:

$$\mu\nabla^2 u + (\lambda+\mu)\nabla(\nabla.u) = 0 \tag{6}$$

Where the Laplacian is given by [11]:

$$\nabla^2 = \frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial z^2} \tag{7}$$

2.4 Boundary Conditions

Uniformly distributed load is supposed on DLC coated piston surface. Boundary condition at the top surface of coating is given by:

$$\sigma_x = q; \quad w = 0.$$

Boundary condition at the sides and bottom is:

$$w = 0; \quad u = 0.$$

The continuity conditions at the interface are:

$$\sigma'_x = \sigma_x$$

Which shows the continuity of vertical stress at the interface

$$\sigma'_{xz} = \sigma_{xz}$$

Which is the continuity of shear stress at the interface.

2.5 Strain-Displacement Relations

The relations enable us to calculate the strains produced due to the elastic surface displacements. The principal strains are calculated as [6, 11]:

$$e_z = \frac{\partial w}{\partial z} \tag{8}$$

$$e_x = \frac{\partial u}{\partial x} \tag{9}$$

The shear strain is calculated as [11]:

$$e_{xz} = \frac{1}{2}\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right) \tag{10}$$

2.6 Stress—Strain Relations

Using Hook's Law, the principal and shear stress components are determined as [6, 12]:

$$\sigma_x = \lambda(e_x + e_z) + 2\mu e_x \tag{11}$$

$$\sigma_z = \lambda(e_x + e_z) + 2\mu e_z \tag{12}$$

$$\sigma_{xz} = 2\mu e_{xz} \tag{13}$$

where μ and λ are lame's constants are [6, 11]:

$$\mu = \frac{E}{2(1+v)} \tag{14}$$

$$\lambda = \frac{Ev}{(1+v)(1-2v)} \tag{15}$$

2.7 Discretization

To solve the Navier's equation numerically a finite difference mesh is generated by using the 2nd order central differencing scheme. An explicit numerical scheme is used to determine the displacements at each node of a 21 300 nodes mesh. The eccentricities $e_t(t)$ and $e_b(t)$ are calculated by solving Eq. (2). It constitutes an initial value problem for the two non-linear second order or the four first order differential equations [10]. The values of e_t, e_b, $\dot{e}_t$ and $\dot{e}_b$ are assumed at the time

step considered initially. Such values are considered as the initial values for the new time step. On the basis of these values the contact geometry between the skirts and the liner is calculated by solving Eq. (3). When the values of the eccentric displacement rates $\dot{e}_t$, $\dot{e}_b$ are satisfied and achieved, the piston position at the end of the current time step is obtained as [9]:

$$e_t(t_i + \Delta t) = e_t(t_i) + \Delta t \dot{e}_t(t_i)$$
$$e_b(t_i + \Delta t) = e_b(t_i) + \Delta t \dot{e}_b(t_i)$$

We compute and satisfy the secondary acceleration terms $\ddot{e}t$, $\ddot{e}_b$. These are satisfied from the solution of velocities $\dot{e}_t$, $\dot{e}_b$ at the previous and present time steps. The simulation of second order differential equations show the elastic surface displacements in the contact zone at the respective time steps or crank angles [12]. The Navier's equation is solved to calculate the distribution of the elastic surface displacements, strains and stresses at each node. The Successive-Over-Relaxation (SOR) iterative numerical scheme is used to solve the equations simultaneously. The simulation results correspond to the 720° crank rotation and show the deformation, strain and stress profiles.

3 Numerical Results and Discussion

The study of an elastic behavior of thick diamond-like carbon (DLC) solid lubricant coating over the surface of the skirts encompasses the incorporation of the secondary dynamics of the piston during the 4-stroke engine cycle. It is essential to ascertain the nature and duration of a physical contact between the skirts and the cylinder liner during the 720° crank rotation cycle. The elastic surface displacements occur when a physical contact gets established with the liner during the cyclic piston motion. The piston dynamics model is developed at a very low initial engine start up speed of 500 rpm and solved numerically to generate the simulation results for the analysis. The 720° crank rotation cycle is divided into four piston strokes of an equal duration as is the case in reality. The induction stroke is shown from 0° to 180°, the compression stroke between 181° and 360°, where as the expansion and exhaust strokes are represented by 361°–540° and 541°–720°, respectively. The simulation results show the profiles of the secondary piston eccentricities, secondary velocities, the elastic surface displacements on the sides and at the interface of the DLC coating and the substrate. The results would show the stresses and the strains produced due to a physical contact between the skirts and the liner during the 720° crank rotation cycle.

3.1 Piston Eccentricities and Secondary Velocities

An internal combustion engine considered in our case is assumed to have a piston-to-bore radial clearance of 100 microns at the time of a low-speed initial start up.

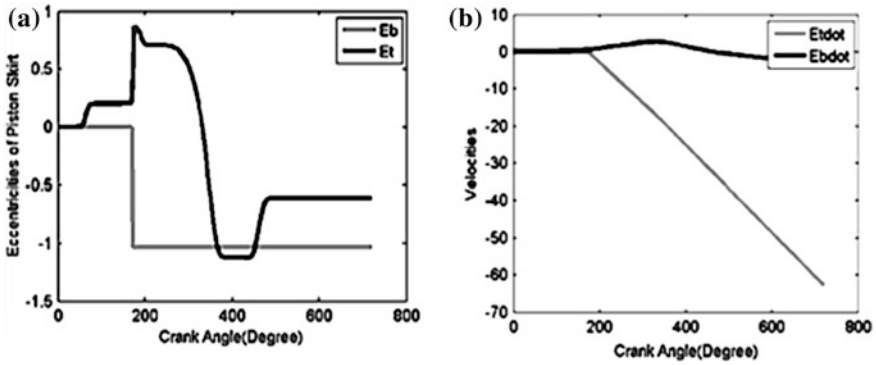


Fig. 2 **a** Piston eccentricities. **b** Secondary velocities versus 720-degree cycle

A relatively large radial clearance and the cyclic nature of the primary piston motion change the nature and duration of the dry physical contact between the skirts and the liner during the 4-stroke cycle. In view of this the secondary eccentric displacements of the piston skirts are shown in Fig. 2a. The secondary displacements of the top and the bottom surface of the skirts or the eccentricities are plotted in the non-dimensional form. The top and the bottom profiles are designated as E_t and the E_b for the top and the bottom sides of the piston skirts, respectively. In Fig. 2a the three horizontal lines at +1, 0 and −1 show the non-thrust side of the liner wall, its center and the thrust side of the cylinder wall, respectively. The piston commences its journey at 0° crank angle in the induction stroke and completes its stroke at 180° crank rotation angle in the 720° cycle. The eccentricity profiles show that the top surface comes close to the non-thrust side of the liner wall at the end of the induction stroke but does not establish a physical contact with it. However, the bottom surface contacts the thrust side of the liner wall at 180°. At the end of the induction stroke the piston changes its direction of travel but the bottom surface remains in physical contact with the liner in the compression, expansion and exhaust strokes. The top surface gets displaced all the way from close to the non-thrust side to the thrust side of the liner wall after experiencing the combustion thrust. Combustion occurs at 372-degree crank angle in the beginning of the expansion stroke. It displaces the top surface, which establishes a physical contact with the liner in the first half of the expansion stroke as the piston accelerates towards the mid-point of the expansion stroke. In the second half of the expansion stroke the piston decelerates towards the bottom dead center (BDC) as the combustion thrust subsides. It results in the physical disengagement of the top surface from the thrust side of the liner. In the exhaust stroke the bottom surface remains in contact with the liner but the top surface avoids a physical contact as shown in Fig. 2a. Figure 2b shows the dimensionless velocity profiles of the top and the bottom surface of the skirts represented as E_{tdot} and E_{bdot}, respectively. It represents the rate of energy transfer that takes place between the skirts and the liner surfaces. In the figure the velocity profiles are plotted in the positive and the negative quadrants. The respective profiles in the positive quadrants show the energy transfer

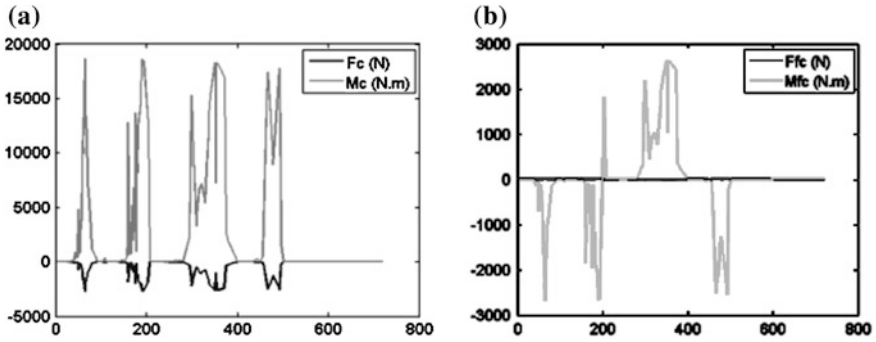


Fig. 3 **a** Normal contact force and moment. **b** Contact friction force and moment versus 720-degree cycle

from the skirts to the liner and vice versa in case of the profiles shown in the negative quadrant. The velocity profile curves explain the magnitude of energy transfer that allows a physical contact between the skirts and the cylinder liner. For the establishment of a physical contact the bottom surface require a relatively low amount of energy transfer during the induction stroke as compared to the case of the top surface. The small amount of energy transfer from the liner to the bottom surface is insufficient to allow a disengagement of the skirts during the remaining three piston strokes. It is not the case when the energy transfer between the top surface and the liner is considered. The rate at which energy is transferred from the liner to the top surface is appreciably high resulting in a surface disengagement in the second half of the expansion stroke.

3.2 Contact and Friction Forces

In view of the secondary piston displacements and the contact profiles the normal friction force and contact friction force acting over the surface are studied. The magnitudes of these forces and the moments generated are estimated and plotted against the 4-stroke cycle, as shown in Fig. 2. The normal force profile in Fig. 3a shows an instantaneous rise and fall at the mid-point and end of an induction stroke. During the process the there is an insignificant amount of energy transfer between the skirts and the liner. It implies that entropy increases as the skirts do not get displaced significantly in the lateral direction. It is also evident from the contact friction force profile shown in Fig. 3b.

An instantaneous increase and reduction of the contact friction force is shown in the negative quadrant. It indicates a low efficiency process implying irreversibility and energy loss. The significant and noticeable changes in the magnitudes of the forces are seen in the second half of the compression stroke and first half of the expansion stroke. The effects are seen in the eccentric displacements of the skirts and the energy transfer in the form of the secondary velocity profiles. There is a

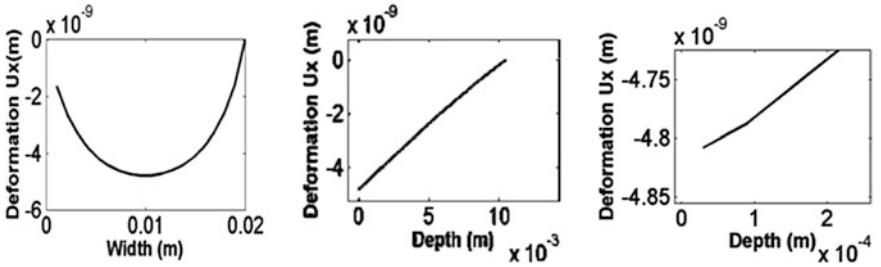


Fig. 4 Axial deformation, **a** along skirts width at coating interface with substrate, **b** in the depth, **c** in the critical depth

negative friction force profile in the second half of the expansion stroke. It affects the lateral eccentricities of the top surface, which gets disengaged from the liner during the process.

3.3 Elastic Surface Displacements

Figure 3 shows the three profiles of the axial surface displacements. On the application of the contact loads there are significant elastic surface displacements generated at the interface of the DLC coating layer and the substrate skirts material. In view of the surface displacements the possibility of plastic flow and wear may be estimated by showing the extent and place of the maximum deformation at the interface. Figure 3a shows the axial deformation at the interface and along the width of the skirts. The parabolic profile of the axial displacement curve shows the maximum deformation at the mid-point and the negligible value at close to the distant side of the skirts. The applied load affects the depth of the coated skirts surface as well. The profile curve giving the extent of the elastic surface displacements is shown in Fig. 3b and c. The elastic displacement is maximum at the surface of the DLC coating and decreases linearly as the depth increases. The displacements at the depth of 70 microns thick coating are linear in nature but the slope differs slightly from that of the substrate, as shown in Fig. 3c. The transverse elastic deformation at the interface of the coating and the substrate is shown in Fig. 4a. The profile shows that the surface deformation increases initially but then decreases sharply till the mid-point of the skirts surface. The point of inflexion at the mid-point allows an increase in the elastic displacement till the two-third of the width of the skirts surface. However, the deformation decreases to a negligible value after approaching the second point of inflexion. Initially the transverse elastic surface deforms linearly followed by the second-order changes experienced by the displacement curve resulting in an energy transfer between the coated surface and the liner. Figures 5 and 6 show the contour and 3-D plots of the axial and transverse deformations, respectively. The contour plots show the relatively high deformation intensities close to the surface (Fig. 7).

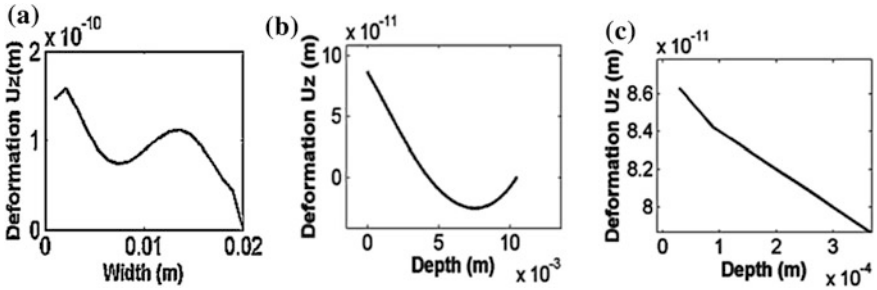


Fig. 5 Transverse deformation, **a** along skirts width at coating interface with substrate, **b** in the depth, **c** in the critical depth

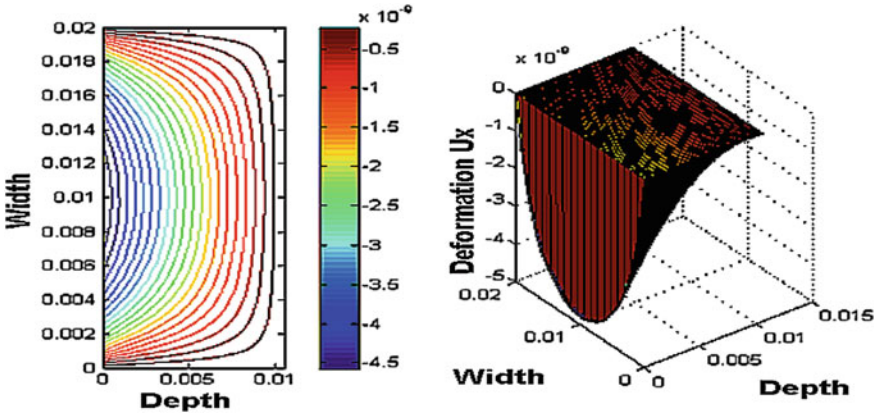


Fig. 6 Axial deformation shown by contour and 3-D plots

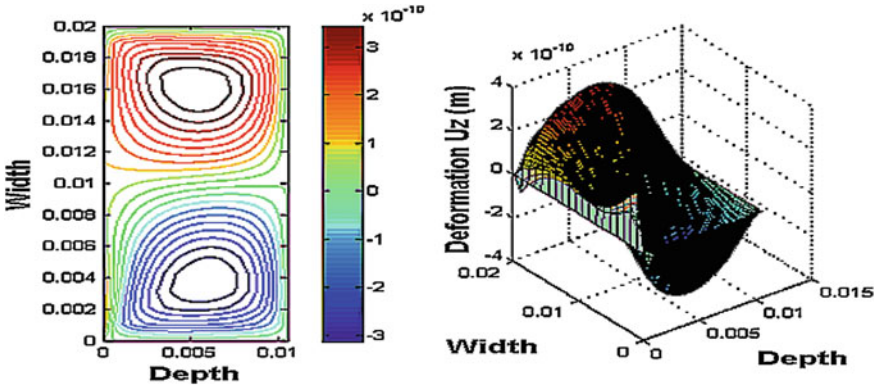


Fig. 7 Transverse deformation shown by contour and 3-D plots

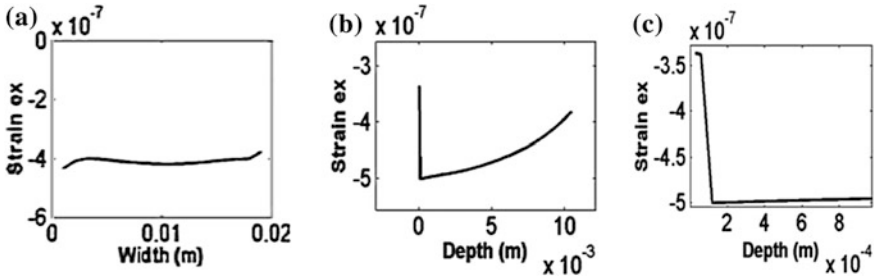


Fig. 8 Axial strain, **a** along skirts width at interface with substrate, **b** in depth, **c** in critical depth

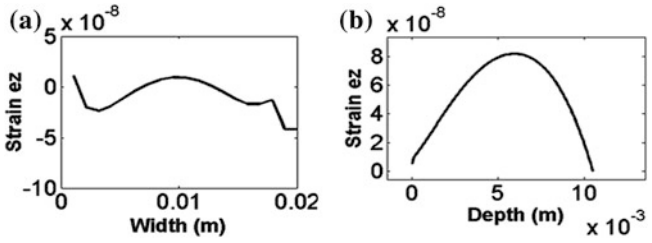


Fig. 9 Transverse strain, **a** along skirts width at coating interface with substrate, **b** in skirts depth

3.4 Axial and Transverse Strains

The axial and transverse elastic surface displacements produce the corresponding strains. The profile curves of the strains generated at the interface and in the depth of the coating and substrate material are plotted and shown in Figs. 8a and 9a, respectively. The axial strain curve at the interface does not vary significantly along the width of the skirts. However, the transverse strain curve shows the cyclic behavior along the surface width. When studying the strains produced in the depth of the coated surface, the results show that the maximum axial strain is produced at the surface of the coating. In contrast, the maximum transverse strain is produced at a depth of 5 cm, which implies that the maximum overall strain is produced over the surface of the coating at 5 cm away from the origin. The contour plots and 3-D fields of axial and transverse strains are plotted and shown in Figs. 10 and 11, respectively.

3.5 Principal and Shear Stresses

On the application of an external load the principal stresses are produced in the axial and transverse directions apart from the shear stress over the coated surface of the skirts. The principal stresses are studied at the interface and the depth of the skirts surface and the results are shown in Figs. 12, 13, 15 and 16 respectively.

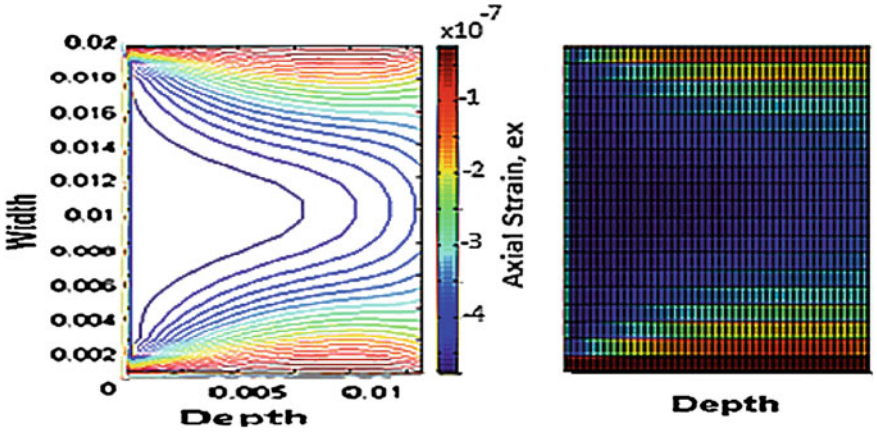


Fig. 10 Contour plot of axial strain

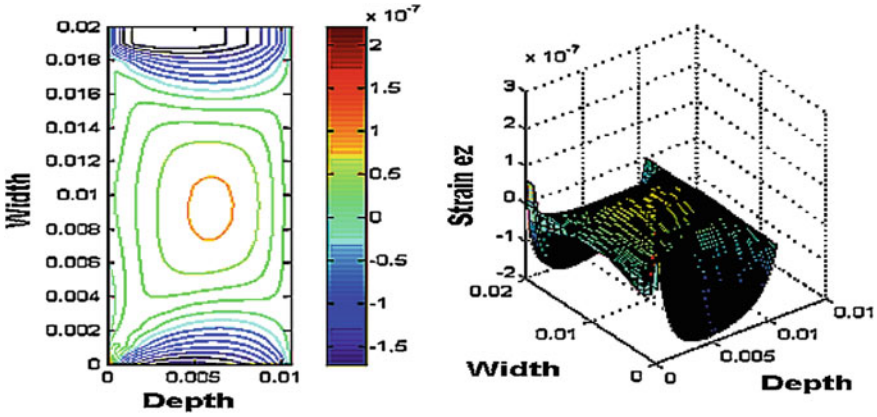


Fig. 11 Contour and 3-D plots of transverse strain

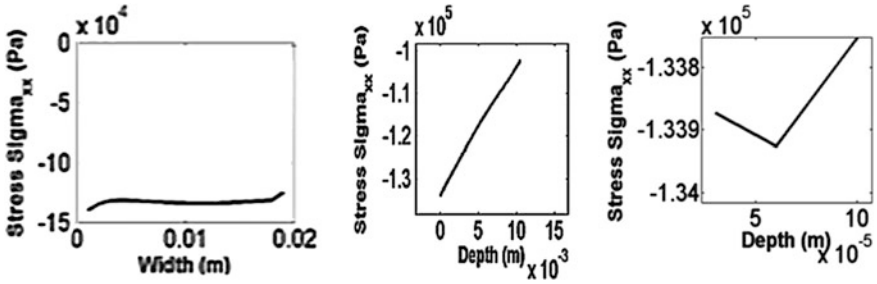


Fig. 12 Principle stress in axial direction, **a** along skirts width at coating interface and substrate. **b** In depth, **c** critical depth

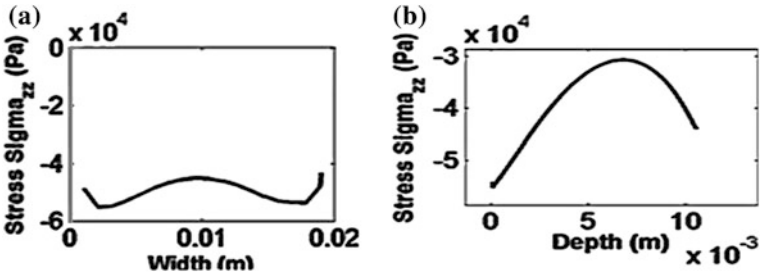


Fig. 13 Principle stress in transverse direction. **a** Along skirts width at coating interface and substrate. **b** In the depth

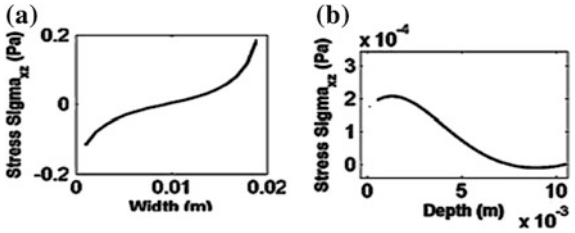


Fig. 14 Shear stress. **a** Along skirts width at coating interface and substrate. **b** In the depth

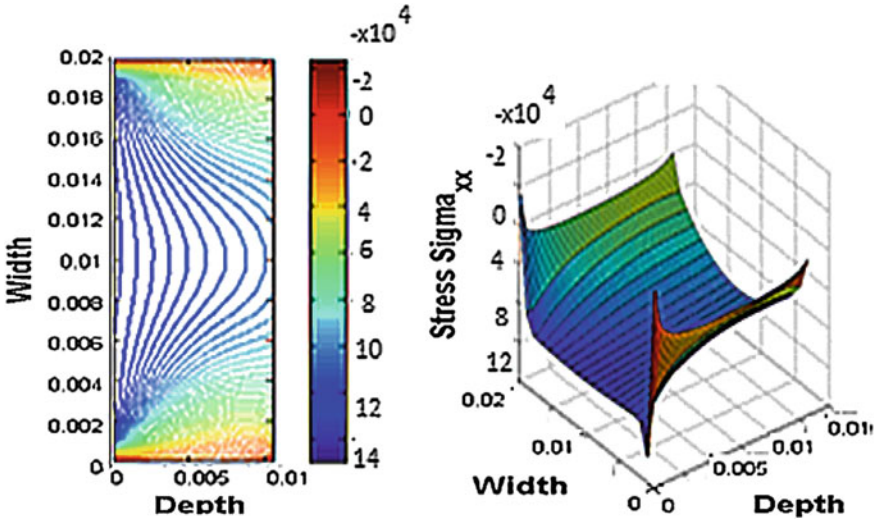


Fig. 15 Contour and 3-D plots of axial stress

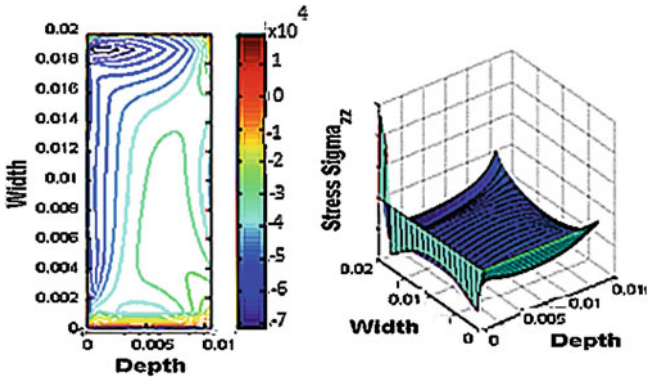


Fig. 16 Contour and 3-D plots of transverse stress

In the axial direction the stress has the nearly uniform magnitude at the interface and along the width of the skirts. When considering the depth the stress increases in the coating, attains the maximum value and then decreases linearly in the substrate. Similarly, in the transverse direction the maximum stress is produced within the coating. The contour plots show the stress intensities. There is maximum shearing of the DLC coating layer as compared to the case of the substrate, as shown in Fig. 14.

3.6 Comparison for Different Coating Thicknesses

Diamond like Carbon (DLC) has high hardness (40 GPa) and elastic modulus above 300 GPa. Study has been carried out by changing DLC coating thicknesses (starting from 60 to 1,000 μm). Figure 17 shows the effect of axial deformation on different coating thickness along width for DLC coating on stainless steel piston. The critical zone of maximum compressive deformation at the mid of skirts width has been shown in circle. It is clear from figure that as the thickness of the coating is increased, the vertical or normal deformation is decreased. The coating with thickness 500 and 1,000 μm experience small deformations. The max normal deformation is placed at the center of the skirt. However there is little difference of deformation between 75 μm thickness ($4.68\ \ 10^{-9}$ m) and 500 μm ($4.68\ \ 10^{-9}$ m). It implies that coating thickness of 75 μm is comparatively cost effective.

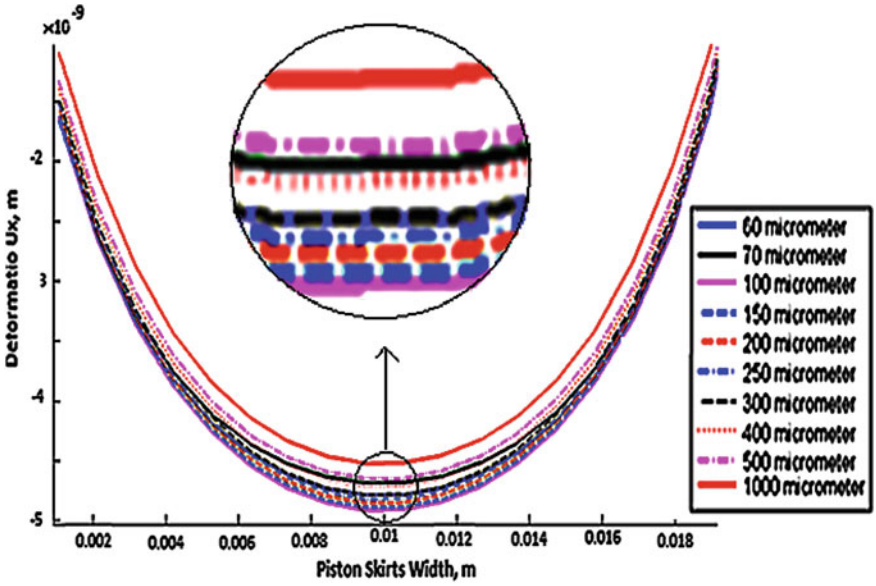


Fig. 17 Axial deformation along the width of piston skirt for different coating thickness

4 Conclusions

The dry contact model incorporating the secondary dynamics of the DLC coated piston skirts is developed at a very low initial engine start up speed. The contact zone was analyzed by determining the magnitude of the contact and friction forces, the elastic deformations, strains and stresses produced during the process. The skirts were in contact with the liner in the compression, expansion and exhaust strokes despite a large radial clearance of 100 microns. The mathematical models considered the different lamination thicknesses ranging from 60 to 1,000 microns. A 70 microns thick DLC coating sustained the maximum loading under the dry contact conditions at low-load and speed conditions. The maximum elastic displacements, strains and stresses were witnessed within the coating thickness. It implies that due to its material properties the relatively thin DLC coating sustains the maximum stresses produced due to the contact of the skirts with the liner. The DLC coating prevents the development of maximum strains and resists the maximum elastic displacements of the substrate material at the low-load and speed conditions in the initial engine start up. The elastic behavior of the thin DLC coating prolongs the life of the substrate material of the skirts even when a physical contact can not be avoided. However, the stresses produced at the interface are reasonably high and the coating thickness need to be optimized to prevent an early plastic flow and delamination of the skirts surface in the initial engine start up.

Acknowledgments This work was sponsored by the National University of Sciences and Technology (NUST), Islamabad, Pakistan. The financial support for this project was provided by the Higher Education Commission of Pakistan.

References

1. Zahid-Ur-Rehman, S. Adnan Qasim, M. Afzaal Malik, in *Modeling Dry DLC Coated Piston Skirts Elastic Behavior at Low-Speed Initial Engine Start Up*. Lecture Notes in Engineering and Computer Science. Proceedings of the World Congress on Engineering 2013 (WCE 2013), 3–5 July 2013, London, U.K., pp. 1723–1728
2. A. Greenwood, J.H. Tripp, in *The Contact of Two Nominally Flat Rough Surfaces.* Proceedings on Institution of Mechanical Engineers (IMechE), UK (1971), pp. 625–633
3. N.G. Demas, R.A. Erck, O.O. Ajayi, G.R. Fenske, Tribological studies of coated pistons sliding against cylinder liners under laboratory test conditions. Lubr. Sci. (2012). doi:10.1002/ls.11752012
4. M.Z. Hossain, S.R. Ahmed, M.W. Uddin, Generalized mathematical model for the solution of mixed-boundary-value elastic problems. J. Appl. Math. Comput. **169**, 1247–1275 (2005)
5. M. Kashtalyan, M. Menshykova, 3-D elastic deformation of a functionally graded coating/substrate system. Int. J. Solids Struct. **44**, 5272–5288 (2007)
6. M. Shakeri, A. Sadough, S.R. Ahmadi, Elastic stress analysis of bi-layered isotropic coatings and substrate subjected to line scratch indentation. J. Mater. Process. Technol. **196**, 213–221 (2008)
7. I.A. Anderson, I.F. Collins, Plain strain stress distributions in discrete and blended coated solids under normal and sliding contact. Wear **185**, 23–33 (1995)
8. D. Zhu, H.S. Cheng, T. Arai, K. Hamai, A numerical analysis for piston skirts in mixed lubrication. ASME J. Tribol **114**(3), 553–562 (1992)
9. S. Adnan Qasim, M. Afzaal. Malik, U. F. Chaudhri, Analyzing viscoelastic effects in piston skirts EHL at small radial clearances in initial engine start up. Tribol. Int. **45**(1), 16–29 (2012)
10. S. Adnan Qasim, M.A. Malik, M.A. Khan, R.A. Mufti, Low viscosity shear heating in piston skirts EHL in the low initial engine start up speeds. Tribol. Int. **44**(10), 1134–1143 (2011)
11. P. Maheshwari, M.N. Viladkar, in *Theory of Elasticity Approach for Strip Footings on Multilayered Soil Media.* Proceedings of 12th International Conference of IACMAG 2008, India, pp. 3464–3472
12. G.W. Stachowiak, A.W. Batchelor, *Engineering Tribology*, 3rd edn. (Elsevier, USA, 2005), pp. 91–325
13. L.F. Ma, A.M. Korsunsky, Fundamental formulation for frictional contact problems of coated systems. Int. J. Solids Struct. **41**, 2837–2854 (2004)

Mineralogical and Physical Characterisation of QwaQwa Sandstones

Mukuna P. Mubiayi

Abstract Mineralogy, texture and sedimentary structure are essential to the study of all sedimentary rock. Sandstones from QwaQwa rural area in South Africa were characterised for their mineralogical and physical properties. Six samples (whitish, greenish, blackish, reddish, yellowish and greyish) were collected and studied. The X-ray Diffraction (XRD) revealed that the samples were quartz based. Scanning Electron Microscopy coupled with energy dispersive X-ray spectroscopy (SEM/EDS) revealed the differences in the morphology and Silicon element was found present in all the samples. The X-Ray Fluorescence (XRF) exhibited that the samples contained 43.14 % of Silicon (whitish). The grain sizes of the samples ranged from coarse, medium to fine grains. The Greyish sample had the highest compressive strength value of 56.74 MPa. Dielectric properties measurements were also conducted on the samples; and the results were temperature dependant. The water absorption by total immersion revealed that the blackish sandstone had the highest percentage of 6.62 %. Furthermore, the greyish sandstone had the highest density (2.7235 g/cm^3) and the reddish sample had the lowest (2.6502 g/cm^3).

Keywords Compressive strength · Density · Dielectric properties · Phase identification · Quartz · Sandstones · Water absorption

1 Introduction

This work is an extension of our previous work [1]. The characterisation and identification of minerals is fundamental in the development and operation of mining and minerals processing systems [2]. Worldwide, sandstones have been

M. P. Mubiayi (✉)
Department of Mechanical Engineering Science, University of Johannesburg, Kingsway Campus, Corner Kingsway and University Road, Auckland Park, P.O. Box 524 Johannesburg 2006, South Africa
e-mail: patrickmubiayi@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_16,

used as construction material for centuries and are still being used for this purpose [3]. Although, sandstones show similar appearances and properties; a geological background may cause differences in colour, mineral composition, granulometric properties, pressure strength and/or weathering behaviour [3]. On the other hand, mineralogical properties of sandstones could predict their mechanical properties such as the uniaxial compressive strength. The inherent parameters of sandstones can be characterised by their petrographical properties [4]. There is a large deposit of sandstones in QwaQwa in the Free State province of South Africa as shown in Fig. 1. The presence of this large deposit of sandstones in the QwaQwa region was the motivation to initiate this study on the characterisation of sandstones. On the other hand, the knowledge on the physical such as compressive strength of QwaQwa sandstones will help the booming construction industry in South Africa by providing the characteristics of sandstones for construction and decorations purposes.

2 Geology of South Africa

The Free State lies in the heart of the Karoo Sequence of rocks, containing mudstones, shales, sandstones and the Drakensburg Basalt forming the youngest capping rocks [5]. The province is high-lying, with almost all the land being 1,000 m above the sea level. Some of the sandstones are resistant to weathering yet easy to work. The geological map of South Africa indicating the QwaQwa region is displayed in Fig. 1.

Sandstones have been used as construction material worldwide and are still used for this purpose [6]. On the other hand, mineralogical properties of sandstones could predict their mechanical properties such as the uniaxial compressive strength. According to Zorlu et al. [4] the uniaxial compressive strength of sandstones is controlled by several inherent and environmental parameters. The inherent parameters can be characterized by the petrographical properties. The mineral composition, the void space, the degree of grain interlocking, the packing density and the grain size is known to be affected by the petrographic characteristics [4]. However, contradictory results have been reported relating the influence of mineral content on the geomechanical properties of sandstones [7]. It has been reported that rocks containing quartz as binding materials are the strongest materials followed by calcite, and ferrous minerals; but rocks clayey binding materials are the weakest ones [4, 8]. Furthermore, the shape of the grains is another petrographical property. The shape of the grains is usually expressed in terms of the roundness or sphericity, roundness being distinct from sphericity in that, it is concerned with the curvature of the corners. In fact, sphericity represents a quantitative means of expressing the departure of a grain from equidimensiality [4, 7]. Shakoor and Bonelli (cited by [4]) found that there is a fairly strong relationship between the uniaxial compressive strength and the percent of angular grains [4]. On the other hand, Ulusay and co-workers (cited by [4]) reported that there is a strong correlation between the uniaxial compressive strength and the

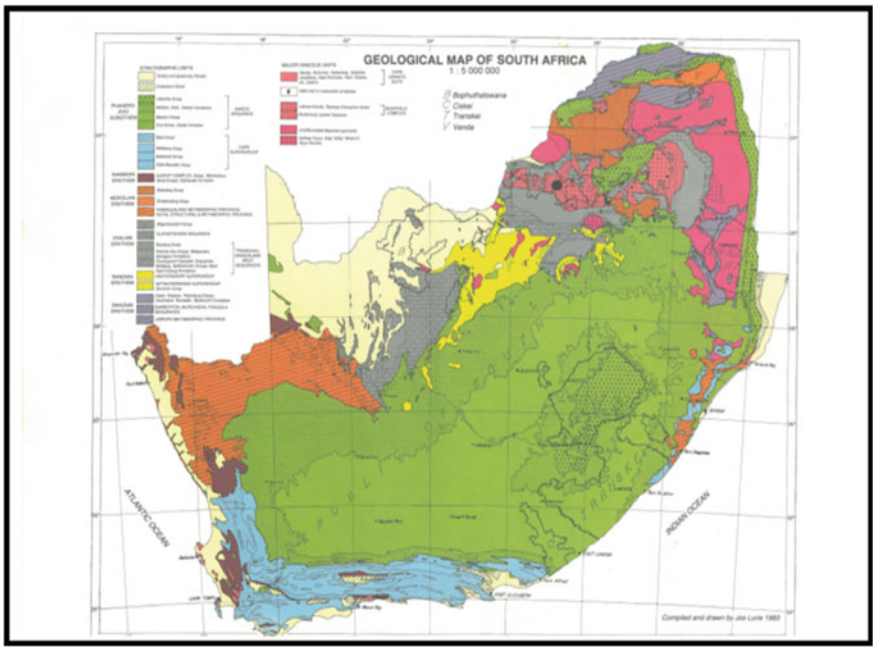


Fig. 1 Geological map of South Africa showing the presence of sandstones in QwaQwa area of Free State [5]

percent of rounded grains [4]. However, Fahy and Guccione cited by [4] obtained no meaningful correlation between the uniaxial compressive strength and the roundness while they found an extremely strong relationship between the uniaxial compressive strength and the sphericity of the grains. A study done by Zorlu et al. [9] did not obtain meaningful correlation between the uniaxial compressive strength and the grain shape parameters.

Within the framework of quantifying the natural stones in the QwaQwa area (South Africa) in term of quality, this paper is focusing on conducting mineralogical and physical characterisation of selected sandstones from QwaQwa rural area in South Africa.

3 Characterisation Techniques

3.1 Mineralogical Studies

Sandstones samples were crushed and milled using a jaw crusher and a milling machine, respectively. The powder obtained was then used for XRD and XRF analysis. The mineralogical studies of the sandstones were conducted using XRD,

XRF and SEM combined with EDS. An X-ray Powder Diffractometer (XRD) Phillips X'pert Model 0993 to determine their mineralogical phases. The elemental composition of the sandstones was analysed using The Philips Magix Pro X-ray Fluorescence spectroscopy. A SEM JEOL JSM-840 combined with EDS instrument was used to analyse the surface morphology and qualitative analysis of the samples respectively. Furthermore, the grain sizes measurements were conducted by using Olympus BX51M optical microscope on the mounted samples.

3.2 Physical Properties

The uniaxial compressive strength was used to measure the capacity of the six types of sandstones to withstand a load. The uniaxial compressive strength result helped to investigate and correlate the use of the six types of sandstones as construction material along with their water absorption and mineralogy. The six sandstones samples were tested in confined uniaxial compression in room-dry condition. The six sandstones samples were cut into cubical shape with an approximately size of $3 \times 3 \times 5$ cm^3. The sandstones samples were then placed in the testing machine and loaded in compression at the Council for Scientific and Industrial Research (CSIR) laboratory (South Africa). An Amsler universal machine was used to determine the uniaxial compressive strength. On the hand, the water absorption analysis was performed using a total immersion method.

The samples were cut in a cubical shape of approximately 70 g. The samples were washed with distilled water in order to eliminate powdered material from their surfaces. Thereafter, the samples were dried in the oven at 60 °C for 24 h. The samples were then placed in a desiccator with dry silica gel to cool. They were removed from the desiccator, measured (weight), put in a container filled up with water until the samples were totally immersed. After 24 and 48 h, the samples were measured (weight) and the water absorption percentage was calculated.

Moreover, the dielectric constant and loss factor of samples placed in a microwave cavity are related to the shift in the resonant frequency and change in the Q-factor or the bandwidth of the resonance. Standard closed-form perturbation models exist for the resonant frequency changes due to a cylindrical dielectric sample in a cylindrical microwave cavity. The cavity used for these measurements resonates at 915 MHz, a frequency allocated for microwave heating purposes. Each sample was placed in a 6 mm ID quartz tube and the microwave properties measured in the dielectric measurement fixture. A conventional high temperature furnace was used to heat the sample in the quartz tube to the required temperature (1,000 °C). It was then rapidly lowered into the microwave test fixture for the measurement of the dielectric properties. This was repeated for each measurement temperature. A ramp and soak heating profile was used, with the temperature increased to the next value over a period of 30 min, with a 30 min soak at that temperature before the measurement was taken. A slow nitrogen purge was bled trough the quartz tube to displace any oxygen present and prevent oxidation at

high temperature. A vector network analyzer (model: Hewlett Packard 8753B) was used to measure the dielectric properties of the sandstones specimens. The dielectric measurements of the samples were carried out at DELPHIUS CIT Laboratory (South Africa). Furthermore, a Micromeritics Accupyc1340 gas pycnometer was used to measure the densities of the different sandstones.

4 Results and Discussion

4.1 Mineralogy and Morphology

A demonstrative fraction of the overall samples were analyzed to establish the starting point for the experiment. The diffractograms (Figs. 2, 3) show the XRD patterns of the reddish and blackish sandstone from QwaQwa rural area in South Africa.

The XRD analysis revealed that the Quartz (SiO_2) was the major mineral present in all the sandstones followed by feldspar minerals such as Illite ((K, H_3O) (Al, Mg, Fe)$_2$(Si, Al)$_4O_{10}$[(OH)$_2$, (H_2O)]), Albite (Na Al Si_3 O_8). Other identified minerals include Glauconite ((K, Na) (Fe^{3+}, Al, Mg)$_2$(Si, Al)$_4O_{10}$(OH)$_2$), Kaolinite ($Al_2Si_2O_5(OH)_4$), Cristobalite (SiO_2) and Orthoclase ($KAlSi_3O_8$) were minor in the composition.

The XRF results are shown in Table 1. The chemical elements for the six types of QwaQwa' sandstones identified were: Aluminium (Al), Calcium (Ca), Iron (Fe), Potassium (K), Magnesium (Mg), Manganese (Mn), Sodium (Na), Phosphorus (P), Silicon (Si) and Titanium (Ti).

The results in Table 1 revealed a significant presence of Silicon in all the sandstones types implying that the QwaQwa' sandstones are silica based. The XRF results correlate to a very large extent with the XRD findings. This is in particular with regards to the dominance of the silicon species followed by Al and Fe. Mn was found in trace level only in the blackish sample whereas P which is also at trace level is evident in three samples: blackish, greyish and greenish as presented in Table 1. In addition, Ti although at negligible percentage is available in the six samples. The presence of these elements in the sandstones could be attributed to the origin of the mineral bearing.

Furthermore, Figs. 4 and 5 show the SEM/EDS micrographs revealing the surface morphology and the chemical composition of the spotted grains for the blackish and greenish sandstones. The correlation between the XRD, XRF results and SEM/EDS is once again evident based on the elemental chemical composition of the spotted grains which reveals Si followed by Al and Fe as the main components. This is in accordance with the chemical composition of phases found using the XRD analysis (Figs. 2, 3) and the chemical composition using XRF displayed in Table 1. Furthermore, the results from the optical microscopy analyses conducted revealed that the samples had a wide range of grain sizes which

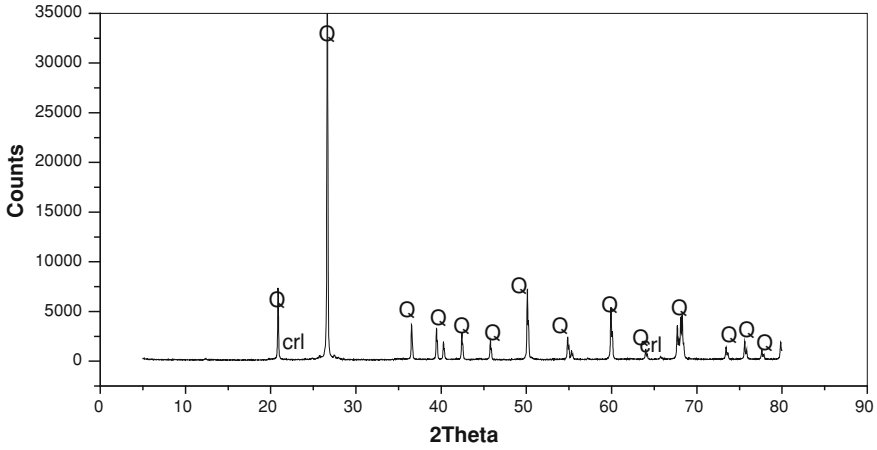


Fig. 2 XRD graph of reddish sandstone showing the minerals identified (*Q* quartz, *crl* cristobalite)

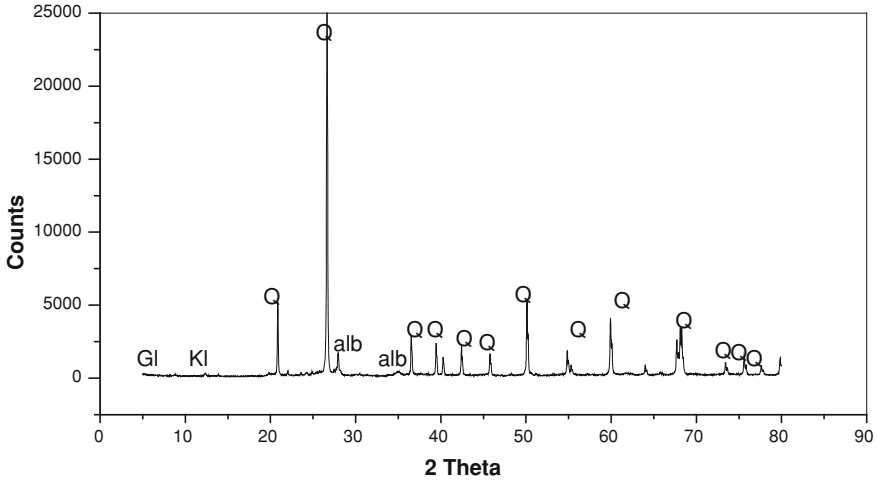


Fig. 3 XRD graph of blackish sandstone showing the minerals identified (*Q* quartz, *alb* albite, *Gl* glauconite, *Kl* kaolinite)

range from coarse, medium to fine grains. The whitish sandstone had coarse grains while the reddish and blackish had medium-coarse grains. On the other hand, the greyish, yellowish and the greenish have medium-fine, fine and fine grains respectively. Figures 6 and 7 show the grain sizes of the whitish and yellowish sandstone respectively.

Table 1 Chemical composition of six sandstones samples using XRF

Elements	Concentration (%)					
	Whitish	Yellowish	Blackish	Reddish	Greyish	Greenish
Al	2.00	4.01	4.62	1.92	4.66	4.81
Ca	–	0.09	0.23	0.03	0.10	0.40
Fe	0.42	0.83	2.04	1.39	1.44	1.32
K	1.01	1.87	0.85	0.59	1.58	1.71
Mg	0.06	0.19	0.38	0.05	0.32	0.28
Mn	–	–	0.30	–	–	–
Na	0.06	1.52	0.71	0.09	0.66	1.62
P	–	–	0.03	–	0.03	0.02
Si	43.14	39.10	38.81	42.76	39.48	38.18
Ti	0.06	0.14	0.26	0.10	0.19	0.16

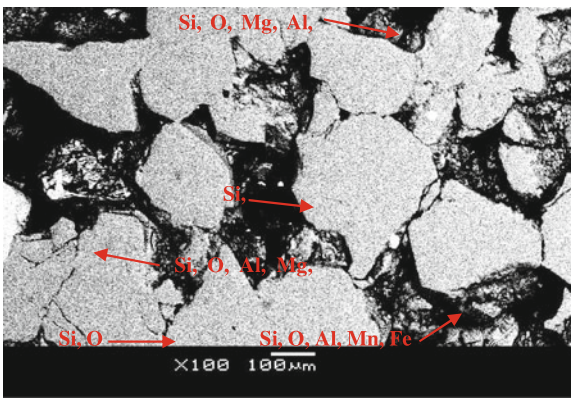


Fig. 4 SEM combined with EDS micrograph of the blackish sandstone showing the surface morphology and the chemical composition of the *spotted grains* respectively

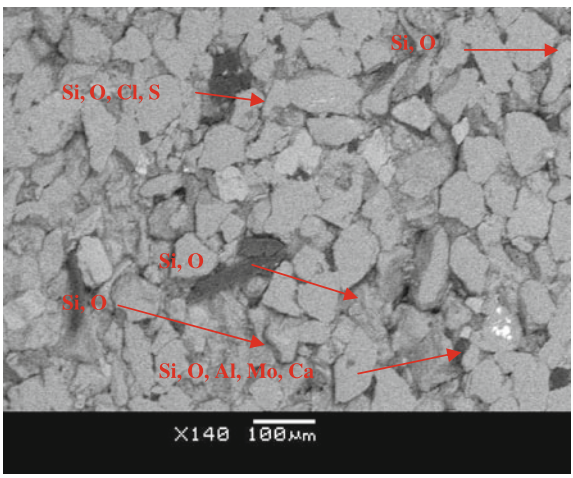


Fig. 5 SEM combined with EDS micrograph of the greenish sandstone showing the surface morphology and the chemical composition of the *spotted grains* respectively

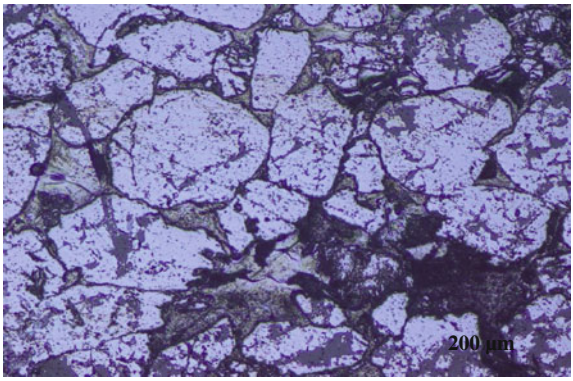


Fig. 6 Optical photomicrograph of the surface of the whitish sandstone showing coarse grains size

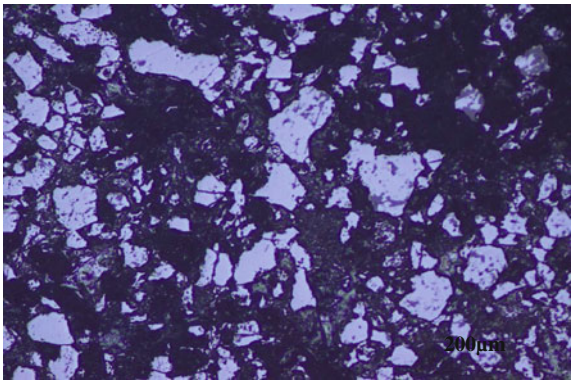


Fig. 7 Optical photomicrograph of the surface of the yellowish sandstone showing fine grains size

4.2 Physical Properties

The uniaxial compressive strength of intact rock is among the main parameters used in almost all engineering projects. The uniaxial compressive strength test requires high quality core samples of regular geometry [4]. Figures 8 and 9 show the graphs of the compressive stress versus compressive extension of the greyish and yellowish sandstones under a load. The loads were recorded and the unconfined compressive strength of the sandstone was determined. The uniaxial compressive strength of the six QwaQwa' sandstones samples recorded were 9.4, 22.8, 26.3, 16, 8.3 and 56.7 MPa for yellowish, reddish, greenish, blackish, whitish and greyish, respectively. The whitish sandstone has the lowest compressive strength (8.3 MPa) whereas the greyish shows the highest value (56.7 MPa). Furthermore, Singh [10] and Zorlu et al. [4] stated that the uniaxial compressive strength also vary with the grain size range. Therefore, it was observed that a good correlation exists between the grain size of the whitish sandstone sample and the compressive strength which was the lowest; the result was in correlation with the results found by Singh [10].

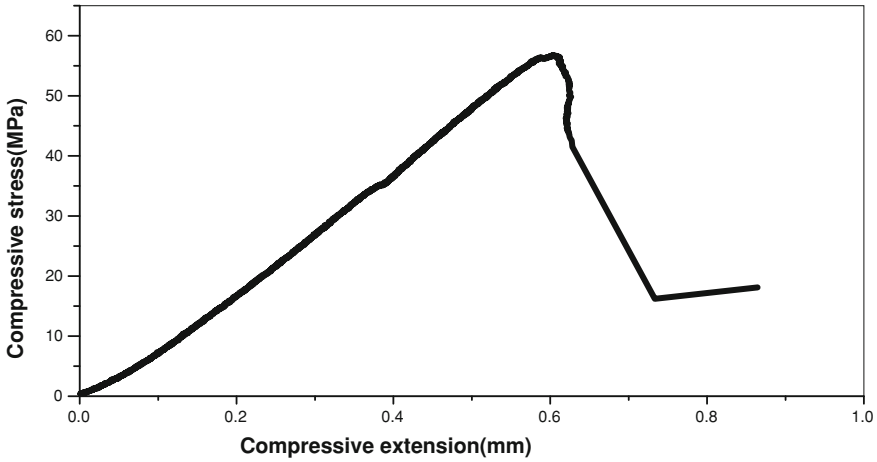


Fig. 8 Compressive stress versus compressive extension of greyish sandstone. The uniaxial compressive stress was measured to be 56.7 MPa

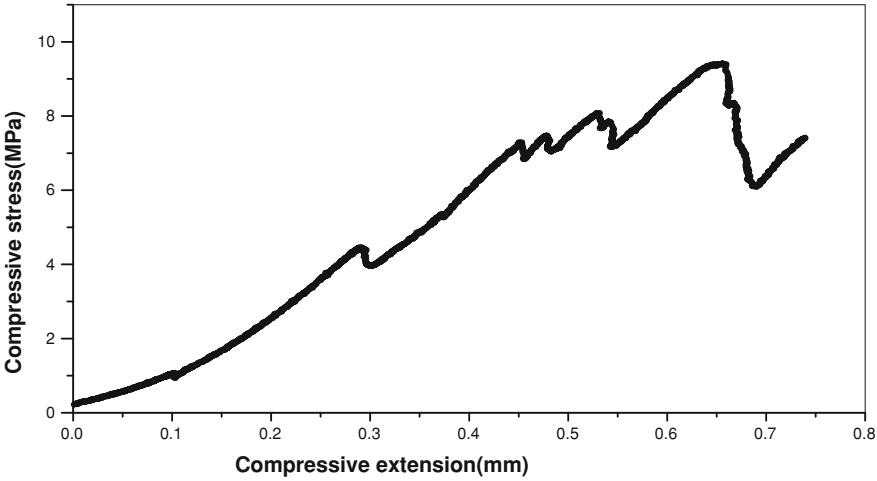


Fig. 9 Compressive stress versus compressive extension of yellowish sandstone. The uniaxial compressive stress was measured to be 9.4 MPa

Dry density is among the most important factors in evaluating the mechanical properties of samples. The density of samples may tell us on the compactness of the grain, which will affect the overall properties of the rock [11]. Table 2 shows the dry density of tested samples.

The water absorption using dried sandstones was performed. The graphs of the water absorption by mass change and percentage are displayed in Figs. 10 and 11. The percentage of water absorption ability of the samples was as followed in a

Table 2 Density results of sandstones

Sandstones	Density (g/cm^3)
Greyish	2.7235
Whitish	2.7032
Blackish	2.6808
Greenish	2.6338
Yellowish	2.6344
Reddish	2.6502

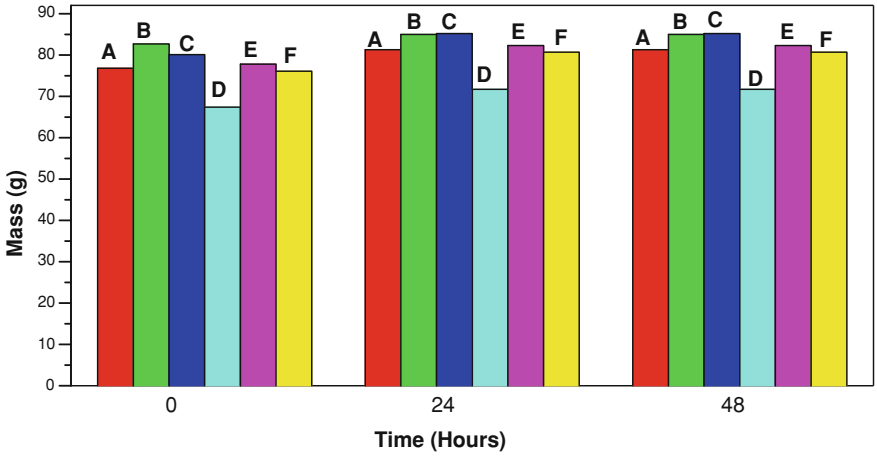


Fig. 10 The water absorption mass change of dried sandstones, samples were dried overnight in an oven and immersed in water for 24 and 48 h (*A* reddish, *B* greyish, *C* blackish, *D* greenish, *E* yellowish, *F* whitish)

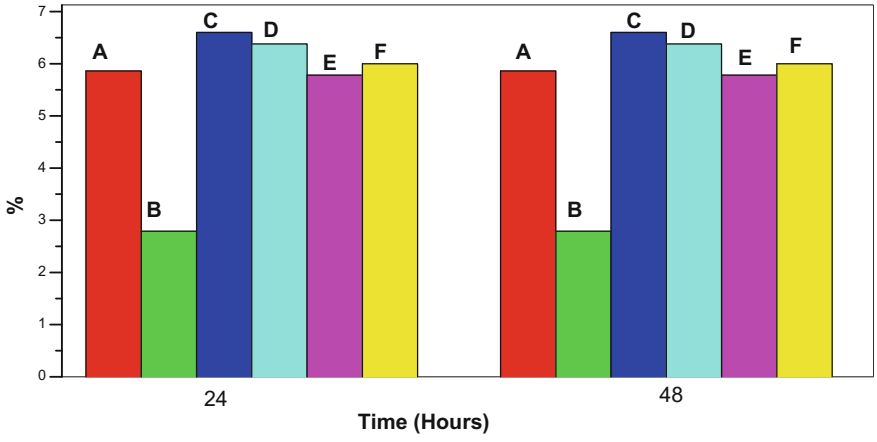


Fig. 11 The water absorption percentage of dried sandstones, samples were dried overnight in an oven and immersed in water for 24 and 48 h (*A* reddish, *B* greyish, *C* blackish, *D* greenish, *E* yellowish, *F* whitish)

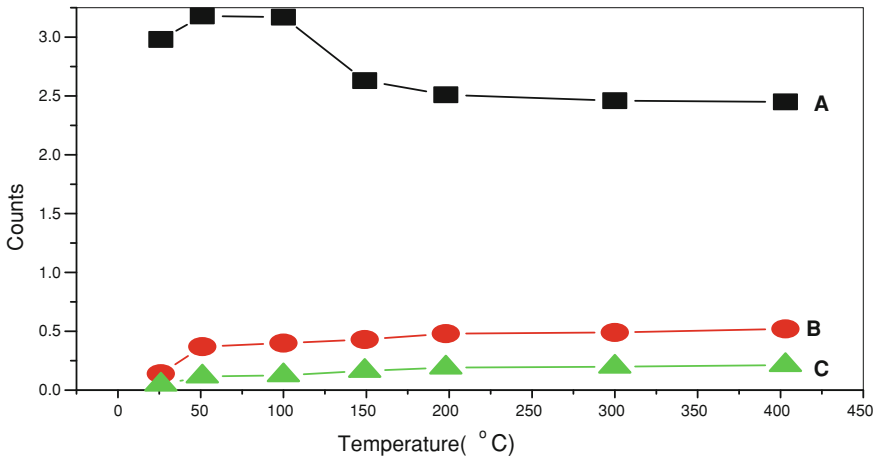


Fig. 12 Dielectric constant (*A*), dielectric loss factor (*B*) and calculated tangent delta (*C*) of greenish showing the temperature dependency

descending order; blackish (6.6 %), greenish (6.4 %), whitish (6.0 %), reddish (5.9 %), yellowish (5.8 %), and greyish (2.7 %) as shown in Fig. 11. Figure 10 shows the water absorption by mass change. The water absorption results suggest that the greyish samples absorbed less water than the other samples. A correlation between the water absorption and the grain sizes for some of the sandstones samples (yellowish, blackish, reddish, whitish, and greyish) was noticed. However, the blackish sample had medium fine grain size. On the other hand, the blackish sample showed the highest water absorption which could be assumingly due to its higher porosity.

The dielectric properties or permittivity of granular or powdered materials is an important parameter in the application of the dielectric heating or sensing moisture content using radio frequency or microwave instruments [12]. Six sandstones samples were used for the microwave dielectric measurements. The microwave dielectric constant and the loss factor measured were found to be temperature dependant. This is shown by the results obtained which varied with the temperature. The dielectric constant and loss factor results ranged from 2.45–3.19, 2.39–3.51, 2.20–2.51, 1.80–2.51, 2.59–2.98, 2.56–2.87 for the greenish, reddish, yellowish, greyish, blackish and whitish sandstones respectively. On other hand, the dielectric loss factor ranged from 0.14–0.52, 0.01–1.01, 0.01–0.19, 0.01–0.18, 0.07–0.37 and 0.001–0.137 for greenish, reddish, yellowish, greyish, blackish and whitish respectively. The dielectric results exhibited in Figs. 12 and 13 are for the greenish and blackish sandstones respectively. Moreover, the tangent delta (tan δ) was calculated using the dielectric constant and loss fact results in the following equation:

$$\tan \delta = \varepsilon'' / \varepsilon' \tag{1}$$

where ε' is the dielectric constant and ε'', the dielectric loss factor.

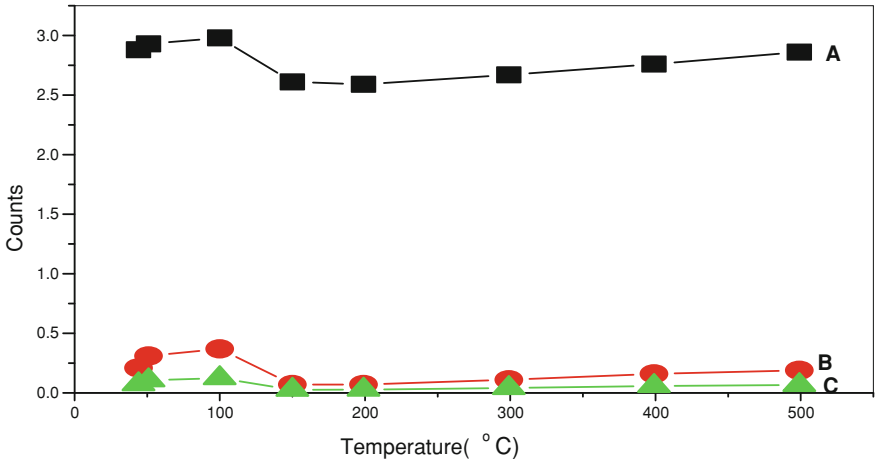


Fig. 13 Dielectric constant (*A*), dielectric loss factor (*B*) and calculated tangent delta (*C*) of blackish showing the temperature dependency

It has been reported that the correlation between the dielectric properties or permittivity and water content of hygroscopic materials for sensing moisture content is important [12]. In our case, the dielectric properties measured were used to simulate the electric field and power absorption distribution using HFSS software for the design of a microwave heating cavity.

5 Conclusion and Future Work

The mineralogy and physical properties of QwaQwa sandstones (South Africa) was successfully achieved. The XRD diffractograms confirmed that the sandstones are quartz based minerals along with traces of Kaolinite, Illite and Albite. Optical microscopy analyses revealed that the QwaQwa sandstones had grain sizes ranging from fine to coarse grains size. Both EDS and XRF results were in good correlation by exhibiting Silicon as the highest chemical element on the sandstones. On the other hand, the compressive strength and water absorption demonstrated a complexity in the ability for the samples to absorb water and to sustain a load. Furthermore, dry density analyses were conducted and results were satisfactory. Thus, the results of the characterisation studies provided good mineralogical and physical properties of different types of sandstones.

It is encouraged that sample collections at different sites in the QwaQwa rural area be carried out for further analyses. This will assist to confirm the findings in this study thus to create a database of identical physical and mineralogical properties results for the whole QwaQwa area which is desirable for use in the construction industry.

Acknowledgments The author acknowledges Prof A. F. Mulaba-Bafubiandi for initiating the project on QwaQwa sandstones. Fikile Moreti sandstones (Mr. Karafu, QwaQwa, South Africa) for providing the sandstone samples. The financial support of the University of Johannesburg is acknowledged.

References

1. M.P. Mubiayi, in *Characterisation of Sandstones: Mineralogy and Physical Properties*. Lecture Notes in Engineering and Computer Science, vol. III. Proceedings of the World Congress on Engineering 2013, WCE 2013, London, U.K., 3–5 July 2013, pp 2171–2176
2. P.A. Olubambi, S. Ndlovu, J.H. Potgieter, J.O. Borode, Mineralogical characterization of Ishiagu (Nigeria) complex sulphide ore. Int. J. Min. Proc. **87**, 83–89 (2008)
3. J. Gotze, H. Siedel, A complex investigation of building sandstones from Saxony (Germany). Mater. Charact. **58**, 1082–1094 (2007)
4. K. Zorlu, C. Gokceoglu, F. Ocakoglu, H.A. Nefeslioglu, S. Acikalin, Prediction of uniaxial compressive strength of sandstones using petrography-based models. Eng. Geol. **96**, 141–158 (2008)
5. J. Lurie, in *South African Geology for Mining, Metallurgical, Hydrological and Civil Engineering*, 7th revised edn. (Lupon Publishing, JHB, 1994)
6. M. Hajpal, A. Torok, Mineralogical and colour changes of quartz sandstones by heat. Environ. Geol. **46**, 311–322 (2004)
7. F.G. Bell, P. Lindsay, The petrographic and geomechanical properties of some sandstones from the newspaper member of the Natal Group near Durban, South Africa. Eng. Geol. **53**, 57–81 (1999)
8. V.S. Vutukuri, R.D. Lama, S.S. Saluja, in *Handbook on Mechanical Properties of Rocks*, vol. 1. (Trans Tech Publications Clausthal, Germany, 1974), p. 280
9. K. Zorlu, R. Ulusay, F. Ocakoglu, C. Gokceoglu, H. Sonmez, Predicting intact rock properties of selected sandstones using petrographic thin-section data. Int. J. Rock Mech. Min. Sci. **41**(1), 93–98 (2004)
10. S.K. Singh, Relationship among fatigue strength, mean grain size and compressive strength of a rock. Rock Mech. Rock Eng. **21**, 271–276 (1988), Springer
11. A. Kassim, E.T. Mohammad, Laboratory study of weathered rock for surface excavation works. VOT 75055, University Teknologi Malaysia (2007)
12. O. Nelson Stuart, in *Useful Relationships Between Dielectric Properties and Bulk Densities of Granular and Powered Materials* (U.S. Department of agricultural, Agricultural Research Service, 2004)

Spatial Prediction of a Pre-curved Bimetallic Strip Under Combined Loading

Geoffrey Dennis Angel, George Haritos and Ian Stuart Campbell

Abstract This work establishes a way of calculating the free end point position, of a pre-curved bimetallic strip, that is subjected to uniform heating. The prediction of the endpoint of a bimetallic strip is required during the design phase of an electronic control circuit sensor switch that uses a sensing/activating unit containing a bimetallic strip. Bimetallic sensors are normally flat at ambient temperature and at the required sensing temperature the strip bends into a radius of curvature, this then displaces the contact on the end of the strip to make or break an electrical circuit. Although the normal, flat type of bimetallic sensor exists, this work concentrates on a pre-curved bimetallic sensor at ambient temperature. A curved bimetallic strip sensor provides a much larger sensing range and displacement at the free end of the strip, per degree of temperature change, than for a straight bimetallic strip. The greater sensing range is due to the arc length of the bimetallic strip being longer which affords a greater flexibility at the activation point when compared to the chord length of an equivalent straight bimetallic strip. Pre-curved bimetallic test samples were subjected to heating whilst the motion of the free end point of the strip was recorded on a metal plate. As the heat applied to the samples was increased, many temperature points were recorded to generate approximate loci of points. The loci of test points compared well to theoretical curve generated by the derived formulae. Therefore the advantages of this work offers a less critical sensing range, it also benefits from a mechanism which can be

G. D. Angel (✉)
Mechanical Group, School of Engineering and Technology, University of Hertfordshire, Hatfield AL10 9AB, England
e-mail: g.d.angel@herts.ac.uk

G. Haritos · I. S. Campbell
Automotive Group, School of Engineering and Technology, University of Hertfordshire, Hatfield AL10 9AB, England
e-mail: g.haritos@herts.ac.uk

I. S. Campbell
e-mail: i.campbell2@herts.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_17,

designed to be much smaller and take less space in the product compared to a comparable flat bimetallic strips sensor.

Keywords Bimetallic · Compact · Curved · Design · Sensor · Spatial · Thermal

1 Introduction

The aim of this paper is to introduce a mathematical method of predicting the end point of a pre-curved bimetallic strip that is being uniformly heated. One end of the curved bimetallic strip is rigidly fixed against displacement and rotation, and the other end, is free to move. By the application of a uniformly distributed heat to the curved strip, the strip will straighten up. If a Cartesian coordinate system is adopted, a formula can be derived to describe the theoretical locus of the end point of the strip relative to an X and Y coordinate axis system. By using Timoshenko's Equation [1], for evaluating the bending of a bimetallic strip under heating conditions, in conjunction with the straightening formulae produced by this paper, it is possible to predict the radius of curvature and displacement of the end point of the free end, as a function of temperature. It will be shown by an actual bimetallic strip straightening test, demonstrating how the locus of test points correlate to the theoretical path. This paper can be used at the design stage of a temperature controlled circuit, whereby it is necessary to know the exact position of the end of a curved bimetallic strip for a given particular temperature. With this paper it will also be possible to specify the geometric and material properties of a curved bimetallic strip necessary to achieve other critical design objectives in a temperature controlled circuit. Additional expressions are introduced to evaluate the externally loaded end point case.

2 Theory

Timoshenko is used to evaluate the radius of curvature of a straight bimetallic strip. With the addition of a correction formula, it is still possible to use Timoshenko to evaluate the straightening of a pre-curved bimetallic strip. With the radius of curvature found and with other formulae derived in this work, it is possible to correlate the theoretical end point position of the free end of the strip, to the actual recorded data results from the test. Consider the curved bimetallic strip that is shown in Fig. 2, it is rigidly fixed at one end and free to move at the other end. When uniformly heated, it will tend to straighten up if the material side of the strip with the higher coefficient of linear expansion $\alpha 2$, lies on the inside surface, see Fig. 1. As the free end of the strip straightens up, it will adhere to a locus predetermined by the initial pre-curved "cold" radius of curvature and material properties and make-up of the bimetallic in question.

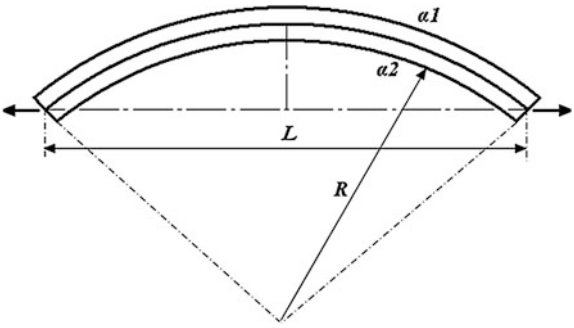


Fig. 1 Curved bimetallic strip with α_2 is on the inside

3 Evaluation of the "Hot" Radius of Curvature R_h

Application of Timoshenko curvature equation to obtain unloaded "hot" radius R_h.

3.1 Assumptions

The pre-curved bimetallic strip is rigidly fixed at one end, and free to move at the other end.
The strip is uniformly heated along the entire length of strip, and the strip remains truly circular.
No external loads are applied during heating.
The material with the higher coefficient of linear thermal expansion α_2 is on the inside radius R.

From Timoshenko [1], the radius of curvature of a bimetallic strip is given by:

$$\rho = \frac{t \cdot \left[3 \cdot (1+m)^2 + (1+m \cdot n) \cdot \left(m^2 + \frac{1}{m \cdot n}\right)\right]}{6 \cdot (\alpha_2 - \alpha_1) \cdot (T_h - T_c) \cdot (1+m)^2} \tag{1}$$

where

ρ is the radius of curvature as function of temperature from an ambient flat strip.
$t = t_1 + t_2$: Total thickness of the strip, t_1, t_2 being the material thicknesses.
$m = \frac{t_1}{t_2}$: Ratio of thicknesses.
$n = \frac{E_1}{E_2}$: Ratio of Young's Modulus.
T_h and T_c Hot and cold temperatures states of the strip.
E_2, E_1 are the linear Modulus of the two separate materials.
α_2 and α_1 are the coefficients of linear thermal expansion for the two metals whereby α_2 is assumed to be numerically larger than α_1.

The R_h correction equation evaluates the radius of curvature of a heated bimetallic strip from an initially pre-curved radius of curvature R_c by subtracting the reciprocals of both radii $\frac{1}{R_c} - \frac{1}{\rho}$.

Thus

$$R_h = \frac{\rho R_c}{\rho - R_c} \tag{2}$$

with R_h established by the application of the Timoshenko formula, the corresponding "hot" chord length L_h can now be found. The general chord length of any arc is generally known to be given by:

$$L = 2R \sin\left(\frac{\theta}{2}\right) \tag{3}$$

where

L is the chord length mm
R is the radius of curvature mm
A is the arc length (in radians) part of a true circle
***θ* = A/R** rad

And thus:

$$L_h = 2R_h \sin\left(\frac{A}{2R_h}\right) \tag{4}$$

3.2 The "Hot" Chord Length of the Straightened Strip

Evaluation of angle θ as a function of hot radius of curvature R_h is by considering the geometry of the pre-curved bimetallic strip. From Fig. 2, two Isosceles triangles exist, **Δoab** and **Δodc**. For both triangles, adding all the angles up to 180°: $2\gamma + \alpha = 180$ and $2\beta + \omega = 180$. The third relationship that can be found in Fig. 2 is $\gamma - \beta = \theta$.

By manipulation and substitution of these sub-formulae, it can be shown that:

$$\theta = \frac{\omega}{2} - \frac{\alpha}{2} \tag{5}$$

with further substitution and manipulation this is equal to:

$$\theta = \frac{A_c}{2}\left(\frac{1}{R_c} - \frac{1}{R_h}\right). \tag{6}$$

Given that $A_c = A_h$ the strip arc changes shape, not its length.

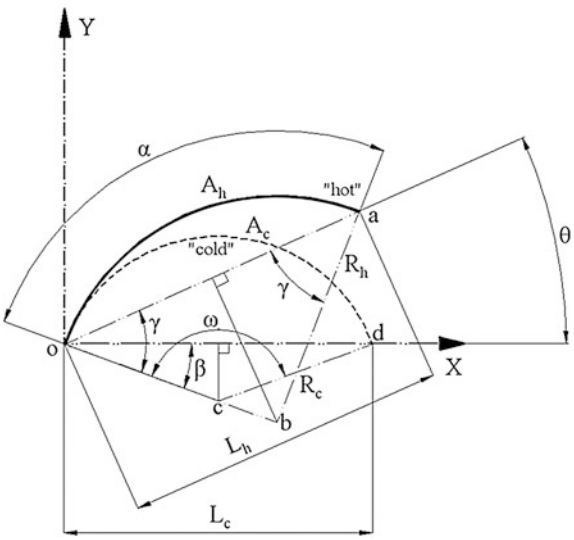


Fig. 2 Curved bimetallic strip geometry

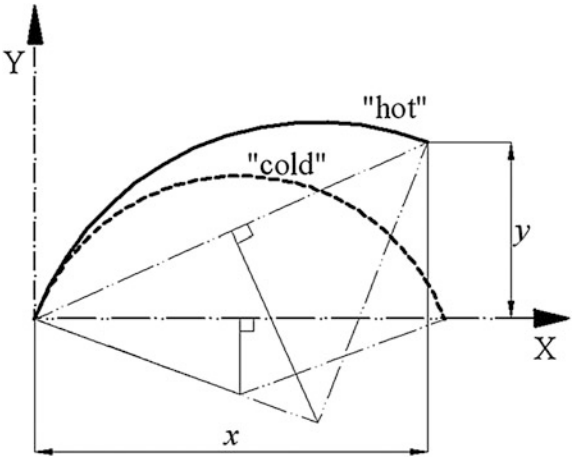


Fig. 3 *x* and *y* position of heated strip

Where

$A_c = A_h$ is the arc length of the curved bimetallic strip
R_c is the cold radius of curvature stated previously and initially known
R_h is the hot radius of curvature calculated by Timoshenko earlier.

Hence the "hot" endpoint position can be now calculated in terms of X, Y coordinate system, see Fig. 3.

$$\chi = L_h \cos\left[\frac{A_c}{2}\left(\frac{1}{R_c} - \frac{1}{R_h}\right)\right] \tag{7}$$

$$\gamma = L_h \sin\left[\frac{A_c}{2}\left(\frac{1}{R_c} - \frac{1}{R_h}\right)\right] \tag{8}$$

From (2) we have the radius of curvature R_h as a function of the temperature change from ambient, with this value entered into (7) and (8) the evaluation the χ, γ end point position of the bimetallic strip is possible. These formulae were used to generate the theoretical curves used later on in this paper.

The preceding case [2] enables the prediction of the free end of the pre-curved bimetallic strip as a function of temperature change from ambient. The output of the calculation was the x and y ordinates of the end point of the free end of the strip. Although this will enable an engineer to pinpoint the position of the end the of the strip quite accurately, if the function of the free endpoint of the strip is to activate a micro-switch or other tripping device, then the reaction load from the switch must be taken into account. When the endpoint of the bimetallic strip is loaded by an external force, the strip will behave as a beam and bend accordingly. The following theory describes the combined loading case whereby the pre-curved bimetallic strip is fixed at "o" and is free to move at point "d". Initially the ambient radius of curvature is R_c, and as the strip is uniformly heated along its entire arc length, it will straighten up as shown previously to reach position "a" with a "hot" radius of curvature R_h. With the application of an externally applied load F_α exerting in an orientation that is perpendicular to the radius of curvature R_h the strip will bend back downwards in accordance with simple beam bending theory. The deflection from point "a" to point "f" changes the radius of curvature from R_h to R_d. To observe the elastic properties of the strip under loading, simple bending theory is employed to evaluate the maximum deflection of the beam "oa" which is being subjected to vector force F_n that is acting upon the free displaced strip perpendicular to its chord line, see Fig. 4.

From the simple bending equation, the radius of curvature of a flat bimetallic strip is taken as:

$$R_b = \frac{E_a}{F_n}\frac{I}{L_h} \tag{9}$$

where

E is the averaged Young's Modulus of both metals making up the bimetallic strip.
I is the second moment of area of the strip.
F_n is the normal force to the chord line of the strip and $F_n = F_a \cos(\varphi)$.
Angle $\varphi = \frac{\pi - \alpha}{2}$.
F_a is the applied input load normal to the strip (N) coming from the application.
$L_h = 2R_h \sin\left(\frac{A}{2R_h}\right)$ chord length of hot free strip defined earlier.

Note R_b is the bend from flat as a function of a normal load F_n as considered to be applied to a straight strip.

But the load F_n is applied to a pre-curved beam with radius of curvature R_h thus a modifier is required to find the radius of curvature R_d which is the deviated radius from the already pre-curved strip R_h.

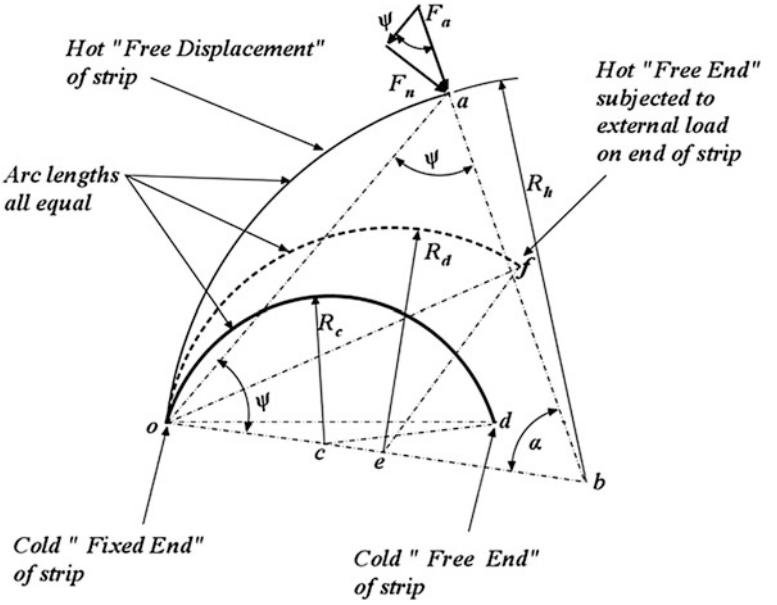


Fig. 4 Combined loading case of fixed pre-curved bimetallic strip

The modified radius of curvature is found by adding the reciprocals of the bend from flat R_b, to the bend from hot radius of curvature R_h;

Thus;

$$R_d = \frac{R_b R_h}{R_b + R_h}. \tag{10}$$

It is very possible to breach the elastic limit of the bimetallic strip by applying an external load that exceeds the permissible bending stress of the strip.

It is normal to get the permissible stress value for most mass produced bimetallic strip from the manufacture's handbook.

Assume that σbp is the maximum permissible bend stress value as supplied by the manufacturer.

The radius of curvature from a flat strip limited by stress is defined as:

$$R_\sigma = \frac{tE}{2\sigma bp} \tag{11}$$

The stress value evaluated as a function of the applied load F_n is given by:

$$\sigma_a = \frac{F_n L_h t}{2I} \tag{12}$$

where $\frac{t}{2}$ equals half the thickness of the strip

If as a function of applied load $F_n$the calculated value of $\sigma_a < \sigma_{ba}$ the manufacturer recommended permissible stress value, then the strip under the external load F_n will remain within the elastic limit, or the limit of proportionality, as defined by Hooke's law.

The radius of curvature as a function of σ_a, which itself is a function of the applied load F_n is therefore:

$$R_{\sigma a} = \frac{tE}{2\sigma_a}. \tag{13}$$

By substitution and simplification:

$$R_{\sigma a} = \frac{EI}{F_n L_h} \tag{14}$$

Thus if $R_{\sigma a}$, the radius of curvature as a function applied load F_n, is substituted for R_b the loaded radius of curvature $R_{d\sigma}$ from the hot free position can be calculated.

Therefore:

$$R_{d\sigma} = \frac{R_{\sigma a} R_h}{R_{\sigma a} + R_h} \tag{15}$$

A check between the calculated stress σ_a coming from the applied load F_n, and the recommended stress level by the manufacture σ_{bp} must be performed to ensure that the applied load and its associated stress level stays within the elastic limits of the bimetallic strip.

$R_{d\sigma}$ is the stress limited radius of curvature as a function of the applied load F_n. Note if the condition $\sigma_\sigma > \sigma_{bp}$ for an applied loading F_n then the geometry of the strip can be modified to reduce the stresses to within the elastic limit of the bimetallic material.

The new chord length as a function of the stress limited $R_{d\sigma}$ is given by:

$$L_{d\sigma} = 2R_{d\sigma} \sin\left(\frac{A}{2R_{d\sigma}}\right) \tag{16}$$

The net loaded position of the end of the strip at point *f is* given by:

$$X_n = L_{d\sigma} \cos\left[\frac{A_b}{2}\left(\frac{1}{R_c} - \frac{1}{R_{d\sigma}}\right)\right] \tag{17}$$

$$Y_n = L_{d\sigma} \sin\left[\frac{A_b}{2}\left(\frac{1}{R_c} - \frac{1}{R_{d\sigma}}\right)\right]. \tag{18}$$

It is normal practice to program the preceding expressions into an electronic spread sheet so that the iterative process can zero in on a working solution that satisfies the combined loading and stress requirements. The tests that follow are for the validation of the unloaded case only.

4 Equipment

Bimetallic strip used in test; Shivalik SBC-206-1 [3] which were initially 202 mm long × 5 mm wide × 0.4 mm thick straight bimetallic strip. Four bimetallic strip test samples were made by gently cold working the strips to form true arcs of a circle equal to D64, D80, D100 and D128 mm. The bimetallic strips were formed with the material side with the highest coefficient of linear expansion on the inner surface. To ensure that the curved bimetallic strips conformed to a true arc during cold working, special formers were produced to check the diameter and roundness see Fig. 4, the formers were held to ±0.25 mm tolerances. The length of each test sample was cut back to equal half the circumference of the former, i.e. 100.53, 125.66, 157.07, 200.06 mm long respectively, within a tolerance of ±0.25 mm. Each test sample was subjected to heat treatment to 350 °C for 2 h before the actual testing took place as according to the Kanthal handbook [4]. This was to normalize the strips from any work hardened induced stresses during the cold forming process.

4-Curved bimetallic test samples, tagged as: D64, D80, D100, D128, Hanna HI 93530 K-Thermocouple Thermometer Digital: Thermocouple: position T1. TES1319 K-Type Thermometer 2 off: Thermocouple: position's T2 and T3. Solex, Digi-Thermo ST 4060 Digital Thermometer recording ambient temperature Ta. Each thermocouple was affixed to each test sample by a spring clip on the outside surface, and shielded by the body of the test sample from direct hot air flow.

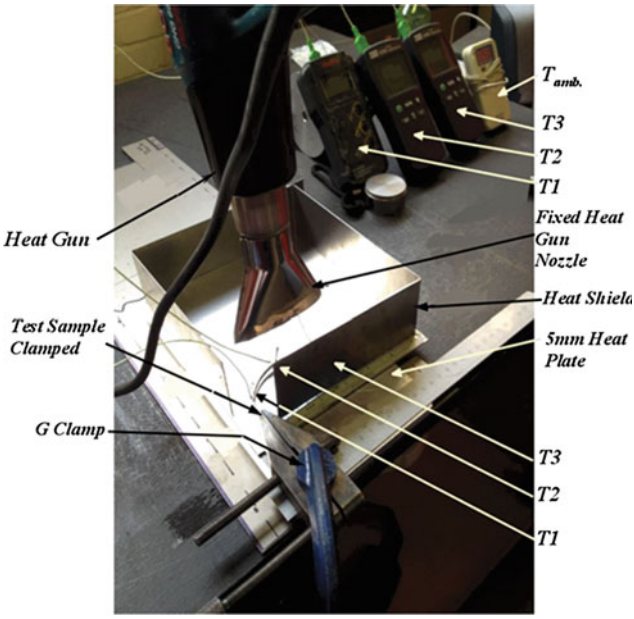
The positioning pattern of the three thermocouples was the same for all test samples, see Fig. 6. Bosch 2.3 kW GHG 660 LCD Professional; variable flow hot air gun; adjustable heat settings in increments from 10 to 600 °C. Fan type nozzle for maximum flow spread along the test samples. 50 °C the lowest temperature setting output of the gun. Hot air flow rate and position of heat gun fixed for all testing.

Heat flow was perpendicular to the Aluminium Base plate for a constant uniform heating environment. A 5 mm heat stabilising plate was placed underneath the Aluminium base plate. A heat stabilising shield was placed around the test pieces during testing.

A 1 mm thick sheet Aluminium base plate was used for recording locus of points during the tests. Test sample holder, clamped to the Aluminium base plate using a workshop "G" clamp. The test samples were clamped in the test sample holder to 1 mm depth for each strip. The test samples were measured to be 1 mm parallel to, and clear of the Aluminium base plate throughout all tests, see Fig. 6. A black fine felt tip pen was used for recording data points on the Aluminium base plate, see Fig. 5.

Fig. 5 Test samples checking for roundness and size

Fig. 6 Test setup



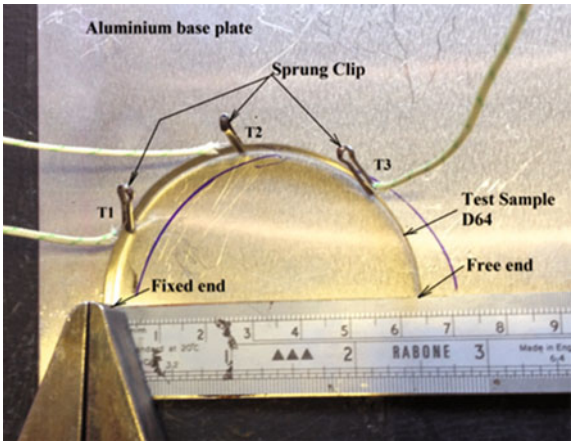


Fig. 7 Test sample setup

5 Test method

Each test sample was clamped parallel to the Aluminum base plate within the heat stabilized zone. One end of the bimetallic strip was rigidly fixed, the other end the strip was free to move 1 mm from the plate, see Fig 6. Each test sample was subjected to uniform heating, and as the strip straightened up, the locus of the free end point was recorded on the Aluminum base plate using the felt tipped pen. At each point, the corresponding thermocouple temperature was recorded, see Fig. 7. The heat from the gun was increased in increments of 20 °C and an identifiable locus of points was produced for each test sample, see Fig. 7.

6 Test Results and Discussion

The result of heating and plotting the loci of points are shown in Fig. 8. The data points were plotted onto the surface of the Aluminum base plate. The range of points were plotted for a thermocouple temperature range of up to 210 °C for an ambient room temperature in the range of 21–24 °C. Sufficient data points were obtained to identify the locus of the free end of the strip for each test sample. For each test sample, 5–7 data points corresponding to the specific thermocouple temperatures, which were measured from the best fit curve in Fig. 8. Sample test points were plotted against the calculated curves generated by the formulae presented in the theory section of this paper, a good correlation was observed for the majority of the test points (Fig. 9).

Generation of the theoretical calculated data curves in Fig. 8 were computed using an Excel program. The properties used to calculate the theoretical values were as follows:

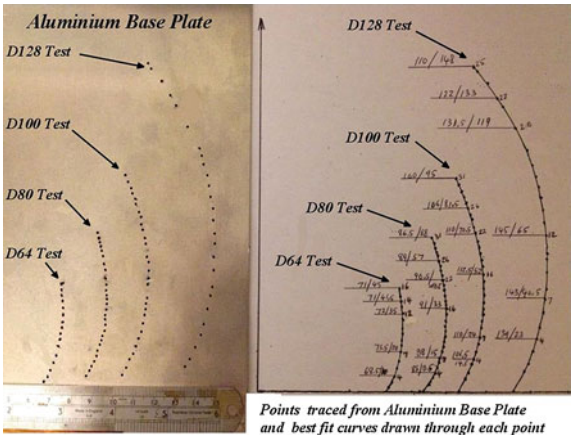


Fig. 8 Full set of test data points transferred to paper and X, Y, distances measured

Thickness of each metal 0.2 mm equal; total thickness = 0.4 mm, Young's Modulus of Steel 210 GPa, Young's Modulus of Invar 36: 145 GPa, source [5].

Coefficient of linear thermal expansion for steel: 20×10^{-6}/K [5].

Coefficient of linear thermal expansion for Invar 36: 1.85×10^{-6}/K.

With the above values and the Timoshenko formula, coupled with the same thermocouple temperatures from the test samples, the theoretical data points were generated.

Despite the manual method of plotting and recording of the data points, a good correlation exists between the theoretically derived curves and the sample data points from the tests, as can be shown in Fig. 8. The best correlation occurs on the smaller test samples, whereby the heat source was the closest to the test samples. The larger the test piece, the further the test sample moved from the fixed direct heat source and thus the scatter of the data points was the greatest. On sample D128, the largest deviation from the theoretical curve was recorded, this was due to the continuous movement of the free end of the bimetallic strip during heating, a phenomena known as hunting. The correlation results are shown in Fig. 9 with an overall average percentage error of the four test samples amounting to X% error 0.35 and the Y% error 5.2 the worst deviation in the D128 sample, Y axis which was 9.32 %, again due to the inaccuracy of recording caused by the hunting of the free end of the test sample. The test results were shown on the whole, to have a good correlation with the theory, thus the equations in this paper can be used with a high confidence as a means of predicting the end point position of the free end of a curved bimetallic strip, when subjected to uniform heating and unloaded from any external forces.

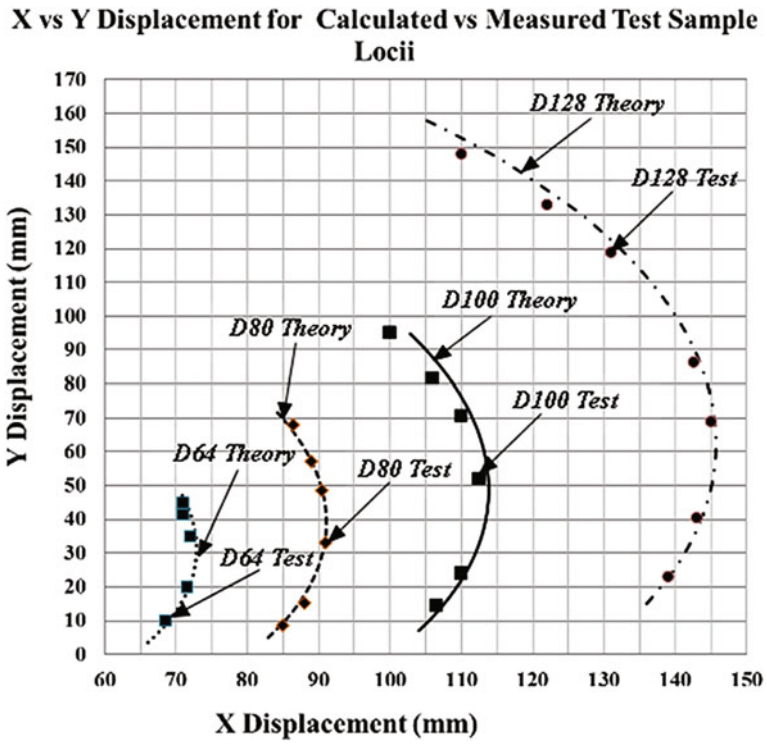


Fig. 9 Comparison of theory to test data

7 Conclusion

This work provides a means of calculating the free end point position of a curved bimetallic strip subject to uniform heating. With the aid of a Microsoft Excel work sheet or other similar electronic worksheet, the major equations can be easily evaluated for any curved bimetallic strip to provide design options in any control circuit using a bimetallic element as the sensing unit. The low overall percentage correlation error between the test data and the theory validates the formulae derived in this paper indicating that they can be applied with high degree of confidence when predicting the movement of the end point of the strip due to heating.

References

1. S. Timoshenko, Analysis of bi-metal thermostats. JOSA **11**, 233–255 (1925)
2. G. Angel, G. Haritos, I. Campbell., in *WCE 2013*: *Straightening locus of a curved bimetallic strip subjected to heating*. Proceedings of The World Congress on Engineering, London. Lecture Notes in Engineering and Computer Science, pp. 2059–2064, 3–5 July 2013

3. B.C. Shivalik, Ltd., *Bimetallic Strip Supplier*, SBC-206-1 [cited 2013 November]. Available from http://www.shivalikbimetals.com
4. A.B. Kanthal, *The Kanthal Thermostatic Bimetal Handbook*, vol. 135. (Hallstahammar: Kanthal, 2008), p. 30.
5. Matweb, Online materials information resource [cited 2013 November]. Available from http://www.matweb.com

Development of a Glass-Fibre Reinforced Polyamide Composite for Rotating Bands

Abdel-Salam M. Eleiche, Mokhtar O. A. Mokhtar and Georges M. A. Kamel

Abstract Projectiles are usually provided with an integral rotating band which serves many purposes. These bands are usually made of copper which causes the wear of the gun barrel steel bore. Hence alternative materials are being thought. In the paper, the Polyamide type 66 is proposed as a base material, and its mechanical and tribological properties are modified by different percentages of glass fiber. Experimental results indicate that 2 % GF content as reinforcement to PA66 resin appears to be the ideal compromise.

Keywords Development scheme · Glass-fiber reinforcement · High-speed friction wear and deformation test · Mechanical properties · Polyamide 66 resin · Rotating band

1 Introduction

A rotating band (RB) is a part of a projectile that serves to engage the rifling on the gun barrel and trap propellant gases at the rear of the projectile. Engaging with the rifling imparts a spin on the projectile, which stabilizes it during its flight.

A.-S. M. Eleiche (✉)
Department of Mechanical Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Kingdom of Saudi Arabia
e-mail: eleichea@hotmail.com

M. O. A. Mokhtar
Department of Mechanical Design and Production Faculty of Engineering, Cairo University, Guiza, Egypt
e-mail: moamokhtar@gmail.com

G. M. A. Kamel
Engineering Consultant, Cairo, Egypt
e-mail: gmakhh75@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_18,

Its secondary function is to hold the projectile in its proper position in the gun after loading and ramming, and to ensure that it will not slip back when the gun is elevated. The band has considerable effect on muzzle velocity, range, accuracy, and the life of the gun.

Rotating bands have been usually made of fine soft copper. At high rate of fire and high muzzle velocities, this leads to the wear or erosion of the gun barrel steel bore. Since the operating variables cannot be changed, the system structure change is the only way, which addresses the selection of a suitable alternative material. In major-caliber projectiles, a small percentage of nickel is added to provide greater strength. In recent designs, some projectiles have been banded with gilding metal (90 % copper, 10 % zinc), to increase strength and reduce the amount of copper deposited on gun barrel bore. Rotating bands have also been made out of other materials including brass, iron, and plastic [1–4].

In considering the design and selection of RB material, it should be remembered that the RB must:

- withstand the shearing and the bearing pressure resulting from normal forces
- be permanently clamped to the projectile body
- plastically deform through the barrel forcing cone (FC) without rupture
- have a minimum effect on wear of the barrel bore.

In the current work, it has been thought to use a polyamide material instead of the copper alloy RB, in an endeavor to minimize wear rates and hence increasing barrel life, while maintaining or even enhancing the main performance characteristics (mechanical and thermal properties, sealing, etc.) Polyamide type 66 (PA 66) was proposed as a base material, and modified by adding different fillers to improve strength, wear resistance, and flame reduction. The adopted test scheme is shown in Fig. 1, which consists of four stages. The current paper, which expands on a previous presentation [5], reports in detail on only the first one of these phases, where different percentages of glass fiber are added to improve strength and wear resistance.

2 Experimental

The primary criterion for the optimum choice of RB material is its minimum wear effect on the barrel under working conditions. Also, this material must withstand all applied stresses without failure, and resist wear for realizing complete sealing to prevent the escape of gases. Therefore, the design of a RB material must be based upon several experiments simulating as much as possible the real function.

In addition to mechanical properties measurements, a special rig is constructed for the friction, wear and deformation testing, in conditions simulating the real ones.

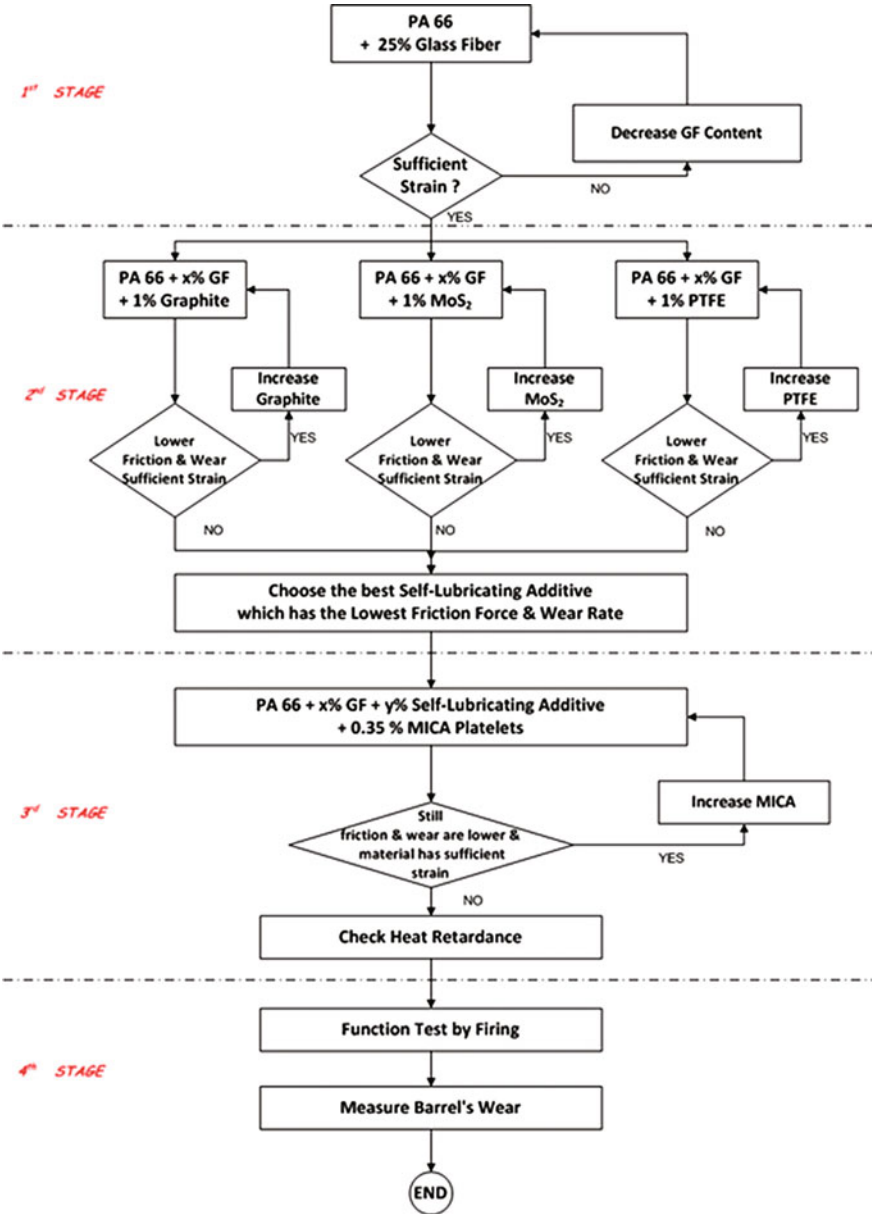


Fig. 1 Full scheme adopted to develop the new reinforced polyamide RB

Table 1 Resin properties

Properties	ULTRAMIT A-5	ULTRAMIT A3EG6
Fiber glass content, %	0	25
Density, gm/cm^3	1.14	1.44
Tensile strength, psi (MPa)	12.5×10^3 (86.2)	18×10^3 (124.1)
Elongation at break, %	40–80	2
Young's modulus	4.3×10^5 (2965)	7×10^5 (4862)
Compressive strength, psi (MPa)	16×10^3 (110.3)	25×10^3 (172.4)
Shear strength, psi (MPa)	9.5×10^3 (65.5)	13×10^3 (89.6)
Izod impact strength, ft.lb/0.5 in notch	1.1	2.1
Melting point (^{o}C)	264	240
Deflection temperature (^{o}C)		
Load 88 psi	190	234
Load 264 psi	75	228

For specimen fabrication, a screw extruder is constructed for laboratory use, allowing the preparation of specimens with different additive contents.

Mechanical properties have been recorded using standard testing machines and test specifications. Both tensile and compression tests were conducted and corresponding stress/strain relations could be identified.

Furthermore, a specially constructed rig has been used to measure the friction and wear under very high rates of loading and sliding speeds, simulating the actual working conditions. Coefficients of friction and wear values could be assessed as functions of the applied loads and sliding speeds.

2.1 Materials

Two types of resin were used:

- Polyamide 66 resin of natural color and for general use acquired in the form of grain from BASF Company. The commercial name of this material is "ULTRAMIT A-5" and has the properties given by the supplier as listed in Table 1.
- Short GF reinforced PA 66 acquired also from BASF. The GF content is 25 % of E-type randomly oriented in the resin. The commercial name is "ULTRAMIT A3 EG6"; properties are listed in Tables 1 and 2.

2.2 Specimens

The RB was molded in place around the projectile body by melt extrusion in a specially designed and constructed laboratory screw extruder. Specimens were also molded to determine the mechanical properties.

Table 2 E-Glass properties

Single fibre tensile strength at 25 °C, psi (MPa)	530,000 (3,654.2)
Young's modulus, psi (MPa)	11.8×10^6 (75,842.3)
Density (g/cm^3)	2.53

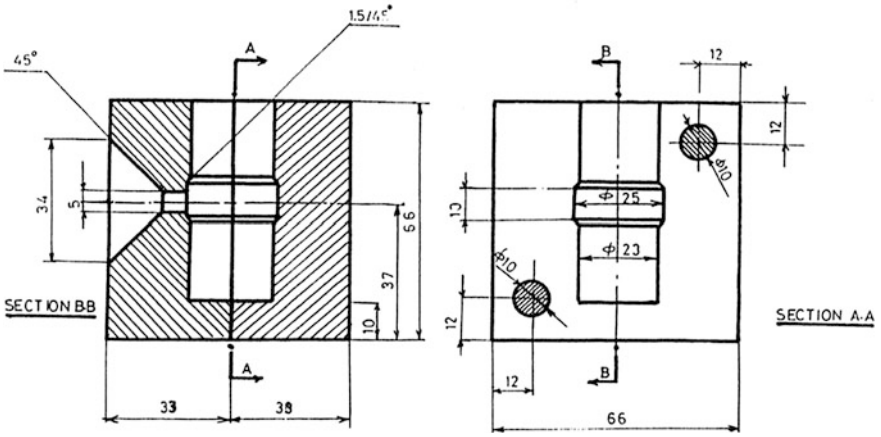


Fig. 2 Special die used for molding the RB around the projectile body

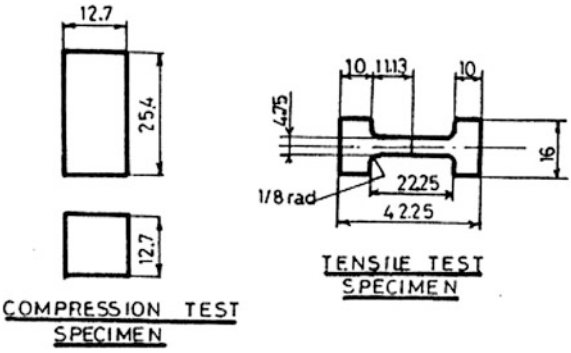


Fig. 3 Tensile and compression test specimens used for material characterization

2.3 Screw Extruder and Special Dies

This consists essentially of a drive, heating cylinder and screw. Detailed drawing of this extruder has been given in [5]. The special die used for molding the RB around the projectile body is shown in Fig. 2.

Specimens used for material characterization are shown in Fig. 3, whereas the dies used to mould them are shown in Fig. 4.

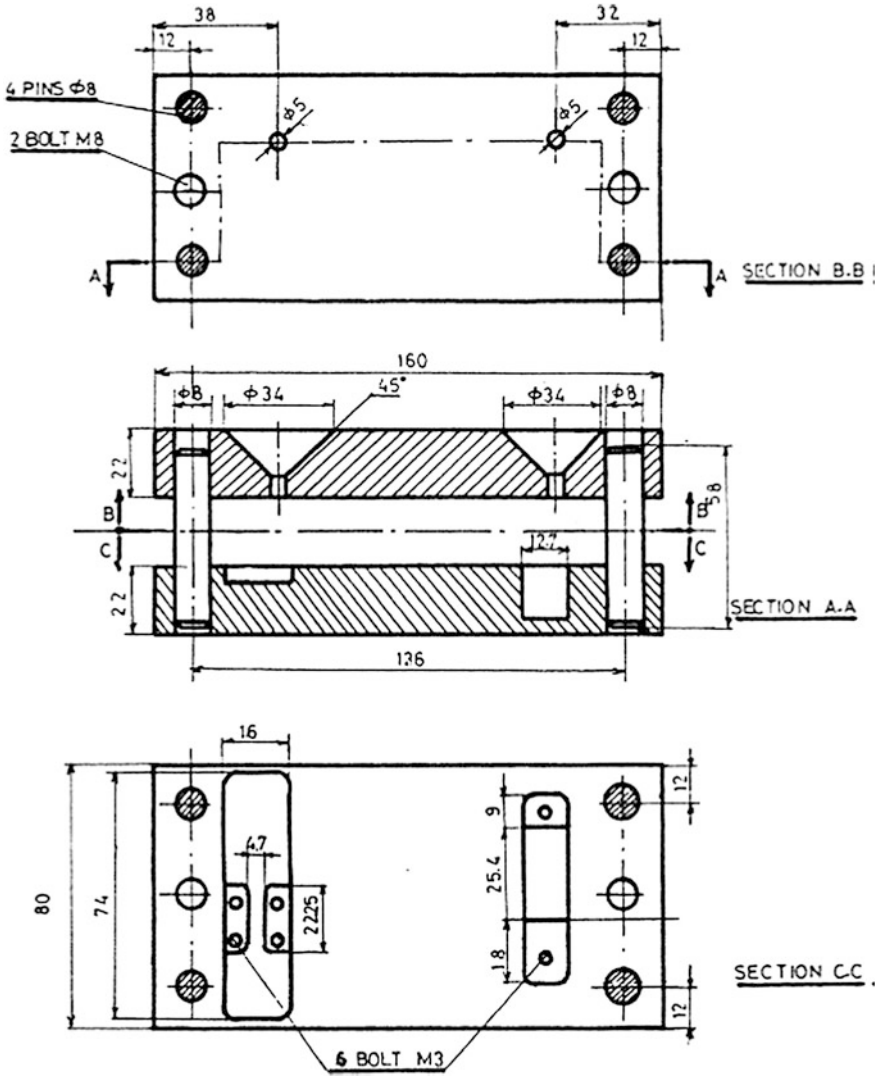


Fig. 4 Special die used for molding the tensile and compression test specimens

2.4 Test Rig for Measuring Friction, Wear and Deformation of a RB at High Sliding Speeds

Montgomery measured the friction and wear of the RB material using a pin-on-disc type testing rig, but with a very short contact time in the range of 100 ms [6]. Experiments were made at a sliding speed of 1.7 m/s, corresponding to real projectile velocities. Wear rate was taken as the loss in pin length during the experiment.

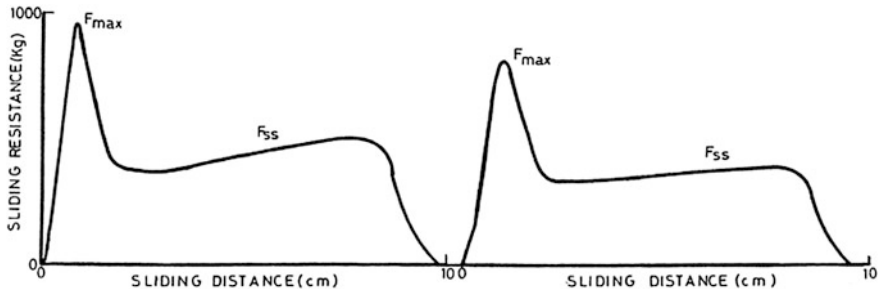


Fig. 5 Typical trace obtained in the static friction, wear and deformation test

In practice, the performance of the RB is checked, after the projectile is produced in the factory, by pressing a sample from each lot inside a portion of a barrel, using a universal testing machine, at a sliding speed of no more than 0.1 m/s. Figure 5 shows the resisting force versus distance traveled at speeds of 0.00055 and 0.1 m/s in this type of static friction, wear, and deformation test. Three parameters are measured: the maximum force reached (F_{max}), the steady-state force (F_{ss}) during the sliding motion of the projectile, and the slope of the initial part of the curve (K).

The method used by Montgomery does not take into account the deformation of the RB, and also neglects the non-uniformity of wear throughout the length of the RB. On the other hand, the method used in production is conducted at very low speeds in comparison with those existing in practice.

Therefore, a test rig is specially designed and constructed in order to simulate the real severe conditions the RB suffers during service (Fig. 6). It consists of a falling weight used to drive the projectile into an actual portion of a real barrel. The RB deforms when going through the forcing cone of the barrel, and then slides over the remaining part of the barrel until it is finally pulled out. The system allows a maximum falling height of four meters, corresponding to a striking velocity of 8.9 m/s. The applied pressure at striking was calculated and estimated at 5 MPa. In this test, the friction force is detected by a compression load cell, and its output recorded by a storage oscilloscope triggered externally by a pulse generator. Traces were properly calibrated on a universal testing machine. Also, the wear of the RB was expressed by recording the dimensional changes after exiting the barrel.

2.5 Mechanical Properties Testing

Tensile and compressive tests were conducted according to ASTM-D-1708 [7] and ASTM-D-695 [8], at 23 °C and at 1–1.3 mm/min and 1 mm/min, respectively. Five specimens were tested for each sample. True stress–true strain curves were

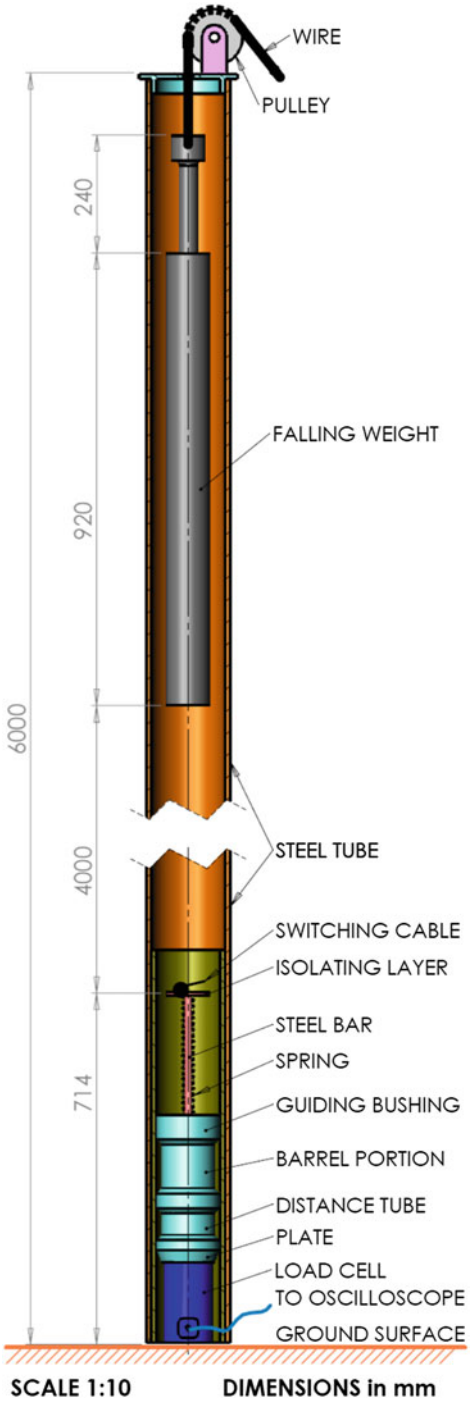


Fig. 6 Test rig for dynamic friction, wear and deformation test

deduced from the load-displacement traces obtained on an X–Y recorder. The following data were calculated: (a) strength at yield and at break; (b) modulus of elasticity; (c) strain at yield and at break.

3 Results and Discussion

3.1 OFHC Copper RB Reference Material

3.1.1 Low-Sliding Speed Friction, Wear, and Deformation

Figure 7 shows typical traces obtained at the sliding speed of 0.00055 m/s. From such traces, average values calculated were:

- Maximum resisting force, $F_{max} = 3{,}400$ kg
- Steady-state resisting force, $F_{ss} = 1{,}800$ kg
- Slope of initial part of the trace, $K = 486$ kg/mm

Dimensional changes were carefully measured. Using a simple analysis [9], the static coefficient of friction could be derived at $\mu_{st} = 0.38$.

3.1.2 High-Sliding Speed Friction, Wear, and Deformation

Figure 8 shows a test trace obtained at the sliding speed of 8.9 m/s using the rig of Fig. 6. From such traces, average values calculated were:

- Maximum resisting force, $F_{max} = 7{,}750$ kg
- Steady-state resisting force, $F_{ss} = 1{,}650$ kg
- Slope of initial part of the trace, $K = 1{,}107$ kg/mm

Also, the coefficient of friction was calculated at $\mu_{dyn} = 0.348$ [9], which is smaller than μ_{st}, but close to the values reported by Montgomery [6].

3.2 Glass Fiber Reinforcement of PA

3.2.1 Compression Test Results

Figure 9 shows that the addition of GF improves the strength and modulus of elasticity of PA 66, but lowers its ductility.

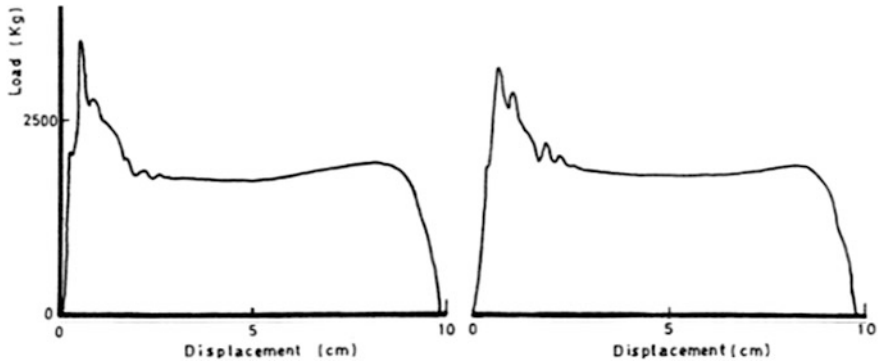


Fig. 7 Typical traces from friction, wear and deformation test on copper at 0.00055 m/s

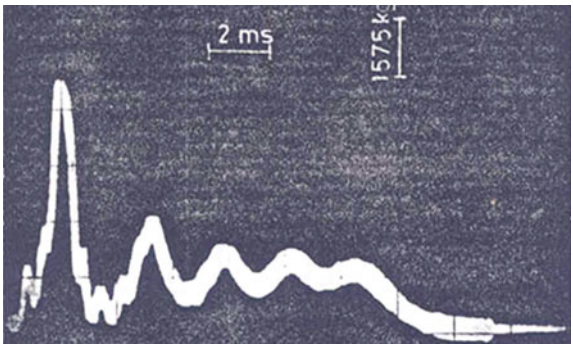


Fig. 8 Typical trace from friction, wear and deformation test on copper RB at 8.9 m/s

3.2.2 Low-Sliding Speed Friction, Wear, and Deformation

GF contents used were: 0, 2, 5, 10, 15, 20, and 25 %, while sliding speeds chosen were: 0.55×10^{-3} and 0.1 m/s. From the recorded traces, results were plotted in Fig. 10. The following remarks can be made:

- Recorded traces were similar to those obtained on copper specimens (Fig. 7)
- F_{max} and F_{ss} increase with GF content for both speeds.
- μ decreases by increasing GF content up to 15 %, then increases slightly; and generally increases with speed.

Visual inspection of the test specimens before and after tests revealed:

- All specimens did not fail, except the specimen with 25 % GF content at 0.1 m/s where cracks were clear inside the RB.
- No debris were found on the barrel bore at all speeds.

Specimen dimensions were measured after tests as reference against which the severe wear expected at high sliding speeds will be compared.

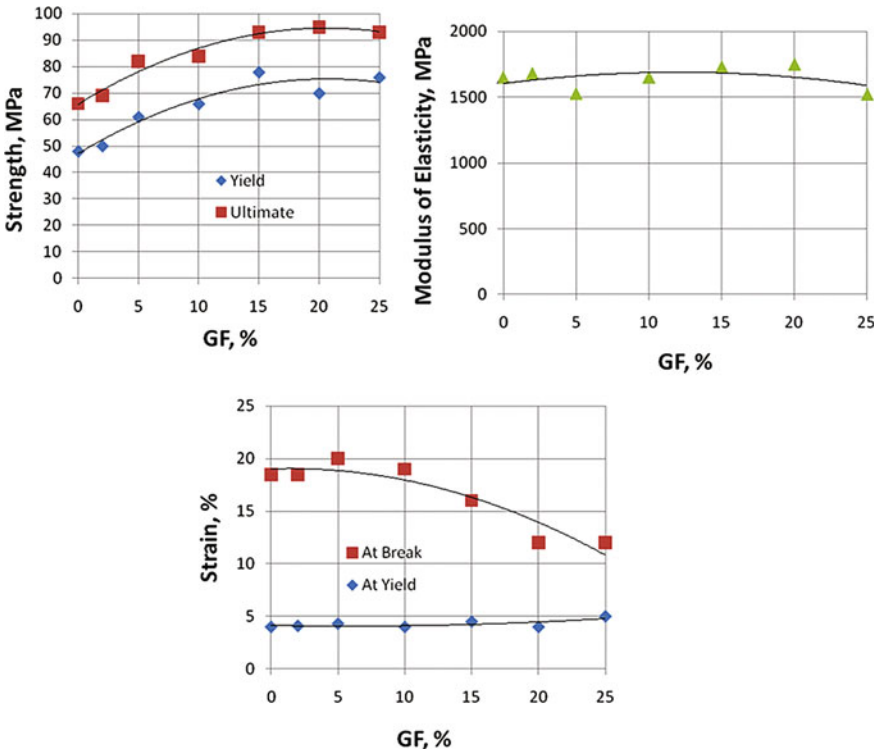


Fig. 9 Effect of GF content on compressive properties

3.2.3 High-Sliding Speed Friction, Wear, and Deformation

GF contents used were: 0, 2, 5, 10, 15, 20, and 25 %, and sliding speeds chosen were: 7.75 and 8.9 m/s. From the recorded traces, sample results for the resisting forces are given in Fig. 11. The following remarks are made:

- Recorded traces are similar to those obtained on copper specimens (Fig. 8). In the SS region, the RB seems to behave as a damped vibrating spring with possible slight stick-slip behavior.
- F_{max} increases with GF content and sliding speed
- At 7.75 m/s, F_{ss} slightly increases with GF content up to 10 %, then slightly decreases; at 8.9 m/s, it continuously decreases.
- μ decreases by increasing GF content up to 15 %, then increases slightly; and generally decreases with speed up to 15 % GF content, then slightly increases (cf. Fig. 10).

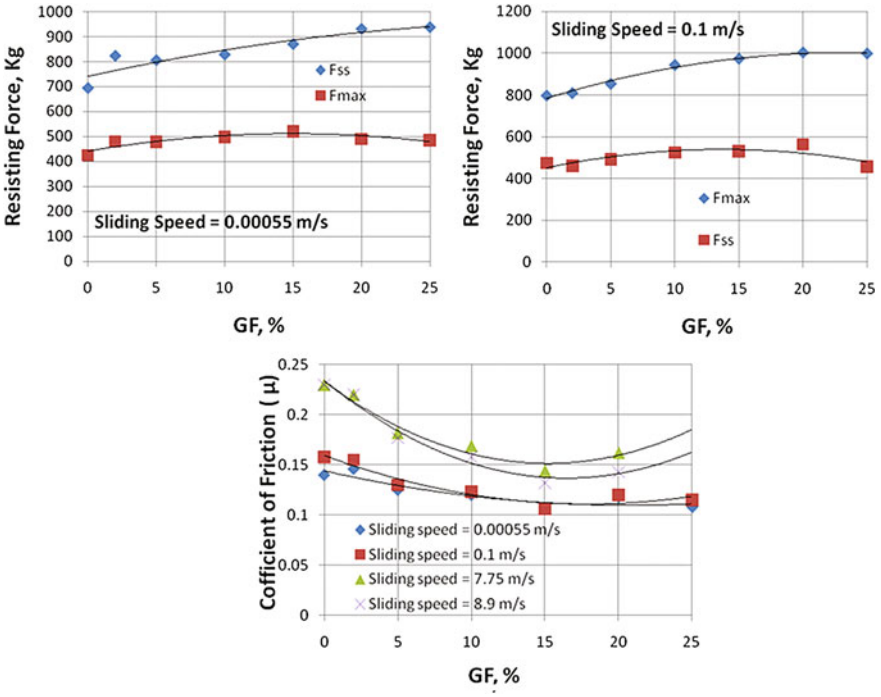


Fig. 10 Effect of GF content and sliding speed on F_{max}, F_{ss}, and μ

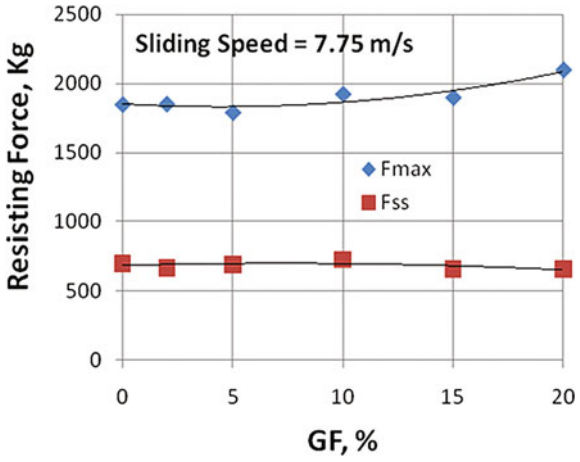


Fig. 11 Effect of GF content on resisting forces at 7.75 m/s

Visual inspection of the test specimens before and after tests revealed:

- At 7.75 m/s, specimens with GF content >5 % break during the test; at 8.9 m/s, specimens break when GF content is greater than 2 %.
- Transferred layers were observed adhering to the bore.

Dimensions of specimens that did not fail were measured after tests and compared to the reference dimensions of specimens tested statically. Relevant observations are:

- Front parts of the RB wear more than the rear parts
- Rear parts of the RB increase in dimensions for the GF filled specimens
- At 8.9 m/s, all dimensions of unfilled PA66 RB decrease in value. The decrease is more than for the case of OFHC copper RB at the same conditions.
- The quantity of wear occurring in the grooves is much greater than that occurring in the lands for unfilled PA66
- Wear resistance is improved by GF reinforcement
- By increasing GF content above 5 % for sliding speed of 7.75 m/s, and above 2 % for sliding speed of 8.9 m/s, it is seen that the RB is completely destroyed.

4 Conclusions

Polyamide type 66 (PA66) is proposed as candidate base material for rotating bands. Various percentages of glass fiber are to PA66 in order to modify its mechanical and tribological properties. In the assessment process, various tests were conducted, including mechanical properties determination, friction, wear and deformation tests at low and high sliding speeds, and dimensional change measurements.

Experimental results indicate that 2 % GF content as reinforcement to PA66 resin appears to be the ideal compromise. In this case, the wear resistance of PA66 resin is improved, and the RB withstands the associated plastic deformation without failure at all sliding speeds.

Acknowledgments A. M. Eleiche would like to thank the staff of the Tribology Laboratory at the Faculty of Engineering, Cairo University for their support in conducting the tests, and King Fahd University of Petroleum and Minerals for providing necessary facilities for the preparation of this paper.

References

1. M. Eig, Evaluation and critique on use of polymeric materials as rotating bands on 20 mm projectiles, Technical Report 4358, Picatinney Arsenal, Dover, NJ, Sept 1972
2. W.S. Larsen, R.B. Steidley, S.J. Bilsbury, O.K. Heiney, Development of a plastic rotating band for high performance projectiles. AFATL-TR-74-106, July 1974

3. R.S. Montgomery, Interaction of copper-containing rotating band metal with gun bores at the environment present in a gun tube. Wear **33**, 109–128 (1975)
4. M.D. Raby, Effects of temperature and humidity on glass-reinforced nylon rotating bands. M. Sc. thesis, Mechanical Engineering, Utah State University, 2010
5. A.M. Eleiche, M.O.A. Mokhtar, G. M. A. Kamel, in Glass-fiber reinforced Polyamide for rotating band application. *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013* (*WCE 2013*), London, UK, 3–5 July 2013, pp. 1761–1765
6. R.S. Montgomery, in The sliding behaviors of copper alloys. *The International Conference on Wear of Materials*, Washington, D.C, April 1983, pp. 465–470
7. ASTM, Standard test methods for tensile properties of plastics by use of micro tensile specimens, 1975 Annual Book of ASTM Standards, Part 35, D-1798, USA
8. ASTM, Standard test methods for compressive properties of rigid plastics, 1975 Annual Book of ASTM Standards, Part 35, D-695, USA
9. M.O.A Mokhtar, A.M. Eleiche, E.S. Nasr, G.M.A. Kamel, in *The Design of a new test for dynamic sliding friction and wear measurements of dynamic plastically deforming materials*. 3rd Tribology Conference of the Egyptian Society of Tribology (EGTRIB-92), Cairo University, Cairo, 1992

Design and Development of the Ultrasound Power Meter with a Three Axis Alignment System for Therapeutic Applications

Sumet Umchid and Kakanumporn Prasanpanich

Abstract The total output power from medical ultrasound devices must be determined and strictly regulated to ensure patient safety and to evaluate the performance of the ultrasound devices. The objectives of this research were to design and develop an ultrasound power meter with a three axis alignment system to measure the total output power from medical ultrasound devices especially for therapeutic applications. The implementation of this work utilizes a radiation force balance technique based on the method recommended in the International Electrotechnical Commission (IEC 61161). An ultrasound therapy machine was used as an ultrasonic source. To verify the performance of the developed system, the total output powers measured from our developed ultrasound power meter were compared with those measured from the commercial ultrasound power meter (UPM) and compared with those measured from the standard ultrasonic power measurement system at the National Institute of Metrology, Thailand (NIMT) at five nominal intensity values (0.5, 1, 1.5, 2, 3 W/cm^2) with three frequencies, 0.86, 2 and 3 MHz, and four output pulse modes; continuous wave (100 % duty cycle), 1:2 (50 % duty cycle), 1:5 (20 % duty cycle) and 1:10 (10 % duty cycle). The correlation coefficients and measuring uncertainty were then calculated. The results show that the developed system is currently able to determine the ultrasonic output power in the power range from 100 mW to approximately 12 W.

Keywords Radiation force balance · Ultrasonic power measurement · Ultrasonic transducer · Ultrasound metrology · Ultrasound power meter · Ultrasound therapy

S. Umchid (✉) · K. Prasanpanich
Department of Industrial Physics and Medical Instrumentation,
King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road,
Wongsawang, Bangsue, Bangkok 10800, Thailand
e-mail: sumetu@kmutnb.ac.th

K. Prasanpanich
e-mail: kanoonkapong@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_19,

1 Introduction

Medical ultrasound has been used extensively in both diagnostic and therapeutic applications during the past few decades [1–5]. For medical equipment that interacts with human tissue, whether invasive or noninvasive, the patient safety is the most important issues. Although medical ultrasound devices do not give ionizing radiation such as X-ray, other possible biological effects associated with medical ultrasound such as thermal or mechanical effects are a concern [6, 7]. For this reason, the ultrasonic power produced at the output of medical ultrasound devices should be determined and strictly regulated [8–10]. Other main reasons to measure the total output power are to ascertain whether the device is performing properly and to ensure the most effective exposure levels used during the treatment of the patient [11].

The ultrasonic output power can be determined by various procedures such as the radiation force balance technique [8–10, 12–15], the use of piezoelectric hydrophones [16], acousto-optic [17], thermo-acoustic [18], calorimetry [19] and ultrasonic power through electro-acoustic efficiency of transducers [20]. However, the radiation force balance method was employed in this work because this method is inexpensive, simple and accurate [8–10].

Radiation force balance is a standard technique to measure the total output power from an ultrasonic transducer. The ultrasonic power is determined by using the time-average force acting on a target in the acoustic field. The radiant power is directly proportional to the total radiation force (weight) on the target. The measurement principle is as follows: the ultrasonic beam to be measured is directed vertically upwards on the target and the radiation force acting on the target is measured by the electrobalance. The ultrasonic power is determined from the difference between the force on the target with and without ultrasonic radiation [8–10, 21–25].

In view of the above, it is clear that there is a well defined need for the ultrasonic power measurement. Consequently, this paper describes the design and development of an ultrasound power meter with the three axis alignment system to obtain faithful results of the ultrasonic power.

2 Methods

The implementation of this work utilizes a radiation force balance technique based on the method recommended in the International Electrotechnical Commission (IEC 61161) [26]. A schematic diagram and a photograph of the developed ultrasound power meter are presented in Figs. 1 and 2, respectively.

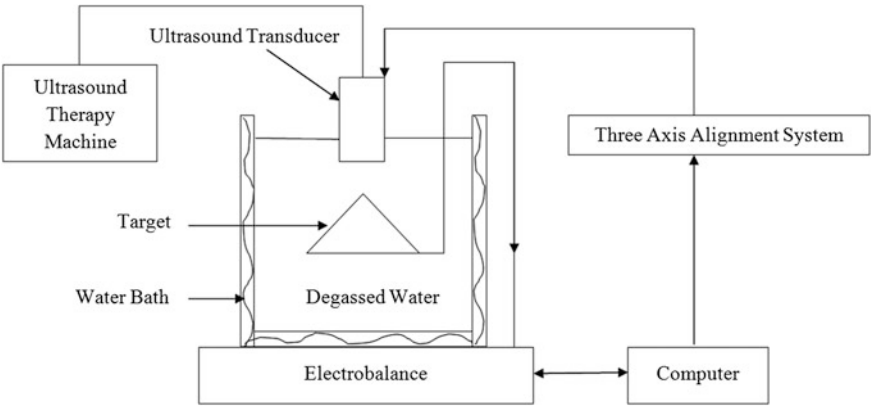


Fig. 1 Schematic diagram of the developed ultrasound power meter

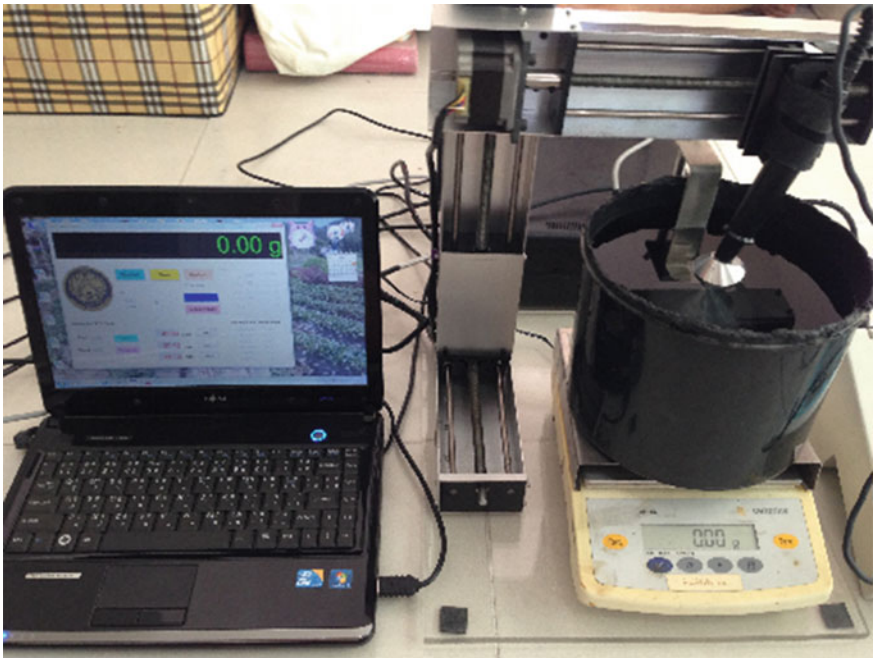


Fig. 2 Photograph of the developed ultrasound power meter with the three axis alignment system

The developed ultrasound power meter consists of the following parts:

1. Ultrasound therapy machine: Ultrasound therapy unit and its transducer, model Ultrasonic S3004 from Diter-elektroniikka, Finland, were used as an ultrasonic source to test the performance of the developed ultrasound power meter. The operating frequencies of this machine are 0.86, 2 and 3 MHz. In addition, the ultrasound therapy unit can be adjusted to five intensity levels (0.5, 1, 1.5, 2, 3 W/cm^2) and four different output pulse modes; continuous wave (100 % duty cycle), 1:2 (50 % duty cycle), 1:5 (20 % duty cycle) and 1:10 (10 % duty cycle). The effective radiating area (ERA) of the face of the transducer is approximately 4.5 $\pm$ 0.5 cm^2. The ultrasonic transducer was connected directly to the ultrasound therapy unit and placed vertically upwards on the target with a transducer holder of the three axis alignment system.
2. Water bath: A water bath was built from plastic and sealed with sound clad (Dinitrol 448, EFTEC Aftermarket GmbH) on the inner surface of the water bath in order to minimize ultrasonic reflections from the surface of the water bath. Its diameter and height are 200 and 155 mm, respectively. The water bath was then filled with degassed water to avoid cavitation. It is also good to note that air bubbles must not be presented on the faces of the ultrasonic transducer or the target during the measurement since it may cause a measurement error.
3. Target: A custom-made reflecting target was made of thin aluminum in an air-backed convex cone shape. It has a diameter of 80 mm. The cone half-angle of this conical reflector was designed to be 45°, so that the reflected waves could leave at right angles to the ultrasound beam axis. The target is directly connected to the electrobalance.
4. Electrobalance: The radiation force was measured by a precision electrobalance model GE2102 (Sartorius, Germany), with 0.01 g of readability and 2,100 g maximum load capacity. The weight measured from the electrobalance was then transferred to a computer via a serial port.
5. Three axis alignment system: The placement of the ultrasonic transducer was controlled by three axis stepper motors. The precision of the XYZ stepper motors of the alignment system is 1 mm per step, which allows the displacement from one position to another position of the transducer very accurately. In this work, the ultrasonic transducer was positioned about 1 cm away above the reflecting target in the degassed water.
6. Data acquisition and control system: The measurement sequence, such as aligning the ultrasonic transducer, obtaining weight data of the target from the electrobalance, and calculating the total output power from the transducer under test, was performed by a custom-made Visual Studio program presented in Fig. 3.

During the power measurement, the ultrasonic beam is directed vertically upwards on the target and the radiation force exerted by the ultrasonic beam will be measured by the electrobalance in gram units. The ultrasonic power (in watt units) can then be determined from the difference between the force measured with

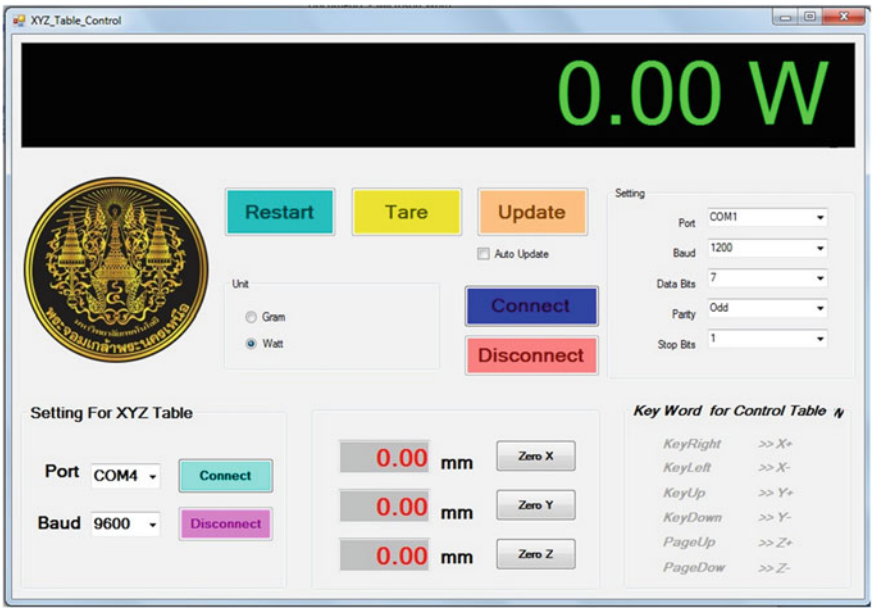


Fig. 3 Custom-made visual studio program used to determine the total output power from the ultrasonic transducer

and without ultrasonic radiation using the help of the theory in [21–25]. For plane waves, the relationship between the measured radiation force (F) and the ultrasonic power (P) can be expressed by the following equation:

$$P = \frac{c(t) \cdot F}{2\cos^2\theta} \tag{1}$$

where P is the ultrasonic power, F is the measured radiation force, $c(t)$ is the velocity of ultrasound waves in water as a function of the water temperature (t) and θ is the angle between the beam direction and the normal of the reflecting surface.

During this work, the measured radiation force (F) is obtained from the multiplication between the deviated weight (Δm) caused by the radiation force and the gravity (g). In addition, the angle between the beam direction and the normal of the reflecting surface is 45° since the cone angle of the target was designed to be 90°. Therefore, the Eq. 1 could be rewritten as Eq. 2.

$$P = \Delta m \cdot g \cdot c(t) \tag{2}$$

To verify the performance of our developed ultrasound power meter, the ultrasonic power measurement results from our developed system were compared with those from the commercial ultrasound power meter (UPM), Model UPM-DT-10 (Ohmic Instruments, Maryland, USA) and compared with those from the standard

ultrasonic power measurement system at the National Institute of Metrology, Thailand (NIMT). The correlation coefficients and measuring uncertainty were then calculated.

It is good to note that all power measurements with the developed ultrasound power meter, the commercial ultrasound power meter and the standard ultrasonic power measurement system at NIMT were repeated for six times by resetting the transducer, the water bath and the target completely to investigate the reproducibility of the measurement system.

3 Results

The comparisons of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the commercial ultrasound power meter (UPM) at five nominal intensity values (0.5, 1, 1.5, 2, 3 W/cm^2) with three different frequencies, 0.86, 2 and 3 MHz, using four different output pulse modes; continuous wave (100 % duty cycle), 1:2 (50 % duty cycle), 1:5 (20 % duty cycle) and 1:10 (10 % duty cycle) are presented together with their correlation coefficients in Figs. 4, 5, 6 and 7, respectively.

In addition, to enhance the verification of our developed ultrasound power meter's performance, the correlation coefficients of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the standard ultrasonic power measurement system at NIMT at five nominal intensity values (0.5, 1, 1.5, 2, 3 W/cm^2) with three different frequencies, 0.86, 2 and 3 MHz, using four different output pulse modes; continuous wave (100 % duty cycle), 1:2 (50 % duty cycle), 1:5 (20 % duty cycle) and 1:10 (10 % duty cycle) are determined and shown in Figs. 8, 9, 10 and 11, respectively.

4 Discussions

In this work, a measuring uncertainty of the developed ultrasound power meter was calculated and it was found to be within $\pm$10 %. To verify the performance of the developed system, the ultrasonic power measurement results from our developed ultrasound power meter were compared with those from the commercial ultrasound power meter (UPM) as shown in Figs. 4a, 5a, 6a and 7a. The difference between the results measured from these two systems is mostly smaller than the measuring uncertainty, so it can be deduced that these two systems of measuring ultrasonic power are in good agreement. However, the huge differences between the measurement results from these two systems at the intensity beyond 2 W/cm^2 using the continuous wave mode in Fig. 4a can be clearly observed and the reasons for these discrepancies are currently being investigated. In addition, the correlation coefficients of our developed system and the commercial UPM were found to be

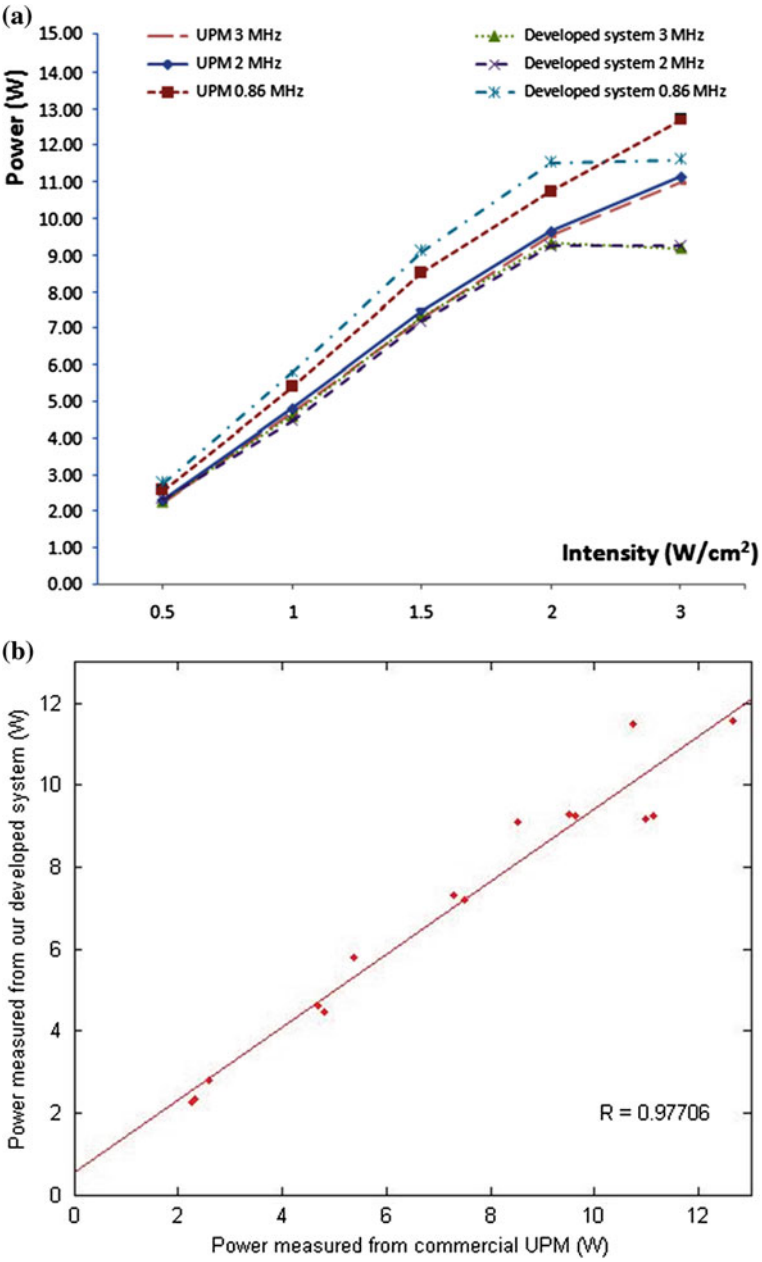


Fig. 4 **a** Comparison of the ultrasonic powers using continuous wave mode (100 % duty cycle) with three different frequencies, 0.86, 2 and 3 MHz, measured by the developed ultrasound power meter and the commercial ultrasound power meter (UPM), and **b** correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the commercial ultrasound power meter (UPM) with three different frequencies, 0.86, 2 and 3 MHz, using continuous wave mode (100 % duty cycle)

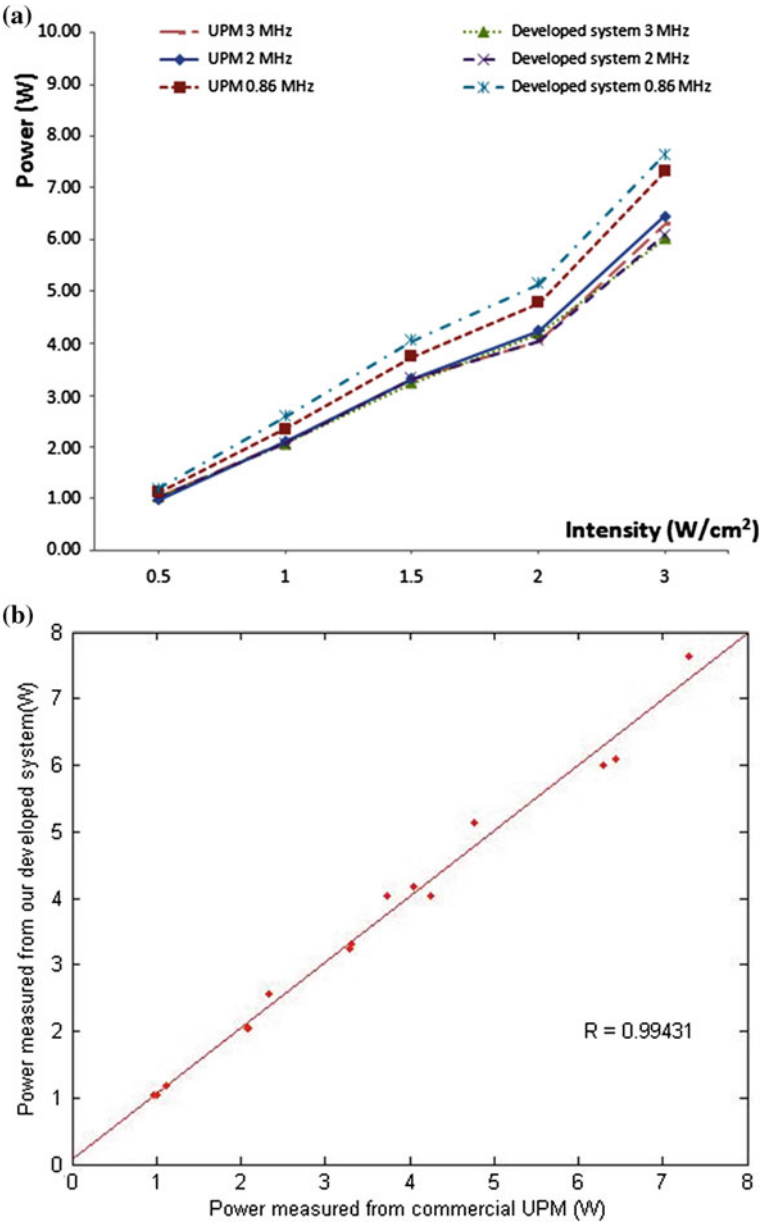


Fig. 5 **a** Comparison of the ultrasonic powers using 1:2 pulse mode (50 % duty cycle) with three different frequencies, 0.86, 2 and 3 MHz, measured by the developed ultrasound power meter and the commercial ultrasound power meter (UPM), and **b** correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the commercial ultrasound power meter (UPM) with three different frequencies, 0.86, 2 and 3 MHz, using 1:2 pulse mode (50 % duty cycle)

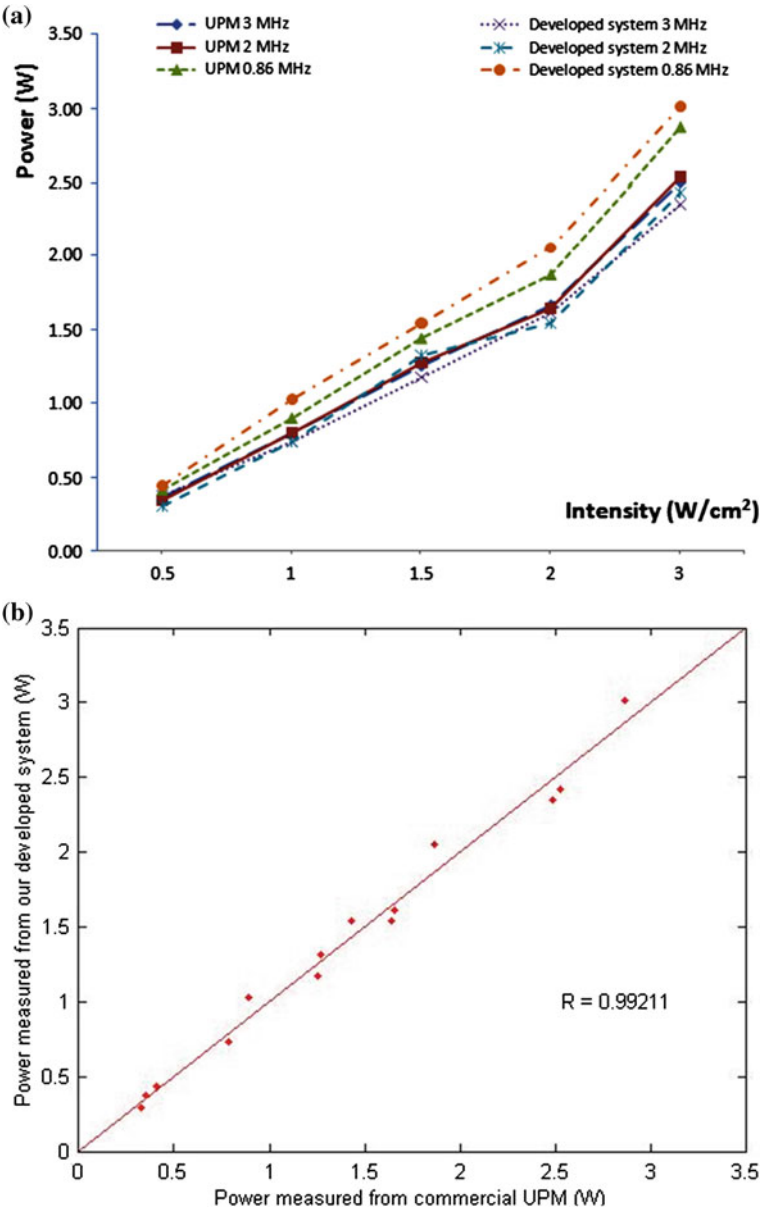


Fig. 6 **a** Comparison of the ultrasonic powers using 1:5 pulse mode (20 % duty cycle) with three different frequencies, 0.86, 2 and 3 MHz, measured by the developed ultrasound power meter and the commercial ultrasound power meter (UPM), and **b** correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the commercial ultrasound power meter (UPM) with three different frequencies, 0.86, 2 and 3 MHz, using 1:5 pulse mode (20 % duty cycle)

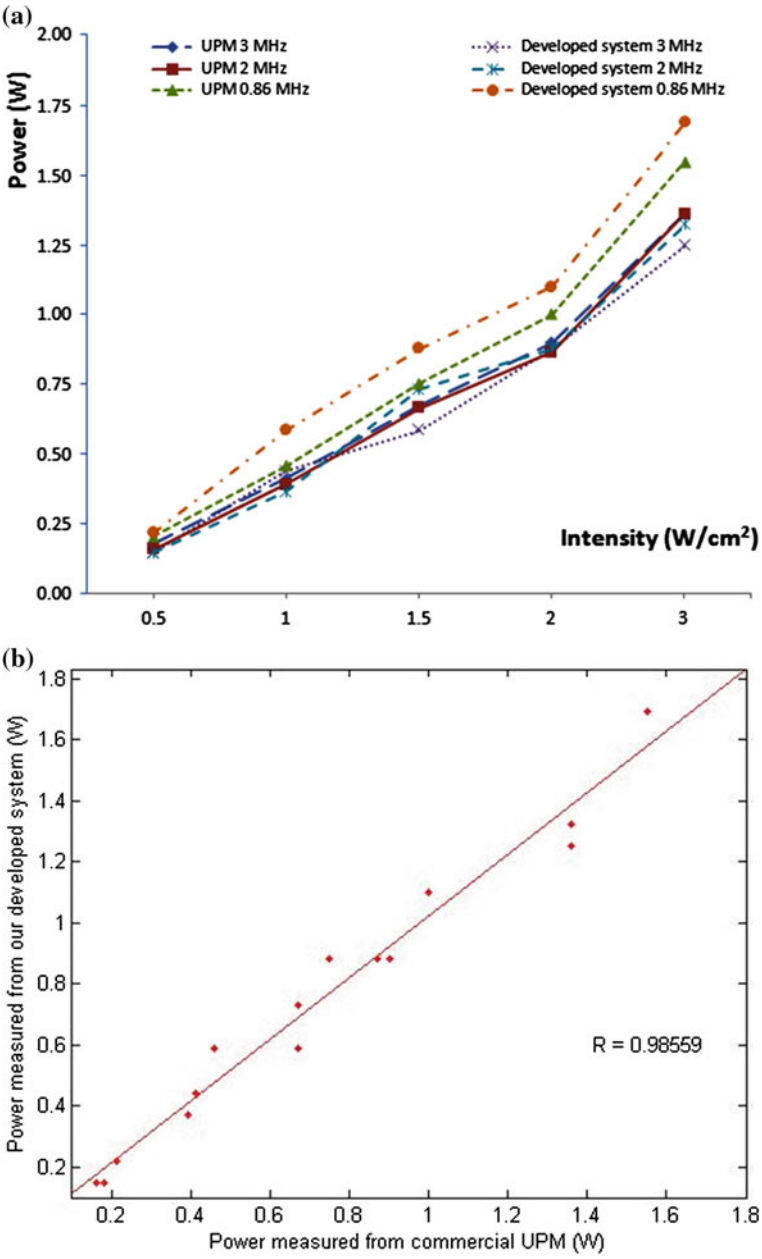


Fig. 7 **a** Comparison of the ultrasonic powers using 1:10 pulse mode (10 % duty cycle) with three different frequencies, 0.86, 2 and 3 MHz, measured by the developed ultrasound power meter and the commercial ultrasound power meter (UPM), and **b** correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the commercial ultrasound power meter (UPM) with three different frequencies, 0.86, 2 and 3 MHz, using 1:10 pulse mode (10 % duty cycle)

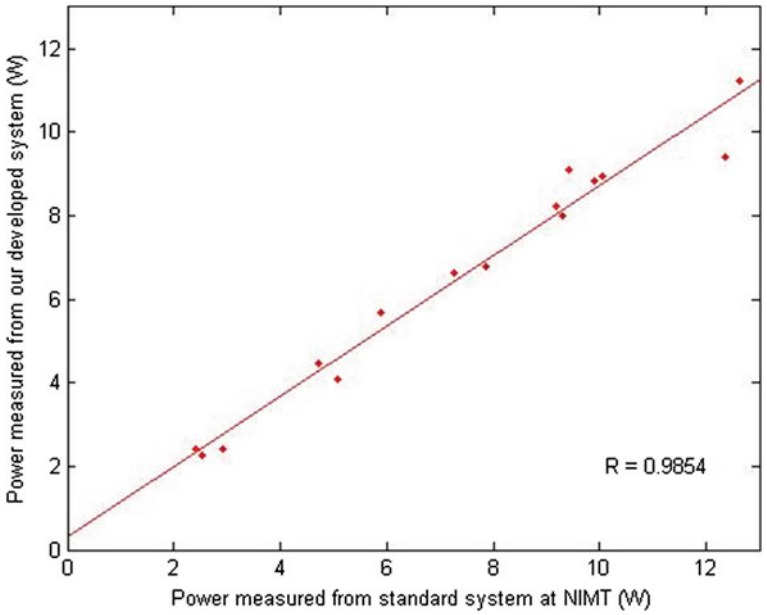


Fig. 8 The correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the standard ultrasonic power measurement system at NIMT with three different frequencies, 0.86, 2 and 3 MHz, using continuous wave mode (100 % duty cycle)

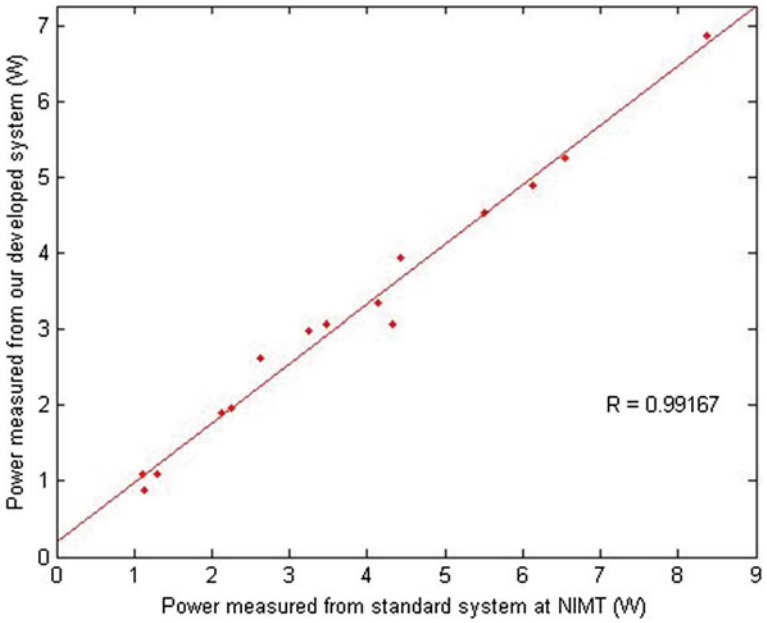


Fig. 9 The correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the standard ultrasonic power measurement system at NIMT with three different frequencies, 0.86, 2 and 3 MHz, using 1:2 pulse mode (50 % duty cycle)

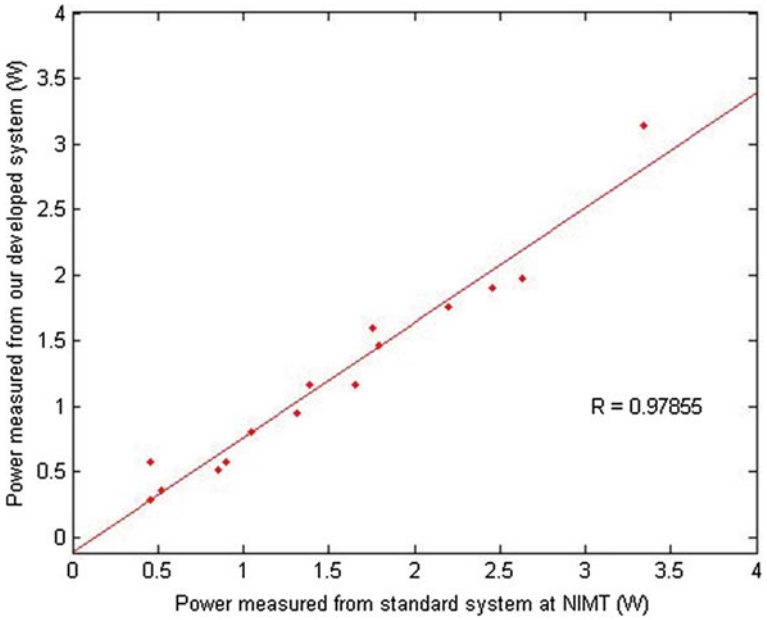


Fig. 10 The correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the standard ultrasonic power measurement system at NIMT with three different frequencies, 0.86, 2 and 3 MHz, using 1:5 pulse mode (20 % duty cycle)

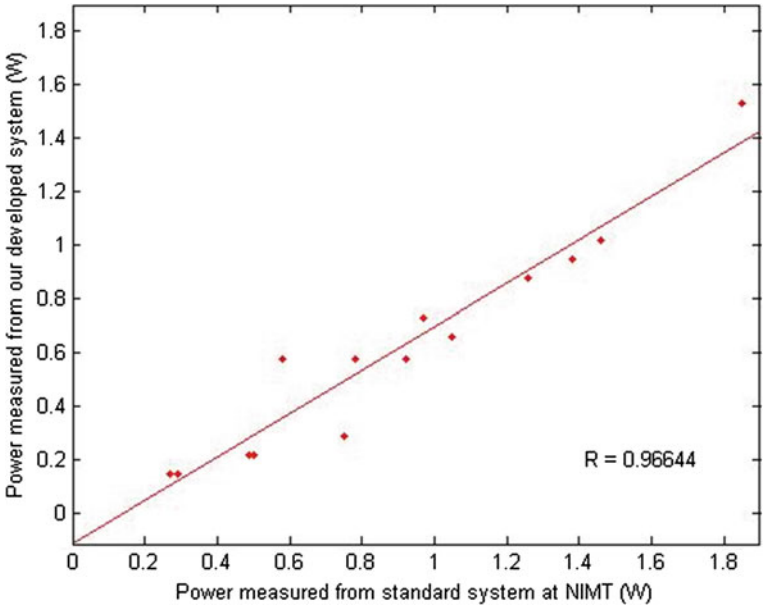


Fig. 11 The correlation coefficient of the ultrasonic powers measured from the developed ultrasound power meter and those measured from the standard ultrasonic power measurement system at NIMT with three different frequencies, 0.86, 2 and 3 MHz, using 1:10 pulse mode (10 % duty cycle)

0.97706 in Fig. 4b for the continuous wave mode, 0.99431 in Fig. 5b for the 1:2 pulse mode, 0.99211 in Fig. 6b for the 1:5 pulse mode and 0.98559 in Fig. 7b for the 1:10 pulse mode.

The measurement results from our developed ultrasound power meter were also compared with those from the standard ultrasonic power measurement system at the National Institute of Metrology, Thailand (NIMT). The correlation coefficients of our system and the standard system at NIMT were calculated to be 0.9854 in Fig. 8 for the continuous wave mode, 0.99167 in Fig. 9 for the 1:2 pulse mode, 0.97855 in Fig. 10 for the 1:5 pulse mode and 0.96644 in Fig. 11 for the 1:10 pulse mode. The results show that the values of the correlation coefficient are all higher than 0.96 (approximately 1) so it can be deduced that the developed system is in an excellent agreement with both the commercial ultrasound power meter (UPM) and the standard ultrasonic power measurement system at the National Institute of Metrology, Thailand (NIMT).

5 Conclusion and Future Work

In conclusion, the ultrasound power meter with a three axis alignment system was successfully developed. Currently, the system is able to determine the ultrasonic output power in the power range from 100 mW to approximately 12 W. This measurement range is normally suitable for the commercial medical ultrasound devices in therapeutic applications. Current efforts are being made to focus on testing the frequency range of the developed system.

Acknowledgment The authors would like to thank the financial support provided by the Coordinating Center for Thai Government Science and Technology Scholarship Students (CSTS), National Science and Technology Development Agency (NSTDA). In addition, we gratefully thank the Acoustics and Vibration Department, National Institute of Metrology, Thailand (NIMT) for the use of the standard ultrasonic power measurement system.

References

1. M. Pong, S. Umchid, A.J. Guarino, P.A. Lewin, J. Litniewski, A. Nowicki, S.P. Wrenn, In vitro ultrasound-mediated leakage from phospholipid vesicles. Ultrasonics **45**, 133–145 (2006)
2. T. Wu, J.P. Felmlee, J.F. Greenleaf, S.J. Riederer, R.L. Ehman, MR imaging of shear waves generated by focused ultrasound. Magn. Reson. Med. **43**, 111–115 (2000)
3. J. Bercoff, M. Tanter, and M. Fink, Supersonic shear imaging: a new technique for soft tissue elasticity mapping. IEEE Trans. Ultrason. Ferroelec. Freq. Control **51** (2004)
4. D. Lertsilp, S. Umchid, U. Techavipoo, and P. Thajchayapong, Improvements in Ultrasound Elastography using Dynamic Focusing, in *IEEE Biomedical Engineering International Conference (IEEE BMEiCON2011)*, Chiang Mai, Thailand (2011), pp. 225–228

5. D. Lertsilp, S. Umchid, U. Techavipoo, P. Thajchayapong, Resolution Improvements in Ultrasound Elastography using Dynamic Focusing, in *IEEE Biomedical Engineering International Conference (IEEE BMEiCON2012)*, Ubon Ratchathani, Thailand and Champasak, Laos (2012)
6. S. Umchid, T. Leeudomwong, Ultrasonic Hydrophone's Effective Aperture Measurements, in *IEEE International Conference on Biomedical Engineering and Biotechnology (IEEE iCBEB2012)*, Macau, China (2012), pp. 1136–1139
7. C. Patton, G.R. Harris, R.A. Philips, Output levels and bioeffects indices from diagnostic ultrasound exposure data reported to the FDA. IEEE Trans. Ultrason. Ferroelec. Freq. Contr **41**, 353–359 (1994)
8. K. Jaksukam, S. Umchid, Development of Ultrasonic Power Measurement Standards in Thailand, in *IEEE 10th International Conference on Electronic Measurement and Instruments*, Chengdu, China (2011), pp. 1–5
9. S. Umchid, K. Prasanpanich, in *Ultrasound Power Meter with a Three Axis Positioning System for Therapeutic Applications*. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE 2013, London, UK, 3–5 July 2013, pp. 1369–1373
10. S. Umchid, K. Prasanpanich, Development of the Ultrasound Power Meter for Therapeutic Applications, in *IEEE Biomedical Engineering International Conference (IEEE BMEiCON2012)*, Ubon Ratchathani, Thailand and Champasak, Laos (2012)
11. F. Davidson, Ultrasonic power balances, in *Output Measurements for Medical Ultrasound*, ed. by R. Preston (Springer, London, 1991), pp. 75–90
12. K. Beissner, Radiation force and force balances, in *Ultrasonic Exposimetry*, ed. by M.C. Ziskin, P.A. Lewin (CRC Press, Boca Raton, 1993), pp. 163–168
13. K. Beissner, Primary measurement of ultrasonic power and dissemination of ultrasonic power reference values by means of standard transducers. Metrologia **36**, 313–320 (1999)
14. K. Beissner, Summary of a European comparison of ultrasonic power measurements. Metrologia **36**, 313–320 (1999)
15. S. Umchid, K. Jaksukam, Development of the Primary Level Ultrasound Power Measurement System in Thailand, in *IEEE International Symposium on Biomedical Engineering (IEEE ISBME2009)*, Bangkok, Thailand (2009)
16. R.T. Hekkenberg, K. Beissner, B. Zeqiri, R. Bezemer, M. Hodnett, Validated ultrasonic power measurements up to 20 W. Ultrasound Med. Biol. **27** (2001)
17. R. Reibold, W. Molkenstruk, K.M. Swamy, Experimental study of the integrated optical effect of ultrasonic fields. Acustica **43**, 253–259 (1979)
18. B. Fay, M. Rinker, P.A. Lewin, Thermoacoustic sensor for ultrasound power measurements and ultrasonic equipment calibration. Ultras. Med. Biol **20**, 367–373 (1994)
19. M.A. Margulis, I.M. Margulis, Calorimetric method for measurement of acoustic power absorbed in a volume of a liquid. Ultrason. Sonochem. **10**, 343–345 (2003)
20. S. Lin, F. Zhang, Measurement of ultrasonic power and electro-acoustic efficiency of high power transducers. Ultrasonics **37**, 549–554 (2000)
21. K. Beissner, The acoustic radiation force in lossless fluids in Eulerian and Lagrangian coordinates. J. Acoust. Soc. Am. **103**, 2321–2332 (1998)
22. T. Kikuchi, S. Sato, Ultrasonic power measurements by radiation force balance method—characteristics of a conical absorbing target. Jpn. J. Appl. Phys. **30**, 3158–3159 (2000)
23. T. Kikuchi, S. Sato, M. Yoshioka, Ultrasonic power measurements by radiation force balance method—experimental results using burst waves and continuous waves. Jpn. J. Appl. Phys. **41**, 3279–3280 (2002)
24. T. Kikuchi, S. Sato, M. Yoshioka, Quantitative Estimation of Acoustic Streaming Effects on Ultrasonic Power Measurement, in *IEEE International Ultrasonics, Ferroelectrics, and Frequency Control Joint 50th Anniversary Conference* Montréal, Canada (2004), pp. 2197–2200

25. T. Kikuchi, M. Yoshioka, S. Sato, Ultrasonic Power Measurement System Using a Radiation Force Balance Method at AIST, in *18th International Congress on Acoustics (ICA2004)* (2004)
26. IEC Standard 61161, Ultrasonics—Power Measurement—Radiation Force Balances and Performance Requirements, in *International Electrotechnical Commission*, Geneva (2006)

Mass Transfer Properties for the Drying of Pears

Raquel Pinho Ferreira de Guiné and Maria João Barroca

Abstract Portugal is among the tropical countries that have a long tradition of drying fruits, such as figs, grapes or pears. The knowledge of the transfer phenomena involved in drying processes is of major importance for the design and optimization of the most adequate operating conditions. In this way, the present work intended to study the mass transfer properties for the convective air drying of pears, based on the diffusion model. The values of the diffusivity and mass transfer coefficient for the drying of pears of the Portuguese native variety D. Joaquina were determined for two drying temperatures, 60 and 70 °C. The results obtained showed an increase of 38 % in diffusivity and an even more pronounced increase in the mass transfer coefficient, 56 %, for a temperature variation of 10 °C. Regarding the dimensionless numbers, Biot number increased 13 % while Dincer number decreased 28 %.

Keywords Biot number · Diffusivity · Dincer number · Drying · Mass transfer coefficient · Mass transfer properties · Pear

1 Introduction

Pear species belong to the genus Pyrus in the subfamily *Maloideae* of the Rosaceae, and the commercial production lies essentially in the species *P. communis* L. and *P. pyrifolia* Burm. The *P. communis* is known as European pear, and is the

R. P. F. de Guiné (✉)
CI&DETS, Instituto Politécnico de Viseu, ESAV, Quinta da Alagoa, Ranhados, 3500-606 Viseu, Portugal
e-mail: raquelguine@esav.ipv.pt

M. J. Barroca
CERNAS-ESAV-IPC/Departamento de Engenharia Química e Biológica, ISEC-IPC, Rua Pedro Nunes, Quinta da Nora 3030-199 Coimbra, Portugal
e-mail: mjbarroca@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_20,

most commonly cultivated in Europe, North America, Northern Africa and temperate regions of the Southern hemisphere [1].

Due to the nutritive properties, pleasant taste and low caloric level, the pears are a much appreciated fruit by the consumers worldwide. Besides, pears are characterized by a high digestibility. They have a low content of protein and lipids and are particularly rich in sugars such as fructose, sorbitol, sucrose, and, in lower amount, glucose [2].

They also possess others nutritional components such as vitamins, minerals and antioxidants as well as bioactive elements, that are important sources of healthy-beneficial compounds [3]. In particular, phenolic compounds attract considerable interest in several fields, like chemistry, food and medicine, greatly due to their promising antioxidant properties [4]. Furthermore, pears are a potential source of dietary fibres, also reported to have important benefits for human health, particularly regarding the pathologies of the intestinal tract and as a preventing agent against some types of cancer [5]. In the particular case of four small varieties of Portuguese pears, the values of dietary fibre ranged between 12 and 15 % (dry mass) [2].

The analysis of the compositions of pears in relation to sugars, organic and fatty acids, amino acids, phenolics, vitamins, volatiles and minerals in different pear cultivars has been reported, being both the nature and concentration of such constituents responsible for their organoleptic characteristics [6]. The flavor of the pears is influenced by the volatile aromatic compounds present and by the contents in sugars and organic acids (mainly citric and malic acids). Their bitter taste is normally associated with the phenolic and polyphenolic compounds. The colour of the peel depends on the amount and type of pigments present, being mainly chlorophyll (green) and carotenoid (yellow) [7].

Pear production represents an important economic activity in Portugal, with an annual production reaching 200 thousand tonnes, being 95 % corresponding to the production of the cultivar Rocha, an exclusive Portuguese variety [8]. However, other cultivars, also native from Portugal, have a local importance [2, 9].

Drying of foods is an important method of preservation and can be applied to a wide range of products. Drying represents a very important way of preserving foods, because the removal of a considerable part of the moisture improves its conservation through the inhibition of microbial growth and enzymatic modifications. Still, the advantages of drying surpass the preservation capacity of the dried products [10].

In countries where the climatic conditions allow it, foods are often dried by open-air sun exposure. However, and despite being a cheap method, it has some important disadvantages, like the dependency on the weather conditions and the problems that may arise from contaminations, infestations and microbial attacks. Additionally, the drying times can be quite long if the quantities to be dried are relatively high [11, 12]. Therefore, in present days, many alternative drying methods have been in use, being undoubtedly air drying the most common drying method employed for foodstuffs [2].

Drying stabilizes food products by decreasing their water activity and moisture content. However, because drying, and particularly air drying, usually implies an exposure to a high temperature for a certain period of time, some of the properties may be affected, both at the physical and chemical levels [13, 14]. In particular, colour and flavour are much sensitive to thermal treatment and influenced by oxidation and the high degree of water loss due to evaporation. Also the polyphenols are sensitive to high temperatures, thus decreasing their amounts in the dried products. Thus, the application of heat treatments can lead to an important reduction both on the phenolic content and antioxidant capacity [15].

The design of driers is often done empirically, by extrapolation of knowledge existing for other cases. For reliable process modelling is very important a profound knowledge of the physical and chemical behaviour of the food, as well as its drying kinetics, which accounts for the mechanisms of water removal [16, 17].

From the engineering point of view it is very important to understand the complex processes that occur during drying, being this achieved through modelling. Many mathematical models have been used to describe drying processes, being quite common the use of the diffusion laws. During drying many changes take place inside the foods [18], and these modifications affect the product mass transfer properties such as the mass diffusion and mass transfer coefficients.

The present work aimed at determining the mass transfer properties of pears of the variety D. Joaquina for hot air drying performed in a convective drier.

2 Experimental

2.1 Materials

Pears of a Portuguese variety, D. Joaquina, which were bought in a local market were used in the present study. The pears of this variety are very sweet and quite small (about 4 to 5 cm diameter maximum) and exhibit good drying features [16].

2.2 Methods

The pears were dried in a drying chamber with ventilation (WTB-Binder) at constant temperature, but with trials done at different temperatures (60 and 70 °C). The air flow was 300 m^3/h, corresponding to an air velocity of 0.5 m/s. The drying times were 225 and 360 min respectively for 70 and 60 °C.

Periodically the samples were removed in order to measure their average water content with a Halogen Moisture Analyzer, model HG53 from Mettler Toledo, which was previously calibrated in terms of optimal operating parameters for this type of food. The parameters used were drying temperature set to 125 °C and speed 3 (in a scale from 1 = very fast to 5 = very slow).

3 Mathematical Modelling

For the mass transfer in one direction in non-steady state the moisture diffusion in the pears, assuming that they can be approximated to spheres, can be expressed by the Fick's second law [18]:

$$\frac{\partial W}{\partial t} = \frac{1}{r}\left\{\frac{\partial}{\partial r}\left(D_e r \frac{\partial W}{\partial r}\right)\right\}, \tag{1}$$

where W(r, t) is the dry basis moisture content in kg water/kg dry solids, t is time in seconds, D_e is effective diffusivity in m^2/s and r is the sphere radius in meters.

Assuming that the initial moisture content is uniform in the whole pear and that the sample presents central symmetry, the initial and boundary conditions are given by the following equations:

$$\text{for t} = 0: \quad W(r, 0) = W_0, \tag{2}$$

$$\text{for r} = 0: \quad \frac{\partial W(0, t)}{\partial r} = 0, \tag{3}$$

$$\text{for r} = \text{R}: \quad -D_e \frac{\partial W(R, t)}{\partial r} = h_m[W - W_a], \tag{4}$$

where W_0 is the initial product moisture content, W_a is the surrounding air moisture content (all dry basis) and h_m is the convective mass transfer coefficient.

In transient conditions, the solution of Fick's Law can be approximated by an infinite series, of the form [18, 19]:

$$MR = \frac{W - W_e}{W_0 - W_e} = \sum_{n=1}^{\infty} \frac{6}{\pi^2} \exp\left[-D_e \frac{\pi^2 t}{r^2}\right], \tag{5}$$

where MR is the moisture ratio, dimensionless, and W, We and W_0 are, respectively, the moisture content at time t, the equilibrium moisture content and the initial moisture content, all expressed in dry basis (g water/g dry solids). Taking into account that the terms of the series after the first are relatively insignificant and can therefore be ignored, then the solution of the Fick's equation is given by:

$$MR = \left(\frac{6}{\pi^2}\right) \exp\left[-D_e t \left(\frac{\pi^2}{r^2}\right)\right], \tag{6}$$

The thin-layer models assume that the moisture evolution along drying is related to some parameters, such as the drying constant, k (1/s), or the lag factor, k_0 (dimensionless), that account for combined effects of various transport phenomena during drying [20]. The Henderson and Pabis model is an example of such models, and is expressed through the following equation [21]:

$$MR = k_0 \exp(-kt). \tag{7}$$

The convective mass transfer coefficient, h_m (m/s), and the diffusivity coefficient, or effective diffusivity, are correlated by the dimensionless Biot number for mass transfer [22]:

$$Bi_m = \frac{h_m r}{D_e}, \tag{8}$$

where r is the sphere diameter (m). Equation (8) is valid for Bi greater than 0.1 [23], and allows the estimation of h_m, if the Bi_m is known.

Dincer and Hussain [23] report the equation that correlates the Biot number with the dimensionless Dincer Number:

$$Bi_m = \frac{24.848}{Di^{0.375}}, \tag{9}$$

with,

$$Di = \frac{u}{kr}, \tag{10}$$

where u is the flow velocity of drying air (m/s), k the drying constant and r the radius.

The determination of the mass transfer properties of the pears was done following the steps:

1. Estimate MR from the experimental drying data for every time t;
2. From a plot ln(MR) = f(t) estimate D_e from the slope through (6) (slope = $-D_e\ \pi^2/r^2$);
3. Estimate k and k_0 by combining (6) and (7);
4. Calculate Di, Bi_m and h_m from (10) (9) and (8), respectively.

4 Results and Discussion

Figure 1 shows how the moisture content of the D. Joaquina pears, expressed as percentage wet basis, varied along drying for the two temperatures tested. At each measurement interval several measurements were made and some variability can be observed for the values of moisture determined for different samples, all taken from the drier at the same time. This is natural, since some variations in the sample volume or even properties, may influence the removal of the water from the food, therefore originating different values for moisture content [24]. Furthermore, the pears before drying also presented some variability in their moisture content, which can be attributed to different slight difference in the ripening state of the pears. The pears were dehydrated to a very high extent, in order to obtain a crispy snack.

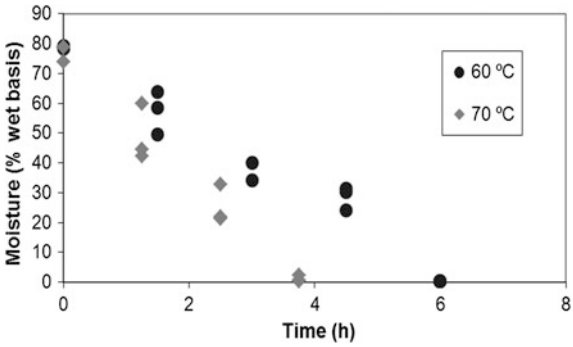


Fig. 1 Variations of moisture content along drying for different drying temperatures

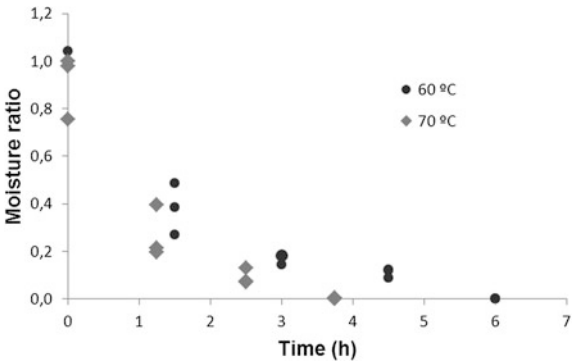


Fig. 2 Decrease in moisture ratio for the temperatures tested

The graph in Fig. 2 reveals the fast diminish in the values of the moisture ratio, as defined in Eq. (5), starting with 1 initially and tending to approach zero as the drying proceeded.

Figure 3 shows the linearization of the functions ln(MR) = f(t) for the two temperatures tested, being the results corresponding to the equations obtained presented in Table 1. In the two cases both the slope and intercept are slightly different, and it is also visible that the fitting was not so good. The values of R^2 found were 0.7782 for the drying at 60 °C and 0.8321 for the drying at 70 °C.

The values of the drying constant and lag factor were calculated from Eq. (7). The values obtained for the drying constant were 3.3970×10^{-4} and 4.6920×10^{-4} s^{-1}, respectively for 60 and 70 °C. These values are higher than those reported by Roberts et al. [25] for the convective hot air drying of grape seeds in the range of temperatures from 40 to 60 °C and with air velocities above 1.5 m/s. This is expected, having in consideration that the temperatures used in the present study are also higher, and the products are different. Furthermore, the values found by Guiné et al. [26] for the solar drying of pears of the variety S. Bartolomeu, ranging between 5.7188×10^{-6} and 1.1037×10^{-5} s^{-1}, are also inferior to those found in the present work, because the drying conditions were

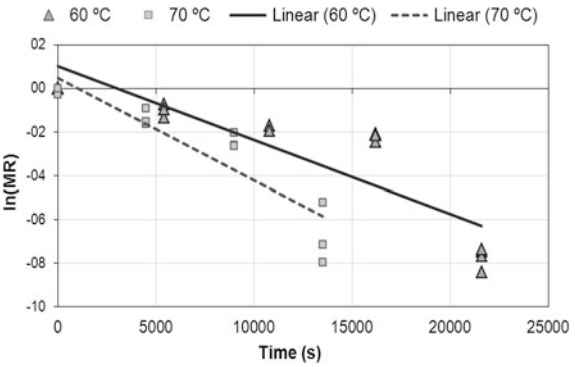


Fig. 3 Linearization of the functions ln(MR) = f(t) for the different temperatures

Table 1 Parameters for the function ln(MR) = f(t) for the two temperatures tested

Drying temperature (°C)	Slope	Intercept	R^2
60	-3.397×10^{-4}	1.038	0.7782
0	-4.692×10^{-4}	0.496	0.8321

variable. Nevertheless, the values found for the two temperatures clearly indicate an increase in the drying constant as the temperature raised from 60 to 70 °C.

As to the values fund for the lag factor, they were 2.8236 and 1.6418 respectively for the temperatures 60 and 70 °C. Dincer and Hussain [22], for the drying of potatoes at 40 °C, with an air velocity of 1 m/s and a characteristic dimension of 0.09 m, report a value for the lag factor of 1.0074, which is just slightly lower than those in the present study. The values found by Guiné et al. [26] for the S. Bartolomeu pears varied from 0.6658 to 2.2538, depending on the drying system, and stand in the same range as those found for the D. Joaquina pears.

Table 2 shows the values obtained using the methodology explained previously for the different mass transfer properties based on the experimental data obtained for the two temperatures. The values of the effective diffusion coefficient or diffusivity, D_e, are 8.6047×10^{-10} and 11.8850×10^{-10} m^2/s, for the temperatures 60 and 70 °C, respectively. These stand in the same range of the values found by Guiné et al. [26], 1.4218×10^{-10}–2.7439×10^{-9} m^2/s, and they are also of the same magnitude of that reported by Dincer and Hussain [22] for the air drying of cylindrical okara, 5.6752×10^{-10} m^2/s. However, the values found for the D. Joaquina pears are inferior to those reported by Dincer and Hussain [22] for the air drying of spherical potatoes, 9.4259×10^{-7} m^2/s, or those of Tripathy and Kumar [20], which vary from 3.28×10^{-8} to 6.09×10^{-8} m^2/s, for cylindrical potato samples for temperatures in the interval 33.74–47.70 °C, or stand in the range 2.43×10^{-8}–4.18×10^{-8} m^2/s for sliced potato samples at temperatures between 35.55 and 49.88 °C.

The values of the Di number in Table 2 decrease from 116,820 at 60 °C to 84,575 at 70 °C. These values are higher than the value presented by Dincer and

Table 2 Mass transfer properties of pears calculated for the two temperatures studied

Drying temperature (°C)	D_e (m^2/s)	Di	Bi	h_m (m/s)
60	8.6047×10^{-10}	116,820	0.3126	5.3795×10^{-8}
70	11.8850×10^{-10}	84,575	0.3528	8.3870×10^{-8}

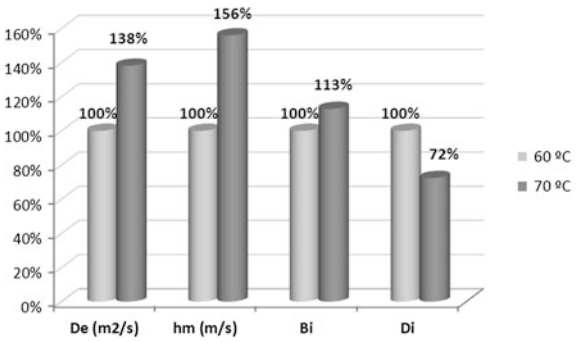


Fig. 4 Variation of the thermo physical properties of pears with drying temperature

Hussain [22], which was 12,356 for potatoes of spherical geometry dried with air at 1 m/s and 40 °C. On the contrary the values in the present work are considerably lower than those found by Guiné et al. [26], varying from 371,350 to 2,317,400 for the drying of S. Bartolomeu pears in different systems.

The Biot numbers are very similar for both temperatures, being 0.3126 for 60 °C and 0.3528 for 70 °C. Both values are higher than 0.1, thus allowing the use of Eq. (8) for the estimation of the mass transfer coefficients. These values are higher when compared to those reported by Guiné et al. [26] ranging between 0.1020 and 0.2026, according to the drying system. However, for the convective drying of spherical potatoes at 40 °C Dincer and Hussain [22] reported a value of 0.3119, which is very similar to those in this work.

The convective mass transfer coefficients, h_m, varied from 5.3795×10^{-8} to 8.3870×10^{-8} m/s, respectively for 60 and 70 °C. These values stand in the same range of the values encountered by Guiné et al. [26] for pears of a different variety and submitted to varied drying systems, 3.5040×10^{-9}–1.1222×10^{-8} m/s, and they are also similar to that for the convective drying of cylindrical okra (1.6098×10^{-8} m/s) at 80 °C, as reported by Dincer and Hussain [22]. On the other hand, the present values are considerably smaller than those found by Dincer and Hussain [22] for the convective drying of spherical potatoes (3.2665×10^{-5} m/s). Tripathy and Kumar [20] determined the convective mass transfer coefficients for potato elements in cylindrical and sliced shapes, and their results lead to h_m ranging from 1.61×10^{-7} to 4.17×10^{-7} m/s in the range of temperatures from 33.74 to 47.70 °C for the cylindrical shape and from 1.70×10^{-7} to 3.21×10^{-7} m/s for temperatures between 35.55 and 49.88 °C for the sliced shape.

Figure 4 shows the effect of the temperature raise on the mass transfer properties of the D. Joaquina pears. Regarding the diffusivity, D_e, the increase of 10 °C originated a raise of 38 % in the effective diffusion coefficient, corresponding to

about 14 m^2/s per °C. The mass transfer coefficient increased very pronouncedly, 56 % for the 10 °C raise. As to the dimensionless numbers, the Biot number increased 13 % while the Dincer number decreased 28 %.

5 Conclusion and Future Work

The values of the diffusion and mass transfer coefficients for the drying of pears of the variety D. Joaquina were estimated in this work for two drying temperatures, 60 and 70 °C. The results obtained enabled to conclude that the raise in temperature originated an important increase in the value of the diffusivity, demonstrating the effect of temperature on the efficiency of the internal mass transfer. Also the mass transfer coefficient suffered a very important increase with temperature, owing to a higher efficiency of the moisture transfer at the surface of the pears.

For future work is planned to extend the present study to more varieties of pears, so as to compare the results, and to study a more wide range of temperatures, from 40 to 90 °C, in order to confirm the effect of temperature over these thermo physical properties.

References

1. W. Brini, M. Mars, J.I. Hormaza, Genetic diversity in local Tunisian pears (*Pyrus communis* L.) studied with SSR markers. Sci. Hortic. **115**(4), 337–341 (2008)
2. M.J. Barroca, R.P.F. Guiné, A. Pinto, F. Gonçalves, D.M.S. Ferreira, Chemical and microbiological characterization of Portuguese varieties of pears. Food Bioprod. Process. **84**(2), 109–113 (2006)
3. H. Silos-Espino, L. Fabian-Morales, J.A. Osuna-Castro, E. Valverde, F. Guevara-Lara, O. Paredes-López, Chemical and biochemical changes in prickly pears with different ripening behavior. Nahrung **47**(5), 334–338 (2003)
4. Z. Cheng, Y. Liand, W. Chang, Kinetic deoxyribose degradation assay and its application in assessing the antioxidant activities of phenolic compounds in a Fenton-type reaction system. Anal. Chim. Acta **478**(1), 129–137 (2003)
5. C.A.C. Martinho, A.C. Correia, F.M. Gonçalves, J.L. Abrantes, R. Carvalho, R.P.F. Guiné, Study about the knowledge and attitudes of the Portuguese population about food fibres. Curr. Nutr. Food Sci. **9**(3), 180–188 (2013)
6. J. Chen, Z. Wang, J. Wu, Q. Wang, X. Hu, Chemical compositional characterization of eight pear cultivars grown in China. Food Chem. **104**(1), 268–275 (2007)
7. K.J. Park, A. Bin, F.P.R. Brod, T.H.K.B. Park, Osmotic dehydration kinetics of pear D'Anjou (*Pyrus communis* L.). J Food Eng. **52**(3), 293–298 (2002)
8. J. Salta, A. Martins, R.G. Santos, N.R. Neng, J.M.F. Nogueira, J. Justino, A.P. Rauter, Phenolic composition and antioxidant activity of Rocha pear and other pear cultivars—a comparative study. J. Funct. Foods **2**(2), 153–157 (2010)
9. R.P.F. Guiné, Drying kinetics of some varieties of pears produced in Portugal. Food Bioprod. Process. **83**(4), 273–276 (2005)
10. R.P.F. Guiné, Pear drying: experimental validation of a mathematical prediction model. Food Bioprod. Process. **86**(4), 248–253 (2008)

11. R.P.F. Guiné, J.A.A.M. Castro, Pear drying process analysis: drying rates and evolution of water and sugar concentrations in space and time. Drying Technol. **20**(7), 1515–1526 (2002)
12. I.T. Togrul, D. Pehlivan, Modelling of drying kinetics of single apricot. J Food Eng. **58**(1), 23–32 (2003)
13. M.A. Coimbra, C. Nunes, P.R. Cunha, R. Guiné, Amino acid profile and Maillard compounds of sun-dried pears. Relation with the reddish brown colour of the dried fruits. Eur. Food Res. Technol. **233**(4), 637–646 (2011)
14. R.P.F. Guiné, Influence of drying method on some physical and chemical properties of pears. Int. J. Fruit Sci. **11**(3), 245–255 (2011)
15. M.H. Ahmad-Qasem, J. Cánovas, E. Barrajón-Catalán, V. Micol, J.A. Cárcel, J.V. García-Pérez, Kinetic and compositional study of phenolic extraction from olive leaves (var. Serrana) by using power ultrasound. Innov. Food Sci. Emerg. Technol. **17**(1), 120–129 (2013)
16. R.P.F. Guiné, D.M.S. Ferreira, M.J. Barroca, F.M. Gonçalves, Study of the drying kinetics of solar-dried pears. Biosyst. Eng. **98**(4), 422–429 (2007)
17. K. Sacilik, Effect of drying methods on thin-layer drying characteristics of hull-less seed pumpkin (*Cucurbita pepo* L.). J. Food Eng. **79**(1), 23–30 (2007)
18. J. Crank, *The Mathematics of Diffusion*, 2nd edn. (Oxford University Press, London, 1975)
19. N.P. Zogzas, Z.B. Maroulis, D. Marinos-Kouris, Moisture deffusivity methods of experimental determination. A review. Drying Technol. **12**(3), 483–515 (1994)
20. P.P Tripathy, S. Kumar, A methodology for determination of temperature dependent mass transfer coefficients from drying kinetics: application to solar drying. J. Food Eng. **90**(2), 212–218 (2009)
21. S. Lahsasni, M. Kouhila, M. Mahrouz, J.T. Jaouhari, Drying kinetics of prickly pear fruit (*Opuntia ficus indica*). J. Food Eng. **61**(2), 173–179 (2004)
22. I. Dincer, M.M. Hussain, Development of a new Bi–Di correlation for solids drying. Int. J. Heat Mass Transf. **45**(15), 3065–3069 (2002)
23. A.Z. Sahin, I. Dincer, B.S. Yilbas, M.M. Hussain, Determination of drying times for regular multi-dimensional objects. Int. J. Heat Mass Transf. **45**(8), 1757–1766 (2002)
24. R.P.F. Guiné, M.J. Barroca, in *Estimation of the Diffusivities and Mass Transfer Coefficients for the Drying of D. Joaquina Pears*. Lecture Notes in Engineering and Computer Science. Proceedings of the World Congress on Engineering 2013, WCE 2013, London, UK, 3–5 July 2013, pp. 132–1323
25. J.S. Roberts, D.R. Kidd, O. Padilla-Zakour, Drying kinetics of grape seeds. J. Food Eng. **89**(8), 460–465 (2008)
26. R.P.F. Guiné, M.J. Barroca, P. Lopes, V. Silva, Mass transfer properties of pears for different drying methods. Int. J. Food Prop. **16**(2), 251–262 (2013)

The Application of Negative Hamaker Concept to the Human Immunodeficiency Virus (HIV)-Blood Interactions Mechanism

C. H. Achebe and S. N. Omenyi

Abstract HIV-blood interactions were studied using the Hamaker coefficient approach as a thermodynamic tool in determining the interaction processes. Application was made of the Lifshitz derivation for van der Waals forces as an alternative to the contact angle approach. The methodology involved taking blood samples from twenty HIV-infected persons and from twenty uninfected persons for absorbance measurement using Ultraviolet Visible Spectrophotometer. From the absorbance data the variables (e.g. dielectric constant, etc.) required for computations were derived. The Hamaker constants $A_{11,}$ $A_{22,}$ A_{33} and the combined Hamaker coefficients A_{132} were obtained. The value of $A_{132abs} = 0.2587 \times 10^{-21}$ J was obtained for HIV-infected blood. A significance of this result is the positive sense of the absolute combined Hamaker coefficient which implies net positive van der Waals forces indicating an attraction between the virus and the lymphocyte. This in effect suggests that infection has occurred thus confirming the role of this principle in HIV-blood interactions. A near zero value for the combined Hamaker coefficient for the uninfected blood samples $A_{131abs} = 0.1026 \times 10^{-21}$ J is an indicator that a negative Hamaker coefficient is attainable. To propose a solution to HIV infection, it became necessary to find a way to render the absolute combined Hamaker coefficient A_{132abs} negative. As a first step to this, a mathematical derivation for $A_{33} \geq 0.9763 \times 10^{-21}$ J which satisfies this condition for a negative A_{132abs} was obtained. To achieve the condition of the stated A_{33} above with possible additive(s) in form of drugs to the serum as the intervening medium will be the next step.

Keywords Absorbance · Dielectric constant · Hamaker coefficient · Human immunodeficiency virus · Lifshitz formula · Lymphocyte · van der Waal

C. H. Achebe (✉) · S. N. Omenyi
Department of Mechanical Engineering, Nnamdi Azikiwe University,
PMB 5025 Awka, Nigeria
e-mail: chinobert2k@yahoo.com

S. N. Omenyi
e-mail: omenyinj@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_21,

1 Theoretical Considerations

1.1 Concept of Interfacial Free Energy

The work done by a force F to move a flat plate along another surface by a distance dx is given, for a reversible process, by (Fig. 1)

$$\delta w = Fdx \tag{1}$$

However, the force F is given by; $F = L\gamma$. Where L is the width of the plate and γ is the surface free energy (interfacial free energy)

Hence; $\delta w = L\gamma dx$

But; $dA = Ldx$

Therefore

$$\delta w = \gamma dA \tag{2}$$

This is the work required to form a new surface of area dA. For pure materials, γ is a function of T only, and the surface is considered a thermodynamic system for which the coordinates are γ, A and T. The unit of γ is Joules. In many processes that involve surface area changes, the concept of interfacial free energy is applicable.

1.2 The Thermodynamic Approach to Particle–Particle Interaction

The thermodynamic free energy of adhesion of a particle P on a solid S in a liquid L at a separation d_0 is given by [1];

$$\Delta F_{pls}^{adh}(d_o) = \gamma_{ps} - \gamma_{pl} - \gamma_{sl} \tag{3}$$

where ΔF^{adh} is the free energy of adhesion, integrated from infinity to the equilibrium separation distance d_o; γ_{ps} is the interfacial free energy between P and S; γ_{pl} is that between P and L and γ_{sl} that between S and L.

For the interaction between the individual components, similar equations can be written also;

$$\Delta F_{ps}^{adh}(d_1) = \gamma_{ps} - \gamma_{pv} - \gamma_{sv} \tag{4}$$

$$\Delta F_{sl}^{adh}(d_1) = \gamma_{sl} - \gamma_{sv} - \gamma_{lv} \tag{5}$$

$$\Delta F_{pl}^{adh}(d_1) = \gamma_{pl} - \gamma_{pv} - \gamma_{lv} \tag{6}$$

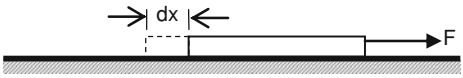


Fig. 1 Schematic diagram showing application of a force on a surface

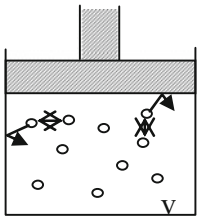


Fig. 2 Attraction of surface molecules by bulk molecules in a container of volume *V* [4]

For a liquid, the force of cohesion, which is the interaction with itself is described by;

$$\Delta F_{ll}^{coh}(d_1) = -2\gamma_{lv} \tag{7}$$

ΔF^{adh} can be determined by several approaches, apart from the above surface free energy approach. The classical work of Hamaker is very appropriate [2].

To throw more light on the concept of Hamaker Constants, use is made of the van der Waals explanation of the derivations of the ideal gas law;

$$PV = RT. \tag{8}$$

It was discovered that the kinetic energy of the molecules which strike the container wall is less than that of the bulk molecules. This effect was explained by the fact that the surface molecules are attracted by the bulk molecules (as in Fig. 2) even when the molecules have no permanent dipoles. It then follows that molecules can attract each other by some kind of cohesive force [3]. These forces have come to be known as van der Waals forces. van der Waals introduced the following corrections to Eq. (8);

$$\left[P + \frac{a}{V^2}\right](V - b) = RT. \tag{9}$$

The correction term to the pressure, $\left(\frac{a}{V^2}\right)$ indicates that the kinetic energy of the molecules which strike the container wall is less than that of the bulk molecules. This signifies the earlier mentioned attraction between the surface molecules and the bulk molecules.

After the development of the theory of quantum mechanics, London quantified the van der Waals statement for molecules without a dipole and so molecular attraction forces began to be known as London/van der Waals forces [5]. London stated that the mutual attraction energy, V_A of two molecules in a vacuum can be given by the equation;

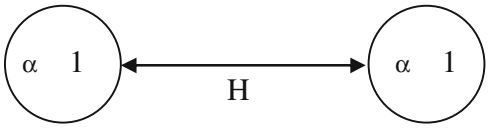


Fig. 3 Interaction of two identical molecules of materials *1* and polarizability α, at a separation *H* [4]

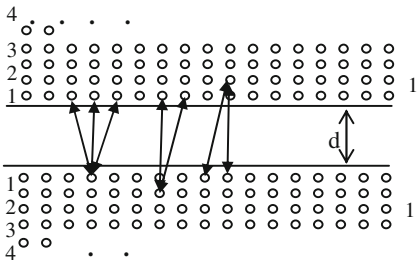


Fig. 4 Interaction of two semi-infinite solid bodies *1* at a separation *d* in vacuum [4]

$$V_A = -\frac{3}{4} h\upsilon_0 \left[\frac{\alpha^2}{H^6}\right] = -\left[\frac{\beta_{11}}{H^6}\right] \tag{10}$$

The interaction of two identical molecules of a material 1 is shown in Fig. 3.

Hamaker made an essential step in 1937 from the mutual attraction of two molecules. He deduced that assemblies of molecules as in a solid body must attract other assemblies. The interaction energy can be obtained by the summation of all the interaction energies of all molecules present as in Fig. 4.

This results in a van der Waals pressure, P_{vdw} of attraction between two semi infinite (solid) bodies at a separation distance, d in vacuum;

$$P_{vdw} = \left[\frac{A_{11}}{6\pi d^3}\right] \tag{11}$$

For a sphere of radius, R and a semi-infinite body at a minimum separation distance, d the van der Waals force, F_{vdw} of attraction is given by;

$$F_{vdw} = -\left[\frac{A_{11}R}{6d^2}\right] \tag{12}$$

where A_{11} is the Hamaker Constant (which is the non-geometrical contribution to the force of attraction, based on molecular properties only) (Fig. 5).

According to Hamaker, the constant A_{11} equals;

$$A_{11} = \pi^2 q_1^2 \beta_{11} \tag{13}$$

where q_1 is the number of atoms per cm^3 and β_{11} is the London/van der Waals constant for interaction between two molecules. Values for β can be obtained in approximation from the ionization potential of the molecules of interest, and so the Hamaker Constant can be calculated. The corresponding van der Waals force

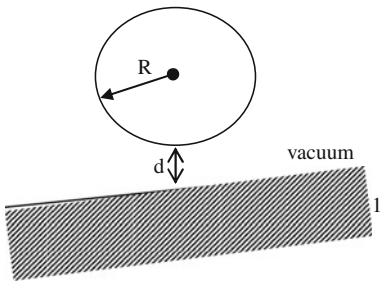


Fig. 5 Interaction of a sphere of radius *R* at a separation *d* from a solid surface of the same material *1* in *vacuum* [4]

between two condensed bodies of given geometry can be calculated provided their separation distance is known.

For combination of two different materials 1 and 2 in approximation:

$$B_{12} \approx \sqrt{\beta_{11}\beta_{22}} \tag{14}$$

Thus;

$$A_{12} = \sqrt{A_{11}A_{22}}. \tag{15}$$

For a combination of three materials when the gap between 1 and 2 is filled with a medium 3, from Hamaker's calculations;

$$A_{131} = A_{11} + A_{33} - 2A_{13} \tag{16}$$

Also;

$$A_{132} = A_{12} + A_{33} - A_{13} - A_{23} \tag{17}$$

Rewriting these equations will give;

$$A_{131} = \left(\sqrt{A_{11}} - \sqrt{A_{33}}\right)^2 \tag{18}$$

And;

$$A_{132} = \left(\sqrt{A_{11}} - \sqrt{A_{33}}\right)\left(\sqrt{A_{12}} - \sqrt{A_{33}}\right) \tag{19}$$

Equation (19) shows that, for a three-component system involving three different materials, 1, 2 and 3, A_{132} can become negative;

$$A_{132} \prec 0 \tag{20}$$

When;

$$\sqrt{A_{11}} \succ \sqrt{A_{33}} \quad and \quad \sqrt{A_{22}} \prec \sqrt{A_{33}} \tag{21}$$

Or;

$$\sqrt{A_{11}} \prec \sqrt{A_{33}} \prec \sqrt{A_{22}} \tag{22}$$

The limitations of Hamaker's approach led Lifshitz et al. to develop an alternative derivation of van der Waals forces between solid bodies [6]. The interaction between solids on the basis of their macroscopic properties considers the screening and other effects in their calculations. Thus the Hamaker Constant A_{132} becomes;

$$A_{132} = \frac{3}{4}\pi\hbar \int_0^\infty \left[\frac{\varepsilon_1(i\zeta) - \varepsilon_3(i\zeta)}{\varepsilon_1(i\zeta) + \varepsilon_3(i\zeta)}\right]\left[\frac{\varepsilon_2(i\zeta) - \varepsilon_3(i\zeta)}{\varepsilon_2(i\zeta) + \varepsilon_3(i\zeta)}\right] d\zeta \tag{23}$$

where, $\varepsilon_1(i\zeta)$ is the dielectric constant of material, j along the imaginary, i frequency axis ($i\zeta$) which can be obtained from the imaginary part $\varepsilon_1''(\omega)$ of the dielectric constant $\varepsilon_1(\omega)$.

The value of A_{11} could be obtained from the relation;

$$A_{11} = 2.5\left[\frac{\varepsilon_{10} - 1}{\varepsilon_{10} + 1}\right]^2 = 2.5\left[\frac{n_1^2 - 1}{n_1^2 - 1}\right]^2 \tag{24}$$

where ε_{10} is the dielectric constant and n_1 the refractive index of the polymer at zero frequency, both being bulk material properties which can easily be obtained. Relationship between Hamaker Coefficients and Free Energy of Adhesion ΔF^{adh}. For all given combinations, it is possible to express ΔF^{adh} in terms of van der Waals energies. For instance, for a flat plate/flat plate geometry;

$$\Delta F_{12}^{adh}(d_1) = -\left[\frac{A_{12}}{12\pi d_1^2}\right] \tag{25}$$

$$\Delta F_{132}^{adh}(d_0) = -\left[\frac{A_{12}}{12\pi d_0^2}\right]. \tag{26}$$

For the four given combinations i.e. Eqs. (4)–(7) the equilibrium separation distances, however, are not necessarily the same. When a gap is a vacuum, the equilibrium separation d_1 probably is but when the gap contains a liquid, a different separation distance d_o may be expected. As a result of this the following becomes true;

$$\left[\frac{A_{132}}{d_0}\right] = \left[\frac{A_{12} - A_{13} - A_{23} + A_{33}}{d_1^2}\right] \tag{27}$$

A detailed study of van der Waals forces revealed that in the case of a three-component system, the corresponding Hamaker Constant A_{132} could attain a negative value given the conditions stated below.

$$A_{11} < A_{33} < A_{22} \text{ or } A_{11} > A_{33} > A_{22} \tag{28}$$

where; A_{11}, A_{22} and A_{33} are the individual Hamaker Constants of components 1, 2 and 3 respectively.

The implication of this is that two adhering bodies 1 and 2 of different composition will separate spontaneously upon immersion in a liquid 3 provided the conditions given by Eq. (28) are fulfilled.

2 Mathematical Model for the Interactions Mechanism

The mutual attraction energy, V_A of two molecules in a vacuum is given by;

$$V_A = -\frac{3}{4} h v_0 \left[\frac{\alpha^2}{H^6}\right] = -\left[\frac{\beta_{11}}{H^6}\right] \tag{29}$$

where;

H = Planck's constant
v_0 = characteristic frequency of the molecule
α = polarizability of the molecule
H = their separation distance

The assemblies of molecules as in a solid body have interaction energy as the summation of all the interaction energies of all the molecules present and the van der Waals pressure, P_{vdw} as follows;

$$P_{vdw} = \left[\frac{A_{11}}{6\pi d^3}\right] \tag{30}$$

For a sphere of radius, R and a semi-infinite body at a maximum separation distance, d the van der Waals force of attraction, F_{vdw} is given as;

$$F_{vdw} = \left[\frac{A_{11} R}{6d^2}\right] \tag{31}$$

where

A_{11} = Hamaker constant

$$A_{11} = \pi^2 q_1^2 \beta_{11} \tag{32}$$

where;

q_1 = number of atoms per cm^3
β_{11} = London-van der Waals constant

Given two dissimilar condensed bodies of given geometry with a separation distance, d, the corresponding van der Waals force between them can be

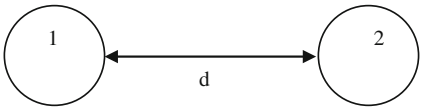


Fig. 6 Interaction of two unidentical molecules of lymphocyte *1* and virus (HIV) *2*, at a separation *d*

determined. For the system under study, the interacting bodies are the lymphocytes, 1 and the virus, 2 (Fig. 6).

The van der Waals force between the lymphocyte, 1 and the virus, 2 is given by the relations;

$$F_{vdw} = -\left[\frac{A_{12}R_{12}}{6d^2}\right] \tag{33}$$

where

$A_{11} = \pi^2 q_1^2 \beta_{11}$ = Hamaker constant for lymphocyte
$A_{22} = \pi^2 q_2^2 \beta_{22}$ = Hamaker constant for the virus (HIV)
$A_{12} = \pi^2 q_{12}^2 \beta_{12}$ = Hamaker constant for both materials (i.e. lymphocyte and the virus)

where; $\beta_{12} = \sqrt{\beta_{11}\beta_{22}}$

Thus the Hamaker constant becomes;

$$A_{12} = \sqrt{\left(\pi^2 q_1^2 \beta_{11}\right)\left(\pi^2 q_2^2 \beta_{22}\right)} \tag{34}$$

$$A_{12} = \sqrt{A_{11}A_{22}} \tag{35}$$

For a combination of the two dissimilar materials (i.e. lymphocyte 1, and the virus 2) with the gap between them filled with plasma or serum as the medium 3 the combined Hamaker coefficient will be given by;

$$A_{132} = \left(\sqrt{A_{11}} - \sqrt{A_{33}}\right)\left(\sqrt{A_{22}} - \sqrt{A_{33}}\right) \tag{36}$$

$$A_{132} = A_{12} + A_{33} - A_{13} - A_{23} \tag{37}$$

A_{33} = Hamaker constant for serum (plasma)
A_{13} = Hamaker constant for both materials (i.e. lymphocyte and plasma)
A_{23} = Hamaker constant for both materials (i.e. the virus and plasma)

$$A_{132abs} = \frac{3}{4}\pi\hbar \int_0^\infty \left[\frac{\varepsilon_1(i\zeta) - \varepsilon_3(i\zeta)}{\varepsilon_1(i\zeta) + \varepsilon_3(i\zeta)}\right]\left[\frac{\varepsilon_2(i\zeta) - \varepsilon_3(i\zeta)}{\varepsilon_2(i\zeta) + \varepsilon_3(i\zeta)}\right] d\zeta \tag{38}$$

Table 1 Comparison of the values of the Hamaker constants A_{11}, A_{22}, A_{33} and Hamaker coefficients A_{132}, A_{232} for the infected and A_{131} for the uninfected blood samples [7–10]

Variable ($\times 10^{-21}$ J)	Infected blood		Uninfected blood	
	Peak value	Absolute value	Peak value	Absolute value
A_{11}	–	–	1.3899	0.9659
A_{22}	1.4049	0.9868	–	–
A_{33}	0.7052	0.2486	0.9369	0.4388
A_{132}	0.7601	0.2587	–	–
A_{131}	–	–	0.9120	0.1026
A_{232}	0.7514	0.2823	–	–

The mean of all the values of the combined Hamaker coefficient, A_{132} gives an absolute value for the coefficient denoted by A_{132abs};

$$A_{132abs} = \frac{\sum_0^N (A_{132})}{N} \tag{39}$$

Table 1 shows the peak and absolute values of Hamaker constants A_{11} for the uninfected blood samples. A_{22} is the Hamaker constant for the virus, here represented by the infected lymphocytes. This is against the backdrop of no known process of isolation of the virus yet. However, it is a very close approximation for the virus owing to the manner of the infection mechanism. The Hamaker constants A_{33} for the plasma reveal greater values for the uninfected samples which invariably indicate a higher surface energy than the infected ones. This could be validated by the surface energy approach using the contact angle method. The higher absolute values of A_{132} and A_{232} as against that of A_{131}, as well as the near zero (i.e. 0.1026×10^{-21} J) value of the absolute combined Hamaker coefficient A_{131abs} for the uninfected samples is once again a clear indication of the relevance of the concept in the HIV infection process.

3 Deductions for the Absolute Combined Hamaker Coefficient A_{132abs}

Applying Lifshitz derivation for van der Waals forces as in Eq. (23) or (38) and getting a mean of all the values of the Hamaker coefficients to obtain a single value known as absolute combined Hamaker coefficient A_{132abs} became necessary and was got as [8];

$$A_{132abs} = 0.2587 \times 10^{-21} \text{ J}.$$

This was done by obtaining a mean of all the values of the Hamaker coefficients for the infected blood over the whole range of wavelength, $\lambda = 230$–890 Å, to obtain a single value of 0.2587×10^{-21} J. This value agrees with those obtained

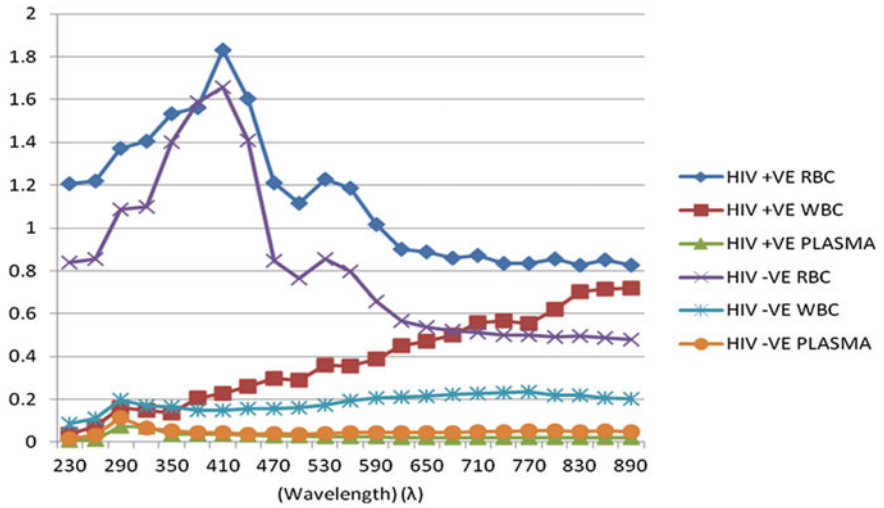


Fig. 7 Combined plot of absorbance versus wavelength for all three blood components [9]

by various authors for other biological processes [8, 10] as have earlier been shown in the literature review (Fig. 7).

The disparity between the peak absorbance values of HIV positive and negative blood components respectively is an indication of how the virus affects the properties of blood. The trend is such that the absorbance values of the HIV positive samples is generally decreased by a significant factor as shown in Table 1. The lymphocytes are of particular interest to this research since the virus attacks this blood component by being attached to the CD4+ cells which serve as receptor cells. As could be seen from the table, the variation between the peak values of absorbance of the various blood components is such that the lymphocytes vary with a magnitude of 0.021–0.034, the red blood cells differ by a factor of 0.049 to >1.168 while the plasma had a difference of between 0.003 and 0.040. The decrease in the absorbance of the HIV infected blood samples reveals the role of the virus in significantly affecting the surface properties of the infected blood cells and specimens [9].

4 Deductions for the Absolute Combined Negative Hamaker Coefficient

To define the condition where the absolute Hamaker coefficient becomes negative will require employing the relations that express that condition. Hence, recalling Eqs. (16)–(22), we can derive a state where the Hamaker Coefficient, A_{132} is less

than zero. This situation could be possible with the following already stated conditions;

$$A_{132} < 0 \tag{40}$$

When;

$$\sqrt{A_{11}} > \sqrt{A_{33}} \quad \text{and} \quad \sqrt{A_{22}} < \sqrt{A_{33}} \tag{41}$$

Or

$$\sqrt{A_{11}} < \sqrt{A_{33}} < \sqrt{A_{22}} \tag{42}$$

The mean of all values of A_{11} and A_{22} could be obtained and substituted into the relation below [i.e. Eq. (43)] in order to derive a value for A_{33} at which A_{132} is equal to zero in agreement with the earlier stated reasons.

$$A_{132} = \left(\sqrt{A_{11}} - \sqrt{A_{33}}\right)\left(\sqrt{A_{22}} - \sqrt{A_{33}}\right) \tag{43}$$

Rearranging Eq. (43) and making A_{33} subject of the formula we obtain;

$$A_{33} = \left[\frac{2\sqrt{A_{11}}\sqrt{A_{22}} - A_{132}}{\sqrt{A_{11}} + \sqrt{A_{22}}}\right]^2 \tag{44}$$

Obtaining a mean of the values of A_{11} and A_{22} to give absolute values of the Hamaker constants yielded the values given below;

$$A_{11} = 0.9659 \times 10^{-21}\ \text{J}$$
$$A_{22} = 0.9868 \times 10^{-21}\ \text{J}$$

Thus, plowing these values into Eq. (44) and rendering $A_{132} \leq 0$ will give the critical value of A_{33C} that satisfies the condition for the combined Hamaker coefficient to be equal to or less than zero. Hence any value of A_{33} greater than the critical would be the desired value necessary to attain a negative combined Hamaker coefficient.

Hence, the critical absolute Hamaker constant A_{33C} for the plasma (serum) which renders the A_{132} negative is given as;

$$A_{33C} = 0.9763 \times 10^{-21}\ \text{J}$$

Thus for negative combined Hamaker coefficient A_{132} of the infected blood to be attained, the combined Hamaker constant of the serum (plasma) as the intervening medium A_{33} should be of the magnitude;

$$A_{33} \geq 0.9763 \times 10^{-21}\ \text{J}$$

Inserting the above value of A_{33} into Eq. (43) would yield a negative value for A_{132} as follows;

$$A_{132} = -0.2809 \times 10^{-25}\ \text{J}\ \left(\text{when } A_{33} = 0.9763 \times 10^{-21}\ \text{J}\right)$$

To obtain a value for the combined Hamaker coefficient A_{131} for the uninfected blood the relation of Eqs. (45) and (46) are employed.

$$A_{131} = A_{11} + A_{33} - 2A_{13} \tag{45}$$

$$A_{131} = \left(\sqrt{A_{11}} - \sqrt{A_{33}}\right)^2 \tag{46}$$

Upon deriving a mean of all values of A_{131} for the twenty uninfected blood samples, an absolute value A_{131abs} was derived as given below;

$$A_{131abs} = 0.1026 \times 10^{-21}\ \text{J}$$

This value is very nearly equal to zero which is a clear indication of the validity of the concept of Hamaker coefficient to the process and progress of human infection with the Human Immunodeficiency Virus (HIV). The near zero value of the A_{131abs} shows the absence of infection in the blood samples thus suggesting the usefulness of the concept of negative Hamaker coefficient in finding a solution to HIV infection. Table 1 shows the comparison of the Hamaker constants and coefficients for the infected and uninfected blood samples.

5 Conclusion

This research work reveals that the interactions of the HIV and the lymphocytes could be mathematically modeled and the ensuing mathematics resolved in a bid to find a solution to the infection. The values of the Hamaker coefficients and constants derived are a proof of the relevance of the concept of Hamaker coefficient to the HIV-blood interactions and by extension to other biological and particulate systems. This study equally reveals the possibility of solving for the value of A_{33C} (i.e. a condition of the serum) which would favour the prevalence of a negative combined absolute Hamaker coefficient A_{132abs}. Such a condition in essence would mean repulsion between the virus and the blood cells and could prove the much desired solution to the HIV jinx. A synergy of Engineers, Pharmacists, Doctors, Pharmacologists, Medical laboratory scientists etc. may well be needed in interpreting the meaning of the A_{33C} values and the medical, biological and toxicity implications of additives in form of drugs that could yield the required characteristics.

References

1. S.N. Omenyi, Attraction and Repulsion of Particles by Solidifying Melts, Ph.D. thesis, University of Toronto (1978), pp. 23, 33, 34
2. H.C. Hamaker, Physica, vol. 4 (1937), p. 1058
3. J.D. van der Waals, Thesis, Leiden, 1873
4. J. Visser, *Advances in Interface Science*, vol. 15 (Elsevier Scientific Publishing Company, Amsterdam, 1981), pp. 157–169
5. F. London, Z. Phys. **63**, 245 (1930)
6. I.E. Dzyaloshinskii, E.M. Lifshitz et al., The general theory of van der Waals forces. Adv. Phys. **10**, 165 (1961)
7. C.H. Achebe, Human Immunodeficiency Virus (HIV)-Blood Interactions: Surface Thermodynamics Approach, Ph.D. Dissertation, Nnamdi Azikiwe University, Awka, 2010
8. C.H. Achebe, S.N. Omenyi, in *WCE 2013*: *Mathematical Determination of the Critical Absolute Hamaker Constant of the Serum* (*as an Intervening Medium*) *Which Favours Repulsion in the Human Immunodeficiency Virus* (*HIV*)*-Blood Interactions Mechanism*. Proceedings of the World Congress on Engineering 2013. Lecture Notes in Engineering and Computer Science. (London, 3–5 July 2013), pp. 1380–1384
9. C.H. Achebe, S.N. Omenyi, The effects of human immunodeficiency virus (HIV) infection on the absorbance characteristics of different blood components. Int. J. Sci. Invent. **2**(5), 53–61 (2013) www.ijesi.org
10. C.H. Achebe, S.N. Omenyi, O.P. Manafa, D. Okoli, in *IMECS 2012*: *Human Immunodeficiency Virus* (*HIV*)*-Blood Interactions*: *Surface Thermodynamics Approach*, Proceedings of The International Multi Conference of Engineers and Computer Scientists 2012. Lecture Notes in Engineering and Computer Science, (Hong Kong, 14–16 March 2012), pp. 136–141

Modeling and Analysis of Spray Pyrolysis Deposited SnO_2 Films for Gas Sensors

Lado Filipovic, Siegfried Selberherr, Giorgio C. Mutinati, Elise Brunet, Stephan Steinhauer, Anton Köck, Jordi Teva, Jochen Kraft, Jörg Siegert, Franz Schrank, Christian Gspan and Werner Grogger

Abstract Metal oxide materials such as tin oxide (SnO_2) show powerful gas sensing capabilities. Recently, the deposition of a thin tin oxide film at the backend of a CMOS processing sequence has enabled the manufacture of modern gas sensors. Among several potential deposition methods for SnO_2, spray pyrolysis deposition has proven itself to be relatively easy to use and cost effective while providing excellent surface coverage on step structures and etched holes. A model for spray pyrolysis deposition using a pressure atomizer is presented and implemented in a Level Set framework. A simulation of tin oxide deposition is performed on a typical gas sensor geometry and the resulting structure is imported into a finite element tool in order to analyze the electrical characteristics and thermo-mechanical stress present in the grown layer after processing. The deposition is performed at 400 °C and the subsequent cooling to room temperatures causes a stress to develop at the material interfaces due to variations in the coefficient of thermal expansion between the different materials.

L. Filipovic (✉) · S. Selberherr
Institute for Microelectronics, Technische Universität Wien, Gußhausstraße 27–29/E360, A–1040 Wien, Austria
e-mail: filipovic@iue.tuwien.ac.at

S. Selberherr
e-mail: selberherr@iue.tuwien.ac.at

G. C. Mutinati · E. Brunet · S. Steinhauer · A. Köck
Molecular Diagnostics, Health and Environment, AIT GmbH, Donau-City-Straße 1, A–1220 Wien, Austria
e-mail: Giorgio.Mutinati@ait.ac.at

E. Brunet
e-mail: Elise.Brunet.fl@ait.ac.at

S. Steinhauer
e-mail: Stephan.Steinhauer.fl@ait.ac.at

A. Köck
e-mail: Anton.Koeck@ait.ac.at

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_22,

Keywords Electrical characterization of tin oxide · FEM simulation · Level set method · Modeling spray pyrolysis · Monte Carlo · Smart gas sensors · Spray pyrolysis deposition · Thermal stress modeling · Tin oxide · Von Mises stress

1 Introduction

The ability to detect harmful and toxic gases in the environment is a subject of extensive research. Usually, the manufacture of gas sensors is incompatible with that of the CMOS process sequence. The miniaturization of electronic devices has proven to be essential, while bulky gas sensors are still lagging behind the overall progress of CMOS and MEMS devices. Metal oxides may serve as a gas sensing layer, when a thin film of the deposited material is exposed to high temperatures (250–400 °C). A material which has been proven to exhibit all the properties required for good gas sensing performance is tin oxide (SnO_2) [5, 20, 24], while others such as zinc oxide (ZnO), indium tin oxide (ITO), CdO, $ZnSnO_4$, NiO, etc. have also been widely studied [4]. This work mainly concerns itself with tin oxide gas sensors and the ability to develop a model which depicts the growth of thin tin oxide layers to act as a gas sensing surface. In addition, the resistance and stress generation through the device after the processing sequence is determined using finite element simulations. The deposition of SnO_2 has been reported to be performed using various standard techniques such as chemical vapor deposition [32], sputtering [3], pulsed-laser deposition [30], sol-gel process [6], and spray pyrolysis (SP) deposition [24].

J. Teva · J. Kraft · J. Siegert · F. Schrank
ams AG, Tobelbaderstraße 30, A–8141 Unterpremstätten, Austria
e-mail: Jordi.Teva@ams.com

J. Kraft
e-mail: Jochen.Kraft@ams.com

J. Siegert
e-mail: Joerg.Siegert@ams.com

F. Schrank
e-mail: Franz.Schrank@ams.com

C. Gspan · W. Grogger
Institute for Electron Microscopy and Fine Structure Research, Graz University of Technology and the Centre for Electron Microscopy Graz, Steyrergasse 17, A–8010 Graz, Austria
e-mail: christian.gspan@felmi-zfe.at

W. Grogger
e-mail: werner.grogger@felmi-zfe.at

The spray pyrolysis deposition technique is gaining traction in the scientific community due to its cost effectiveness and relative ease of integration at the back end of a standard CMOS process. The technique is used to grow crystal powders [19], which can then be further annealed for use in gas sensors, solar cells, and other applications. The technique is cost-effective as it involves relatively inexpensive equipment and raw materials and is simple to perform. For these reasons the spray pyrolysis deposition of SnO_2 thin films is further explored. A model for spray pyrolysis deposition, which can be incorporated onto a sensor device after a standard CMOS process simulator is desired [11].

Spray pyrolysis requires no vacuum and provides high flexibility in terms of material composition. In order to optimize this technology for the heterogeneous integration of gas sensing layers with CMOS fabricated micro-hotplate chips [15], a complete understanding of the spray pyrolysis deposition process by modeling is a challenging issue. It was our goal to develop and incorporate a model for the growth of ultrathin SnO_2 layers into a traditional CMOS process simulator using the Level Set framework [8]. Due to the temperature required for this process (400 °C), the subsequent cooling to room temperature can cause the arise of stress through the device due to the varying coefficients of thermal expansion (CTE) for the different materials. This is explored with finite element simulations.

1.1 Smart Gas Sensor Devices

Different variants of metal oxide based gas sensors, which rely on changes of electrical conductance due to the interaction with the surrounding gas, have been developed. However, today's gas sensors are bulky devices, which are primarily dedicated to industrial applications. Since they are not integrated in CMOS technology, they cannot fulfill the requirements for smart gas sensor applications in consumer electronics. A powerful strategy to improve sensor performance is the implementation of very thin nanocrystalline films, which have a high surface to volume ratio and thus a strong interaction with the surrounding gases. SnO_2 has been one of the most prominent sensing materials and a variety of gas sensor devices based on SnO_2 thin films has been realized so far [13, 31] as depicted in Fig. 1. The growth of the ultrathin SnO_2 layers on semiconductor structures requires a deposition step which can be integrated after the traditional CMOS process [12, 20]. This alleviates the main concern with today's gas sensor devices and their bulky nature, namely high power consumption. The sensing mechanism of SnO_2 is related to the chemisorption of gas species over the surface, leading to charge transfer between the gas and surface molecules and changes in the electrical conductance.

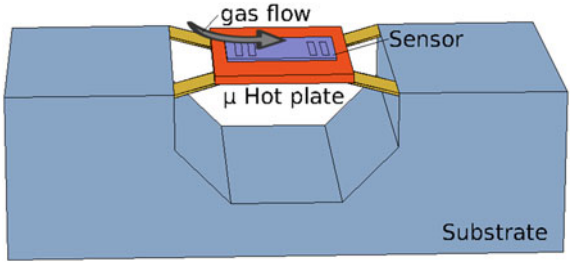


Fig. 1 Principle view of the gas sensor on a micro hot plate

1.2 Level Set Method

Since the introduction of the Level Set Method by Osher and Sethian [23], it has developed into a favorite technique for tracking moving interfaces. The presented simulations and models function fully within the process simulator presented in [7]. The software also supports memory parallelization during execution [9]. The Level Set method is utilized in order to describe the top surface of a semiconductor wafer as well as the interfaces between different materials. The method describes a movable surface $S(t)$ as the zero Level Set of a continuous function $\Phi(\mathbf{x}, t)$ defined on the entire simulation domain,

$$S(t) = \{\mathbf{x} : \Phi(\mathbf{x}, t) = 0\}. \tag{1}$$

The continuous function $\Phi(\mathbf{x}, t)$ is obtained using a signed distance transform

$$\Phi(\mathbf{x}, t = 0) := \begin{cases} -\min\limits_{\mathbf{x}' \in S(t=0)} \|\mathbf{x} - \mathbf{x}'\| & \text{if } \mathbf{x} \in M(t = 0) \\ +\min\limits_{\mathbf{x}' \in S(t=0)} \|\mathbf{x} - \mathbf{x}'\| & \text{else,} \end{cases} \tag{2}$$

where M is the material described by the Level Set surface $\Phi(\mathbf{x}, t = 0)$. The implicitly defined surface $S(t)$ describes a surface evolution, driven by a scalar velocity $V(\mathbf{x})$, using the Level Set equation

$$\frac{\partial \Phi}{\partial t} + V(\mathbf{x}) \|\nabla \Phi\| = 0. \tag{3}$$

In order to find the location of the evolved surface, the velocity field $V(\mathbf{x})$, which is a calculated scalar value, must be found. For the case of spray pyrolysis deposition, this scalar value is derived using Monte Carlo techniques and ray tracing.

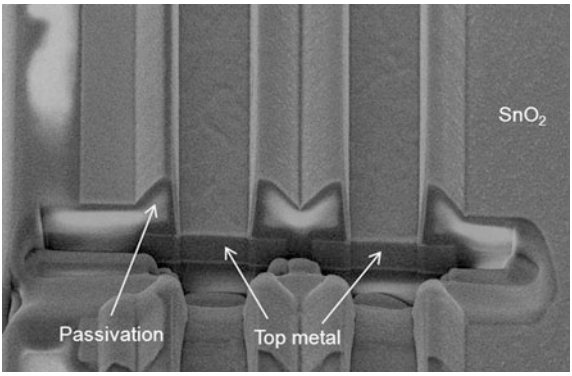


Fig. 2 Three-dimensional view of the electrode locations on the substrate

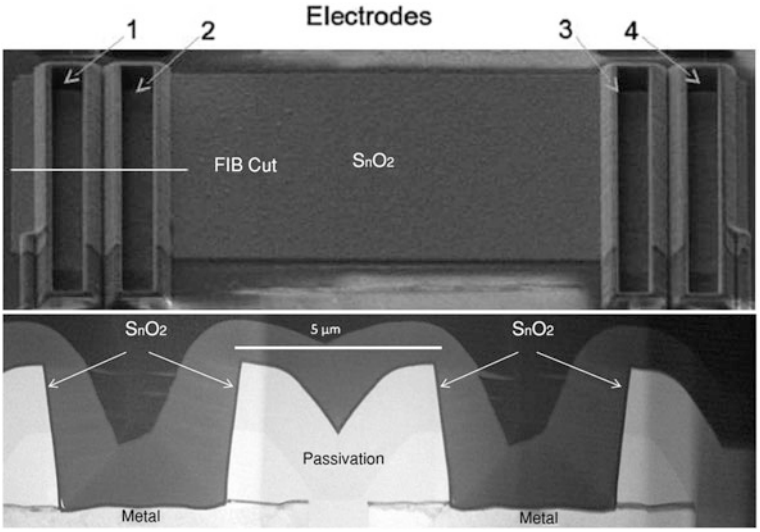


Fig. 3 View of the electrode locations on the substrate and the side view through the *FIB cut* line

2 Tin Oxide Based Gas Sensor

A smart gas sensor device has been manufactured using a 50 nm SnO_2 thin film as the sensing material. The device has four electrodes, shown in Figs. 2 and 3, which are stationed above a heat source. The substrate is a CMOS chip with four contact electrodes which are coated with SnO_2, as depicted in Fig. 3.

The sensor has been tested in a H_2 environment at concentrations down to 10 ppm and the results are depicted in Fig. 4. The sensor itself operates on top of a micro-sized hot plate [29] which heats the sensor locally to 250–400 °C in order to

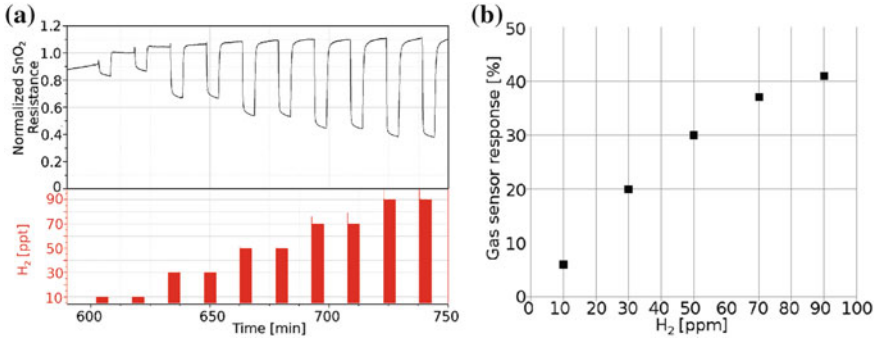


Fig. 4 **a** SnO_2 electrical resistance change. **b** Gas sensor response (%) as H_2 in varying concentration is pulsed into a humid synthetic air (RH = 40 %) environment. The sensing temperature is set to 350 °C

detect humidity and harmful gases in the environment, such as CO, CH_4, H_2, CO_2, SO_2, and H_2 [5].

The gas measurements are performed in an automatic setup to test the functionality of the described SnO_2 structure, while heated up to 350 °C. The electrical resistance change is monitored, while pulses of H_2 at different concentrations (10–90 ppm) are injected in the gas chamber. Humid synthetic air (RH = 40 %) is used as the background gas. In Fig. 4a, the normalized resistance of the SnO_2 structure is plotted for various H_2 gas concentrations. In Fig. 4b the gas response, defined as the relative resistance difference in percentage, is shown. The gas sensor results appear to be functional in the entire H_2 concentration range investigated.

3 Spray Pyrolysis Deposition

During the last several decades, coating technologies have garnered considerable attention, mainly due to their functional advantages over bulk materials, processing flexibility, and cost considerations [21]. Thin film coatings can be deposited using physical methods or chemical methods. The chemical methods can be split according to a gas phase deposition or a liquid phase deposition. Spray pyrolysis is a technique which uses a liquid source for thin film coating. The main advantages of spray pyrolysis over other similar deposition techniques are:

- Cost effectiveness.
- Possible integration after a standard CMOS process.
- Substrates with complex geometries can be coated.
- Uniform process, which can be scaled over larger areas.
- CMOS-compatible temperatures (<400 °C) can be used for processing.

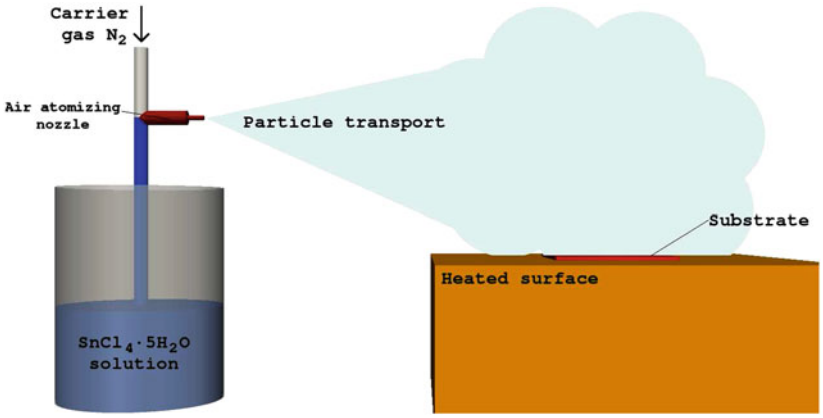


Fig. 5 Schematic of the spray-pyrolysis deposition process, as set up for SnO_2 deposition

Spray pyrolysis is increasingly being used for various commercial processes, such as the deposition of a transparent layer on glass [18], the deposition of a SnO_2 layer for gas sensor applications [17], the deposition of a YSZ layer for solar cell applications [25], anodes for lithium-ion batteries [22], and optoelectronic devices [2]. The setup, shown in Fig. 5 is simple and inexpensive when compared to other deposition alternatives.

The three steps which describe the processes taking place during spray pyrolysis deposition are summarized by:

1. Atomization of the precursor solution.
2. Aerosol transport of the droplet.
3. Decomposition of the precursor to initiate film growth.

3.1 Experimental Setup

When depositing a thin film using spray pyrolysis, an ultrasonic, electrical, or gas pressure atomizing nozzle can be used [26]. For smart gas sensor applications, a gas pressure nozzle is ideal due to its ease of use and its ability to create very small droplets which deposit evenly on a desired surface. The retardant forces experienced by droplets during their transport include the Stokes force and the thermophoretic force, while gravity is the only accelerating force, when a gas pressure nozzle is used. An electrical force is included in models which depict ultrasonic or electrical atomizing nozzles [10]. Figure 5 shows a simplified schematic for spraying a specific precursor solution onto the substrate, which is placed on top of a hotplate.

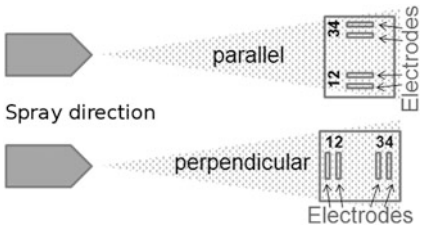


Fig. 6 Spray direction during deposition

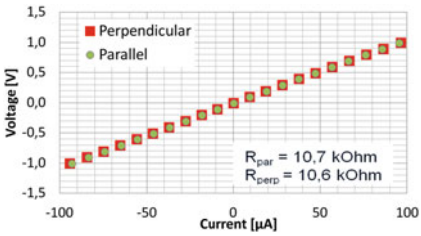


Fig. 7 V-I curves between electrode 1 and electrode 4 for the chips depicted in Fig. 6

The atomizer has also been adjusted so that the droplets exiting the nozzle are relatively small in volume with an average diameter of approximately 5–10 μm. This ensures that a majority of the droplets reaching the heated wafer surface will evaporate prior to impact, allowing for a uniform deposition of the vapor, which results in a subsequent uniform film growth. In order to ensure that the initial direction of the solution, as it exits the atomizer, does not affect the deposition process, two spray directions have been tested, as shown in Fig. 6.

3.2 Experimental Observations

After performing spray pyrolysis deposition, while directing the spray in parallel and perpendicular to the electrodes, the V-I curves of the two chips are visualized in Fig. 7. It was noted that regardless of the initial spray direction and the direction of the droplets as they leave the atomizing nozzle, the thickness of the grown thin film and its electrical properties remain unchanged.

This is likely due to the nozzle being placed at a distance of approximately 30 cm away from the substrate surface, giving enough time for the Stokes retardant force to effectively remove any influence of the droplets' initial horizontal velocity. Therefore, directionality plays no role in the film deposition process. This also helps to eliminate the potential of large droplets splashing onto the substrate surface. Large droplets generally do not deposit uniformly on the desired surface, but rather impact the heated wafer while in liquid form, leaving behind a powder residue with weak sticking properties to the silicon.

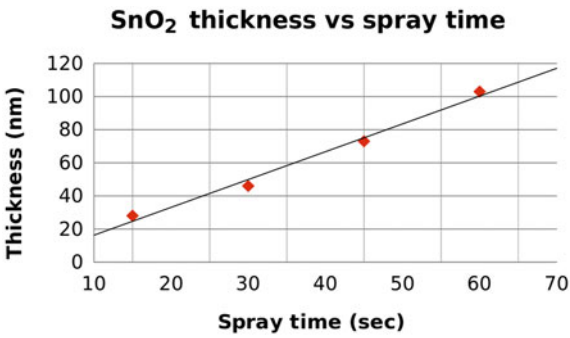


Fig. 8 The influence of spray time on SnO_2 thickness, with temperature set to 400 °C

The thickness of the tin oxide layer depends on the spray temperature and the spray time. With the heatpad temperature set to 400 °C, the thickness of the SnO_2 layer is plotted against the spray time in Fig. 8. A linear relationship is evident. The thickness of the grown film does not change, when the deposition takes place on a step structure, as shown in Fig. 9a, suggesting that the deposition is a result of a vapor interaction and not a process which alludes to a direct interaction between the deposition surface and the liquid droplets. Figure 9a shows a resulting SnO_2 thin film after a spray pyrolysis deposition step lasting 30 s with the substrate heated to 400 °C. The resulting film thickness is approximately 50 nm. We conclude that the droplets, which carry the depositing material, interact with the substrate surface as a vapor and then deposit in a process analogous to CVD.

4 Modeling Spray Pyrolysis

There are two main approaches to modeling spray pyrolysis. One deals with tracking of the trajectories of individual droplets, resulting in their direct impact with the wafer. The second model assumes that the droplets vaporize prior to impact resulting in a CVD-like uniform deposition step.

4.1 Modeling the Uniform Deposition Process

The experimental data shows a linear dependence on spray time and a logarithmic dependence on wafer temperature for the growth rate of the deposited SnO_2 layer. A good agreement is given by the Arrhenius expression

$$d_{SnO_2}(t,T) = A_1 t e^{(-E/k_B T)}, \tag{4}$$

where the thickness is given in µm, $A_1 = 3.1\,\mu m/s$, t is the time in seconds, T is the temperature in Kelvin, and E is 0.427 eV.

4.2 Spray Pyrolysis as a Vapor Deposition Process

The growth model given in (4) relates the thickness of the deposited material to the applied time and temperature. However, this representation is only valid, when no complex geometries such as deep wells are present. In order to model deposition on a deep well structure, a simulation which considers more than a single deposition rate is required. Since the droplets fully evaporate prior to depositing on the surface, a non-linear simulation model analogous to CVD is used. The implementation requires the combination of the Monte Carlo method within the Level Set framework. A single particle species is considered during deposition. As the simulation is initiated, multiple particles are generated in the simulation space with an average direction perpendicular to and moving towards the wafer. The particles are represented in terms of individual fluxes, given by

$$\Gamma_{src} = \Gamma_{src}(\mathbf{x}; \mathbf{w}, E) \quad \mathbf{x} \in P, \tag{4}$$

where P is the surface which divides the region above the wafer (reactor-scale) and the region comprising the wafer (feature-scale). The flux of particles Γ_{src} is described in terms of particles which are moving in direction $\mathbf{w}$ arriving with an energy E per unit area. This energy E depends on the time and temperature of the simulation space, as it is the deciding factor in the speed of the film growth. The distribution of particles stems from the idea that their transport is characterized by the mean free path of a gas $\bar{\lambda}$, which for our process is in the range of 9 mm. In this range, the particle velocities follow the Maxwell-Boltzmann distribution and the flux has a cosine-like directional dependence

$$\Gamma_{src}(\mathbf{x}; q, \mathbf{w}, E) = F_{src} \frac{1}{\pi} \cos\theta \quad \cos\theta = \mathbf{w} \cdot \mathbf{n}_P, \tag{6}$$

where F_{src} is uniform, $\mathbf{n}_p$ is the normal vector to P pointing towards the wafer surface.

In the feature-scale region particles may also be reflected from the wafer walls, modeled with a sticking coefficient, resulting in their deposition elsewhere on the wafer or them leaving the simulation space entirely. Reflected particles follow the Knudsen cosine law [14]. The total flux is then composed of source particles with flux Γ_{src} and re-emitted particles, which stick on recursive impact Γ_{re}. For the spray pyrolysis deposition process, it was found that a sticking coefficient of 0.01 has the best fit to experimental data and was therefore used for the model. The motion of reflected or re-emitted particles is then tracked with their sticking probability reduced after each surface impact. The tracking of a single particle is deemed concluded, when its sticking probability reaches 0.1 % of the original sticking coefficient.

Using the presented model, a simulation was performed for 30 s at 400 °C with the result shown in Fig. 9b. The deposited film has an evenly distributed thickness of approximately 50 nm, as expected from the measured thickness in Fig. 9a. The model is also applied to half of the complete sensor geometry, including two electrodes, as shown in Fig. 3, with the result shown in Fig. 10.

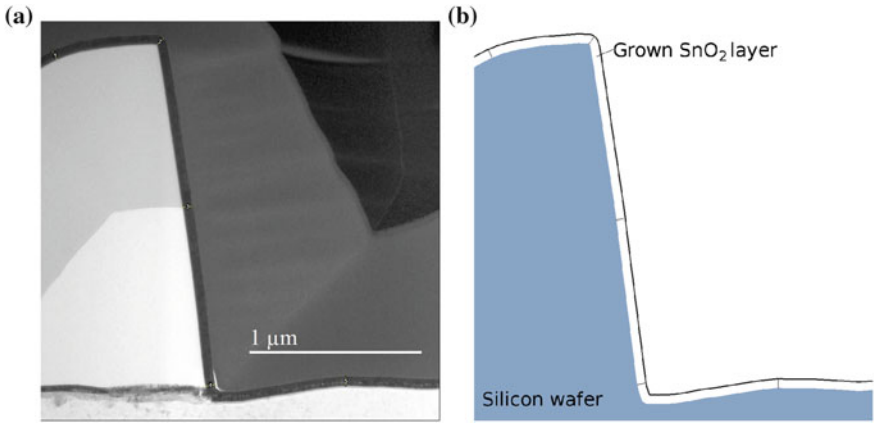


Fig. 9 Experimental and simulated SnO_2 deposition on a step structure. **a** TEM image of a FIB cut. **b** Simulation of **a**

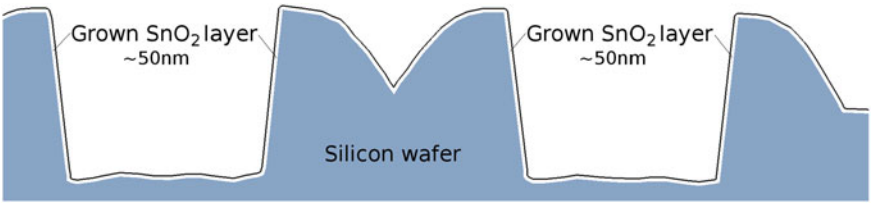


Fig. 10 The simulation of SnO_2 deposition, performed with the heated substrate at a temperature T = 400 °C for a time t = 30 s

4.3 Sample Trench Simulations

With the presented model one can simulate the resulting coverage of various geometries, required for gas sensor manufacturing. For devices which need a large surface area in order to function adequately, it may be more affordable to deposit a layer on trench and well geometries rather than utilizing expensive chip surface area. Therefore, several analysis are performed to estimate at which trench width to depth ratio and at which sidewall angle is the process no longer appropriate. First, we investigate the trench sidewall angle of 5.7° in Fig. 11 and note that for a trench of height 1,000 nm, the separation between the walls must be at least 150 nm in order for no void to form. However, a separation greater than 200 nm is desired in order to ensure no void is formed even when potential variation is introduced in the process.

Similarly, an analysis is performed for a structure with the same dimensions, but with the sidewall angles more vertical at 2°. It was observed that, even at a trench width of 150 nm, a void has formed. Therefore, although vertical walls

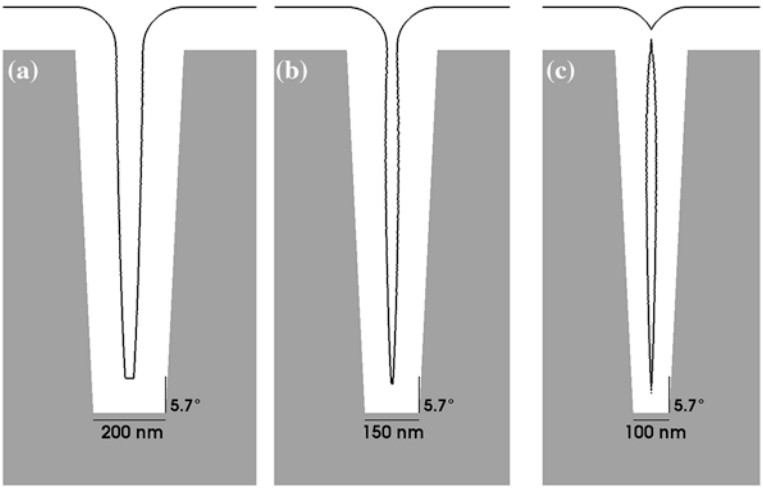


Fig. 11 Sample SnO_2 simulation showing a 60 s spray pyrolysis deposition at 400 °C with a trench depth of 1,000 nm and width ranging between **a** 200 nm **b** 150 nm, and **c** 100 nm. Sidewall angles are set to 5.7°

provide more surface area for the tin oxide to deposit, one must be careful in ensuring enough separation is provided at the top of the trench, where the deposited material tends to accumulate.

In addition, the implemented model can inherently be used to analyze three-dimensional surface coverage, fully integrated in a standard CMOS simulator using a combination of Monte Carlo methods within a Level Set framework.

5 Electrical and Stress Analysis of the Sensor Structure

The resulting simulated structure is imported into a finite element tool in order to determine the resistivity and stress distribution in the structure after the processing step. The half-structure is shown in Fig. 12 and the material parameters used are listed in Table 1.

5.1 Electrical Resistivity of the Tin Oxide

The I-V curve for the sensor was experimentally measured in Fig. 7, resulting in a full sensor resistance of 10.7 kΩ. The resistivity of the deposited material is determined using finite element methods. A total resistance of 5.35 kΩ is assumed for the half-sensor structure, which corresponds to a tin oxide resistivity of

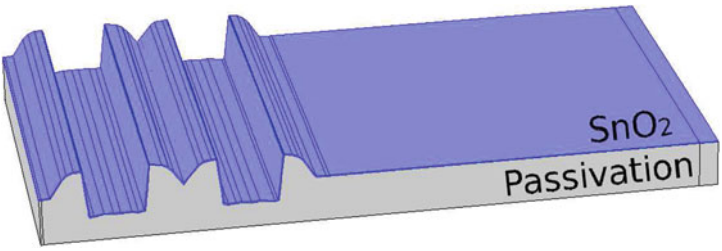


Fig. 12 The three-dimensional half-sensor structure showing the deposited SnO_2 layer, which is imported into the finite element tool for electrical and stress analysis

Table 1 Electrical and thermal properties of tin oxide

SnO_2 parameter	Value	Units	References
Density	6.95	$g.cm^{-3}$	[1]
Electrical resistivity	1.233×10^{-2}	Ω cm	This study, [16, 28]
Thermal conductivity	0.4	$W{\cdot}cm^{-1}{\cdot}K^{-1}$	[27]
Coefficient of thermal expansion	4×10^{-6}	K^{-1}	[1]

1.233×10^{-2} Ω cm. This value corresponds to the general range of published values in [16] and [28] for tin oxide. The sensor, when operating at a current of 2 μA [5], experiences a voltage drop of 21.4 mV.

5.2 Thermo-Mechanical Stress

After spray pyrolysis deposition and annealing is performed at 400 °C, the structure is cooled to room temperature. This temperature difference can cause stresses between the passivation and the tin oxide material due to their different coefficients of thermal expansion. This stress can lead to material cracking and delamination when the tin oxide does not stick to the passivation layer properly. The stress is determined using the von Mises stress (σ_{VMS}) as a benchmark, which is defined using a combination of the three principal stresses as

$$\sigma_{VMS} = \sqrt{\frac{1}{2} \cdot (\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2} \tag{7}$$

The tin oxide layer experiences a maximum stress of approximately 1GPa and an average stress of 210 MPa. The interface between the tin oxide layer and the passivation layer experiences a maximum stress of 440 MPa and an average of 120 MPa. The stress distribution in the device is shown with two-dimensional slices through the material in Fig. 13. During the gas sensor operation, the device is frequently heated to temperatures ranging from 200 to 400 °C using the micro

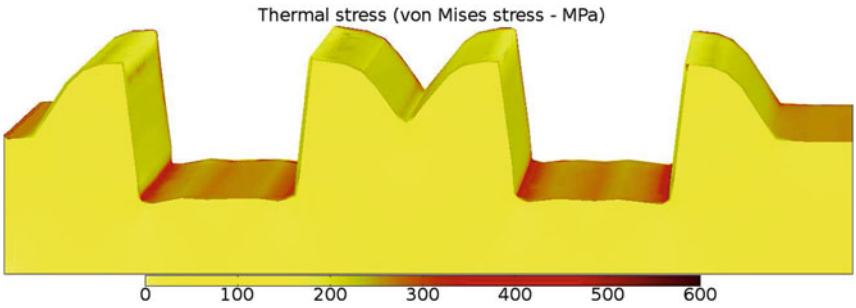


Fig. 13 The von Mises stress distribution through the passivation and SnO_2 layers after processing at 400 °C and cooling to room temperature

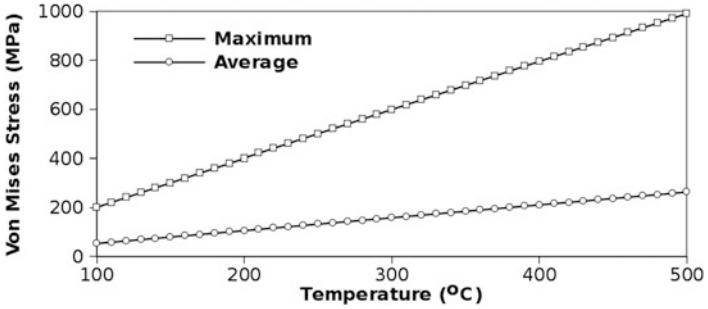


Fig. 14 The dependence of the maximum and average von Mises stress (MPa) in the SnO_2 layer as a result of temperature variation in the sensor

hotplate shown in Fig. 1. This thermal cycling results in further stresses in the tin oxide layer. The effects on the maximum and average stress in the SnO_2 layer when cycling the temperature is depicted in Fig. 14. A linear relationship is noted between the applied temperature and generated stress σ_{VMS}.

6 Conclusion

While the general trend for CMOS and MEMS devices has been aggressive miniaturization, gas sensors still generally remain bulky devices the manufacturing of which is difficult to integrate with CMOS processing. The capability of thin metal oxides to detect harmful gases in the environment has lead to the development of smart gas sensors. These oxides can be deposited at the back end of the CMOS process using the spray pyrolysis technique. In this work a model for spray pyrolysis deposition, when an air atomizer is used, has been developed and implemented in a Level Set framework. The model is based on the observation that the material is deposited in a vapor form, analogous to CVD. The deposition of tin

oxide on one half of a typical sensor structure has been simulated and the resulting geometry has been extracted and imported into a finite element tool. The properties of the grown tin oxide has been further analyzed, including the material resistivity and the stress developed in the region as a consequence of the processing step. The deposition is performed at 400 °C and the subsequent cooling to room temperature causes stress development between different layers due to the varying coefficients of thermal expansion between the SnO_2 and surrounding passivation.

Acknowledgements This work has been partly performed in the COCOA-CATRENE European project and in the project ESiP. In this latter the Austrian partners are funded by the Austrian Research Promotion Agency (FFG) under project no. 824954 and the ENIAC Joint Undertaking.

References

1. M. Batzill, U. Diebold, The surface and materials science of tin oxide. Prog. Surf. Sci. **79**(2), 47–154 (2005)
2. G. Blandenet, M. Court, Y. Lagarde, Thin layers deposited by the pyrosol process. Thin Solid Films **77**(1–3), 81–90 (1981)
3. J. Boltz, D. Koehl, M. Wuttig, Low temperature sputter deposition of SnO_x: Sb films for transparent conducting oxide applications. Surf. Coat. Technol. **205**(7), 2455–2460 (2010)
4. A. Bouaoud, A. Rmili, F. Ouachtari, A. Louardi, T. Chtouki, B. Elidrissi, H. Erguig, Transparent conducting properties of Ni doped zinc oxide thin films prepared by a facile spray pyrolysis technique using perfume atomizer. Mater. Chem. Phys. **137**(3), 843–847 (2013)
5. E. Brunet, T. Maier, G.C. Mutinati, S. Steinhauer, A. Köck, C. Gspan, W. Grogger, Comparison of the gas sensing performance of SnO_2 thin film and SnO_2 nanowire sensors. Sens. Actuator B **165**, 110–118 (2012)
6. D.M. Carvalho, J.L. Maciel Jr, L.P. Ravaro, R.E. Garcia, V.G. Ferreira, L.V. Scalvi, Numerical simulation of the liquid phase in SnO_2 thin film deposition by sol-gel-dip-coating. J. Sol-Gel. Sci. Technol. **55**(3), 385–393 (2010)
7. O. Ertl, Numerical methods for topography simulation. Dissertation, Technischen Universität Wien, Fakultät für Elektrotechnik und Informationstechnik, http://www.iue.tuwien.ac.at/phd/ertl/ (2010)
8. O. Ertl, S. Selberherr, A fast level set framework for large three-dimensional topography simulations. Comput. Phys. Commun. **180**(8), 1242–1250 (2009)
9. L. Filipovic, O. Ertl, S. Selberherr, Parallelization strategy for hierarchical run length encoded data structures. In *Proceedings of IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN) 2011*, 15–17 Feb 2011, Innsbruck, Austria, pp. 131–138 (2011)
10. L. Filipovic, S. Selberherr, G.C. Mutinati, E. Brunet, S. Steinhauer, A. Köck, J. Teva, J. Kraft, J. Siegert, F. Schrank, Modeling spray pyrolysis deposition. In *Lecture Notes in Engineering and Computer Science: Proceedings of World Congress on Engineering 2013*, 3–5 July 2013, London, UK, pp. 987–992 (2013)
11. L. Filipovic, S. Selberherr, G.C. Mutinati, E. Brunet, S. Steinhauer, A. Köck, J. Teva, J. Kraft, J. Siegert, F. Schrank, C. Gspan, W. Grogger, Modeling the growth of thin SnO_2 films using spray pyrolysis deposition. In *Proceedings of International Conference on Simulation of Processes and Devices (SISPAD) 2013*, 3–5 Sept 2013, Glasgow, UK, pp. 208–211
12. L. Filipovic, S. Selberherr, G.C. Mutinati, E. Brunet, S. Steinhauer, A. Köck, J. Teva, J. Kraft, J. Siegert, F. Schrank, A method for simulating spray pyrolysis deposition in the level set framework. Eng. Lett. **21**(4), 224–240 (2013)

13. W. Göpel, K. Schierbaum, SnO_2 sensors: current status and future prospects. Sens. Actuator B **26**(1–3), 1–12 (1995)
14. J. Greenwood, The correct and incorrect generation of a cosine distribution of scattered particles for Monte-Carlo modelling of vacuum systems. Vacuum **67**(2), 217–222 (2002)
15. C. Griessler, E. Brunet, T. Maier, S. Steinhauer, A. Köck, J. Teva, F. Schrank, M. Schrems, Tin oxide nanosensors for highly sensitive toxic gas detection and their 3D system integration. Microelectron. Eng. **88**(8), 1779–1781 (2011)
16. J. Joseph, V. Mathew, J. Mathew, K. Abraham, Studies on physical properties and carrier conversion of SnO_2:Nd thin films. Turkish J. Phys. **33**, 37–47 (2009)
17. G. Korotcenkov, V. Brinzari, J. Schwank, M. DiBattista, A. Vasiliev, Peculiarities of SnO_2 thin film deposition by spray pyrolysis for gas sensor application. Sens. Actuator B **77**(1–2), 244–252 (2001)
18. S. Major, A. Banerjee, K. Chopra, Highly transparent and conducting indium-doped zinc oxide films by spray pyrolysis. Thin Solid Films **108**(3), 333–340 (1983)
19. G.L. Messing, S.C. Zhang, G.V. Jayanthi, Ceramic powder synthesis by spray pyrolysis. J. Am. Ceram. Soc. **76**(11), 2707–2726 (1993)
20. G. Mutinati, E. Brunet, S. Steinhauer, A. Köck, J. Teva, J. Kraft, J. Siegert, F. Schrank, E. Bertagnolli, CMOS-integrable ultrathin SnO_2 layer for smart gas sensor devices. Procedia Eng. **47**, 490–493 (2012)
21. A. Nakaruk, C. Sorrell, Conceptual model for spray pyrolysis mechanism: fabrication and annealing of titania thin films. J. Coat. Technol. Res. **7**(5), 665–676 (2010)
22. S.H. Ng, J. Wang, D. Wexler, S.Y. Chew, H.K. Liu, Amorphous carbon-coated silicon nanocomposites: a low-temperature synthesis via spray pyrolysis and their application as high-capacity anodes for lithium-ion batteries. J. Phys. Chem. C **111**(29), 11131–11138 (2007)
23. S. Osher, J.A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
24. G. Patil, D. Kajale, V. Gaikwad, G. Jain, Spray pyrolysis deposition of nanostructured tin oxide thin films. ISRN Technol. **2012**(1–5), 275872 (2012)
25. D. Perednis, L.J. Gauckler, Solid oxide fuel cells with electrolytes prepared via spray pyrolysis. Solid State Ion. **166**(3–4), 229–239 (2004)
26. D. Perednis, L.J. Gauckler, Thin film deposition using spray pyrolysis. J. Electroceram. **14**(2), 103–111 (2005)
27. C. Poulier, D. Smith, J. Absi, Thermal conductivity of pressed powder compacts: tin oxide and alumina. J. Eur. Ceram. Soc. **27**(2), 475–478 (2007)
28. K. Shamala, L. Murthy, K.N. Rao, Studies on tin oxide films prepared by electron beam evaporation and spray pyrolysis methods. Bull. Mater. Sci. **27**(3), 295–301 (2004)
29. M. Siegele, C. Gamauf, A. Nemecek, G.C. Mutinati, S. Steinhauer, A. Kock, J. Kraft, J. Siezert, F. Schrank, Optimized integrated micro-hotplates in CMOS technology. In *Proceedings of IEEE International New Circuit and Systems Conference (NEWCAS) 2013*, 16–19 June 2013, Paris, France, pp. 1–4 (2013)
30. S. Sinha, R. Bhattacharya, S. Ray, I. Manna, Influence of deposition temperature on structure and morphology of nanostructured SnO_2 films synthesized by pulsed laser deposition. Mater. Lett. **65**(2), 146–149 (2011)
31. A. Tischner, T. Maier, C. Stepper, A. Köck, Ultrathin SnO_2 gas sensors fabricated by spray pyrolysis for the detection of humidity and carbon monoxide. Sens. Actuators B **134**(2), 796–802 (2008)
32. I. Volintiru, A. de Graaf, J. Van Deelen, P. Poodt, The influence of methanol addition during the film growth of SnO_2 by atmospheric pressure chemical vapor deposition. Thin Solid Films **519**(19), 6258–6263 (2011)

SISO Control of TITO Systems: A Comparative Study

Yusuf A. Sha'aban, Abdullahi Muhammad, Kabir Ahmad and Muazu M. Jibrin

Abstract This paper presents a brief study of concepts used in control of two-input and two-output systems. A novel decentralised model predictive control (DMPC) for two-input and two-output (TITO) processes is presented. To reduce the computational load, shifted input sequence is used to cater for loop interactions. The proposed scheme is applied to a coupled system to demonstrate its performance. Model predictive control (MPC) and decentralised proportional, integral and derivative (PID/PI) controllers were also applied for comparison purposes. The proposed controller has a performance similar to MPC but outperforms the decentralised PID/PI controllers.

Keywords Decentralised control · Decoupling · FOPDT · MPC · PID · Predictive control · SISO · TITO

1 Introduction

Two-input and two-output (TITO) systems form a large class of industrial multivariable processes.These systems are often characterised by loop coupling and interactions; making the design of efficient controllers challenging [1]. PID

Y. A. Sha'aban (✉) · A. Muhammad · K. Ahmad · M. M. Jibrin
Ahmadu Bello University, Zaria, Nigeria
e-mail: yashaaban@abu.edu.ng

A. Muhammad
e-mail: masani@abu.edu.ng

K. Ahmad
e-mail: kahmad@abu.edu.ng

M. M. Jibrin
e-mail: mjmusa@abu.edu.ng

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_23,

controllers are the most popular in the industry; accounting for over 90 % of all industrial controllers [2]. These PID controllers are either implemented in a multi-loop or a decentralised fashion using decouplers. The tuning of multi-loop PID is challenging; as the loops cannot be tuned independently. Controllers are therefore loosely tuned to ensure system stability [3, 4]. This leads to inefficient performance and even marginal stability in some cases. Decentralised PID controllers on the other hand are easier to tune; the use of decouplers allow for SISO design. This allows for tighter tuning of controllers, hence resulting to more efficient performance. Significant work has been done in both multi-variable and decentralised PID controllers [5–7].

MPC is the only advanced control strategy that has had impact on the industry [8]. On multi-variable systems, MPC is traditionally implemented as a multi-variable control strategy. It deals with loop interactions systematically. The difficulty in control of large scale and networked systems has led to development of decentralised/distributed MPC (DMPC) mainly to mitigate the difficulty associated with maintenance. Most existing industrial control loops are already configured in a SISO manner. As such practitioners are generally more comfortable with SISO design. Exploring the deployment of DMPC on TITO systems may motivate more implementation of MPC on loops in which benefits are possible. Recent studies in the area of DMPC for large scale and networked systems include [9–11].

The purpose of this paper is to propose a decentralised MPC scheme for TITO systems and then compare its performance with decentralized PID and traditional MPC. The paper is therefore organised as follows: In Sect. 2, control loop interaction and coupling is discussed. MPC and a modified distributed MPC are presented in Sect. 3. PID control and decentralised PID is discussed in Sect. 4; here some methods of decoupling found in the literature are briefly highlighted. The simulation results are given in Sect. 5 and the paper is concluded in Sect. 6.

2 Control Loop Interaction and Pairing

An important characteristic feature of multi-variable systems is *process interaction*. It is used to describe the effect of each manipulated variable (MV) on all other controlled variables (CV). Various methods have been proposed to measure the degree of interaction and therefore determine the best MV–CV pairing. Consider the 2 × 2 system described by (1). The output Y_1 depends on both U_1 and U_2. Similarly, Y_2 depends on both inputs.

$$\begin{bmatrix} Y_1(s) \\ Y_2(s) \end{bmatrix} = \begin{bmatrix} g_{11}(s)e^{-s\theta_{11}} & g_{12}(s)e^{-s\theta_{12}} \\ g_{21}(s)e^{-s\theta_{21}} & g_{22}(s)e^{-s\theta_{22}} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} \tag{1}$$

where

$$g_{11}(s) = \frac{K_{11}}{\tau_{11}s + 1} \quad g_{12}(s) = \frac{K_{12}}{\tau_{12}s + 1}$$
$$g_{21}(s) = \frac{K_{21}}{\tau_{21}s + 1} \quad g_{22}(s) = \frac{K_{22}}{\tau_{22}s + 1}$$

Two commonly used methods for determining the best pairing of CV and MV are the relative gain array (RGA) and singular value decomposition (SVD) methods. More on SVD can be found in [3]. The RGA Λ is a normalised dimensionless matrix array that can be computed as follows [3]:

$$\Lambda = K \otimes H = \left[\lambda_{ij}\right] \tag{2}$$

where, $\otimes$ is the Schur product i.e. element by element multiplication, K is the gain matrix and $H = (K^{-1})^T$. So that $\lambda_{ij} = K_{ij}H_{ij}$.

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}; \quad \Lambda = \begin{bmatrix} \lambda & 1-\lambda \\ 1-\lambda & \lambda \end{bmatrix} \tag{3}$$

For a TITO system, Λ is symmetrical and can be completely specified by λ as shown in (3). The sum of the RGA elements is unity for each column and row. When $\lambda \geq 0.5$, then Y_1 should be paired with U_1(direct pairing or 1–1/2–2 pairing) otherwise it should be paired with Y_2 (reverse pairing or 1–2/2–1 pairing) [3]. However, when λ is large, then it is impossible to control both outputs independently [12] and when $\lambda < 0$, the loops may counteract each other; often leading to closed loop instability [13]. Once the loop pairing has been successfully achieved the next step involves deciding the control scheme to be used i.e. multi-variable, decentralised or advanced such as MPC.

3 Model Predictive Control

Model predictive control is a matured technology with most recent research focused on the state space formulation [8]. In MPC, the trajectory of manipulated variables is computed to optimise the future behaviour of the plant. This is achieved with the aid of predictions using plant model. These predictions are compared with actual plant outputs and adjustments made to compensate for errors. The concept of MPC is depicted in Fig. 1. In this work, the state space formulation in [14] is used. Other formulations can be found in [8, 15]. The model given in (1) can be converted to discrete state space and the augmented velocity formats (4) and (6) respectively [14].

$$\begin{aligned} x_p(k+1) &= A_p x_p(k) + B_p u(k) \\ y(k) &= C_p x_p(k) \end{aligned} \tag{4}$$

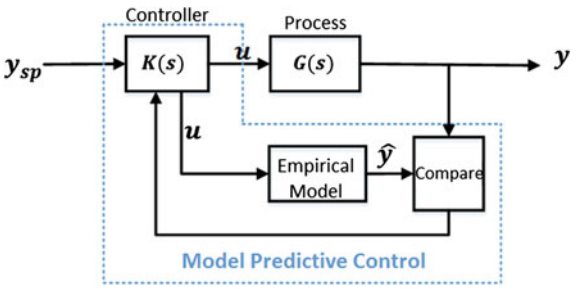


Fig. 1 Block diagram depicting the concept of model predictive control

then

$$\begin{aligned} x_p(k) &= A_p x_p(k-1) + B_p u(k-1) \\ x_p(k+1) - x_p(k) &= A_p\left[x_p(k) - x_p(k-1)\right] + B_p[u(k) - u(k-1)] \\ \Delta x_p(k+1) &= A_p \Delta x_p(k) + B_p \Delta u(k) \\ y(k+1) &= C_p x_p(k+1) \\ &= C_p\left[x_p(k) + A\Delta x_p(k) + B_p \Delta u(k)\right] \end{aligned} \tag{5}$$

Let an extended state $x(k)$ be defined as:

$$x(k) = \begin{bmatrix} \Delta x_p(k) \\ y_p(k) \end{bmatrix}, \quad then\ x(k+1) = \begin{bmatrix} \Delta x_p(k+k) \\ y_p(k+1) \end{bmatrix}$$

Then,

$$\begin{aligned} x(k+1) &= \begin{bmatrix} A_p & 0_n^T \\ C_p A_p & I \end{bmatrix} \begin{bmatrix} \Delta x_p(k) \\ y_p(k) \end{bmatrix} + \begin{bmatrix} B_p \\ C_p B_p \end{bmatrix} \Delta u(k) \\ y(k) &= \begin{bmatrix} 0_{n_p} & I_{n_{out}} \end{bmatrix} \begin{bmatrix} \Delta x_p(k) \\ y_p(k) \end{bmatrix} \end{aligned}$$

Finally, the state equations are summarized as:

$$\begin{aligned} x(k+1) &= Ax(k) + B\Delta u(k) \\ y(k) &= Cx(k) \end{aligned} \tag{6}$$

where

$A = \begin{bmatrix} A_p & 0_n^T \\ C_p A_p & I \end{bmatrix}$; $B = \begin{bmatrix} B_p \\ C_p B_p \end{bmatrix}$; $C = [0_{n_p} \quad I_{n_{out}}]$; $x(k)^T = \begin{bmatrix} \Delta x_p(k)^T & y_p(k)^T \end{bmatrix}$

and $\Delta x_p(k) = x_p(k) - x_p(k-1)$.

Considering the effects of measured disturbance $d(k)$, (4) and (6) become (7) and (8) respectively:

$$\begin{aligned} x_p(k+1) &= A_p x_p(k) + B_p u(k) + B_d d(k) \\ y_p(k) &= C_p x_p(k) \\ x(k+1) &= Ax(k) + B\Delta u(k) + B_D \Delta d(k) \end{aligned} \tag{7}$$

$$y(k) = Cx(k) \tag{8}$$

where $B_D = \begin{bmatrix} B_p \\ C_p B_p \end{bmatrix}$. A formulation of the cost function which penalizes the tracking error as well as the change in control manipulated variable is given in (9).[1]

$$J = \sum_{i=1}^{p} \|r(k+1) - y(k+i)\|_q^2 + \sum_{i=1}^{M} \|\Delta u\|_{r_w}^2 \tag{9}$$

If the initial state $x(k_1)$ is denoted by x_0, and the vectors X, ΔU and Y are defined as:

$$X = \begin{bmatrix} x(k+1) \\ x(k+2) \\ \vdots \\ x(k+N_p) \end{bmatrix} \quad \Delta U = \begin{bmatrix} \Delta u(k) \\ \Delta u(k+1) \\ \vdots \\ \Delta u(k+N_c-1) \end{bmatrix} \quad Y = \begin{bmatrix} y(k+1) \\ y(k+2) \\ \vdots \\ y(k+N_p) \end{bmatrix} \tag{10}$$

With ΔD defined in a similar manner as ΔU, the prediction equations can be written in compact form as:

$$\begin{aligned} X &= F_1 x_0 + \Phi_1 \Delta U + \Phi_{d1} \Delta D \\ Y &= F x_0 + \Phi \Delta U + \Phi_d \Delta \end{aligned} \tag{11}$$

where

$$F = \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^P \end{bmatrix} \quad \Phi = \begin{bmatrix} CB & 0 & \dots & 0 \\ CAB & CB & \dots & 0 \\ \vdots & & & \\ CA^{P-1}B & CA^{P-2}B & \dots & CA^{P-M}B \end{bmatrix}.$$

The matrix Φ_d is obtained by substituting $B = B_d$ in the definition of Φ. The cost function (9) can be written in compact form as:

$$J = (S - Y)^T \bar{Q} (S - Y) + \Delta U^T \bar{R} \Delta U \tag{12}$$

where $S^T = r(k)[\,1 \quad 1 \quad \dots \quad 1\,]$, $\bar{Q} > 0 \in R^{pP \times pP}$ is a block diagonal output weighting matrix. $\bar{R} \geq 0$ is a block diagonal input weighting matrix defined in (14). Substituting (11) in (12) gives an expression for the cost function. The optimal unconstrained control trajectory is obtained by differentiating the cost function and equating to zero:

[1] $\|x\|_P^2 = x^T P x$.

$$\Delta U = -\left(\Phi^T \bar{Q} \Phi + \bar{R}\right)^{-1} \Phi^T \bar{Q}(Fx_0 + \Phi_d \Delta D - S) \tag{13}$$

$$\bar{Q} = \begin{bmatrix} q & 0 & \dots & 0 \\ 0 & q & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & q \end{bmatrix} \quad \bar{R} = \begin{bmatrix} r_w & 0 & \dots & 0 \\ 0 & r_w & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & r_w \end{bmatrix} \tag{14}$$

The constrained MPC problem can be written as:

$$\begin{aligned} &\min_{\Delta U} \Delta U^T \left(\Phi^T \bar{Q} \Phi + \bar{R}\right) \Delta U + 2\Delta U^T \Phi^T \bar{Q}(Fx_0 + \Phi_d \Delta D - S) + constant \\ &\quad : M\Delta U \leq N \end{aligned} \tag{15}$$

Detailed derivation of prediction equations and cost function and general MPC formulation can be found in [8, 14].

3.1 Decentralised Model Predictive Control

The process $G(s)$ in (1) can be partitioned into two subsystems:

$$\begin{aligned} G_1(s) &= \begin{bmatrix} g_{11}(s)e^{-\tau_{11}(s)} & g_{12}(s)e^{-\tau_{12}(s)} \end{bmatrix} \\ G_2(s) &= \begin{bmatrix} g_{21}(s)e^{-\tau_{21}(s)} & g_{22}(s)e^{-\tau_{22}(s)} \end{bmatrix} \end{aligned} \tag{16}$$

These sub-systems can be converted to discrete state space format with the second input as a measured disturbance and first input as a measured disturbance in the first and second subsystems respectively. The resulting subsystems represented by (17) are only coupled through their inputs i.e. state couplings do not exist. Note that it is always possible to bring any system to this format [10].

$$\begin{aligned} x_{pi}(k+1) &= A_{p_i} x_{p_i}(k) + Bp_i u_i(k) + B_{d_i} u_j(k) \\ y_i(k) &= c_{p_i} x_{p_i} \\ i,j &= 1,2, \quad i \neq j \end{aligned} \tag{17}$$

The velocity augmented form model as formulated in (6) can then be formed as:

$$\begin{aligned} x_i(k+1) &= A_i x_i(k) + B_i \Delta u_i(k) + B_{D_i} \Delta u_{d_j}(k) \\ y_i(k) &= C_i x_i(k) \end{aligned} \tag{18}$$

The prediction equations then become:

$$X_i = F_{1i} x_{i0} + \Phi_{1i} \Delta U_i + \Phi_{d_i} \Delta D_i \tag{19}$$

$$Y_i = F_i x_{i0} + \Phi_i \Delta U_i + \Phi_{D_i} \Delta D_i \tag{20}$$

With all parameters defined as in MPC formulation earlier presented in this section and ΔD_i defined as follows:

$$\Delta D_i = \begin{bmatrix} \Delta u_j(k) \\ \Delta u_j(k+1) \\ \vdots \\ \Delta u_j(k+M-1) \end{bmatrix} \tag{21}$$

The prediction equation and MPC law for each of the sub-systems can be derived with Δu_2 and Δu_1 as disturbances in subsystems 1 and 2 respectively. Iteration is used to obtain optimal solutions. However, due to time requirements, the system only converges to the *pareto-optima* or *Nash* equilibrium [16]. To prevent iteration as with traditional distributed MPC, the computed input trajectory at sampling step k is defined as:

$$\Delta \hat{D}_i(k) = \begin{bmatrix} \Delta \hat{u}_j(k) \\ \Delta \hat{u}_j(k+1) \\ \vdots \\ \Delta \hat{u}_j(k+M-1) \end{bmatrix} \tag{22}$$

Then since at sampling step $k+1, \Delta \hat{D}_i$ for $i = 1, 2$ will not be available, it will be assumed that the value computed at k is still optimal. Hence, the sequence is shifted and the value at the end of the control horizon repeated[2] i.e.

$$\Delta \hat{D}_i(k+1) = \begin{bmatrix} \Delta \hat{u}_j(k+1) \\ \vdots \\ \Delta \hat{u}_j(k+M-2) \\ \Delta \hat{u}_j(k+M-1) \\ \Delta \hat{u}_j(k+M-1) \end{bmatrix} \tag{23}$$

So that (20) is written as:

$$Y_i = F_i x_{i0} + \Phi_i \Delta U_i + \Phi_{Di} \Delta \hat{D}_i(k) \tag{24}$$

The problem then becomes that of solving two quadratic programming problems, one for each subsystem:

$$\begin{aligned} &\min_{\Delta U_i} \left\{ \Delta U_i^T \left(\Phi_i^T \bar{Q}_i \Phi_i + R_i \right) \Delta U_i + 2 \Delta U_i^T \Phi_i^T \bar{Q}_i \left(F_i x_{i0} + \Phi_{Di} \Delta \hat{D}_i - S \right) \right\} \\ &\quad : M_i \Delta U_i \leq N_i : i = 1, 2 \end{aligned} \tag{25}$$

In this formulation, iteration is not required as in other DMPC implementations. This will reduce the computational and convergence requirements needed by other

[2] In MPC, it is assumed that the MV remains constant at the end of the horizon.

decentralised MPC formulations. However, when loop interaction is strong the performance of the scheme decreases significantly.

4 PID

PID control is the most popular in the industry. Different structures of the controller are available. In this work the ideal (or parallel structure) is used. For proportional, integral and derivative gains of k_p, k_i and k_d respectively, the parallel PID is given in Eq. (26)

$$k(s) = k_p + \frac{k_i}{s}s + k_d s \tag{26}$$

4.1 Decoupling Control

Decouplers are commonly used to reduce loop interactions; thereby reducing the effect of set-point change for one controlled variable on other controlled variables. Several decoupling techniques have been presented in the literature. The TITO system presented in (1) can be written as:

$$\begin{aligned} Y(s) &= G(s)U(s) \\ \begin{bmatrix} Y_1(s) \\ Y_2(s) \end{bmatrix} &= \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} \end{aligned} \tag{27}$$

For a 1–1/2–2 pairing, the aim of the decoupler is to effectively cancel the off-diagonal dynamics. So that the effective transfer function $Q(s)$ after decoupling using a decoupling matrix $D(s)$ is given in (28).

$$\begin{aligned} Q(s) &= G(s)D(s) \\ &= \begin{bmatrix} q_{11} & 0 \\ 0 & q_{22} \end{bmatrix} \end{aligned} \tag{28}$$

$$D(s) = \begin{bmatrix} 1 & d_{12}(s) \\ d_{21}(s) & 1 \end{bmatrix} \tag{29}$$

Hence, the parameters d_{12} and d_{21} can be computed using (27)–(29) as follows:

$$d_{12} = -\frac{G_{12}}{G_{11}} \qquad d_{21} = -\frac{G_{21}}{G_{22}} \tag{30}$$

The decoupler parameters d_{12} and d_{21} are not always physically realisable. For simplicity, *static decoupling* [17] is often used. This involves using steady state model i.e.

$$d_{12} = -\frac{K_{12}}{K_{11}} \quad d_{21} = -\frac{K_{21}}{K_{22}} \tag{31}$$

When loop dynamics are not similar, static decoupling has the disadvantage that loop interactions occur during transients. After decoupling, the system $Q(s)$ can then be controlled using a decentralised PID controller $K(s)$. To improve the transient performance, more advanced decouplers are used. Two of such presented in [4, 18] are briefly presented.

$$K(s) = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix} \tag{32}$$

where $k_1(s)$ and $k_2(s)$ are of PID/PI type in (26).

4.1.1 PID with Lead-Lag Decoupler (*Wang-PID*)

An auto tuned PID was presented in [18]. This uses a lead-lag decoupler with parameters d_{12} and d_{21} in (29) obtained follows:

$$d_{12} = -\frac{K_{12}}{K_{11}} \frac{1 + \tau_{11} s}{1 + \tau_{12} s} e^{-(\theta_{12} - \theta_{11})} \tag{33}$$

$$d_{21} = -\frac{K_{21}}{K_{22}} \frac{1 + \tau_{22} s}{1 + \tau_{21} s} e^{-(\theta_{21} - \theta_{22})} \tag{34}$$

The terms d_{12} and d_{21} are not physically realisable when $(\theta_{12} - \theta_{11}) < 0$ or $(\theta_{21} - \theta_{22}) < 0$. However, they can be made realizable by multiplying the corresponding columns by $e^{(\theta_{21} - \theta_{22})}$ and $e^{(\theta_{12} - \theta_{11})}$ respectively [18]. The lead-lag decoupling matrix is therefore given as:

$$D(s) = \begin{bmatrix} e^{v(\theta_{22} - \theta_{21})} & -\frac{g_{12}}{g_{11}} e^{v(\theta_{12} - \theta_{11})} \\ -\frac{g_{21}}{g_{22}} e^{v(\theta_{21} - \theta_{22})} & e^{v(\theta_{22} - \theta_{21})} \end{bmatrix} \tag{35}$$

where

$$v(\theta) = \begin{cases} 1 & \text{if } \theta \geq 0 \\ 0 & \text{if } \theta < 0 \end{cases}$$

Another factor which affects the performance is the existence of right half poles and zeroes. The next section considers the modification of $D(s)$ to cater for the existence or otherwise of RHP poles/zeroes.

4.1.2 PID with Non-dimensional Tuning (*NDT-PID*)

Three cases are identified and associated decouplers designed [4]. These include:

1. *Case I:* This is when the off-diagonal elements of the plant model have no RHP-poles and the diagonal elements have no RHP-zeros:

$$D(s) = \begin{bmatrix} w_1(s) & d_{12}(s)w_2(s) \\ d_{21}(s)w_1(s) & w_2(s) \end{bmatrix} \tag{36}$$

then,

$$w_1(s) = \begin{cases} 1 & \text{if } \theta_{21} \geq \theta 22 \\ e^{(\theta_{21}-\theta_{22})} & \text{if } \theta_{21} < \theta_{22} \end{cases} \tag{37}$$

$$w_2(s) = \begin{cases} 1 & \text{if } \theta_{12} \geq \theta 11 \\ e^{(\theta_{12}-\theta_{11})} & \text{if } \theta_{12} < \theta_{11} \end{cases} \tag{38}$$

$$\begin{aligned} d_{12}(s) &= -\frac{g_{12}}{g_{11}} e^{-(\theta_{12}-\theta_{11})} \\ d_{21}(s) &= -\frac{g_{21}}{g_{22}} e^{-(\theta_{21}-\theta_{22})} \end{aligned} \tag{39}$$

   This corresponds to the lead-lag decoupler presented in (35).

2. *Case II:* This is when there are no RHP-poles in diagonal and no RHP-zeros in the off-diagonal elements of the plant model:

$$D(s) = \begin{bmatrix} d_{11}(s)w_3(s) & w_3(s) \\ w_4(s) & d_{22}(s)w_4(s) \end{bmatrix} \tag{40}$$

$$w_3(s) = \begin{cases} 1 & \text{if } \theta_{22} \geq \theta 21 \\ e^{(\theta_{22}-\theta_{21})} & \text{if } \theta_{22} < \theta_{21} \end{cases} \tag{41}$$

$$w_4(s) = \begin{cases} 1 & \text{if } \theta_{11} \geq \theta 12 \\ e^{(\theta_{11}-\theta_{12})} & \text{if } \theta_{11} < \theta_{21} \end{cases} \tag{42}$$

$$\begin{aligned} d_{11}(s) &= -\frac{g_{22}}{g_{21}} e^{-(\theta_{22}-\theta_{21})} \\ d_{22}(s) &= -\frac{g_{11}}{g_{12}} e^{-(\theta_{11}-\theta_{12})} \end{aligned} \tag{43}$$

3. *Case III:* This is when both the diagonal and non-diagonal elements of $G(s)$ have RHP-zeros. Then it is not possible to obtain a stable decoupler using Eq. (39) or (43). The solution to this is beyond the scope this work, details can be found in [4].

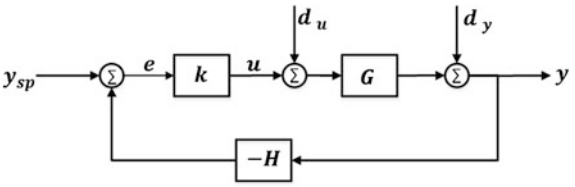


Fig. 2 Block diagram of control problem showing controller, k, plant model, G, inputs and outputs

5 Simulation and Results

In this work, an example is used to compare performance of controllers in terms of set-point y_{sp} and output disturbance d_y. Refer to Fig. 2 for block diagram representation of the control problem; in this work, the input load disturbance d_u is assumed to be zero and the sensor has a transfer function of $H = 1$. To evaluate the performance of the proposed decentralised MPC, it is applied to a well studied process model. The Wood-Berry binary distillation column process is a TITO system that has been studied extensively [4, 18, 19]. The model is given as [3]:

$$\begin{bmatrix} X_D(s) \\ X_B(s) \end{bmatrix} = \begin{bmatrix} \frac{12.8e^{-s}}{16.7s+1} & \frac{-18.9e^{-3s}}{21s+1} \\ \frac{6.6e^{-7s}}{10.9s+1} & \frac{-19.4e^{-3s}}{14.4s+1} \end{bmatrix} \begin{bmatrix} R(s) \\ S(s) \end{bmatrix} \tag{44}$$

where, $X_D(s)$ and $X_B(s)$ are the overhead and bottom composition respectively, $R(s)$ and $S(s)$ are the reflux flow rate and steam flow respectively. The system is strongly coupled, simultaneous control of the compositions is therefore challenging. The relative gain array (RGA) shows that the process is interacting:

$$\begin{aligned} \Lambda &= K \otimes H \\ &= \begin{bmatrix} 2.0094 & -1.0094 \\ -1.0094 & 2.0094 \end{bmatrix} \end{aligned}$$

$\Lambda_{11} > 1$ indicates coupling. The values of decoupling matrices and PID/PI controllers obtained by the discussed methods are given. For this example, the same decoupler $D(s)$ is used for both Wang-PID and NDT-PI.

$$D(s) = \begin{bmatrix} 1 & \frac{1.477(16.7s+1)e^{-2s}}{21s+1} \\ \frac{0.34(14.4s+1)e^{-4s}}{10.9s+1} & 1 \end{bmatrix} \tag{45}$$

The controllers are given as [4]:

$$K_{NDT} = \begin{bmatrix} 0.41 + \frac{0.074}{s} & 0 \\ 0 & -0.12 - \frac{0.024}{s} \end{bmatrix} \tag{46}$$

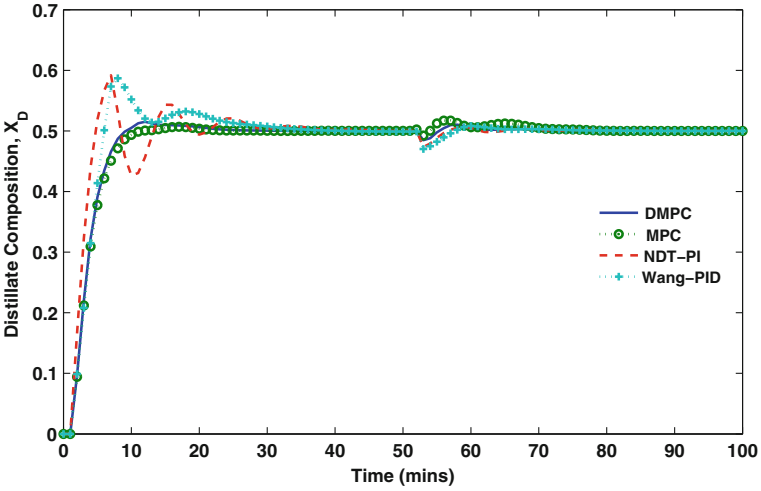


Fig. 3 Set-point response of distillate composition DMPC (*solid*), MPC (*dotted-circle*) and NDT-PI (*dashed*) and Wang-PID (*dashed-cross*)

$$K_{Wang} = \begin{bmatrix} 0.216 + \frac{0.076}{s} + 0.017s & 0 \\ 0 & -0.068 - \frac{0.019}{s} - 0.064s \end{bmatrix} \tag{47}$$

Model predictive control was also implemented on the process. A sampling period of $T_s = 1$, prediction horizon of $P = 20$, and a control horizon of $M = 4$ were used. An input weighting matrix of $r_w = diag\{\ 10 \quad 100\ \}$ was also used. The proposed DMPC was also implemented using the following parameters; a sampling period of $T_s = 1$, a prediction horizon of $P = 20$, a control horizon of $M = 4$ for both loops. The input weightings used were $r_{w_1} = 10$ and $r_{w_2} = 100$. The designed controllers were then implemented to compare their set-point tracking and output disturbance rejection performance. A step reference of 0.5 was applied to the first loop at time $t = 0$ and the second loop at time $t = 50$ min. The results of these are given in Figs. 3 and 4. Output step disturbance of 0.5 was also applied to the first and second loops at times $t = 0$ and $t = 50$ min respectively. The resulting plots are also given in Figs. 5 and 6.

The mean squared error between the set-point and the output obtained with the various controllers are given in Table 1. For set-point tracking, the loop interaction for DMPC and MPC in the first loop is smaller than that of the PID/PI controllers. In the second loop the interaction is more pronounced in the DMPC and MPC. This is due to the weighting on the second loop which is deliberately made larger. Typically in an industrial setting, the purity of the distillate is more important. It can be seen from Table 1 that the MSE of both loops is lower for the DMPC as well as MPC; both have a similar performance.

For output disturbance rejection. The MSE for the distillate shows that DPMC and MPC outperforms NDT-PI and Wang-PID. MPC outperform all the others in

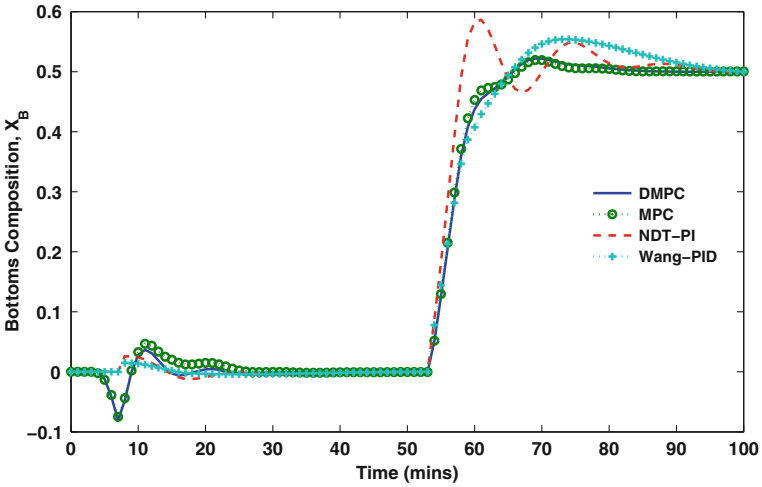


Fig. 4 Set-point response of bottoms composition for DMPC (*solid*), MPC (*dotted-circle*) and NDT-PI (*dashed*) and Wang-PID (*dashed-cross*)

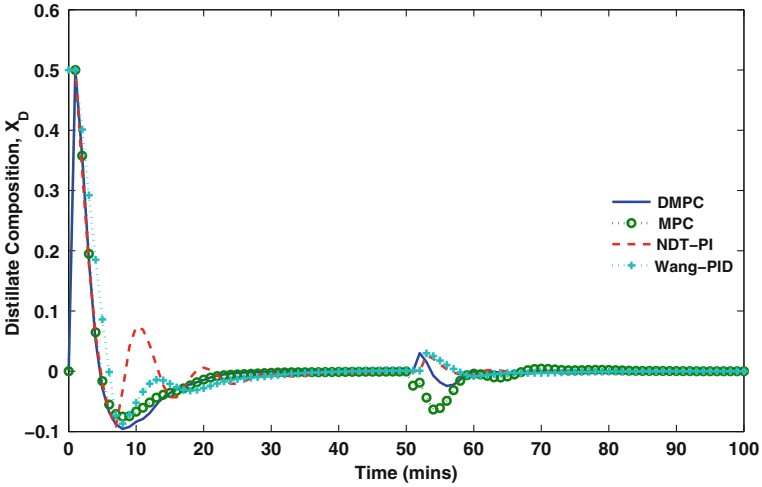


Fig. 5 Output disturbance response of distillate composition for DMPC (*solid*), MPC (*dotted-circle*) and NDT-PI (*dashed*) and Wang-PID (*dashed-cross*)

bottoms, with Wang-PID having the highest value of MSE. Results indicate that the TITO DMPC performs at least better than the PID/PI controllers used in this problem. Improved performance is expected when process dead times are larger and when constraints are imposed on the process. The total variance (TV) is used to measure the smoothness of control; smaller values indicate lower control

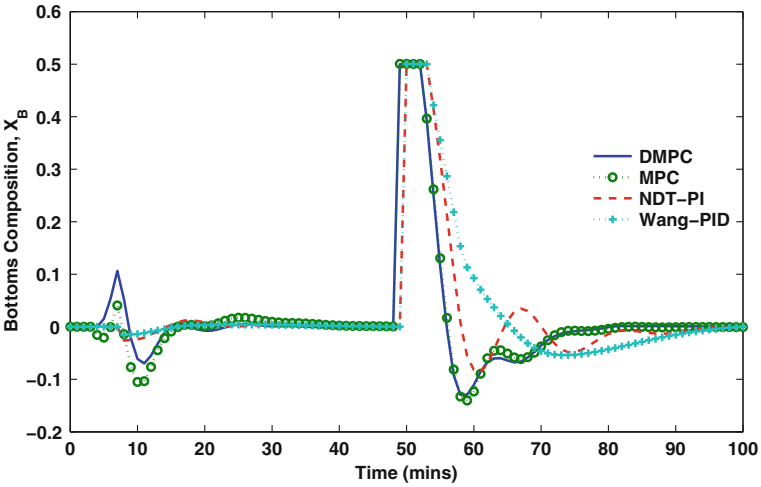


Fig. 6 Output disturbance response of bottoms composition DMPC (*solid*), MPC (*dotted-circle*) and NDT-PI (*dashed*) and Wang-PID (*dashed-cross*)

Table 1 Mean squared error (MSE) for both set-point tracking and disturbance rejection

Controller	Set-point response		Disturbance rejection	
	X_D	X_B	$X_D \times 10^{-3}$	$X_B \times 10^{-2}$
MPC	0.2414	0.1087	4.7006	1.3415
DMPC	0.2411	0.1088	4.6867	1.3647
NDT-PI	0.2438	0.1170	6.7597	1.3616
Wang-PID	0.2454	0.1131	8.1202	1.5143

Table 2 Total variance (TV) for both set-point tracking and disturbance rejection

Controller	Set-point response		Disturbance rejection	
	X_D	X_B	X_D	X_B
MPC	0.3954	0.1801	0.7079	0.3127
DMPC	0.3824	0.1076	0.6103	0.3041
NDT-PI	0.6228	0.1806	0.6228	0.1806
Wang-PID	0.2647	0.1105	0.2647	0.1185

activity. It can be observed from Table 2 that for set-point tracking, Wang-PID has the smallest value while NDT-PI has the largest value. MPC and DMPC have similar values. For disturbance rejection, PID/PI controllers have slightly lower values. These higher values of TV in MPC/DMPC can be justified by the improved performance. TV is computed using (48)

$$TV = \sum_{k=0}^{\infty} \|u(k+1) - u(k)\| \tag{48}$$

6 Conclusions

A DMPC for TITO processes is proposed. Shifted input sequence from the previous step is used to cater for loop interactions thereby avoiding iterations. The performance of the proposed scheme was compared with MPC and PID/PI by applying to a coupled processes. It was found that the proposed controller has a performance similar to that of existing methods.

References

1. Y.A. Sha'aban, A. Muhammad, K. Ahmad, M.M. Jibrin, in A comparative study of SISO control for TITO systems, *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013, WCE 2013*, London, U.K., pp. 1024–1028, 3–5 July 2013
2. K.J. Åström, T. Hägglund, in *Automatic Tuning of PID Controllers*. Instrument Society of America (Research Triangle Park, North Carolina, 1988)
3. D.E. Seborg, *Process Dynamics and Control* (Wiley, London, 2011)
4. S. Tavakoli, I. Griffin, P.J. Fleming, Tuning of decentralised PI (PID) controllers for TITO processes. Control Eng. Practice **14**, 1069–1080 (2006)
5. J. Garrido, F. Vázquez, F. Morilla, Centralized multivariable control by simplified decoupling. J. Process Control **22**, 1044–1062 (2012)
6. P. Nordfeldt, T. Hägglund, Decoupler and PID controller design of TITO systems. J. Process Control **16**, 923–936 (2006)
7. Q.-G. Wang, B. Zou, T.-H. Lee, Q. Bi, Auto-tuning of multivariable PID controllers from decentralized relay feedback. Automatica **33**, 319–330 (1997)
8. J.M. Maciejowski, *Predictive Control: With Constraints* (Pearson Education, Upper Saddle River, 2002)
9. A. Alessio, D. Barcelli, A. Bemporad, Decentralized model predictive control of dynamically coupled linear systems. J. Process Control **21**, 705–714 (2011)
10. I. Alvarado, D. Limon, D.M. de la PeAa, J. Maestre, M. Ridao, H. Scheu, W. Marquardt, R. Negenborn, B.D. Schutter, F. Valencia, J. Espinosa, A comparative analysis of distributed MPC techniques applied to the HD-MPC four-tank benchmark. J. Process Control **21**, 800–815 (2011)
11. B.T. Stewart, A.N. Venkat, J.B. Rawlings, S.J. Wright, G. Pannocchia, Cooperative distributed model predictive control. Syst. Control Lett. **59**, 460–469 (2010)
12. S. Skogestad, I. Postlethwaite, *Multivariable Feedback Control: Analysis and Design* (Wiley, London, 2005)
13. F.G. Shinskey, *Process Control Systems: Application, Design, and Tuning* (McGraw-Hill, New York, 1996)
14. L.A. Wang, Tutorial on model predictive control using a linear velocity-form model. Develop. Chem. Eng. Mineral Process. **12**, 573–614 (2004)
15. S. Qin, T.A. Badgwell, A survey of industrial model predictive control technology. Control Eng. Practice **11**, 733–764 (2003)

16. R. Scattolini, Architectures for distributed and hierarchical model predictive control. A Rev. J. Process Control **19**, 723–731 (2009)
17. T.J. McAvoy, Steady-state decoupling of distillation columns. Ind. Eng. Chem. Fundam. **18**, 269–273 (1979)
18. Q.-G. Wang, B. Huang, X. Guo, Auto-tuning of TITO decoupling controllers from step tests. ISA Trans. **39**, 407–418 (2000)
19. R. Wood, M. Berry, Terminal composition control of a binary distillation column. Chem. Eng. Sci. **28**, 1707–1717 (1973)

Fuzzy-Logic Based Computation for Parameters Identification of Solar Cell Models

Toufik Bendib and Fayçal Djeffal

Abstract The identification of the electrical parameters of the organic solar cells, such as the series resistance, the shunt resistance, the diode saturation current and the diode ideality factor, is an important task to improve their models behavior and the time simulation for photovoltaic applications. The conventional extraction methods using the optimization and measurement techniques, which are based on calculating derivatives, are quite computationally expensive and difficult to code. Therefore these parameters are required new optimization and modeling methods that capture the effect of each model parameter accurately and efficiently. In the present work, a new, fast and accurate organic solar cell extraction technique using Fuzzy-Logic-based computation is presented. This approach is based on fuzzy control techniques. These techniques allow using knowledge about the model behavior into the parameter extraction method, thus simplifying the task. The procedure is applied to extract the different parameters of a single-diode solar cell model for which results show good performances. The encouraging results have indicated the applicability of the developed approach to be incorporated in solar cell simulator tools for photovoltaic applications.

Keywords Accuracy · Circuit · Extraction · Fuzzy logic · Identification · Inference · Solar cell

T. Bendib · F. Djeffal (✉)
LEA Department of Electronics, University of Batna, 05000 Batna, Algeria
e-mail: faycaldzdz@hotmail.com; faycal.djeffal@univ-batna.dz

T. Bendib
e-mail: bendib05.t@gmail.com; toufikdzdz@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_24,

1 Introduction

In recent years, the strong demand for renewable energy has increased interest in solar cells as a long-term, exhaustless, environmentally friendly and reliable energy technology [1, 2]. Various intensive research efforts have been devoted to develop new materials and modeling techniques for solar cells are being made in order to produce new photovoltaic devices with improved electrical performances.

During the last few years, the production of the organic solar cells is rising in photovoltaic domain; it is due to the simple production technology such as: roll-to-roll or printing process. This fact by itself can reduce the cost production significantly [3–7]. Moreover, the importance of these kinds of solar cells with lower thermal price and less stringent requirements in comparison with conventional inorganic semiconductor technology leads to more investigate the organic solar cells and thinking to improve their electrical characteristic. Therefore, the tailoring of molecular proprieties and the versatility of production methods of the organic solar cell make it as a promising candidate for fabrication the future generation of solar cells for photovoltaic applications.

Compared to other inorganic counterparts, it shows low efficiency. Thus, more study needs to be carried out on the organic solar cells, while keeping an eye on improving their conversion efficiencies. Accurate extraction and optimization of organic solar cell parameters are very important in improving the solar cell quality during fabrication process and in device simulation and modeling [5]. Extraction of the organic solar cell parameters is a complex task, because of the huge number of parameters in recent models. These parameters are, usually, the series resistances R_s, the shunt resistances R_{sh}, the cell-generated photocurrent I_{ph}, the diode saturation current I_0 and the diode ideality factor n witch describe the nonlinear electrical model of the organic solar cell.

Most of these parameters are correlated, and requires global extraction and optimization methods [8, 9]. One way to simplify this task is to use direct extraction methods for some parameters. This eases the entire extraction procedure and reduces the iteration time in case of optimization, because these values can be used as initial values. Once the parameters have been extracted, most of the direct extraction methods need a second step to take into account the interactions among the different parameters. This leads to the use of global methods (Genetic algorithms (GA), Particle Swarm Optimization (PSO), Multi-Objective-Genetic Algorithms (MOGAs), etc.) to find the set of values that can best fit the experimental data [8–10]. Numerous authors are studied and validated the applicability of these techniques for photovoltaic applications [11, 12]. Therefore, the accuracy of these techniques is limited by the high computational time caused by global optimization processing and the constraints functions used for parameter extraction [8, 9]. The other techniques for parameter extraction rely on the use of fitting algorithms to determine the solar cell parameters. Their accuracy depends on the

applied fitting algorithm [8]. There are other methods such as finding parameters through complicated numerical analysis using Lambert's *W* function [13, 14]. However, these methods seem complicated processes for accurate extraction of parameters from a non linear solar cell characteristic. These models are obtained by simplifications of the full model of the organic solar cell which lead to the applicability limitations and a questionable accuracy. Zhang et al. [14] proposed a fitting method using Lambert's *W* function to study the properties of solar cells. However, their study is validated only on single diode model, where this method cannot be extended to study the double diode model. Moreover, this method is used to extract only three parameters n, R_s and R_{sh}. The other parameters I_0 and I_{ph} are deduced using analytic expressions. Therefore, the development of new approaches to overcome these limitations should be developed in order to study the solar cells for wide illumination range, where other parameters and double diode modeling should be taken into account.

But from the photovoltaic circuit design point of view even 2-D solution of numerical Lambert's *W* function is an overkill approach in term of both complexity and computational cost [13–15]. Moreover, the organic solar cell introduces challenges to analytical parameters extraction using the *I–V* characteristics. In addition, in most studies, the *I–V* curve obtained from extracted parameters was not compared with a measured *I–V* curve, making it difficult to demonstrate the accuracy of the methods used. Accurate parameters extraction methods are required to be utilized in photovoltaic system simulators and circuit design tools.

Fuzzy logic is the widely used technique in control applications [16]. In comparison with the other techniques, the main advantage of this technique is that: we don't need an accurate description of the behaviour of the model, but we are only required to know the effect of each parameter on the behaviour. For example, we only need to know that the shift of the *I–V* curve to the right in the voltage region is performed by increasing the value of the shunt resistance and vice versa.

In the present work, we present a simple and accurate method to extract the different electrical parameters of the organic solar model that uses fuzzy logic, based on a behavioural computation in order to reduce the human trial and error efforts. The method is applied to a one-diode solar cell model, in order to show the feasibility of the extraction method. In this context, in order to achieve the required accuracy and method simplicity, in this work we present the applicability of Fuzzy Logic (FL) computation approach to extract the organic solar cell parameters. The correlation between the *I–V* curves drawn using final obtained parameter values and the measured *I–V* curves was confirmed by the least-squares computation. With this ability, we can use our proposed approach as the interface between the experimental setup and photovoltaic simulator (like PC1D and Silvaco) [17], in order to achieve high computation speed and improve the simulation time for photovoltaic simulation.

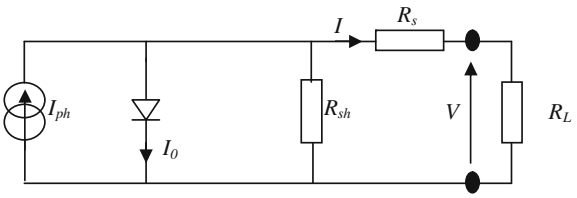


Fig. 1 Equivalent circuit model of the organic solar cell

2 Fuzzy Logic Computation Methodology

2.1 Solar Cell Model

The model of the organic solar cell can be represented by the simplified equivalent circuit as shown in Fig. 1, and expressed by the lumped parameter of a one-diode model solar cell as Eq. 1 which has been suggested in early 1980s [7].

For a given incident light intensity, at a given temperature, the implicit *I–V* analytical model is as follows:

$$I = I_{ph} - \frac{V + R_s I}{R_{sh}} - I_0[exp(\beta(V + R_s I)) - 1] \tag{1}$$

where $\beta = q/n_1 kT$.

I_{ph} represents the total current generated by the cell for a given lighting and temperature conditions in Amperes, k is the Boltzmann constant, q is the electron charge, T is the temperature, R_s is the series resistance, R_{sh} is the shunt resistance, I_0 represents the reverse saturation current; n is the diode ideality factor.

The processes associated with various compounds in the equivalent circuit of an organic solar cell are: The photo-current source generates current which is equal to number of dissociated excitons/s (number of free electron/hole pairs per second) [18]. The shunt resistance is due to recombination of charge carriers near the dissociation site (e.g. donor/acceptor interface) and it may also include recombination farther away from the dissociation site (e.g. near electrode). The series resistance reflects conductivity, i.e. mobility of specific charge carrier in the respective transport medium, where the mobility is affected by space charges and traps or other barriers (hopping) [18]. It is to note that the analysis of the processes associated with various elements of the equivalent circuit allows, to the designer, the control of each design parameter to develop the desired organic solar cell. The purpose of this work is to use the Fuzzy Logic (FL) technique to extract the parameters of organic solar cells. Furthermore, the method validation is done by calculating the errors in the curves obtained using the estimated parameters with respect to those curves obtained experimentally.

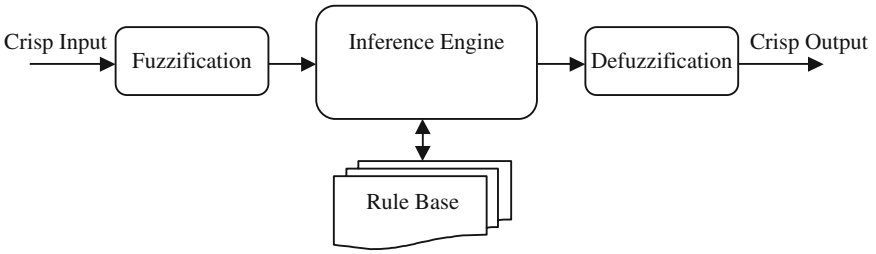


Fig. 2 Fuzzy inference system

2.2 Fuzzy Logic Architecture

This section introduces the basic concept and the operations for fuzzy logic system that will be needed in the following section.

The Fuzzy Logic tool was introduced in 1965, also by Zadeh [19], and is a mathematical tool for dealing with uncertainty. It offers to a soft computing partnership the important concept of computing with words.

Fuzzification, decision-making, fuzzy rules base and defuzzification are the four steps of fuzzy system as shown in Fig. 2.

During fuzzification, crisp inputs are transformed into degrees of matching linguistic values and are related to the input linguistic variables. Subsequently, as the fuzzification process is completed, the inference engine which is the core unit and is also known as decision-making refers to the fuzzy rule base containing fuzzy IF-THEN rules to deduct the linguistic values for the intermediate and output linguistic variables. It is important to note that the database defines the membership functions of the fuzzy sets which is used in the fuzzy rules. Finally, when the output linguistic measures are obtainable, the defuzzifier produces the final crisp values from the output linguistic values (fuzzy results).

In general, based on the past known behavior of a target system, we can design a fuzzy inference system to be expected in order to reproduce the behavior of the target system. For example, the case of our work, the target system is the experimental *I–V* characteristic, then the fuzzy inference system becomes a fuzzy logic parameters controller that can regulate and fit the calculated *I–V* characteristic of the organic solar cell.

2.3 Fuzzy Logic Implementation

The fuzzy logic provides an inference structure that enables appropriate human reasoning capabilities. On the contrary, the traditional binary set theory describes crisp events, events that either do or do not occur. It uses probability theory to explain if an event will occur, measuring the chance with which a given event is

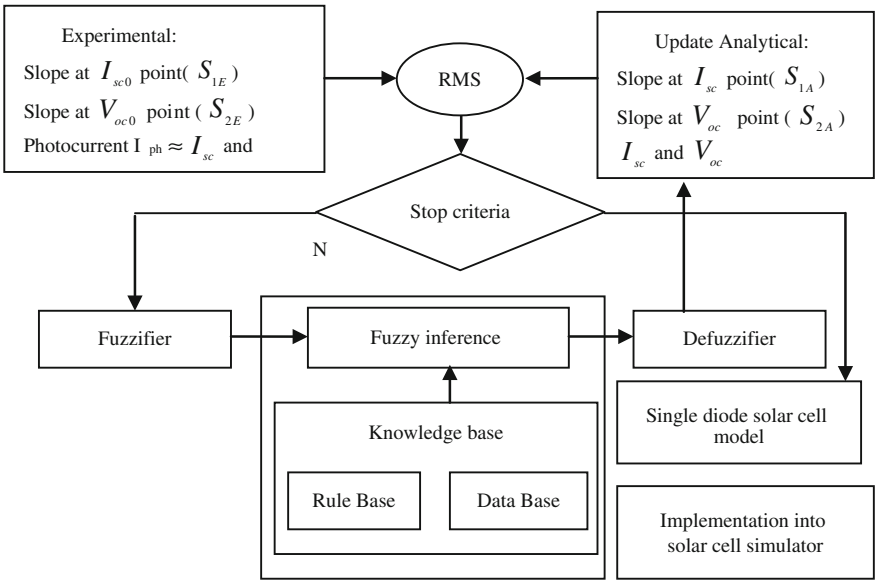


Fig. 3 Fuzzy Logic controller block diagram of our computation approach

expected to occur [20]. Several published papers offer extensive explanations to the fundamentals of FL and its wide range of applications [19–22]. Due to its simple mechanism and high performance for behavior modeling, FL can be applied to study the *I–V* organic solar behavior in order to extract their electrical parameters [23], which is the main objective of this work. The flowchart of our proposed approach is detailed in Fig. 3.

To define the fuzzy associative memory, we need some knowledge about the how each of the electrical parameters affects the *I–V* curve behavior. *I–V* curves were drawn using MATLAB software. Changes in the shape of the I–V curve due to changes in parameter values were displayed in real time by the continuous execution of MATLAB program, allowing easy verification of visual effects of specific parameters on the shape of *I–V* curve. First, the shape of the *I–V* curve in the voltage region is depressed horizontally with a gradual increase in the value of R_s from zero. When R_{sh} decreases from infinity, the shape of the *I–V* curve in the current region is depressed leftward. The distinct change in the shape of the *I–V* curve due to changes in the ideality factor and reverse saturation current is that the open circuit voltage of the solar cell changes. Therefore, if parameters are extracted while assuming the ideality factor to be 1, there is significant error in the *I–V* curve drawn using the parameters [13].

As it is explained, previously, the *I–V* curve behavior will be used to define the fuzzy associative memory using some knowledge about the effect of each parameter on the *I–V* behavior:

- R_s describes the curvature of *I–V* curve in the voltage region and shifts the short circuit current (I_{sc}) value.
- R_{sh} describes the curvature of *I–V* curve in the current region and shifts the open circuit voltage (V_{co}) value.
- n and I_0 shift the open circuit voltage (V_{co}) value.

Based on the effect of each parameter on the *I–V* curves, we can define the following rules in order to develop our knowledge Base (Fig. 3) for each parameter, which will be extracted:

- If the calculated *I–V* characteristic is more curved than the experimental data in the voltage region, then increase R_s, and vice versa.
- If the calculated curve *I–V* is under the experimental data at I_{sc} point, then decrement R_s, and vice versa.
- If the calculated *I–V* characteristic is more curved than the experimental data in the current region, then increase R_{sh}, and vice versa.
- If the calculated curve *I–V* is to the right of the experimental data in the voltage region, then decrement R_{sh}, and vice versa.
- If the calculated curve *I–V* is to the right of the experimental data in the voltage region, then decrement n, and vice versa.
- If the calculated curve *I–V* is to the right of the experimental data in the voltage region, then increment I_0, and vice versa.

It is to note that the curvature of the *I–V* characteristics for both regions, current and voltage regions, can be estimated from the computation of the slopes of the calculated *I–V* characteristic at the short circuit current point for the first region and at the open circuit voltage point for the second one (Fig. 3). So, the first and second fuzzy rules, about R_s andR_{sh}, can be replaced by:

- If the slope of the calculated *I–V* curve (S_{2A}) at V_{oc} point is too small than the experimental data (S_{2E}), then decrease R_s, and vice versa.
- If the short circuit current of the calculated *I–V* curve (I_{sc}) is too small than the experimental short circuit current (I_{sc0}), then decrease R_s, and vice versa.
- If the slope of the calculated *I–V* curve (S_{1A}) at I_{sc} point is too small than the experimental data (S_{1E}), then decrease R_{sh}, and vice versa.
- If the open circuit voltage of the calculated *I–V* curve (V_{oc}) is too small than the experimental open circuit voltage (V_{oc0}), then increase R_{sh}, and vice versa.

In order to implement the above rules, we have used the Gaussian fuzzy sets, while defuzzification is done through the method of centre of area. The input and output parameters are normalized by using the experimental database as reference. The linguistic variables chosen for our FL-based approach is the mean square error (MSE) for each parameter of the *I–V* model. These errors are the input linguistic variables and the *I–V* model is the finale output linguistic variable. The error of each parameter represents the current deviation affected by the solar cell parameter, which will be extracted using the proposed approach. The fuzzy system is then expected to be able to reproduce the behavior of the experimental curves.

Table 1 Fuzzy associate memory table (FAM) for the fuzzy R_{sh}, R_s,n and I_0 controllers, respectively

		Input1 ΔS_1							
		NL	*NM*	*NS*	*Z*	*NS*	*PM*	*PL*	
Input2 ΔV_{oc}	*N*	*NL*	*NM*	–	*NS*	–	*PS*	*PM*	Output R_{sh}
	Z	*NM*	*NM*	–	*Z*	–	*PM*	*PM*	
	P	*NM*	*NS*	–	*PS*	–	*PM*	*PL*	
		Input1 ΔS_2							
		NL	*NM*	*NS*	*Z*	*NS*	*PM*	*PL*	
Input2 ΔI_{sc}	*N*	*NL*	*NM*	*NS*	*Z*	*NS*	*PM*	*PL*	Output R_s
	Z	*PL*	*PM*	–	*PS*	–	*NS*	*NM*	
	P	*PM*	*PM*	–	*Z*	–	*NM*	*NM*	
		Input1 ΔV_{oc}							
		NL	*N*	*NS*	*Z*	*NS*	*PM*	*PL*	
		PL	*PM*	*PS*	*Z*	*NL*	*NM*	*NL*	Output1 n
		NL	*NM*	*NS*	*Z*	*NS*	*PM*	*PL*	Output2 I_0

Each of the input and output fuzzy variables is assigned seven linguistic fuzzy subsets varying from negative large (NL) to positive big (PL). Each subset is associated with a Gaussian membership function to form a set of seven membership functions for each fuzzy variable. The linguistic terms chosen for this controller are seven. They are negative large (NL), negative medium (NM), negative small (NS), zero (Z), positive small (PS), positive medium (PM) and positive large (PL). After assigning the input, output ranges to define fuzzy sets, mapping each of the possible four input fuzzy values of the calculated error. Table 1 shows the fuzzy associate memory of the designed fuzzy controllers, with:

$$\Delta R_s = R_s - R_{s0} \tag{2a}$$

$$\Delta R_{sh} = R_{sh} - R_{sh0} \tag{2b}$$

$$\Delta V_{oc} = V_{oc} - V_{oc0} \tag{2c}$$

$$\Delta n = n - n_0 \tag{2d}$$

$$\Delta I_0 = I_0 - I_{00} \tag{2e}$$

$$\Delta S_1 = S_{1A} - S_{1E} \tag{2f}$$

$$\Delta S_2 = S_{2A} - S_{2E} \tag{2g}$$

$$\Delta I_{sc} = I_{sc} - I_{sc0} \tag{2h}$$

where R_{s0}, R_{sh0}, V_{oc0}, n_0, I_{00} represent the slope of the experimental *I–V* curve at V_{oc} point, the slope of the experimental *I–V* curve at I_{sc} point, the experimental open circuit voltage, the experimental ideality factor and the experimental diode saturation current respectively (Fig. 3).

3 Results and Discussion

3.1 Parameter Extraction

The experimental current–voltage (I–V) data were taken from Conde et al. [24] for the organic solar cell. The photocurrent has been taken directly as the short circuit current according to the approximation $I_{sc} \approx I_{ph}$.

The extracted parameters obtained using our method will be carried out from an initial point, guest parameter values, until a stopping condition is found (accuracy or iterations number). New parameter values, for each iteration, will be estimated by the fuzzy controller. The global RMS (Root Mean Square) error between the experimental and the calculated results, considering all points on the database, will be updated. The stopping condition used in our approach is the global RMS error. This latter should be less than 5 %. The values for initialization of the FL-based computation approach for the investigated solar cell are: $[R_s, R_{sh}, n, I_0] = [2, 50, 1, 10^{-11}]$. Figure 4 shows good agreement between experimental and predicted results of the I–V characteristic for the investigated organic solar cell.

3.2 Comparison with Other Methods

Table 2 shows the comparison between the extracted parameters using our FL-based computation and the different extraction methods [24]. In order to find the model that best describes the organic solar cell, we have constructed simulated *I–V* plots based on the analytical model given in (1) according to the extracted parameters of each method, and compared the simulated plots to the experimental results [24], as it is shown in Fig. 4, where a good agreement between the experimental and our results is found. The statistical accuracy of our FL-model indicates that the best fit of *I–V* characteristic is obtained with our FL-computation method.

Table 3 gives a comparison of the CPU (central processing unit) time requirements, the accuracy and the model limitations for the organic solar cell parameters extraction with those obtained from numerical and manual computations [14, 24, 25], where the proposed FL-based computation time should be compared to the orders of magnitude increase in computation time, accuracy and model limitations for more rigorous parameters extraction techniques, such as those based on the manual and numerical parameters computation.

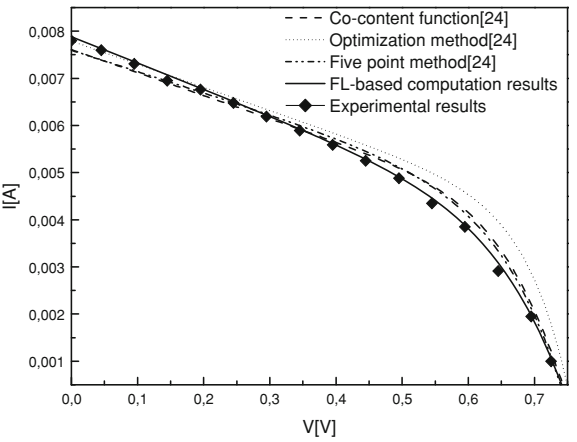


Fig. 4 Comparison between *I–V* characteristics of the investigated organic solar cell

Table 2 Extracted parameters using our FL-based computation and other approaches, respectively

Parameters	Co-content function [24]	Optimization method [24]	Five-point method [24]	FL-method
R_s (Ω)	8.59	0.804	1.0277	5.22
R_{sh} (Ω)	197.23	203.25	221.73	174.3
I_{ph} (mA/cm^2)	7.94	7.82	7.63	7.88
I_0 (nA/cm^2)	13.6	13.6	957	421
n	2.31	2.32	3.45	3.20

Table 3 Comparison between the various approaches used for the solar cell parameters extraction

Model	Performances		
	Computation time	Accuracy	Model complexity
FL-based model	Seconds	Accurate	Simple
Numerical-based model [24, 25]	Hours	Accurate	Complex
Zhang et al. [14]	Seconds	Accurate	Complex
Manual-based model	Days	Less accurate	Complex

In this context, the proposed FL-based approach can be extended to study the double diode solar cell model by including new measurements and experimental data. The obtained results can be explained by the fact that the FL-based techniques are characterized as behavioural modeling approaches, in which the computational models are based on parallel distributed processing of data. Hence, the fuzzy computation provides practical insight into solar cells modeling and identification to design photovoltaic panels without the uncertain accuracy or

meticulous tuning effort that face more rigorous solar cell models. The FL-based computation is a step towards a new generation of simulation tools that will allow solar cells and photovoltaic panel engineers to explore new classes of photovoltaic devices.

4 Conclusion

In this contribution, we have proposed a new simple and powerful approach based on Fuzzy Logic computation for organic solar cells parameters extraction. It was found that the proposed technique is applicable to study the organic solar cell behaviour. The obtained results have confirmed the accuracy and the rapid convergence to the solution with high consistency, with no need for user intervention during the search. In addition, the proposed approach is useful for the accurate and easy extraction of solar cell parameters from a measured *I–V* characteristic. It is to note that the proposed approach can be extended to include other design and environmental parameters like: temperature effect, organic transport mechanisms, metal contact resistances and double-diode effect. However, new measurements and complex compact models which include these parameters should be developed. The comparisons with the experimental results and the various extraction methods has demonstrated that our proposed approach offers higher efficiency encouraging its implementation in the development of an accurate solar cell simulator to study the photovoltaic systems. In addition, the proposed F-L- based approach does not only benefit the modeling and study of organic solar cells but can also be extended for other real-world applications. The similar methodology can be used to study the molecular and inorganic semiconductor devices.

References

1. Y. Chen, X. Wang, D. Li, R. Hong, H. Shen, Parameters extraction from commercial solar cells I–V characteristics and shunt analysis. Appl. Energy **88**(6), 2239–2244 (2011)
2. X. Han, Y. Wang, L. Zhu, Electrical and thermal performance of silicon concentrator solar cells immersed in dielectric liquids. Appl. Energy **88**(12), 4481–4489 (2011)
3. M.A. Green, K. Emery, K. Buecher, D.L. King, S. Igari, Solar cell efficiency tables (Version 9). Prog. Photovoltaics: Res. Appl. **5**(1), 51–54 (1997)
4. H.J. Lewerenz, H. Jungblut, *Photovoltaik Grundlagen und Anwendungen* (Springer, Berlin, 1995)
5. M. Tivanov, A. Patryn, N. Drozdov, A. Fedotov, A. Mazanik, Determination of solar cell parameters from its current–voltage and spectral characteristics. Sol. Energy Mater. Sol. Cells **87**(1–4), 457–465 (2005)
6. E. Radziemska, Dark I-U–T measurements of single crystalline silicon solar cells. Energy Convers. Manag. **46**(9–10), 1485–1494 (2005)

7. T. Jeranko, H. Tributsch, N.S. Sariciftci, J.C. Hummelen, Patterns of efficiency and degradation of composite polymer solar cells. Sol. Energy Mater. Sol. Cells **83**(2–3), 247–262 (2004)
8. J.A. Jervase, H. Bourdoucen, A. Al-Lawati, Solar cell parameter extraction using genetic algorithms. Meas. Sci. Technol. **12**(11), 1922–1925 (2001)
9. M. Ye, S. Zeng, Y. Xu, An extraction method of solar cell parameters with improved particle swarm optimization. In China Semiconductor Technology International Conference 2010 (CSTIC 2010), 18–19 Mar 2010, Shanghai, China
10. A. Askarzadeh, A. Rezazadeh, Artificial bee swarm optimization algorithm for parameters identification of solar cell models. Appl. Energy **102**(Special), 943–949 (2013)
11. T. Bendib, A.Maoucha, F. Djeffal, N. Lakhdar, D. Ararand, M.A. Abdi, A multi-objective optimization-based approach to improve the organic solar cell efficiency. In 1st International Conference on Renewable Energies and Vehicular Technology, REVET 2012, Hammamet, Tunisia, 26–28 Mar 2012, pp. 425–429 (2012)
12. F. Djeffal, T. Bendib, D. Arar, Z. Dibi, An optimized metal grid design to improve the solar cell performance under solar concentration using multiobjective computation. Mater. Sci. Eng. B **178**(9), 574–579 (2013)
13. W. Kim, W. Choi, A novel parameter extraction method for the one-diode solar cell model. Sol. Energy **84**(6), 1008–1019 (2010)
14. C.F. Zhang, J.C. Zhang, Y. Hao, Z. Lin, C. Zhu, A simple and efficient solar cell parameter extraction method from a single current–voltage curve. J. Appl. Phys. **110**(6), 064504–064510 (2011)
15. A. Maoucha, F. Djeffal, D. Arar, N. Lakhdar, T. Bendib, M.A. Abdi, An accurate Organic solar cell parameters extraction approach based on the illuminated I–V characteristics for double diode modeling, International Conference on Renewable Energies and Vehicular Technology, IEEE-REVET' 2012, Hammamet, Tunisia, 26–28 Mar 2012, pp. 74–77 (2012)
16. T. Bendib, F. Djeffal, D. Arar, Z. Dibi,, A. Ferdi, Fuzzy-logic-based approach to study the electrons mobility in nanoscale double gate MOSFETs. IOP Conference Series: Materials Science and Engineering, vol. 41, p. 012016 (2012)
17. ATLAS: 2D Device Simulator, SILVACO International 2008
18. A. Cheknane, H.S. Hilal, F. Djeffal, B. Benyoucef, J.-P. Charles, An equivalent circuit approach to organic solar cell modelling. Microelectron. J. **39**(10), 1173–1180 (2008)
19. L.A. Zadeh, Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
20. R.R. Yager, L.A. Zadeh, *An introduction to fuzzy logic applications in intelligent systems* (Kluwer, Norwell, 1991)
21. K. Mishra, I.G. Sarma, K.N. Swamy, Performance evaluation of two fuzzy-logic-based homing guidance schemes. J. Guidance Control Dyn. **17**(6), 1389–1391 (1994)
22. M. Esposito, G.D. Pietro, An ontology-based fuzzy decision support system for multiple sclerosis. Eng. Appl. Artif. Intell. **24**(8), 1340–1354 (2011)
23. T. Bendib, F. Djeffal, D. Arar, M. Meguellati, Fuzzy-logic-based approach for organic solar cell parameters extraction. In *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013*, WCE 2013, 3–5 July 2013, London, UK, pp 1182–1185 (2013)
24. A. Ortiz-Conde, F.J. Garcia Sanchez, J. Muci, New method to extract the model parameters of solar cells from the explicit analytic solutions of their illuminated I–V characteristics. Sol. Energy Mater. Sol. Cells **90**(3), 352–361 (2006)
25. J. Nocedal, S.J. Wright, *Numerical Optimization*, 2nd edn. Springer Series in Operations Research (Springer, Berlin, 2006)

An ANFIS Based Approach for Prediction of Threshold Voltage Degradation in Nanoscale DG MOSFET Devices

Toufik Bentrcia and Fayçal Djeffal

Abstract Nowadays, the tremendous shrinking of electronic devices has reduced their sizes to very low scales. However, this process has been accompanied unavoidably with many well-recognized reliability challenges basically for MOSFET devices. Our objective, in this work, is the proposition of an Adaptive Network based Fuzzy Inference System (ANFIS) to study the threshold voltage behavior caused by the interface traps generated by the ageing mechanism phenomenon. The consideration of a nanoscale DG MOSFET device makes the application of compact modeling tools a very hard task to carry out, due to approximations made during model development. The obtained results are in a good agreement with numerical simulations based on Atlas 2D-simulator. In addition, the comparison with a feed-forward artificial neural network shows that our fuzzy system provides higher accuracy performances. The proposed approach can be incorporated in Integrated Circuit modeling frameworks in order to support more complex degradation situations.

Keywords DG MOSFET · Functions · Hot carrier · Learning algorithm · Membership · Quantum confinement · Short channel · Threshold voltage

T. Bentrcia
Physics Department, University of Batna, 05000 Batna, Algeria
e-mail: toufikmit@yahoo.com

F. Djeffal (✉)
Electronics Department, University of Batna, 05000 Batna, Algeria
e-mail: faycaldzdz@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_25,

1 Introduction

The actual trend in the microelectronics industry requires the fabrication of components with very small dimensions. The Metal Oxide Semiconductor Field Effect Transistor (MOSFET) subject to a continuous downscaling has resulted in many challenging drawbacks that are reflected by the increase of some undesirable quantities such as the leakage current [1]. In addition, as the channel length is scaled down to increase both the operation speed and the integration density, it can reach the same order of magnitude as the depletion layer widths of the source and the drain junctions and at this level the so-called short channel effects (SCEs) occur. This situation mainly induces the deterioration of the gate controllability efficiency over the channel. Therefore, more attentions should be focused on the proposition and evaluation of new proposed architectures [2–4]. As stated by the International Technology Roadmap of Semiconductors (ITRS), a 10-nm gate length with fully depleted silicon on insulator technology and a 7-nm gate length with double gate devices are predicted in the near future. Various enhancements have been also conducted at the level of fabrication materials, where the use of high-k materials and strained Silicon has been proved to be beneficial in improving performances of the multi-gate MOSFETs [5]. In these devices, the gate electrode is wrapped around a silicon nanowire, forming a multigate structure with excellent control of the channel potential allowing the full depletion of the channel region.

In addition to the SCEs, another phenomenon that is closely linked to the reduction of the channel length consists in the hot carrier injection effect leading to the alteration of the device function. In fact, such reliability concern is the result of the interface trap generation because of the increase of the maximum electric field experienced by carriers in the channel near the drain side, since the associated power supply voltages is scaled at a slower rate compared to the channel length [6]. An impact ionization process is initiated when carriers moving from the source to the drain sides acquire enough kinetic energy in the high field region of the drain junction. A fraction of carriers penetrates the Si-SiO_2 potential barrier and becomes trapped in the gate oxide. From a phenomenological view point, hot carriers can be identified by two main characteristics: (i) violation of the thermal equilibrium with the lattice and (ii) energy gain higher than the thermal energy. With the accumulation of injected carriers along the Si-SiO_2 interface, an interface trap buildup and the trapping of carriers in the dielectric occur, initiating in turn the degradation of many device parameters such as the transconductance [7]. At advanced stages of the process, the energy gained by the carriers in the high field region of the Silicon substrate induces the break of bonds associated with extrinsic or intrinsic defects in the oxide, and such rearrangement in its atomic structure is the generator factor of the device instabilities observed during the hot-carrier injection.

In order to account for ultrathin Silicon body and short channel length in symmetrical DG MOSFET modeling, two or three dimensional solutions of the coupled Schrodinger/Poisson equations should be considered. The resolution

methodologies are based mostly on numerical methods, which are very time consuming due to the intrinsic nature of the physical effects governing the electrical properties under such constraints [8]. The development of compact analytical formulations that are both supporting quantum aspects and valid in all operating regions is an intractable, even an impossible task to realize [9]. As a result, soft computing based approaches have been proposed to include short channel and quantum effects in the same framework for DG MOSFET performance modeling [10–13].

Our main aim in this paper is to investigate more in details the capabilities of Adaptive Network based Fuzzy Inference Systems (ANFISs) in predicting the relative degradation of DG MOSFET threshold voltage due to the hot carrier injection effect. Both the short channel and quantum effects are taken into consideration, where two geometrical parameters are selected as input variables to the fuzzy system. For the ANFIS methodology, several types of membership functions (MFs) are tested. The best one in term of the Mean Squared Errors (MSE) and correlation coefficient (R) criteria for the training and testing phases is selected. The efficiency of the proposed approach is validated through comparison with an Artificial Neural Network (ANN), where superior performances are recorder.

The organization of the paper is as follows. Firstly, some analytical models dealing with the degradation of the threshold voltage as a result of miniaturization effects are presented. Then, we illustrate the principal steps towards the elaboration of an ANFIS methodology. After, the device design and the generation of the training database used by the fuzzy system are investigated. Results and their discussion are depicted in the fifth section. Finally, we terminate with some concluding remarks and further research directions.

2 Threshold Voltage Behavior Under Downscaling Effects

The threshold voltage of a MOSFET device defines the applied gate voltage corresponding to the switch of the device to the ON-state. Such value is strongly affected by various miniaturization effects [14]. Accurate expressions for the modeling of the threshold voltage are highly recommended so that the correct behavior of new designed circuits could be predicted appropriately.

2.1 Hot Carrier Effect

The MOSFET device ageing due to the long duration of function is reflected by damages in the form of interface traps created by hot carriers at the interface between the channel and the oxide regions [15]. Many studies have been reserved to the analysis of MOSFET performances when subject to the interface trap degradation. Various analytical compact models have been proposed (see for

example [16, 17]). The existence of interface charge densities leads to an increase or a decrease in the threshold voltage depending on the sign of the trapped charges. An additional term in the model of the fresh device is included as defined in [18],

$$\Delta V_{th} = -\frac{Q_0 + Q_{it}}{C_{ox}} \tag{1}$$

where Q_0 is the trapped charge density in the oxide, Q_{it} is the interface trap charge density and C_{ox} is the oxide capacitance. It can be concluded that the device becomes less sensitive to biasing and higher gate voltage values are needed to make the device switching from the OFF-state to the ON-state.

2.2 Reduction of Geometrical Parameters Effect

In the case of dimensional parameters, experimental measurements indicate that the threshold voltage tends to decrease by decreasing the channel length. This monotone reduction becomes more noticeable when the channel length becomes comparable to the source and the drain depletion widths. The change in the threshold voltage due to the SCE is given by [19],

$$\Delta V_{th} = \frac{Q_b}{C_{ox}} \frac{X_j}{L} \left(\sqrt{1 + \frac{2X_{dm}}{X_j}} - 1 \right) \tag{2}$$

where Q_b is the bulk charge per unit area, X_j is the junction depth and X_{dm} is the maximum depletion width.

For the more general case where both device parameters (the channel width and length) are of the same order of magnitude as the depletion width (small geometry device), the change in the threshold voltage caused by the small geometry effect can be estimated as the sum of the short channel and narrow width effects as [20],

$$\Delta V_{th} \approx \Delta V_{th,L} + \Delta V_{th,W} \tag{3}$$

2.3 Quantum Confinement Effect

If we consider the quantization of energy levels, the distribution of accumulated or inverted carriers at the interface is different from the classical prediction [21]. The model of any surface potential based parameter such as the threshold voltage or the drain induced barrier lowering will be modified. Since the quantum effects are significant for oxide and channel thicknesses less than 5 nm, the inclusion of quantum corrections in the available up-to-date DG MOSFET models cannot be

ignored. The threshold voltage shift resulted from the quantum confinements effects can be expressed by [22],

$$\Delta V_{th} = \frac{\beta t_{ox} N_A}{2\varepsilon_{ox}} \left(\frac{kT\varepsilon_{si}}{n_i} \right)^{1/2} \tag{4}$$

where β is a fitting parameter, t_{ox} is oxide thickness, N_A is the channel doping concentration and n_i represents the intrinsic carrier concentration.

3 ANFIS Based Framework

In electronics application field, the backbone of almost artificial intelligence-oriented models is based on input-output mapping. Tuning parameters are included within the model so that a good concordance between the predicted and the target outputs is obtained, where a learning algorithm is used for this purpose [23]. However, it is worth saying that there is no guarantee regarding the global optimality of results obtained by running such algorithms, and therefore a testing stage for these models can be considered as a mandatory requirement.

3.1 ANFIS Architecture

The architecture of ANFIS can be identified to that of a multilayer feed forward network where the weighting coefficients express the parameters of the membership function in addition to the linear model. In this context, it can be considered as a combination of fuzzy logic and artificial neural networks. Thus, ANFIS takes the advantage of the neural network in term of learning algorithms and fuzzy inference systems in term of uncertain knowledge support. This approach has known successful application in many fields like time series prediction and diagnosis. In the pioneer work of Jang [24], it was proved that such approach is a universal competitive approximator when compared to other available methods.

The ANFIS methodology is based on the idea of dividing the input space into many local subregions with the possibility of simultaneous activation of many of them by a single input. These subregions are obtained by partitioning the input space using adequate fuzzy membership functions [25]. In this paper, we assume that the input variables of our fuzzy system are limited only to the channel length and thickness since they influence significantly the short channel, the hot carrier, and the quantum effects. In Fig. 1, we highlight the general structure of an ANFIS model.

We have five layers in the system with the role of each layer in the network described below [26],

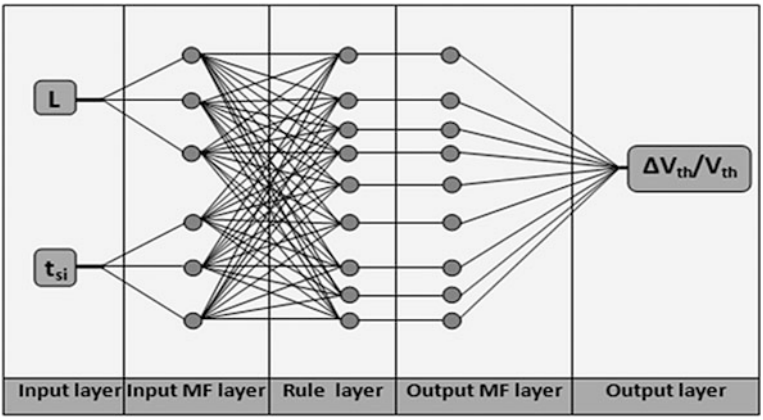


Fig. 1 ANFIS architecture used to predict the degradation in the threshold voltage

- Input layer: the output of each node gives the input variable membership grade.
- Input MF layer: the firing strength associated with each rule is calculated.
- Rule layer: the calculation of the relative weight of each rule is achieved.
- Output MF layer: the multiplication of normalized firing strength by first order of Sugeno fuzzy rule is realized.
- Output layer: one node is composed and all inputs of the node are added up.

It should be noticed the existence of only two adaptive layers in the whole network, the first one includes adjustable parameters belonging to the input membership functions called the premise parameters. The second one represented by the fourth layer contains adjustable parameters known as the consequent parameters contained in the linear model [27]. We denote by n and m the number of fuzzy sets attached to each input. Thus, the number of Takagi-Sugeno fuzzy IF-THEN rules is n × m with the generic expression of such rules in our study given by,

$$IF(L\ is\ A_i)\ AND(t_{si}\ is\ B_j)\ THEN\left(\left(\frac{\Delta V_{th}}{V_{th0}}\right)_l = p_l L + q_l t_{si} + r_l\right) \tag{5}$$

where A_i and B_j are the linguistic terms of the precondition part with membership functions $\mu_{Ai}(L)$ and $\mu_{Bj}(t_{si})$ respectively. The parameters p_l, q_l and r_l represent the consequent parameters.

3.2 Learning Algorithm

In ANFIS, a sort of hybrid learning algorithm is used, which combines the gradient method with the least square method to update the parameters. A generic parameter α formed by the union of the premise and the consequent parameters is updated using the formula [28],

$$\Delta\alpha = -\eta \frac{\partial E}{\partial \alpha} \tag{6}$$

where E is the overall error and η a learning rate calculated according to,

$$\eta = \frac{\delta}{\sqrt{\sum_{\alpha} \left(\frac{\partial E}{\partial \alpha}\right)^2}} \tag{7}$$

with δ is the step-size.

The generic parameter is modified at each training epoch in the sense that the consequent parameters are updated first by using a least square algorithm and the premise parameters are then adjusted by backpropagating the errors. Despite that the hybrid learning algorithm remains the most widely used, other alternative methods have been adopted such as the genetic algorithm and particle swarm optimization [29, 30].

4 Device Design and Elaboration of the Training Database

The DG MOSFET relative degradation in presence of quantum confinement and short channel effects can be assessed by parsing the channel length and thickness values with fixed steps over intervals. This allows the correct estimation of the influence range that may be more interesting for analysis. Atlas 2-D simulator is used to obtain the threshold voltage associated to a structure with specified dimensions [31]. The geometrical and electrical configuration parameters of the simulated DG MOSFET device are expressed by the set of values as indicated in Table 1.

The associated two-dimensional scheme is characterized by the following properties: The device has uniform doping concentrations in the channel and source/drain regions. Two carrier types are used in the simulation, the drift-diffusion model without impact ionization, doping concentration-dependant carrier mobility and electric field-dependant carrier model have been also considered in the models section. To take into account the leakage current, SRH recombination/generation is included in the simulation. A cross-sectional view of the DG MOSFET device used in this study is illustrated in Fig. 2.

We depict in Fig. 3 the relative disposition of the threshold voltage curves as a function of the channel length for fresh and damaged cases. For both cases, the variation law is monotonic with the channel length until the parameter reaches a saturation value. It can be observed that the damaged device has higher threshold voltage due to the increase of the applied gate voltage needed to turn on the device. Another interesting feature related to the discrepancy of curves, which is more important for short channel length values compared to long channel lengths. Such situation can be interpreted by the strong correlation existing between the hot carrier degradation and the SCEs.

Table 1 Common geometrical and electrical configuration parameters

Parameter	Notation	Value
Oxide thickness	t_{ox}	1.5 nm
Drain/Source doping	$N_{D/S}$	1×10^{20} cm^{-3}
Channel doping	N_{Ch}	1×10^{15} cm^{-3}
Work function	ϕ	4.55 eV
Interface trap density	N_f	5×10^{12} cm^{-2}
Drain voltage	V_{ds}	0.1 V
Gate voltage	V_{gs}	0.7 V

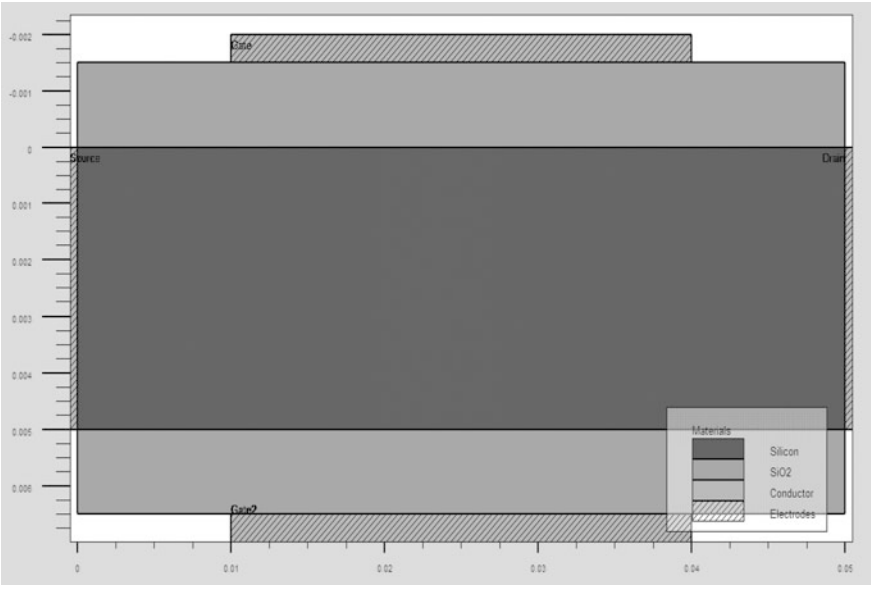


Fig. 2 Cross-sectional view of the simulated DG MOSFET device

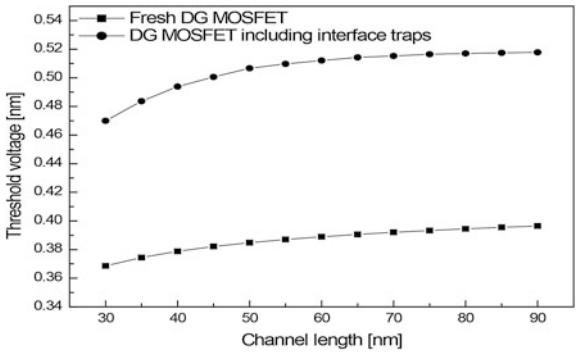


Fig. 3 Threshold voltage variation as a function of the channel length for fresh and damaged devices ($t_{si} = 2.5$ nm)

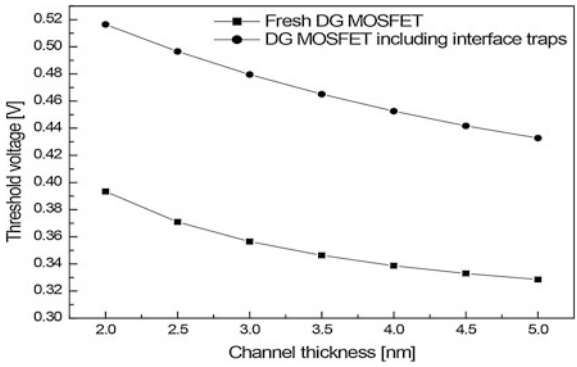


Fig. 4 Threshold voltage variation as a function of the channel thickness for fresh and damaged devices (L = 75 nm)

As shown in Fig. 4, the threshold voltage as a function of the channel thickness has a decreasing tendency and it is more pronounced for law values of the channel thickness with and without interface traps.

5 Simulation Results

The performance of an ANFIS based approach is strongly depending on the database used for learning, which should be as exhaustive as possible to cover a wide range of values in the whole input-output space of data. The lack of data in the selected set reduces the ANFIS prediction ability when an unknown pattern is encountered. The data set used for the training of our fuzzy system is obtained by using Atlas 2-D Simulator. As the accuracy of the trained ANFIS depends on the correctness and the effective representation of the data used during the learning procedure, a total of 91 observations are obtained by sampling the channel length and the channel thickness ranges with a step of 5 and 0.5 nm, respectively. These observations are divided into the training and the checking sets (77 and 14 observations for each set), where the later is used to avoid the overfitting problem leading to bad prediction performances. The testing set is composed of 12 observations used to validate the prediction ability of the resulted fuzzy system. Because of the strong influence of membership function types on the quality of the obtained decision system, it is indispensable to select these functions in an optimal manner. The only solution is trial-error approach due to the absence of any deterministic method that permits the specification of membership functions in the ANFIS. In this paper, six different types of widely used membership functions are tested (Gaussian, Gaussian combination, sigmoid difference, generalized bell, Pi, and sigmoid product shaped MFs). We find that the best membership function leading to superior results in the calculation of the threshold voltage relative degradation is the Gaussian combination shaped membership function. The criteria of selection are based on the best values of the mean square error and the

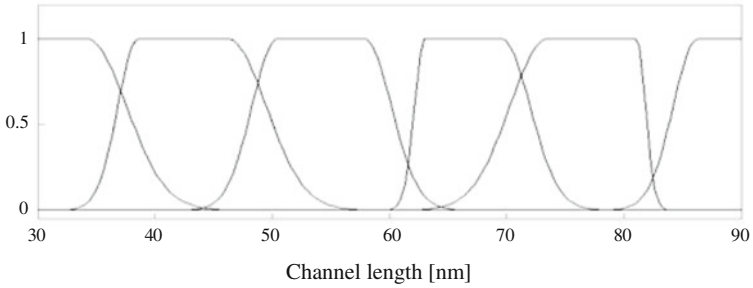


Fig. 5 Partition of the channel length range using Gaussian combination shaped membership functions

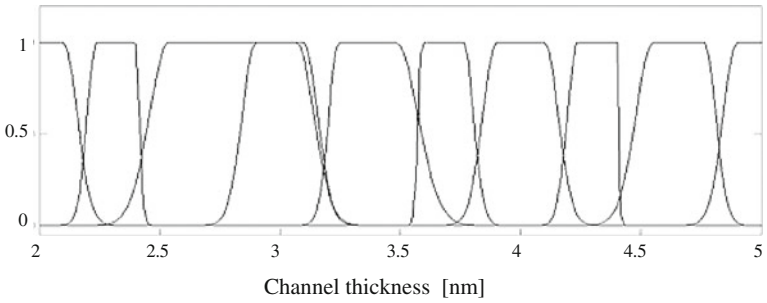


Fig. 6 Partition of the channel thickness range using Gaussian combination shaped membership functions

correlation coefficient relative to the training phase. The Gaussian combination shaped membership function is given by,

$$\mu(x;\sigma_1,\sigma_2,c_1,c_2) = e^{-\frac{(x-c_1)^2}{2\sigma_1^2}} + e^{-\frac{(x-c_2)^2}{2\sigma_2^2}} \tag{8}$$

where $\sigma_{i=1,\ 2}$ and $c_{i=1,\ 2}$ are the Gaussian function parameters.

The number of membership functions for both parameters has been tested over values comprised between 2 and 15. The optimal values are found to be equal to 6 for the channel length and to 10 for the channel thickness. The number of the resulted fuzzy IF-THEN rules for the inference system is equal to 60 (6 × 10). The examination of the prediction ability of these fuzzy rules for testing confirms well that the use of the Gaussian combination shaped membership overpasses other shaped MFs in term of accuracy. The partition of the input intervals according to the selected number and type of the membership functions for the channel length and thickness is presented in Figs. 5 and 6, respectively.

The output surface of the obtained ANFIS is visualized in Fig. 7, the predicted relative degradation of the threshold voltage along the variation ranges of inputs is plotted as a function of the channel length and thickness. As deduced from this

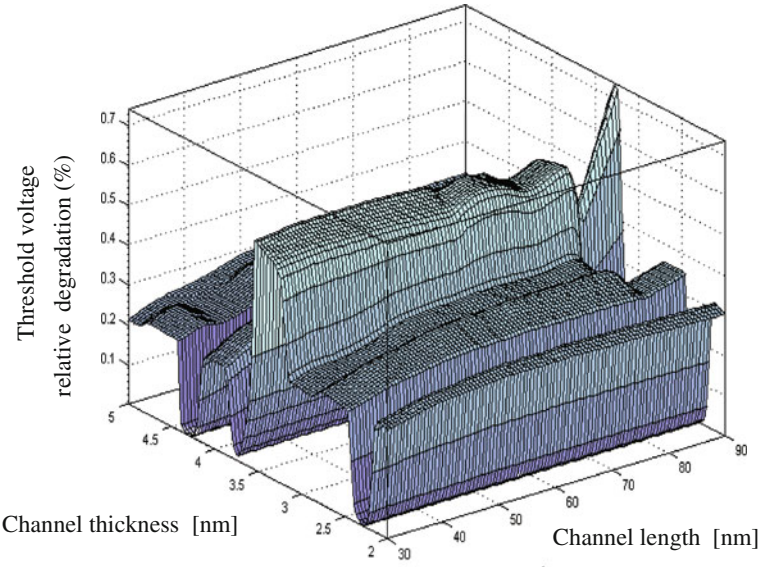


Fig. 7 Response surface of obtained ANFIS over the input parameter ranges

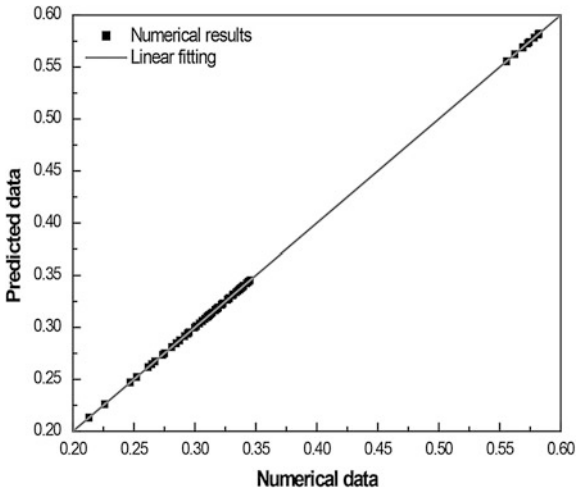


Fig. 8 Regression plot of the threshold voltage relative degradation in the training dataset

graph, the highest degradation is located at the vicinity of a value equals to 3.5 nm for the channel thickness. The variation of the relative degradation is characterized by the presence of many local optima, making the analysis of the device immunity against the hot carrier degradation a very hard task especially when additional constraints are considered within the device.

The regression curves of the numerical and predicted data of the threshold voltage relative degradation are shown in Figs. 8 and 9 for the training and testing

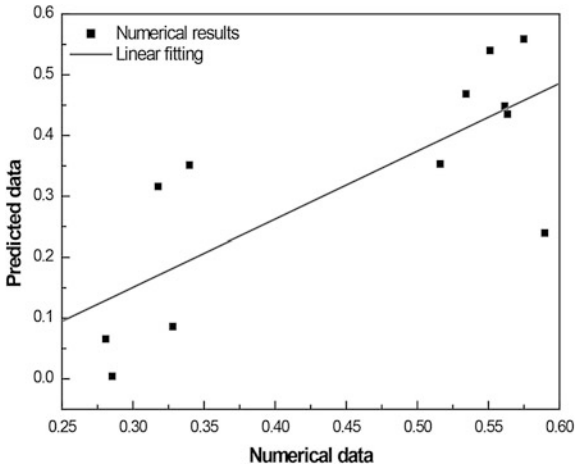


Fig. 9 Regression plot of the threshold voltage relative degradation in the testing dataset

Table 2 Comparison of ANN and ANFIS performance criteria

Criterion	ANN		ANFIS	
	Training	Testing	Training	Testing
Mean squared errors	3×10^{-3}	0.13	2.43×10^{-8}	3×10^{-2}
Correlation coefficient	1	0.67	1	0.77

datasets. It is easy to note that a sufficient agreement is satisfied between the predicted and numerical results especially in the training stage.

A more rigorous validation methodology can be conducted by comparing the performance criteria of our ANFIS based approach to other frameworks. A two layer feed-forward artificial neural network with sigmoid hidden and linear output layers has been used for this purpose. The number of neurons in the hidden layer is fixed to a value of 50 neurons, whereas the learning algorithm used to train the network is Levenberg–Marquard backpropagation algorithm. The statistical criteria for both formalisms are calculated and compared based on the following expressions,

$$MSE = \frac{1}{n}\sum_{1}^{n}\left[\left(\frac{\Delta V_{th}}{V_{th}}\right)_{Num} - \left(\frac{\Delta V_{th}}{V_{th}}\right)_{Calc}\right]^2 \tag{9}$$

$$R = \frac{Cov\left[\left(\frac{\Delta V_{th}}{V_{th}}\right)_{Num}, \left(\frac{\Delta V_{th}}{V_{th}}\right)_{Calc}\right]}{\sigma_{\left(\frac{\Delta V_{th}}{V_{th}}\right)_{Num}} \sigma_{\left(\frac{\Delta V_{th}}{V_{th}}\right)_{Calc}}} \tag{10}$$

A summary of the main results is provided in Table 2. It is clearly indicated that our proposed approach offers higher efficiency compared to the employed artificial

neural network. This can be seen as a natural consequence of integrating hybrid learning algorithms with fuzzy logic tools.

Furthermore, since the correlation coefficient for the testing set in the case of ANFIS is close to 0.8, we can say that we have a strong correlation between the numerical and the predicted relative degradation values of the threshold voltage. Consequently, it can be claimed that the performances of our fuzzy system are satisfactory and can be adopted for further analyses of more complicated integrated circuits.

6 Concluding Remarks

In this paper, an ANFIS based approach has been developed for the prediction of an ageing measurement criterion of DG MOSFETs including hot carrier, short channel and quantum confinement effects. The ageing phenomenon is expressed by the relative degradation in the threshold voltage caused by the working time of the device. The input parameters have been limited to the channel length and thickness since they have the major impact on the device behavior. The elaboration of the training, validation and testing benchmarks has been made easy thanks to ATLAS-2D simulator, where nanoscale value ranges have been attached to input parameters during the simulation. A comparison with an artificial neural network simulation has been conducted in term of the main performance criteria. The obtained performance has demonstrated that the developed approach offers higher efficiency encouraging its implementation in electronics device simulators to study the nanoscale CMOS circuits including ageing phenomenon. Therefore, fuzzy logic can be considered as a prominent tool that can be beneficial in alleviating drawbacks related to many effects that are inevitable in new generations of nanoscale devices. It should be noted also that additional device parameters such as the oxide thickness parameter can be included to extend the proposed ANFIS-based approach so that more accuracy can be gained in the device behavior modeling.

References

1. V.A. Sverdlov, T.J. Walls, K.K. Likharev, Nanoscale silicon MOSFETs: a theoretical study. IEEE Trans. Electron. Devices **50**, 1926–1933 (2003)
2. M.A. Abdi, F. Djeffal, M. Meguellati, D. Arar, Two-Dimensional Analytical Threshold Voltage Model for Nanoscale Graded Channel Gate Stack DG MOSFETs, in *16th IEEE International Conference on Electronics*, *Circuits*, *and Systems*, *ICECS 2009*, 13–16 Dec 2009, Hammamet, Tunisia (2009), pp. 892–895
3. E. Chebaki, F. Djeffal, T. Bentrcia, Two-dimensional numerical analysis of nanoscale junctionless and double gate MOSFETs including the effect of interfacial traps. Phys. Status Solidi C **09**, 2041–2044 (2012)

4. T. Bentrcia, F. Djeffal, M. Chahdi, An analytical two dimensional subthreshold behavior model to study the nanoscale GCGS DG MOSFET including interfacial trap effects. Microelectron. Reliab. **53**, 520–527 (2013)
5. The International Technology Roadmap for Semiconductors (Online). Available: http://public.itrs.net
6. M.P. Pagey, Hot-carrier reliability simulation in aggressively scaled MOS transistors. PhD dissertation, Electrical Engineering Department, Vanderbilt University, Nashville, Tennessee, USA, 2003
7. F. Djeffal, T. Bentrcia, T. Bendib, An analytical drain current model for undoped GSDG MOSFETs including interfacial hot-carrier effects. Phys. Status Solidi C **08**, 907–910 (2011)
8. F. Prégaldiny, C. Lallement, D. Mathiot, Accounting for quantum mechanical effects from accumulation to inversion in a fully analytical surface-potential-based MOSFET model. Solid-State Electron. **48**, 781–787 (2004)
9. T.P. Wen, A.K. Singh, A comprehensive analytical study of an undoped symmetrical double-gate MOSFET after considering quantum confinement parameter. Microelectron. J. **41**, 162–170 (2010)
10. F. Djeffal, Z. Dibi, M.L. Hafiane, D. Arar, Design and simulation of a nanoelectronic DG MOSFET current source using artificial neural networks. Mater. Sci. Eng. C **27**, 1111–1116 (2007)
11. T. Bentrcia, F. Djeffal, E. Chebaki, ANFIS-based approach to studying subthreshold behavior including the traps effect for nanoscale thin-film DG MOSFETs. J. Semiconductors **34**, 084001:1–084001:9 (2013)
12. T. Bentrcia, F. Djeffal, M. Meguellati, D. Arar, New approach based on ANFIS computation to study the threshold voltage behavior including trap effects for nanoscale DG MOSFETs, in *Proceedings of The World Congress on Engineering 2013, WCE 2013*, 3–5 July 2013. Lecture Notes in Engineering and Computer Science, London, UK (2013), pp. 1150–1155
13. T. Bentrcia, F. Djeffal, D. Arar, M. Meguellati, ANFIS-based computation to study the nanoscale circuit including the hot-carrier and quantum confinement effects, in *Proceedings of The 5th International Conference on Modeling, Simulation and Applied Optimization 2013*, *ICMSAO'2013*, 28–30 April 2013, Hammamet, Tunisia (2013), pp. 01–05
14. H.M. Abdelhamid, Compact modeling of multiple gate MOS devices, PhD dissertation, Department of Electronic, Electrical and Automatic Engineering, University of Rovira i Virgili Tarragona-Spain, 2007
15. T. Bentrcia, F. Djeffal, *Compact modeling of multi-gate MOSFET including hot-carrier effects. CMOS Technology: Electrical Engineering Developments Series*, Chap. 4 (Nova Publishers, New York, 2011), pp. 135–158
16. T. Bentrcia, F. Djeffal, M.A. Abdi, M. Chahdi, N. Boukhennoufa, An accurate two dimensional threshold voltage model for nanoscale GCGS DG MOSFET including traps effects, in *Proceedings of the 3rd International Conference on Signals, Circuits and Systems 2009, SCS'2009*, 06–08 Nov 2009, Djerba, Tunisia (2009) pp. 01–06
17. F. Djeffal, T. Bentrcia, M.A. Abdi, T. Bendib, Drain current model for undoped gate stack double gate (GSDG) MOSFETs including the hot-carrier degradation effects. Microelectron. Reliab. **51**, 550–555 (2011)
18. S. Naseh, M.J. Deen, C.-H. Chen, Hot-carrier reliability of submicron NMOSFETs and integrated NMOS low noise amplifiers. Microelectron. Reliab. **46**, 201–2012 (2006)
19. N. Arora, *MOSFET modeling for VLSI simulation theory and practice* (World Scientific Publishing, Singapore, 2007)
20. C. Galup-Montoro, M.H. Schneider, *MOSFET modeling for circuit analysis and design* (World Scientific Publishing, Singapore, 2007)
21. M. Balaguer, J.B. Roldan, L. Donetti, F. Gamiz, Inversion charge modeling in n-type and p-type double-gate MOSFETs including quantum effects: the role of crystallographic orientation. Solid-State Electron. **67**, 30–37 (2012)
22. H.C. Morris, H. Abebe, A compact quantum surface potential model for a MOSFET device. Math. Comput. Modell. **51**, 893–900 (2010)

23. F. Djeffal, S. Guessasma, A. Benhaya, M. Chahdi, An analytical approach based on neural computation to estimate the lifetime of deep submicron MOSFETs. J. Semiconductor Sci. Technol. **20**, 158–164 (2005)
24. J.S.R. Jang, Neuro-fuzzy modeling: architectures, analyses and applications, PhD dissertation, Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA, 1992
25. S.M. Aminossadatia, A. Kargarb, B. Ghasemi, Adaptive network based fuzzy inference system analysis of mixed convection in a two-sided lid-driven cavity filled with a nanofluid. Int. J. Therm. Sci. **52**, 102–111 (2012)
26. M. Singh, Adaptive network-based fuzzy inference systems for sensorless control of PMSG based wind turbine with power quality improvement features, PhD dissertation, University of Quebec, Montreal, Canada, 2010
27. J.S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man and Cybern. **23**, 665–685 (1993)
28. L.-C. Ying, M.-C. Pan, Using adaptive network based fuzzy inference system to forecast regional electricity loads. Energy Convers. Manag. **49**, 205–211 (2008)
29. N. Nariman-Zadeh, A. Darvizeh, M.H. Dadfarmai, Design of ANFIS networks using hybrid genetic and SVD methods for the modelling of explosive cutting process. J. Mater. Process. Technol. **155**, 1415–1421 (2004)
30. H.M. Jiang, C.K. Kwong, W.H. Ip, T.C. Wong, Modeling customer satisfaction for new product development using a PSO-based ANFIS approach. Appl. Soft Comput. **12**, 726–734 (2012)
31. Atlas user manual: Device simulation software, 2008

A Novel Feedback Control Approach for Networked Systems with Probabilistic Delays

Magdi S. Mahmoud

Abstract The networked control system (NCS) design for continuous-time systems with probabilistic delays is discussed in this paper. The delay is assumed to follow a given probability density function. A novel design scheme for the output feedback controller is developed to render the closed-loop networked system exponentially mean-square stable with H_∞ performance requirement. A numerical example is provided to show the advantages of the proposed technique.

Keywords Exponential mean-square stability · Networked control systems · Output feedback control · Packet dropout · Probabilistic delays · Varying sample interval

1 Introduction

In many modern complex and distributed control systems, systems with remotely located sensors, actuators, controllers and filters are often connected over a sharing communication network. Such architectures are often called networked-control systems (NCS), which bring a lot of advantages such as low cost, simple installation and maintenance, increased system agility and so on [1]. The sharing network however makes the analysis and synthesis of such network-based systems challenging. Recently, NCS has attracted much research interest [2]. So far, there has been considerable research work appeared to address modelling, stability analysis, control and filtering problems for NCSs, [3]. Most of the studies on NCSs have

M. S. Mahmoud (✉)
Systems Engineering Department, KFUPM, P.O. Box 5067Dhahran 31261, Saudi Arabia
e-mail: msmahmoud@kfupm.edu.sa; magdim@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_26,

concentrated on state feedbacks [4], and the commonly investigated systems have been discrete-time models, sampled-data models, continuous-time models through sampled-data feedback controls. Upon unavailable state information, observer-based feedbacks have to be performed to achieve prescribed control purposes [5–13].

As has been mentioned above that it is difficult to deal with the NCS with long time-varying or random delays, and one aspect of the difficulties lies in providing an appropriate modeling method for such NCSs. Since the delay may be larger than one sampling period, more than one control signals may arrive at the actuator during one sampling interval. Moreover, the numbers of the arriving control signals vary over different sampling intervals, thus the dynamic model of the overall closed-loop NCS varies from sampling period to sampling period [14]. So, the closed-loop NCS is naturally a switched system with the subsystems describing various system dynamics on the different sampling intervals. The switched system model has been used to describe the NCS with delays [15]. However, it is assumed in most of the existing results that the delay is smaller than one sampling period. In [16], the switched system model was used to describe the NCS with long time-varying delays. However, the arbitrary switching scheme was used, which may be conservative and infeasible when some subsystems of the NCS are unstable. Recently, the observer-based feedback controls have been further studied for discrete-time NCSs with random measurements and time delays. In [17], the closed-loop system was transformed into a delay-free model, and an observer-based $\mathscr{H}_\infty$ control design scheme was given in terms of a linear matrix inequality (LMI) to render the closed-loop systems exponentially mean-square stable.

Motivated by the above observations, in this paper, we provide a generalized approach to treating NCSs with probabilistic delays. Specifically, we build on [18] and extend it further to study the problem of the exponential stability of NCSs with probabilistic time-varying delay. By adopting a Lyapunov–Krasovskii functional (LKF) approach and linear matrix inequalities (LMIs), new criteria for the exponential stability of such NCSs are derived in the form of feasibility testing of LMIs, which can be readily solved by using standard numerical software based on inner-minimization methods. We also adopt an appropriate free-weighting matrix method [19] suitable for the derivation of the main results for our considered problem. Numerical example is provided to illustrate that when the variation probability of the time delay is given, the upper bound of the time delay could be much larger than that when only the variation range of the time delay is known.

Notation: We use I and 0 to denote, respectively, the identity matrix and the zero matrix with compatible dimensions; the superscripts T and '-1' stand for the matrix transpose and inverse, respectively; $W > 0$ means that W is a real symmetric positive definite matrix; $\|\cdot\|$ is the spectral norm; $\mathbf{E}\{\cdot\}$ denotes the expectation and $\mathbf{Pr}\{\cdot\}$ means the probability; $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ denote, respectively, the maximum eigenvalue and the minimum eigenvalue of a matrix. In symmetric block matrices, we use the symbol $\bullet$ to represent a term that is induced by symmetry.

2 Problem Formulation

Consider a continuous-time system described by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + B_{xw}w(t), \\ z(t) &= A_z x(t) + B_z u(t) + B_{zw}w(t), \\ y(t) &= Cx(t) + C_z u(t) + B_{yw}w(t) \end{aligned} \tag{1}$$

where $x(t) \in R^n$, $u(t) \in R^m$, $z(t) \in R^p$, and $w(t) \in R^q$ are the state, the control input, the controlled output and the disturbance input belonging to $\mathscr{L}_2[0, \infty)$, respectively. $A, B, B_{xw}, B_z, B_{zw}, C, C_z, B_{yw}$ and C_z are known constant real matrices with appropriate dimensions. The pair (A, B) is assumed stabilizable. The measured output $y(t) \in R^r$ frequently experiences sensor delay, and it can be described by two random events:

$$\begin{cases} \textit{Event}\,1: \ y(t) & \textit{dose not experience sensor delay}, \\ \textit{Event}\,2: \ y(t) & \textit{experience sensor delay}, \end{cases}$$

Recall from the theory of functional differential equations that a continuous and piecewise differentiable initial condition guarantees the existence of the solutions. Assume that the measurement delay $\tau(t)$ from sensor to controller is a random variable whose density function is given by $p(\tau;\, \pi(t))$, where $\pi(t)$ is a vector of parameters of p. In this paper, we assume that the experience sensor delay distribution is stationary, that is, $\pi(t) = \pi$, where π is a given vector. For example, if p is the normal density function, then $\pi(t) = \{\mu(t),\, \sigma(t)\}$, where $\mu(t)$ and $\sigma(t)$ are the mean and variance of $\tau(t)$. If the support of p contains values that the experience sensor delay cannot attain such as negative values, one could truncate the density function p to have a specified range $[0, \vartheta]$. In this case, the truncated distribution, p_T is given by

$$f_T(\tau; \pi(t)) = \frac{f(\tau; \pi(t))}{\int_\alpha^\beta f(r; \pi(t))dr}, \quad \rho_1 \le \tau(t) \le \rho_2 \tag{2}$$

Next, consider partitioning the range [α, β] into n mutually exclusive partitions whose end points are: $[\tau_0, \tau_1][\tau_1, \tau_2] \ldots [\tau_{n-2}, \tau_{n-1}][\tau_{n-1}, \tau_n]$ where $\tau_0 = \rho_1, \tau_n = \rho_2$. Let $\rho_j = \mathbf{Pr}\big(\tau_{j-1} \le \tau(t) \le \tau_j\big)$. Define the indicator functions $\varphi_j(t)$ as follows

$$\varphi_j(t) = \begin{cases} 1: & \tau_{j-1} \le \tau(t) \le \tau_j, \\ 0: & \textit{otherwise}, \end{cases} \tag{3}$$

Further we introduce the time-varying sensor delay $\tau_j(t)$, $j = 1, \ldots, n$ where $\tau_{j-1} \le \tau_j(t) \le \tau_j$. We will consider the application where the sensor delay $\tau(t)$ is stationary, that is, $\mu(t) = \mu$ and $\sigma(t) = \sigma$, for all t. Observe that

$$\begin{aligned}&\mathbf{Pr}(\varphi_j = 1) = \mathbf{Pr}(\tau_{j-1} \le \tau(t) \le \tau_j) = \rho_j,\\ &\mathbf{Pr}(\varphi_j = 0) = 1 - \rho_j\end{aligned} \tag{4}$$

$$\mathbf{E}(\varphi_j) = \rho_j, \quad \mathbf{Var}(\varphi_j) = \rho_j(1 - \rho_j) \tag{5}$$

In this paper, we consider that the time–delay $\tau(t)$ satisfies

$$\rho_1 \le \tau(t) \le \rho_2, \quad \dot{\tau}(t) \le h, \quad 0 \le \rho_1 < \rho_2 \tag{6}$$

Let the full-order dynamic observer-based feedback control be

$$\dot{\hat{x}}(t) = K_a\hat{x}(t) + K_c y(t), \qquad u(t) = K_b\hat{x}(t), \tag{7}$$

where $\hat{x} \in R^n$ is the observer state, and the feedback gains K_a, K_b and K_c are to be designed. Denote $\delta(t) = \left[x(t)^T \hat{x}(t)^T\right]^T$ and $\rho = \text{diag}\{\rho_1, \ldots, \rho_n\}$. Then the closed-loop system of (1) with (4) and (7) is described by

$$\begin{aligned}\dot{\delta}(t) &= M\delta(t) + M_\tau\delta(t - \tau(t)) + B_{\delta w}w(t)\\ &\quad + \sum_{j=1}^{n} \big(\varphi(t) - \rho_j\big)[N\delta(t) + N_\tau\delta(t - \tau(t))]\\ z(t) &= M_z\delta(t) + B_{zw}w(t),\end{aligned} \tag{8}$$

where

$$\begin{aligned}&M = \begin{bmatrix} A & BK_b \\ \rho K_c C & K_a \end{bmatrix}, \quad N = \begin{bmatrix} 0 & 0 \\ K_c C & 0 \end{bmatrix}, \quad B_{\delta w} = \begin{bmatrix} B_{xw} \\ K_c B_{yw} \end{bmatrix},\\ &M_\tau = \begin{bmatrix} 0 & 0 \\ (I - \rho)K_c D & 0 \end{bmatrix}, \quad N_\tau = \begin{bmatrix} 0 & 0 \\ -K_c D & 0 \end{bmatrix}, \quad M_z = [\, C_z \quad B_z K_a \,]\end{aligned} \tag{9}$$

Here, although the dynamic of the closed-loop system requires only initial values of $\hat{x}(0)$, $w(0)$ and $x(t) = \phi(t)(t \in [-\rho_2, 0])$, for later convenience, we extend the range of the definition of $\phi(t)$ from $[-\rho_2, 0]$ to $[-2\rho_2, 0]$ and define a continuous function $\hat{\phi}(t)$ on $[-2\rho_2, 0]$ such that $\hat{\phi}(t) = \hat{x}(t)$. So, we have $\xi = \left[\phi(t)^T \hat{\phi}(t)^T\right]^T$ for $t \in [-\rho_2, 0]$. We also define $w(t) = 0$ for $t \in [-\tau_0, 0)$. In the sequel, we let

$$\begin{aligned}f(\delta, t) &:= M\delta(t) + M_\tau\delta(t - \tau(t)) + B_{\delta w}w(t),\\ g(\delta, t) &:= N\delta(t) + N_\tau\delta(t - \tau(t)).\end{aligned} \tag{10}$$

Since $f(\delta, t)$ and $g(\delta, t)$ in (8) satisfy the local Lipschitz condition and the linear growth condition, the existence and uniqueness of solution to (8) is guaranteed [19]. Moreover, under $v(t) = 0$ for $t \in [-\tau_0, 0)$, it admits a trivial solution (equilibrium) $\delta \equiv 0$. In this work we will follow the definitions of stochastic stability and H_∞ performance requirements.

Definition 1 System (8) is said to be exponentially mean-square stable (EMS) if there exist constants $a > 0$ and $b > 0$ such that

$$\mathbf{E}\left\{\|\delta(t)\|^2\right\} \leq ae^{-bt} \sup_{\sigma \in [-2\rho_2, 0]} \mathbf{E}\left\{\|\delta(\sigma)\|^2\right\} \tag{11}$$

Definition 2 Given $\eta > 0$, system (8) is said to be EMS with H_∞ performance (EMS-η) if under zero-initial conditions, it is EMS and satisfies

$$\int_0^\infty \mathbf{E}\left\{\|z(t)\|^2\right\} dt \leq \eta^2 \int_0^\infty \mathbf{E}\left\{\|w(t)\|^2\right\} dt \tag{12}$$

Controller for system (8) to be EMS-η will be designed.

3 Main Results

Due to the special structure of matrices M_τ and N_τ in system (8), one may choose $[I_n \; 0]\delta = x$ to construct certain terms of Lyapunov functionals in order to establish stability conditions [20]. In this work, the full information of δ is used to construct a suitable functional $J(\delta_t, t)$ and a similar type Lyapunov functional $V(\delta_t, t)$ in our study. In details, motivated by recent construction type for retarded systems in [20], we chose the following type of functionals suitable for system (8) to investigate the H_∞ performance analysis:

$$J(\delta_t, t) = J_1(\delta_t, t) + J_2(\delta_t, t) + J_3(\delta_t, t) \tag{13}$$

where $\delta_t = \delta(t + \sigma), \tau \in [-2\rho_2, 0]$ and

$$J_1(\delta_t, t) = \delta^T(t) P \delta(t), \; J_2(\delta_t, t) = \int_{t-\tau(t)}^{t} \delta^T(s) Q \delta(s) ds + \sum_{i=1}^{2} \int_{t-\tau_i}^{t} \delta^T(s) Q_i \delta(s) ds$$

$$\begin{aligned} J_3(\delta_t, t) = &\int_{-\rho_2}^{0} \int_{t+\theta}^{t} \begin{bmatrix} f(\delta, s) \\ \varphi_0 g(\delta, s) \end{bmatrix}^T Z \begin{bmatrix} f(\delta, s) \\ \varphi_0 g(\delta, s) \end{bmatrix} ds d\theta \\ &+ \int_{-\rho_2}^{-\rho_1} \int_{t+\theta}^{t} \begin{bmatrix} f(\delta, s) \\ \varphi_0 g(\delta, s) \end{bmatrix}^T Z_1 \begin{bmatrix} f(\delta, s) \\ \varphi_0 g(\delta, s) \end{bmatrix} ds d\theta \end{aligned} \tag{14}$$

in which $\hat{\rho}_j = \sqrt{\rho_j(1 - \rho_j)}$, $j = 0, \ldots, n$, $\varphi_0 = \text{diag}\{\hat{\rho}_1, \ldots, \hat{\rho}_n, 0_n\}$, and $P > 0$, $Q > 0, Q_1 > 0, Q_2 > 0, Z > 0$ and $Z_1 > 0$ are to be determined. For system (8)

with $w(t) = 0$, we use the following Lyapunov functional to obtain EMS conditions:

$$V(\delta_t, t) = V_1(\delta_t, t) + V_2(\delta_t, t) + V_3(\delta_t, t), \tag{15}$$

where $V_i(\delta_t, t) = J_i(\delta_t, t)$ with $w(t) = 0$, $i = 1, 2, 3$. Moreover, we use $\mathscr{L}V$ to denote the infinitesimal operator of V [20], which is defined by

$$\mathscr{L}V(\delta_t, t) = \lim_{\Delta \to 0^+} \frac{1}{\Delta} [\mathbf{E}\{V((\delta_{t+\Delta}, t + \Delta)|(\delta_t, t))\} - V(\delta_t, t)] \tag{16}$$

The following lemma shows that certain condition could ensure system (8) to be EMS.

Lemma 1 *Suppose that $K_a, K_b, K_c, P > 0, Q > 0, Q_i > 0, Z > 0$ and $Z_1 > 0$ are given, and $V(\varphi_t, t)$ is chosen as in (15). If there exists a constant $c > 0$ such that*

$$\boldsymbol{E}\{\mathscr{L}V(\delta_t, t)\} \leq -c\boldsymbol{E}\{\delta(T)\}$$

holds for all $t \geq 0$, then system (8) is EMS.

Proof By Definition 1, the proof is similar to [19]. □

The next lemma will be used to establish the analytical result for EMS-η.

Lemma 2 *Let $\Sigma, \Sigma_1 \in R^{p \times p}$ be symmetric constant matrices. Then,*

$$\Sigma + \tau(t)\Sigma_1 < 0$$

holds for all $\tau(t) \in [\rho_1, \rho_2]$ if and only if the following two inequalities hold:

$$\Sigma + \rho_1\Sigma_1 < 0, \quad \Sigma + \rho_2\Sigma_1 < 0$$

If this is the case, for any $z(t) \in R^p$, the following is true

$$z(t)^T(\Sigma + \tau(t)\Sigma_1)z(t) \leq \max\{\lambda_{max}(\Sigma + \rho_1\Sigma_1), \lambda_{max}(\Sigma + \rho_2\Sigma_1)\}\|z(t)\|^2$$

Proof For any $\tau(t) \in [\rho_1, \rho_2]$, there exists an $\alpha_t \in [0, 1]$ such that $\tau(t) = \alpha_t\rho_1 + (1 - \alpha_t)\rho_2$. This gives $\Sigma + \tau(t)\Sigma_1 = \alpha_t(\Sigma + \rho_1\Sigma_1) + (1 - \alpha_t)(\Sigma + \rho_2\Sigma_1) < 0$. Then

$$\begin{aligned} z(t)^T(\Sigma + \tau(t)\Sigma_1)z(t) &\leq \alpha_t\lambda_{max}(\Sigma + \rho_1\Sigma_1)\|z(t)\|^2 + \lambda_{max}(\Sigma + \rho_2\Sigma_1)\|z(t)\|^2 \\ &\leq \max\{\lambda_{max}(\Sigma + \rho_1\Sigma_1), \lambda_{max}(\Sigma + \rho_2\Sigma_1)\}\|z(t)\|^2 \end{aligned}$$

Theorem 1 *Given $\eta > 0$, the closed-loop system (8) is EMS-η if there exist $2n \times 2n$ matrices $P > 0$, $Q > 0$, $Q_1 > 0$ and $Q_2 > 0$, $4n \times 4n$ matrices $Z > 0, Z_1 > 0, L_1 > 0, L_2 > 0$ and $L_3 > 0$, $(8n + q) \times 2n$ matrices F, G and H, such that*

$$\begin{bmatrix} \Theta+\Theta_0 & \sqrt{\rho_1}F[I,I] & \sqrt{\rho_1-\rho_2}H[I,I] \\ \bullet & -L_1 & 0 \\ \bullet & \bullet & -L_3 \end{bmatrix} < 0 \tag{17}$$

$$\begin{bmatrix} \Theta+\Theta_0 & \sqrt{\rho_2}F[I,I] & \sqrt{\rho_1-\rho_2}G[I,I] \\ \bullet & -L_1 & 0 \\ \bullet & \bullet & -L_2 \end{bmatrix} < 0 \tag{18}$$

$$E_u L_1 E_u + E_l L_1 E_l - Z \le 0 \tag{19}$$

$$E_u L_2 E_u + E_l L_2 E_l - Z_1 \le 0, \quad E_u L_3 E_u + E_l L_3 E_l - Z - Z_1 \le 0 \tag{20}$$

$$\begin{bmatrix} \Xi_{11} & \Xi_{12} & \sqrt{\rho_1}\tilde{F}[I,I] & \sqrt{\rho_2-\rho_1}\tilde{H}[I,I] & \Xi_{15} & \Xi_{16} & \Xi_{17} \\ \bullet & \Xi_{22} & 0 & 0 & \Xi_{25} & 0 & 0 \\ \bullet & \bullet & -\tilde{L}_1 & 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & -\tilde{L}_3 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet & -\kappa_1 I & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet & -\kappa_1^{-1} I & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & -I \end{bmatrix} < 0 \tag{21}$$

$$\begin{bmatrix} \Xi_{11} & \Xi_{12} & \sqrt{\rho_2}\tilde{F}[I,I] & \sqrt{\rho_2-\rho_1}\tilde{G}[I,I] & \Xi_{15} & \Xi_{16} & \Xi_{17} \\ \bullet & \Xi_{22} & 0 & 0 & \Xi_{25} & 0 & 0 \\ \bullet & \bullet & -\tilde{L}_1 & 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & -\tilde{L}_2 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet & -\kappa_2 I & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet & -\kappa_2^{-1} I & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & -I \end{bmatrix} < 0 \tag{22}$$

where

$$\begin{aligned} \Theta = {} & [I_{2n} 0_{2n\times(6n+q)}]^T P \tilde{M} + \tilde{M}^T P [I,\ 0] + \tilde{M}_z^T \tilde{M}_z + F[I,-I,0] + [I,-I,0]^T F^T \\ & + \operatorname{diag}\{Q + Q_1 + Q_2, (h-1)Q, -Q_1, -Q_2, -\eta^2 I_q\} + G[0,-I,I,0] + [0,-I,I,0]^T G^T \\ & + H[0,I,0,-I,0] + [0,I,0,-I,0]^T H^T \end{aligned}$$

$$\begin{aligned} \Theta_0 = {} & [\tilde{M}^T, \varphi_0 \tilde{n}^T](\rho_2 Z + (\rho_2-\rho_1) Z_1)[\tilde{M}^T, \varphi_0 \tilde{n}^T]^T, \\ \tilde{M} = {} & [M, M_\tau, 0_{2n\times 4n}, B_{\delta w}], \quad \tilde{M}_z = [M_z, 0, B_{zw}], \quad \tilde{N} = [N, N_\tau, 0_{2n\times(4n+q)}], \\ E_u = {} & \operatorname{diag}\{I,\ 0\}, \quad E_l = \operatorname{diag}\{0,\ I\}, \end{aligned}$$

Given K_a, K_b, K_c and $\eta > 0$, the conditions of *Theorem* 1 are in terms of strict LMIs which could be easily solved using existing LMI solvers. Note that our purpose is to design LMI schemes to seek these feedback gains K_a, K_b and K_c. The maximum tolerant delay bound for ρ_2 can be searched and the minimum level of η can be computed simultaneously.

Theorem 2 *Given the delay-interval bounds $\rho_1 > 0$, $\rho_2 > 0$ and $\eta > 0$ the closed-loop system (8) is EMS-η if there exist $n \times n$ matrices $X > 0$ and $Y > 0$, $2n \times 2n$ matrices $\tilde{Q} > 0$, $\tilde{Q}_1 > 0$ and $\tilde{Q}_2 > 0$, $4n \times 4n$ matrices $\tilde{Z} > 0$, $\tilde{Z}_1 > 0$, $\tilde{L}_1 > 0$, $\tilde{L}_1 > 0$ and $\tilde{L}_1 > 0$, $(8n+q) \times 2n$ matrices $\tilde{F}, \tilde{G}$ and $\tilde{H}$, $n \times n$ matrix Υ_a, $m \times n$ matrix Υ_b and $n \times r$ matrix Υ_c, such that the following LMIs hold for some scalars $\kappa_1 > 0$ and $\kappa_2 > 0$ (21);*

$$E_u \tilde{L}_1 E_u + E_l \tilde{L}_1 E_l - \tilde{Z} < 0, \tag{23}$$

$$E_u \tilde{L}_2 E_u + E_l \tilde{L}_2 E_l - \tilde{Z}_1 < 0, \quad E_u \tilde{L}_3 E_u + E_l \tilde{L}_3 E_l - \tilde{Z} - \tilde{Z}_1 < 0 \tag{24}$$

$$\Xi_{11} = \begin{bmatrix} YA + A^T Y & \Xi_{11a} & 0 & 0 & 0 & YB_{xw} \\ & \Xi_{11c} & (1-\rho)\Upsilon_c D & (1-\rho)\Upsilon_c D & 0 & XB_{xw} + \Upsilon_c B_{yw} \\ * & 0 & 0 & 0 & 0 & \\ * & * & 0 & 0 & 0 & \\ * & ** & 0_{4n} & 0 & & \\ & ** & ** & 0_q & & \end{bmatrix}$$
$$+ \text{diag}\{\tilde{Q} + \tilde{Q}_1 + \tilde{Q}_2, (h-1)\tilde{Q}, -\tilde{Q}_1, -\tilde{Q}_2, -\eta^2 I_q\}$$
$$+ \tilde{F}[I_{2n}, -I_{2n}, 0_{2n\times(4n+q)}] + [I_{2n}, -I_{2n}, 0_{2n\times(4n+q)}]^T \tilde{F}^T$$
$$+ \tilde{G}[0_{2n}, -I_{2n}, I_{2n}, 0_{2n\times(2n+q)}] + [0_{2n}, -I_{2n}, I_{2n}, 0_{2n\times(2n+q)}]^T \tilde{G}^T$$
$$+ \tilde{H}[0_{2n}, I_{2n}, 0_{2n}, -I_{2n}, 0_{2n\times q}] + [0_{2n}, I_{2n}, 0_{2n}, -I_{2n}, 0_{2n\times q}]^T \tilde{H}^T,$$
$$\Xi_{11a} = A^T X + YA + \rho C^T \Upsilon_c^T + \Upsilon_a^T, \; \Xi_{11c} = XA + A^T X + \rho \Upsilon_c C + \rho C^T \Upsilon_c^T$$
$$\Xi_{12} = \begin{bmatrix} A^T Y & \left(A^T X + \rho C^T \Upsilon_c^T + \Upsilon_a^T\right) & 0 & \varphi_0 C^T \Upsilon_c^T \\ A^T Y & \left(A^T X + {}^T \Upsilon_c^T\right) & 0 & \varphi_0 C^T \Upsilon_c^T \\ 0 & (1-\rho) D^T \Upsilon_c^T & 0 & -\varphi_0 D^T \Upsilon_c^T \\ 0 & (1-\rho) D^T \Upsilon_c^T & 0 & -\varphi_0 D^T \Upsilon_c^T \\ 0_{4n\times n} & 0_{4n\times n} & 0_{4n\times n} & 0_{4n\times n} \\ B_{xw}^T Y & B_{xw}^T X + B_{yw}^T \Upsilon_c^T & 0 & 0 \end{bmatrix}$$
$$\Xi_{22} = -2\text{diag}\left\{ \begin{bmatrix} Y & Y \\ Y & X \end{bmatrix}, \begin{bmatrix} Y & Y \\ Y & X \end{bmatrix} \right\} + \rho_2 \tilde{Z} + (\rho_2 - \rho_1)\tilde{Z}_1,$$

$$\Xi_{15} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \; \Xi_{25} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \; \Xi_{16} = \begin{bmatrix} \Upsilon_b^T B^T \\ 0 \end{bmatrix}, \; \Xi_{17} = \begin{bmatrix} C_z^T + \Upsilon_b^T B_z^T \\ C_z^T \\ 0 \\ B_{zw}^T \end{bmatrix} \tag{25}$$

where E_u and E_l are as in Theorem 1. In this case, the feedback gains K_a, K_b and K_c are given by

$$K_a = U^{-1}(\Upsilon_a - XB\Upsilon_b)Y^{-1}W^{-T}, \; K_b = \Upsilon_b Y^{-1} W^{-T}, \; K_c = U^{-1}\Upsilon_c \tag{26}$$

where U and W are two invertible matrices satisfying $UW^T = I - XY^{-1}$.

Remark 1 Note that Theorem 2 provides an LMI method towards solving the matrix inequalities in (17–20), and hence presents controller designs of the form (7) to make the closed-loop system (8) EMS-η. The novelty of the result mainly lies in that an LMI design scheme is proposed for NCSs in continuous-time system settings with random measurements and time delays. Furthermore, the derivation is proceeded using appropriate Lyapunov functionals and matrix decoupling techniques.

4 Illustrative Example

To illustrate the theoretical developments, we consider a chemical reactor. The linearized model can be described by the following matrices:

$$A = \begin{bmatrix} -4.931 & -4.886 & 4.902 & 0 \\ -5.301 & -5.174 & -12.8 & 5.464 \\ 6.4 & 0.347 & -11.773 & -1.04 \\ 0 & 0.833 & 11.0 & -3.932 \end{bmatrix}, \quad B^t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad B_z = [1 \quad 0],$$

$$C = D = [10 \quad 0 \quad 0 \quad 0], \; C_z = [0.8 \quad 1 \quad 0.1 \quad 0.2]^T, \; B_{zw} = 0.4$$
$$A_z = [1.921 \quad 1.915 \quad 0 \quad 1.908], \; B_{xw} = [0.8 \quad 1 \quad 0.1 \quad 0.2]^T, \; B_{yw} = 0.01$$

Using the LMI toolbox in MATLAB, the ensuing results are summarized by:

$$X = \begin{bmatrix} 0.1448 & -0.0020 & 0.0005 & 0.0002 \\ -0.0020 & 0.1442 & -0.0005 & 0.0009 \\ 0.0005 & -0.0005 & 0.1420 & 0.0001 \\ 0.0002 & 0.0009 & 0.0001 & 0.1463 \end{bmatrix},$$

$$Y = \begin{bmatrix} 0.2142 & -0.0560 & 0.0421 & -0.0153 \\ -0.0560 & 0.0383 & -0.0138 & 0.0515 \\ 0.0421 & -0.0138 & 0.0292 & 0.0435 \\ -0.0153 & 0.0515 & 0.0435 & 0.2597 \end{bmatrix},$$

With $W = I$ and $U = I - XY^{-1}$, the corresponding feedback gains

$$K_a = \begin{bmatrix} 0.7573 & 0.7142 & 0.3973 & 0.8391 \\ 0.2138 & 8.2185 & 13.8882 & -3.4177 \end{bmatrix},$$
$$K_b = \begin{bmatrix} 0.3144 & -0.7983 & -3.8703 & 1.7806 \\ -0.6559 & 7.2776 & 14.8651 & -6.8895 \end{bmatrix},$$
$$K_c = \begin{bmatrix} 0.2634 & -0.1587 & -3.1912 & 1.5713 \\ -1.0803 & 8.5794 & 12.2875 & -3.9658 \end{bmatrix}$$

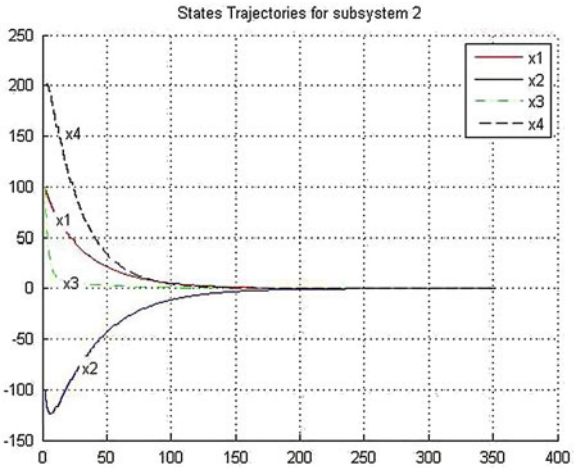


Fig. 1 Closed-loop state trajectories

Simulation of the closed-loop system is performed and the ensuing state trajectories are presented in Fig. 1. It is evident that the the closed-loop system is EMSS-η.

5 Conclusion

An LMI method has been presented for observer-based H_∞ control of NCSs in continuous-time system settings with random measurements and probabilistic time delays. Improved schemes have been shown for the design method. It has been established that these conditions reduce the conservatism by considering not only the range of the time delays, but also the probability distribution of their variation. A numerical simulation example has been presented to show the merits and advantages of the proposed techniques.

Acknowledgements The author would like to thank the deanship for scientific research (DSR) at KFUPM for financial support through research group project **RG-1316-1**.

A.1 6 Appendix

*Proof of Theorem 1*The proof is twofold: we first choose a functional J of the form (13) to show that the H_∞ performance requirement (12) is satisfied, and then use the Lyapunov functional V of the form (15) to prove the EMS property. Denote

$$\chi(t) := [\delta(t)^T, \delta_\tau^T, \delta_1^T, \delta_2^T, w(t)^T]^T, \quad \delta_\tau := \delta(t-\delta_t), \quad \delta_i := \delta(t-\delta_i), \quad i = 1, 2. \tag{27}$$

From the Newton–Leibniz formula $0 = \delta(t) - \delta_\tau - \int_{t-\tau(t)}^{t} \dot{\delta}(s)ds$, we have that

$$\begin{aligned}
\psi_1(t) :&= 2\chi(t)^T F\left[\delta(t) - \delta_\tau - \int_{t-\tau(t)}^{t} \dot{\delta}(s)ds\right] = 0,\\
\psi_2(t) :&= 2\chi(t)^T G\left[\delta_1 - \delta_\tau - \int_{t-\tau(t)}^{t-\rho_1} \dot{\delta}(s)ds\right] = 0,\\
\psi_3(t) :&= 2\chi(t)^T G\left[\delta_\tau - \delta_2 - \int_{t-\rho_2}^{t-\tau(t)} \dot{\delta}(s)ds\right] = 0
\end{aligned} \tag{28}$$

hold for any $(8n+q)\times 2n$ matrices F, G and H. Let the functional $J(\delta_t; t)$ be chosen as in (13), then, from (16), $\mathscr{L}J$ for the evolution of J is given by [20]

$$\begin{aligned}
\mathscr{L}J(\delta_t;t) &= 2\delta(t)^T P f(\delta,t) + \delta(t)^T(Q+Q_1+Q_2)\delta(t)\\
&\quad - (1-\dot{\tau}_t)\delta_\tau^T Q\delta_\tau - \sum_{i=1}^{2}\delta_i^T Q_i\delta_i\\
&\quad + \begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T(\rho_2 Z + (\rho_2-\rho_1)Z_1)\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix} - \int_{t-\rho_2}^{t}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T Z\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds\\
&\quad - \int_{t+\rho_2}^{t+\rho_1}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T Z_1\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds\\
&= 2\delta(t)^T P f(\delta,t) + \delta(t)^T(Q+Q_1+Q_2)\delta(t) - (1-\dot{\tau}_t)\delta_\tau^T Q\delta_\tau - \sum_{i=1}^{2}\delta_i^T Q_i\delta_i\\
&\quad + \psi_1(t)+\psi_2(t)+\psi_3(t) + \begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T(\rho_2 Z + (\rho_2-\rho_1)Z_1)\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}\\
&\quad - \int_{t-\tau(t)}^{t}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T Z\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds - \int_{t-\tau(t)}^{t-\rho_1}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T Z_1\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds\\
&\quad - \int_{t-\rho_2}^{t-\tau(t)}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T(Z+Z_1)\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds
\end{aligned} \tag{29}$$

Note that, in $\psi_1(t)$, the following inequality holds for any $4n \times 4n$ matrix $L > 0$, *thus* :

$$-2\chi(t)^T F \int_{t-\tau(t)}^{t} \dot{\delta}(s)ds = -2\chi(t)^T F[I_{2n}, I_{2n}] \int_{t-\tau(t)}^{t} \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix} ds$$
$$\leq \tau(t)\chi(t)^T F[I_{2n}, I_{2n}]L_1^{-1}[I_{2n}, I_{2n}]^T F^T \chi(t)$$
$$+ \int_{t-\tau(t)}^{t} \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix}^T L_1 \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix} ds$$

Similarly, in $\psi_i(t)$ $(i = 2, 3)$, the following inequalities hold for any $4n \times 4n$ matrices $L_i > 0$:

$$-2\chi(t)^T G \int_{t-\tau(t)}^{t-\rho_1} \dot{\delta}(s)ds \leq (\tau(t)-\rho_1)\chi(t)^T G[I, I]L_2^{-1}[I, I]^T G^T \chi(t)$$
$$+ \int_{t-\tau(t)}^{t-\rho_1} \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix}^T L_2 \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix} ds$$

and

$$-2\chi(t)^T H \int_{t-\tau(t)}^{t-\rho_1} \dot{\delta}(s)ds \leq (\rho_2-\tau(t))\chi(t)^T H[I, I]L_3^{-1}[I, I]^T H^T \chi(t)$$
$$+ \int_{t-\rho_2}^{t-\tau(t)} \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix}^T L_3 \begin{bmatrix} f(\delta,s) \\ (\varphi(s)-\rho)g(\delta,s) \end{bmatrix} ds$$

Considering (6, 8, and 10) and taking the expectation on (29), we have

$$I\,E\{\mathscr{L}V(\delta_t,t) + \|z(t)\|^2 - \eta^2\|w(t)\|^2\} \leq I\,E\{\chi(t)^T(\Theta + \Theta_0 + \tau(t)\Theta_1 + (\tau(t)-\rho_1)\Theta_2 + (\rho_2-\tau(t))\Theta_3)\chi(t)\} + \psi_4(t) \tag{30}$$

and

$$\Theta_1 = F[I, I]L_1^{-1}[I, I]^T F^T, \quad \Theta_2 = G[I, I]L_2^{-1}[I, I]^T G^T,$$
$$\Theta_3 = H[I, I]L_3^{-1}[I, I]^T H^T,$$
$$\psi_4(t) = \int_{t-\tau(t)}^{t} \begin{bmatrix} f(\delta,s) \\ \varphi_0 g(\delta,s) \end{bmatrix}^T (E_u L_1 E_u + E_l L_1 E_l - Z) \begin{bmatrix} f(\delta,s) \\ \varphi_0 g(\delta,s) \end{bmatrix} ds$$

$$+\int_{t-\tau(t)}^{t-\rho_1}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T(E_uL_2E_u+E_lL_2E_l-Z_1)\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds$$

$$+\int_{t-\rho_2}^{t-\tau(t)}\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}^T(E_uL_3E_u+E_lL_3E_l-Z-Z_1)\begin{bmatrix} f(\delta,s)\\ \varphi_0 g(\delta,s)\end{bmatrix}ds$$

Applying the Schur complement, conditions (17) and (18) are equivalent to

$$\tilde{\Theta}_1=\Theta+\Theta_0+\rho_1\Theta_1+(\rho_2-\rho_1)\Theta_3<0, \tag{31}$$

$$\tilde{\Theta}_2=\Theta+\Theta_0+\rho_2\Theta_1+(\rho_2-\rho_1)\Theta_2<0, \tag{32}$$

From (19–20, 31, 32) and Lemma 2, we deduce from (30) that

$$\mathbf{E}\{\mathscr{L}J(\delta_t,t)+\|z(t)\|^2-\eta^2\|w(t)\|^2\}\le\max\{\lambda_{max}(\tilde{\Theta}_1),\lambda_{max}(\tilde{\Theta}_2)\}\ \ \mathbf{E}\{\|\chi\|^2\}\le 0 \tag{33}$$

Under zero-initial conditions and noticing $J(\delta_T;T)\ge 0$ for any $T>0$, integrating (33) from 0 to ∞ yields that the H_∞ performance requirement (12) is satisfied. With a procedure similar to the above, we can arrive under the given conditions and by virtue of Lemma 2 that,

$$\mathbf{E}\{\mathscr{L}V(\delta_t,t)\}\le\max\{\lambda_{max}(\tilde{\Theta}_1),\lambda_{max}(\tilde{\Theta}_2)\}\ \ \mathbf{E}\{\|\delta(t)\|^2\}$$

Hence, system (8) is EMS from *Lemma* 1.

*Proof of Theorem 2*It can be seen from (21) or (22) that

$$\begin{bmatrix} Y & Y\\ Y & X\end{bmatrix}>0$$

which gives $XY>0$, implying that $I--XY^{-1}$ is invertible. Now let U and W be any invertible matrices satisfying $UW^T=I-XY^{-1}$. Choose

$$P=\begin{bmatrix} X & U\\ U^T & *\end{bmatrix}>0,\ P^{-1}=\begin{bmatrix} Y^{-1} & W\\ W^T & *\end{bmatrix}>0 \tag{34}$$

where each ellipsis $*$ denotes a positive definite matrix block that will not influence the subsequent development (of course it makes $PP^{-1}=I$). In the sequel, we show that if (21, 24) are satisfied, then (17, 20) hold with $P>0$ chosen as in (39), and thus the result follows immediately from Theorem 1. Define

$$S=\begin{bmatrix} I & I\\ W^TY & 0\end{bmatrix} \tag{35}$$

which is invertible and produces

$$S^T P = \begin{bmatrix} Y & 0 \\ X & U \end{bmatrix}, \quad S^T P S = \begin{bmatrix} Y & Y \\ Y & X \end{bmatrix}. \tag{36}$$

We first show that (21) implies (17). By Schur complement, the matrix inequality (17) holds if and only if (37)

$$\begin{bmatrix} \Theta & [\tilde{M}^T, \varphi_0 \tilde{N}^T]\mathrm{diag}\{P,P\} & \sqrt{\rho_1}F[I,I] & \sqrt{\rho_2-\rho_1}H[I,I] \\ \bullet & -\Pi & 0 & 0 \\ \bullet & \bullet & -L_1 & 0 \\ \bullet & \bullet & \bullet & -L_3 \end{bmatrix} < 0 \tag{37}$$

$$\Pi = \mathrm{diag}\{P,P\}(\rho_2 Z + (\rho_2-\rho_1)Z_1)^{-1}\mathrm{diag}\{P,P\}$$

In view of

$$\begin{aligned} &(\rho_2 Z + (\rho_2-\rho_1)Z_1 - \mathrm{diag}\{P,P\})(\rho_2 Z + (\rho_2-\rho_1)Z_1)^{-1} \\ &times(\rho_2 Z + (\rho_2-\rho_1)Z_1 - \mathrm{diag}\{P,P\}) > 0 \end{aligned} \tag{38}$$

we obtain

$$\begin{aligned} &-\mathrm{diag}\{P,P\}(\rho_2 Z + (\rho_2-\rho_1)Z_1)^{-1}\mathrm{diag}\{P,P\} \le -2\mathrm{diag}\{P,P\} \\ &+\rho_2 Z + (\rho_2-\rho_1)Z_1 \end{aligned} \tag{39}$$

we have that (37) holds if (40) holds

$$\begin{bmatrix} \Theta & [\tilde{M}^T, \varphi_0 \tilde{N}^T]\mathrm{diag}\{P,P\} & \sqrt{\rho_1}F[I,I] & \sqrt{\rho_2-\rho_1}H[I,I] \\ \bullet & -2\mathrm{diag}\{P,P\} + \rho_2 Z + (\rho_2-\rho_1)Z_1 & 0 & 0 \\ \bullet & \bullet & -L_1 & 0 \\ \bullet & \bullet & \bullet & -L_3 \end{bmatrix} < 0 \tag{40}$$

Now, applying the congruence transformation $\mathrm{diag}\{S,S,S,S,I_q,S,S,S,S,S,S\}$ to (40) and setting

$$\begin{aligned} &\tilde{Q} = S^T Q S, \quad \tilde{Q}_1 = S^T Q_1 S, \quad \tilde{Q}_2 = S^T Q_2 S, \\ &\tilde{Z} = \mathrm{diag}\{S,S\}^T Z \mathrm{diag}\{S,S\}, \quad \tilde{Z}_1 = \mathrm{diag}\{S,S\}^T Z_1 \mathrm{diag}\{S,S\}, \\ &\tilde{L}_i = \mathrm{diag}\{S,S\}^T L_i \mathrm{diag}\{S,S\}, \quad i = 1,2,3 \\ &\tilde{F} = \mathrm{diag}\{S,S,S,S\}^T FS, \quad \tilde{G} = \mathrm{diag}\{S,S,S,S\}^T GS, \\ &\tilde{H} = \mathrm{diag}\{S,S,S,S\}^T HS, \quad \Upsilon_a = XBK_b W^T Y + UK_a W^T Y, \\ &\Upsilon_b = K_b W^T Y, \quad \Upsilon_c = UK_c \end{aligned} \tag{41}$$

we obtain that (40) is equivalent to

$$\tilde{\Xi} + \tilde{Y}\tilde{K} + \tilde{K}^T\tilde{Y}^T < 0, \tag{42}$$

$$\tilde{\Xi} = \begin{bmatrix} \Xi_{11} + \Xi_{17}^T\Xi_{17} & \Xi_{12} & \sqrt{\rho_1}\tilde{F}[I,I] & \sqrt{\rho_2 - \rho_1}\tilde{H}[I,I] \\ \bullet & \Xi_{22} & 0 & 0 \\ \bullet & \bullet & -\tilde{L}_1 & 0 \\ \bullet & \bullet & \bullet & -\tilde{L}_3 \end{bmatrix}$$

$$\tilde{Y} = \left[\Xi_{15}^T, \Xi_{25}^T, 0_{n\times 8n}\right]^T, \quad \tilde{K} = \left[\Xi_{16}^T, 0_{n\times 12n}\right]$$

Inequality (42) holds if the following is true for any $\kappa_1 > 0$,

$$\tilde{\Xi} + \kappa_1^{-1}\tilde{Y}\tilde{Y}^T + \kappa_1\tilde{K}^T\tilde{K} < 0, \tag{43}$$

which is equivalent to

$$\begin{bmatrix} \tilde{\Xi} & \tilde{Y} & \tilde{K}^T \\ \tilde{Y}^T & -\kappa_1 I_n & 0 \\ \tilde{K} & 0 & -\kappa_1^{-1} I_n \end{bmatrix} < 0 \tag{44}$$

The above inequality is, by Schur complement again, exactly that of (21), and we conclude that this implies (17).

Next we show that (22) implies (18). This can be done by using a procedure analogous to the above. As for the verification of other inequalities, applying the congruence transformation diag{S, S} to (19, 20) and setting matrix variables as in (41), it is seen that (19, 20) are equivalent to (23, 24). So far, we have proven that (21– 24) ensure (17–20) and thus the closed-loop system (8) is EMS-η. In this case, from (41), the feedback gains are computed as in (26). This completes the proof.

References

1. L.Q. Zhang, S. Yang, T.W. Chen, B. Huang, A new method for stabilization of networked control systems with random delays. IEEE Trans. Automat. Control **50**(8), 1177–1181 (2005)
2. J.P. Hespanha, Y.G. Xu, A survey of recent results in networked control systems. Proc. IEEE **95**(1), 138–162 (2007)
3. W. Heemels, A. Teel, N. van de Wouw, D. Nesic, Networked control systems with communication constraints: tradeoffs between transmission intervals, delays and performance. IEEE Trans. Automat. Control **55**(8), 1781–1796 (2010)
4. M.S. Mahmoud, *Robust Control and Filtering for Time-delay Systems* (Marcel-Dekker, New York, 2000)
5. F. Yang, Z. Wang, Y.S. Hung, M. Gani, $\mathscr{H}_\infty$ control for networked systems with random communication delays. IEEE Trans. Automat. Control. **51**(3), 511–518 (2007)

6. D. Yue, E. Tian, Z. Wang, J. Lam, Stabilization of systems with probabilistic interval input delays and its applications to networked control systems. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **39**(42), 939–945 (2009)
7. M.G. Rivera, A. Barreiro, Analysis of networked control systems with drops and variable delays. Automatica **43**(9), 2054–2059 (2007)
8. J. Xiong, J. Lam, Stabilization of linear systems over networks with bounded packet loss. Automatica **43**(1), 80–87 (2007)
9. L. Zhang, Y. Shi, T. Chen, B. Huang, A new method for stabilization of networked control systems with random delays. IEEE Trans. Automat. Control **50**(8), 1177–1181 (2006)
10. M.S. Mahmoud, N. Hazem, Nounou, Y. Xia, Dissipative control for internet-based switching systems. J. Franklin Inst. **347**(1), 154–172 (2011)
11. B. Liu, Y. Xia, M.S. Mahmoud, H. Wu, New predictive control scheme for networked control systems. J. Circuits, Syst. Signal Process. **31**(2), 945–960 (2012)
12. M.S. Mahmoud, A.W. Saif, Robust quantized approach to fuzzy networked control systems. IEEE J. Emerg. Sel. Top. Circuits Syst. (JETCAS) **2**(1), 71–81 (2012)
13. M.S. Mahmoud, Improved networked control systems approach with communication constraint. IMA J. Math. Control Inf. **29**(2), 215–233 (2012)
14. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M.I. Jordan, S.S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Automat. Control **49**, 1453–1464 (2004)
15. H.S. Zhang, L.H. Xie, *Control and Estimation of Systems with Input/Output Delays* (Springer, Berlin, 2007)
16. L. Hetel, J. Daafouz, C. Lung, Analysis and control of LTI and switched systems in digital loops via an event-based modelling. Int. J. Control **81**(7), 1125–1138 (2008)
17. Y. Shi, B. Yu, Robust mixed $\mathscr{H}_2/\mathscr{H}_\infty$ control of networked control systems with random time delays in both forward and backward communication links. Automatica **47**(4), 754–760 (2011)
18. M.S. Mahmoud, in *WCE 2013*: *Output-Feedback Control for Networked Systems with Probabilistic Delays*. Proceedings of the World Congress on Engineering 2013. Lecture Notes in Engineering and Computer Science, (London, 3–5 July 2013), pp. 1070–1077
19. M.S. Mahmoud, A.Y. Al-Rayyah, Efficient parameterization to stability and feedback synthesis of linear time-delay systems. IET Control Theor. Appl. **3**(8), 1107–1118 (2009)
20. M.S. Mahmoud, S.Z. Selim, P. Shi, Global exponential stability criteria for neural networks with probabilistic delays. IET Control Theor. Appl. **4**(11), 2405–2415 (2011)

A Probabilistic Method for Optimal Power Systems Planning with Wind Generators

Maryam Dadkhah and Bala Venkatesh

Abstract Radial Distribution Systems (RDS) connect a large number of renewable generators that are inherently uncertain. From being unidirectional power flow systems, RDS now enable bi-directional power flow. Depending upon availability of power from renewables, they receive or feed power to the connected transmission system. RDS optimal power flow (OPF), is an important tool in this new era for utilities, to minimize losses and operate efficiently. With large scale integration of wind generators to distribution systems, they must be appropriately represented using probabilistic models capturing their intermittent nature in these OPF algorithms. This paper proposes characterizing the solution of a Probabilistic Optimal Power Flow (P-OPF) for RDS using the Cumulant Method. This method makes it possible to linearly relate the probabilistic parameters of renewables at the optimal solution point to the state of the RDS. To assess the accuracy of the proposed P-OPF Cumulant Method, wind generators and system probabilistic data are incorporated in a 33-bus and 129-bus test system. The results are compared with those of Monte Carlo simulations (MCS). It is shown that the proposed method possesses high degree of accuracy, is significantly faster and more practical than an MCS approach.

Keywords Cumulant method · Optimal power flow · Radial distribution systems · Reactive power optimization · Stochastic optimization · Wind energy

M. Dadkhah · B. Venkatesh (✉)
Centre for Urban Energy, Ryerson University, 350 Victoria Street, Toronto, M5B 2K3, Canada
e-mail: bala@ryerson.ca

M. Dadkhah
e-mail: maryam.dadkhah@ryerson.ca

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_27,

1 Introduction

OPF (optimal power flow) is a versatile tool used for electric transmission systems for a variety of purposes. The most common amongst them are: (a) real power OPF where real power output of generators are scheduled such that the total cost of generation is minimized [1], and (b) reactive power OPF where generators voltages, reactive power compensation and settings of transformers taps are set to route reactive power optimally such that real power transmission losses are the least and all the voltages are within prescribed limits [2]. OPF has not been easily extended to distribution systems as their system Jacobian is ill-conditioned owing to higher R/X ratio of their lines [3]. In the recent past, numerous Jacobian based OPF methods have been researched and published [4]. Today, with a rush to integrate wind generators to electric power systems, largely to distributions systems, distribution systems OPF must account for wind generators as well.

In essence, an OPF for distribution systems must contend with the challenge that it must account for wind generators that are uncertain in their output and their near term forecasts can be best represented by a normal distribution with mean and variance values [5]. Further to understand the effect of probabilistic nature of loads and availability of wind on the OPF solution, such as the optimal values of transformer taps and capacitor settings, it is necessary to propose an efficient probabilistic OPF method that includes the load and generator probabilistic models. The ultimate goal is to determine the probability density function of typical variables such as voltage and power flow that form a part of OPF solution.

Uncertainties of the power systems components have been addressed with many researchers by adapting probabilistic techniques in the Power Flow solution in transmission systems since 1960s [6]. Later, the probabilistic methods were applied to the optimal dispatch [7] and for the first time the term P-OPF was used in [8]. However, in contrast to the transmission system case, distribution systems have not been studied to the same extent. In [9] the authors proposed a probabilistic optimal capacitor planning method using Cumulant technique to find the probabilistic information of the size of newly installed capacitor banks in the distribution systems with high penetration of wind generations.

This paper uses the idea given in [9] to propose and construct a distribution system OPF using a set of 3N equations such that the Jacobian is robust [10]. The objective of the OPF is to minimize losses in the distribution system by optimally scheduling all the reactive power sources and ensuring that voltages are within the prescribed limits. Then, it proposes to use Cumulant Method (CM) to directly relate probabilistic values of the loads and output power of wind generators to the optimal settings of the distribution system [11]. This approach has been reported for uncertainty without specific application to wind generators and in transmission system by Schellenberg et al. [12].

This paper is an extension to the Ref. [13] and outlined as follows. In Sects. 2 and 3, the model of the radial distribution system for OPF solution is presented and the Cumulant method is described respectively. Section 4 presents the numerical

results of the method as tested on the 33-bus IEEE test system with three wind generators and a 129-bus test system with nine wind generators. Section 5 concludes the paper.

2 System Model and Probabilistic OPF

2.1 Problem Formulation

This subsection uses the radial distribution system (RDS) model from [10]. Figure 1 shows a single-line representation of a tree-like distribution systems structure.

Consider the ith bus in Fig. 1. It has a wind turbine connected to it that injects only real power equal to PW_i. Its bus load is represented by $SD_i = PD_i + j \cdot QD_i$. The total power injected into this bus is $SB_i = SD_i - PW_i$. It is the difference between generation and load at that bus. Consider the *l*th line/transformer between buses i − 1 and i. The tap setting of this transformer/line is represented by T_l and it has an impedance of $Zl = Rl + j \cdot Xl$. The total apparent power reaching the downstream end of this line equals ST_l. The real power loss on this line equals:

$$PL_l = R_l \cdot |ST_l|^2 \cdot V_i^{-2} \tag{1}$$

The total real power loss in all feeders of the system equals:

$$TPL = \sum_{l=1}^{nl} R_l \cdot |ST_l|^2 \cdot V_i^{-2} \tag{2}$$

where V_i is the bus voltage magnitude, *nb* is the number of buses in the system and nl is the number of lines/transformers.

In Fig. 1, the complex power balance at the *i*th can be expressed as:

$$ST_i = SD_i + \left[\sum_{(l,k)=(kl3,k3)}^{(kl3,k3)} Z_l \cdot |ST_l|^2 \cdot V_k^{-2} + ST_k \right] - PW_i - j \cdot QS_i \tag{3}$$

Where QS_i is the reactive power injected into the *i*th bus. Equation (3) is a complex equation and yields a set of 2(NB − 1) equations. Writing the voltage drop equation across line *l* gives:

$$V_i^4 + 2 \cdot V_i^2 \cdot \left[PT_l \cdot R_l + QT_l \cdot X_l - \frac{1}{2} \cdot \frac{V_{i-1}^2}{T_l^2} \right] - |Z_l|^2 \cdot |ST_l|^2 = 0 \tag{4}$$

Equations (3) and (4) provide 3(NB − 1) equations that completely model a RDS.

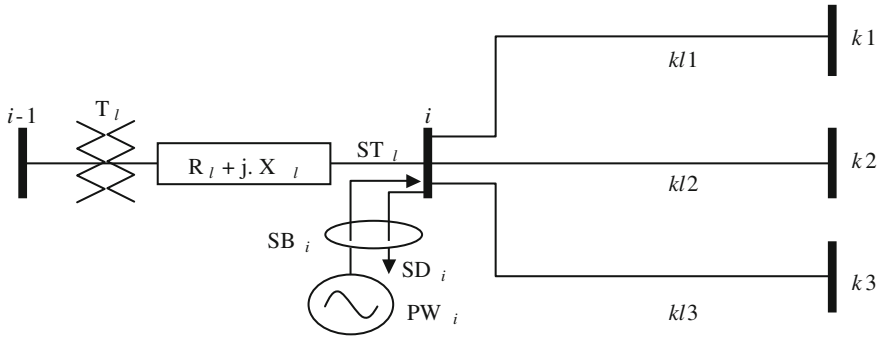


Fig. 1 A tree-like distribution system with wind generator

2.2 OPF for Radial Distribution System

The objective of Radial Distribution System OPF is to minimize the total real power loss. By referring to the set of Eqs. (2)–(4), one may construct an optimal power flow formulation for a radial distribution system as below:

Objective Function:

$$\text{Minimize: TPL} = \sum_{l=1}^{\text{nl}} \text{R}_l \cdot |\text{ST}_l|^2 \cdot \text{V}_\text{i}^{-2} \tag{5}$$

Constraints:

$$\text{ST}_\text{i} = \text{SD}_\text{i} + \left[\sum_{(l,\text{k})=(\text{k}l1,\text{k}1)}^{(\text{k}l3,\text{k}3)} \text{Z}_l \cdot |\text{ST}_l|^2 \cdot \text{V}_\text{k}^{-2} + \text{ST}_\text{k} \right] - \text{PW}_\text{i} - \text{j} \cdot \text{QS}_\text{i} \tag{6}$$

$$\text{V}_\text{i}^4 + 2 \cdot \text{V}_\text{i}^2 \cdot \left[\text{PT}_l \cdot \text{R}_l + \text{QT}_l \cdot \text{X}_l - \frac{1}{2} \cdot \frac{\text{V}_{\text{i}-1}^2}{\text{T}_l^2} \right] - |\text{Z}_l|^2 \cdot |\text{ST}_l|^2 = 0 \tag{7}$$

$$\text{U}_\text{MIN} < \text{U} < \text{U}_\text{MAX} \tag{8}$$

$$\text{V}_\text{MIN} < \text{V} < \text{V}_\text{MAX} \tag{9}$$

where the decision vector is U = [QS, T] and dependent vector is Y = [V, P, Q]. Equations (6) and (7) are equality constraints which correspond to the complex power balance equation and the voltage drop equation across line *l*, respectively. Equations (8) and (9) limit the control and dependent vectors. The optimization problem described by (5)–(9) is solved by using the Logarithmic-Barrier Interior Point Method (LBIPM) [14].

2.3 Optimal Solution

The formulation (5)–(9) is solved using the Lower Bound Interior Point Method [14]. This yields the optimal solution of decision and dependent vectors U and Y. In addition, the Hessian of the Lagrangian formed from the optimization formulation (5)–(9) is evaluated H (U, Y, λ). It provides a linear relation between incremental changes of dependent vector in terms of the decision vector.

3 Cumulant Technique and P-OPF

In probability theory, Cumulants and moments are two sets of quantities of a random variable which are mathematically equivalent. However, in some cases preference is to use Cumulants due to their simplicity over using moments [15]. In this section some properties of Cumulants used to adapt the Cumulant Technique to the radial distribution system P-OPF are presented.

Consider a linear combination of 'n' independent random input variables α used to create a new random output variable β as follows [12]:

$$\beta = c_1 \cdot \alpha_1 + c_2 \cdot \alpha_2 + c_3 \cdot \alpha_3 + \cdots + c_n \cdot \alpha_n \tag{10}$$

where ci is the *i*th coefficient in the linear combination. The above expansion can be written in terms of the moment generation function of random variable β, i.e., $\Phi_\beta(s)$, as

$$\Phi_\beta(s) = E\left[e^{s\beta}\right] = E\left[e^{s\left(c_1\alpha_1 + c_2\alpha_2 + \cdots + c_n\alpha_n\right)}\right] \tag{11}$$

where s is the Laplace operator. Assuming that α_1, α_2, α_3, …, α_n are independent, the above relationship can be written as

$$\begin{aligned}\Phi_\beta(s) &= E[e^{s\beta}] = E[e^{sc_1\alpha_1}] \cdot E[e^{sc_2\alpha_2}] \cdots E[e^{sc_n\alpha_n}] \\ &= \Phi_{\alpha 1}(sc_1) \cdot \Phi_{\alpha 2}(sc_2) \cdots \Phi_{\alpha n}(sc_n)\end{aligned} \tag{12}$$

The Cumulant generating function $\Psi_x(s)$ can be written in terms of the moment generating function Φ_x as [15]

$$\Psi_\beta(s) = \ln\left(\Phi_\beta(s)\right). \tag{13}$$

By taking the natural logarithm at both sides of the (12) and using (13), (12) is written in the terms of cumulant generating function as below:

$$\Psi_\beta(s) = \Psi_{\alpha 1}(sc_1) + \Psi_{\alpha 2}(sc_2) + \cdots + \Psi_{\alpha n}(sc_n). \tag{14}$$

To obtain the different orders of cumulants, we can set $s = 0$ to compute the different order derivatives of the cumulant generating function. A general equation for the mth order cumulant of Ψ is

$$\Psi_{\beta}^{(m)}(0) = c_1^m \cdot \Psi_{\alpha 1}^{(m)}(0) + c_2^m \Psi_{\alpha 2}^{(m)}(0) + \cdots + c_n^m \Psi_{\alpha n}^{(m)}(0) \tag{15}$$

$$K_{\beta,m} = c_1^m \cdot K_{\alpha 1,m} + c_2^m \cdot K_{\alpha 2,m}(0) + \cdots + c_n^m K_{\alpha n,m}(0). \tag{16}$$

where $K_{\beta,m}$ is a vector containing the mth order cumulants of the system unknown variables and $K_{\alpha,m}$ is a vector containing the mth order cumulants of the random bus generation and loading.

3.1 Adaptation to P-OPF

By applying Newton method to the Lagrangian function L (U, Y, λ) for (5)–(9), the following system is obtained:

$$\nabla L(U, Y, \lambda) + H(U,Y,\lambda) \times \begin{bmatrix} \Delta U \\ \Delta Y \\ \Delta \lambda \end{bmatrix} = 0 \tag{17}$$

where ∇L(U, Y, λ) and ∇H(U, Y, λ) are the gradient and the Hessian of the Lagrangian respectively. Rearranging (16) and replacing ∇L(U, Y, λ) with a vector of change in the bus power injections for uncertain wind power, ΔSB, the vector of changes can be linearly mapped with ΔSB by using the inverse of the Hessian:

$$\begin{bmatrix} \Delta U \\ \Delta Y \\ \Delta \lambda \end{bmatrix} = -H(U,Y,\lambda)^{-1} \times \begin{bmatrix} \nabla_{U,Y} L(U,Y,\lambda) \\ \Delta SB \end{bmatrix} \tag{18}$$

By replacing ΔSB with a vector containing the nth order cumulants of loads and generation, the Cumulants of system variables, ΔSB can obtain using the inverse of the Hessian as follow:

$$K_{(U,Y,\lambda),n} = \left(-H^{-1}\right)^{(n)} \times \begin{bmatrix} 0 \\ K_{SB,n} \end{bmatrix} \tag{19}$$

where $K_{(U,Y,\lambda),n}$ is a vector of nth-order cumulants for the optimal settings of the distribution system and $K_{SB,n}$ is a vector of nth-order Cumulants for the random bus power injections. Consequently, the Hessian contains the constant multipliers. Once the cumulants of the random variables of the OPF solution are computed from the input random variables, PDFs are recreated by using Gram-Charlier/ Edgeworth Expansion theory [16].

4 Numerical Results

This section provides the results based on applying the Cumulant method to the 33-bus and 129-bus test systems. In both systems, the loads and the power output of wind generators are considered Gaussian random variables with the mean values set to the nominal bus loading and mean capacity of wind generators respectively. The standard deviation is such that the 99 % confidence interval is equal to ±15 % of the nominal loading value. In order to show the efficiency and accuracy of the Cumulant method, the results have been compared with MCS with 5,000 samples.

4.1 33-Bus Test System Case Study

The method is firstly applied to the 33-bus, 32-branch IEEE test system described in [17]. The system, however, is modified to accommodate the probabilistic data of the loads and wind generators. The modified system and loads data can be found in the Appendix. Three wind turbines are connected to buses 3, 17 and 32 with a mean capacity equal to 500 kW each. The results of both mean and standard deviation values obtained by comparing with those of 5,000 sample points MCS, are discussed as follows.

(1) **Mean Values:** From Table 1 it can be seen that the percentage errors of the voltage mean values are very small with the maximum value equal to 0.0077 % which occurs at bus 18. The maximum percentage error of the real power and reactive power mean values is equal to 3.44 and 7.97 % at bus 29 and 5 respectively. The corresponding maximum error for capacitor value occurs at bus 18 and is as low as 0.1901 %. These results show a small difference between two methods in systems variables mean values which can be seen more clearly in Fig. 2.

(2) **Variance Value:** The maximum percentage errors for the system variables variance values are presented in Table 2. These values for the voltage, active power, reactive power and capacitor value variance are equal to 1.55, 1.91, 2.13 and 1.85 %, which occur at buses 24, 9, 11 and 30 respectively. These small error values for the variance of the systems variables are shown in Fig. 3.

In summary, the percentage errors of the mean and variance values for the system variables are well below 8 %, which implies a close match between two methods. Also, it is worth noting that largest errors (8 %) occur for reactive power and voltage magnitude variables due to inherent nonlinearity. This nonlinearity for voltage and reactive power usually happens in the buses with capacitor banks connected to them.

Table 1 Maximum error of mean values of the system variables

	Bus no.	CM (per unit)	MCS (per unit)	Error (%)
Voltage	18	0.98	0.98	0.0077
Active power	29	0.03	0.03	3.44
Reactive power	5	−0.0045	−0.0049	7.97
Capacitor value	18	0.15	0.15	0.19

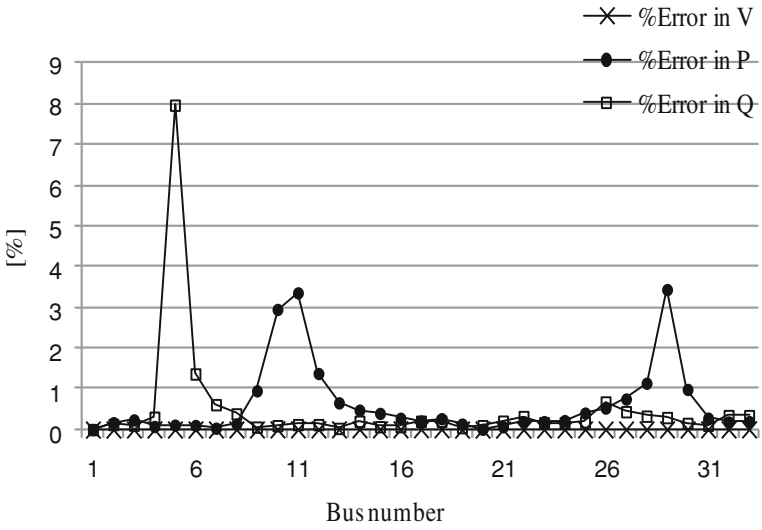


Fig. 2 Error in mean of the output variables using the cumulant method and MCS technique—33-bus system

Table 2 Maximum error of variance values of the system variables

	Bus no.	CM	MCS	Error (%)
Voltage	24	0.0013	0.0012	1.55
Active power	9	0.079	0.0811	1.91
Reactive power	11	0.0062	0.0061	2.13
Capacitor value	30	0.091	0.093	1.85

The analysis using Cumulant method is captured in graphs of Figs. 4 and 5 wherein mean capacitor and bus voltage magnitude values are shown with potential spread using corresponding 3σ values. This analysis and graphing can be rapidly completed using the proposed method.

Table 3 presents the time comparison between Cumulant method and Monte Carlo Simulation technique. It is evident that to identify this spread in values of

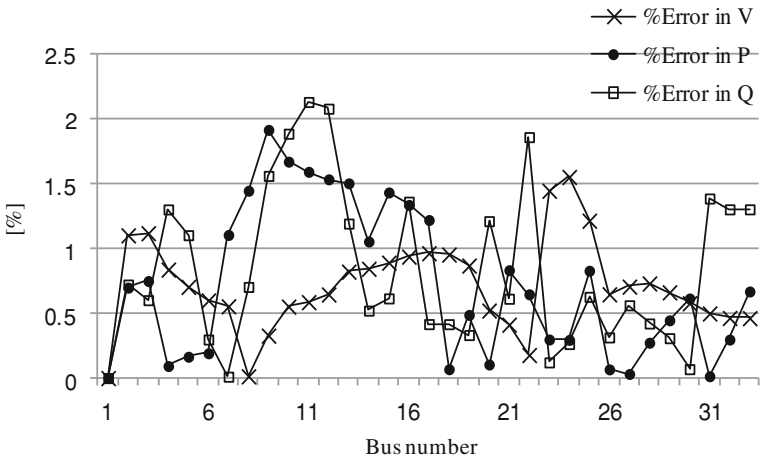


Fig. 3 Error in standard deviation of the output variables using the cumulant method and MCS technique—33-bus system

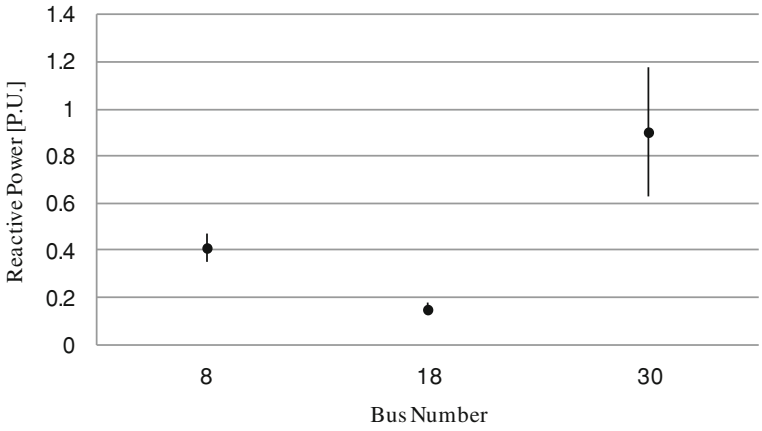


Fig. 4 Mean capacitor values are shown with potential spread using corresponding 3σ values

capacitor settings and voltages at buses, though being important information, is difficult to obtain using the conventional Monte Carlo Simulation technique due to long solution time. However, using the proposed cumulant method, using a few additional steps such as computation of the Hessian of the Lagrangian, yields the variance values of the optimized control (capacitor) and dependent (voltage) variables.

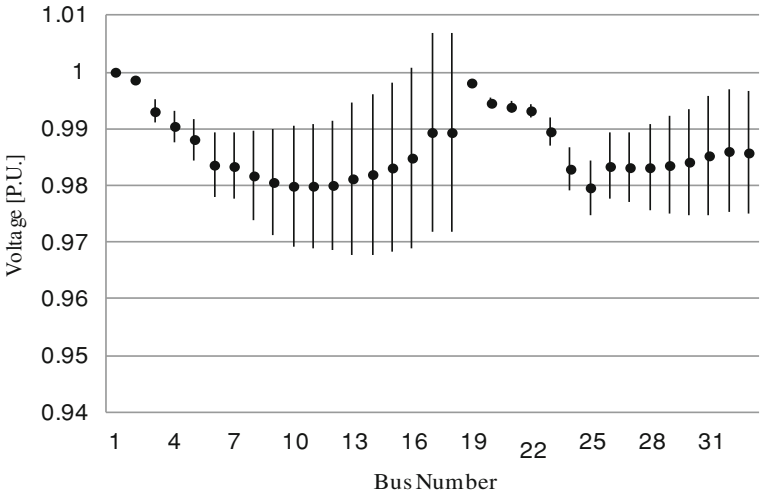


Fig. 5 Mean bus voltage magnitude values are shown with potential spread using corresponding 3σ values

Table 3 Time comparison between cumulant method and Monte Carlo simulation technique

Execution time (s)	CM	MCS
System #1	4.44 s	2888.05 s

This is an important benefit for the RDS operator as s/he would like to how far the capacitor values and voltage solutions might travel from the mean forecasted and anticipate/plan to avoid violations.

4.2 The 129-Bus Test System Case Study

To show the accuracy and efficiency of the Cumulant method, it also has been tested on a large system of 129-bus with nine wind generators each having a mean capacity equal to 500 kW. The results have been compared with 5,000 sample points of MCS.

Table 4 shows the mean absolute percentage error of mean and standard deviation of the problem variables (Table 5).

The method works well for small and reasonably sized systems.

Table 4 Mean absolute percentage error of mean and standard deviation of the problem variables

Variable	Mean	Standard deviation
Voltage	0.0025	1.4285
Real power	0.2206	0.8180
Reactive power	0.4580	1.2150
Capacitor size	0.0589	1.2537

Table 5 Time comparison between cumulant method and Monte Carlo simulation technique

Execution time (seconds)	CM	MCS
System #2	12.79 s	6,504.87 s

5 Conclusion

This paper describes the Cumulant method for probabilistic optimization of radial distribution systems. Random variables used in distribution system are the output of wind generators and bus loads. Both are modeled using Gaussian distribution function. While a Logarithmic Barrier Interior Point Method (LBIPM) based solver is used to solve the probabilistic nonlinear optimization problem, the cumulants of the system variables are easily computed using the inverse of the Hessian of the Lagrangian function. The method was implemented and tested on a 33-bus and 129-bus IEEE test system. In order to illustrate efficiency and accuracy of the cumulant method, the resultant data using cumulant method are benchmarked with those obtained from Monte Carlo Simulation Technique with 5,000 samples. The errors were found well below 8 %. An execution time comparison demonstrates superiority of the Cumulant method. Less computational burden and complexity makes the proposed method very practical and advantageous.

This method can be advantageously used by RDS operators to know possible swings in optimal capacitor settings and voltage solution giving them additional insight into operation of the system and anticipate potential operational challenges.

Acknowledgments Financial Support: This work was supported in part by the NSERC Discovery and Wind Energy Strategic Network grants to Bala Venkatesh.

6 Appendix

Table 6 presents the mean values used for the load demands, and Table 7 presents the feeder data.

Table 6 33 bus RDS—mean value of the loads

Bus no.	Mean	
	Real power (kW)	Reactive power (kW)
1	0	0.0
2	0.1	0.06
3	−0.5	0.0
4	0.12	0.08
5	0.06	0.03
6	0.2	0.1
7	0.2	0.1
8	0.2	0.1
9	0.06	0.02
10	0.06	0.02
11	0.045	0.03
12	0.06	0.035
13	0.06	0.035
14	0.06	0.04
15	0.06	0.01
16	0.09	0.04
17	−0.5	0.0
18	0.09	0.04
19	0.09	0.04
20	0.09	0.04
21	0.09	0.04
22	0.09	0.04
23	0.09	0.05
24	0.42	0.2
25	0.42	0.2
26	0.06	0.025
27	0.06	0.025
28	0.06	0.02
29	0.12	0.07
30	0.2	0.6
31	0.15	0.07
32	−0.5	0.1
33	0.06	0.04

Table 7 33 bus RDS—line data

From bus	To bus	R (Ω)	X (Ω)	Rating (MVA)	System voltage (kV)
1	2	0.0922	0.047	100	12.66
2	3	0.493	0.2511	100	12.66
3	4	0.3662	0.1864	100	12.66
4	5	0.3811	0.1941	100	12.66
5	6	0.819	0.707	100	12.66
6	7	0.1872	0.6188	100	12.66
7	8	1.7114	1.2351	100	12.66
8	9	1.03	0.74	100	12.66
9	10	1.044	0.74	100	12.66
10	11	0.1966	0.065	100	12.66
11	12	0.3744	0.1238	100	12.66
12	13	1.468	1.155	100	12.66
13	14	0.5416	0.7129	100	12.66
14	15	0.591	0.526	100	12.66
15	16	0.7463	0.545	100	12.66
16	17	1.289	1.721	100	12.66
17	18	0.732	0.574	100	12.66
18	19	0.164	0.1565	100	12.66
19	20	1.5042	1.3554	100	12.66
20	21	0.4095	0.4784	100	12.66
21	22	0.7089	0.9373	100	12.66
22	23	0.4512	0.3083	100	12.66
23	24	0.898	0.7091	100	12.66
24	25	0.896	0.7011	100	12.66
25	26	0.203	0.1034	100	12.66
26	27	0.2842	0.1447	100	12.66
27	28	1.059	0.9337	100	12.66
28	29	0.8042	0.7006	100	12.66
29	30	0.5075	0.2585	100	12.66
30	31	0.9744	0.963	100	12.66
31	32	0.3105	0.3619	100	12.66
32	33	0.3410	0.5302	100	12.66

References

1. J.A. Momoh, M.E. El-Hawary, R. Adapa, A review of selected optimal power flow literature to 1993 part I: non-linear and quadratic programming approaches. IEEE Trans. Power Syst. **14**(1), 96–104 (1990)
2. K.R.C. Mamandur, R.D. Chenoweth, Optimal control of reactive power flow for improvements in voltage profiles and for real power loss minimization. IEEE Trans. Power Appar. Syst. **PAS-100**(7), 3185–3194 (1981)
3. S.C. Tripathy, D. Prasad, O.P. Malik, G.S. Hope, Load flow solution for ill conditioned power systems by Newton like method. IEEE Trans. Power Appar. Syst. **101**, 3648–3657 (1982)
4. M. Ponnavsikko, K.S. Prakasa Rao, Optimal distribution systems planning. IEEE Trans. Power Appar. Syst. **PAS-100**(6), 2969–2977 (1981)

5. F. Vallee, J. Lobry, O. Deblecker, Impact of the wind geographical correlation level for reliability studies. IEEE Trans. Power Syst. **22**(4), 2232–2239 (2007)
6. M. Schilling, A.L. da Silva, R. Billinton, Bibliography on power system probabilistic analysis (1962–1988). IEEE Trans. Power Syst. **5**(1), 1–11 (1990)
7. G. Vivian, G. Heydt, Stochastic energy dispatch. IEEE Trans. Power Appar. Syst. **100**(7), 3221–3227 (1981)
8. M. El-Hawary, G. Mbamalu, A comparison of probabilistic perturbation and deterministic based optimal power flow solutions. IEEE Trans. Power Syst. **6**(3), 1099–1105 (1991)
9. M. Dadkhah, B. Venkatesh, Cumulant based stochastic reactive power planning method for distribution systems with wind generators. IEEE Trans. Power Syst. **27**, 2351–2359 (2012)
10. A. Dukpa, B. Venkatesh, L. Chang, An accurate voltage solution method of radial distribution system. Can. J. Elect. Comput. Eng. **34**(1/2), 69–74 (2009). (Win/Spr 2009)
11. A. Papoulis, S. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th edn. (McGraw-Hill, New York, 2002)
12. A. Schellenberg, W. Rosehart, J. Aguado, in Cumulant based probabilistic optimal power flow (P-OPF). *Proceedings of the 2004 International Conference on Probabilistic Methods Applied to Power Systems*, pp. 506–511, Sept 2004
13. M. Dadkhah, B. Venkatesh, in Probabilistic—optimal power flow for radial distribution systems with wind generators. *Proceedings of the World Congress on Engineering 2013, WCE 2013*, London, U.K., pp. 1038–1043, 3–5 July
14. G.O.G. Torres, V. Quintana, An interior-point method for nonlinear optimal power flow using voltage rectangular coordinates. IEEE Trans. Power Syst. **13**(4), 1211–1218 (1998)
15. A. Schellenberg, W. Rosehart, J. Aguado, Introduction to cumulant-based probabilistic optimal power flow (P-OPF). IEEE Trans. Power Syst. **20**(2), 1184–1186 (2005)
16. P. Zhang, S.T. Lee, Probabilistic load flow computation using the method of combined cumulants and Gram–Charlier expansion. IEEE Trans. Power Syst. **19**(1), 676–682 (2004)
17. M. A. Kashem, V. Ganapathy, G. B. Jasmon, M. I. Buhari, in *A novel method for loss minimization in distribution networks*. International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (2000)

Four Quadrant Operation of Field Weakened FOC Induction Motor Drive Using Sliding Mode Observer

G. K. Nisha, Z. V. Lakaparampil and S. Ushakumari

Abstract Field Weakening (FW) is applied in order to allow operation of variable speed induction motor drives at high speeds. Field Oriented Controlled (FOC) induction machine has the capability for easy field weakening and the full utilization of voltage and current rating of the inverter to obtain a wide dynamic speed range. In a sensorless FOC induction machine, the estimation of rotor speed is difficult in the high speed region. Model Reference Adaptive System (MRAS) based techniques are one of the best methods to estimate the rotor speed due to its better performance and unsophisticated stability approach. MRAS scheme based on Sliding Mode (SM) technique provides accurate speed estimation during operation in the FW region. In this chapter, operation of FOC induction motor with and without sensor using Sinusoidal Pulse Width Modulation (SPWM) and Space Vector Modulation (SVM) inverters are compared and assessed in terms of their performance in the FW region. Further, the four possible combinations of polarities of torque and speed in four quadrant operation of induction machine are analyzed. The drive system with the proposed adaptive mechanism is simulated by MATLAB/Simulink to verify the performance of the drive system.

Keywords Field oriented control · Field weakening · Model reference adaptive system · Sinusoidal pulse width modulation · Sliding mode observer · Space vector modulation

G. K. Nisha (✉) · S. Ushakumari
Department of Electrical Engineering, College of Engineering Trivandrum, Trivandrum, Kerala, India
e-mail: nishacharu@gmail.com

S. Ushakumari
e-mail: ushalal2002@gmail.com

Z. V. Lakaparampil
Centre for Development of Advanced Computing (C-DAC), Trivandrum, Kerala, India
e-mail: zvlakapara@cdac.in

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_28,

1 Introduction

Power electronics has deeply expanded the use of induction machine in automation applications such as the traction machine of the electric train, which requires high torque at starting and low torque at high speed. In order to achieve these requirements, the induction machine is to be operated in FW region. The dynamic behavior of a field weakened FOC induction motor is dependent on sophisticated speed control [1, 2].

Sensorless speed control of induction motor drive by the elimination of the rotor speed sensor without affecting performance is the foremost trend in advanced drive technology for a wide speed range. MRAS methods are generally accepted as better solution for high sensorless performance because of its simplicity. The adaption mechanism for MRAS can be taken care of the overall stability of the system and to ensure that the estimated speed will converge to the desired value with satisfactory dynamic characteristics [3–6].

Discrepancy between the actual plant and its mathematical model is generally due to the presence of external disturbances, plant parameters and unmodelled dynamics and it is necessary to design a controller for minimizing the error. Designing the control laws that provide the desired performance to the control system in presence of these disturbances is challenging, this leads to the development of robust control methods. One particular approach to robust controller design is the SM control technique. SM control is considered to be the appropriate methodology for the robust nonlinear control of induction motor drives due to its order reduction, disturbance rejection, strong robustness and simple implementation by means of power converter [7, 8].

The FW approaches can be categorized as: (1) variation of stator flux in inverse proportion to the rotor speed ($1/\omega_r$); (2) feed forward reference flux generation on machine equations or machine models and (3) closed loop control of the stator voltage or voltage detection model. The first approach as presented in [9] is the most frequently used method in FW control, in which the flux is inversely proportional to the rotor speed. Although the method is simple, it is justified only when considering the machine as a linear magnetic circuit. The method thus cannot produce maximum output torque for the available current and the full utilization of DC-link voltage. The second approach, as presented in [10] relies on the nonlinear equations of machine model and the constraints of voltage and current, which makes it parameter dependent. Thus the method can provide accurate results only if magnetic saturation is considered with known machine parameters of sufficient accuracy. The third approach as described in [11–18], maximum available inverter voltage is utilized to produce maximum torque in FW region when the excitation level is adjusted by closed loop control of the machine voltage. Although it is not dependent on motor parameters and DC link voltage, it demands an additional outer loop which is to be tuned and requires intensive computation. On comparing the above three approaches, the method based on machine model seems to be a more practical approach giving reasonable results.

The major problem of the machine model approach in the FW region is the substantial variation of magnetizing inductance which is considered constant in the base speed range. In the FW region, the rotor flux is getting reduced below its rated value due to the increase of rotor speed than the base speed. The variable level of the main flux saturation in the machine causes the variation of magnetic inductance [19]. Therefore, in model based approach, accurate speed estimation is possible only if the speed estimation algorithm is modified to account the variation of magnetic inductance in the FW region.

In [20–27], mathematical models for speed estimation of sensorless FOC induction motor drive with SVM inverter using MRAS scheme based on SM observer have been developed. The electric machine for the traction application can operate at four quadrants in a torque-speed plane, in this chapter the four simulation models: (1) FOC with sensor using SPWM inverter, (2) FOC with sensor using SVM inverter, (3) sensorless FOC using SPWM inverter and (4) sensorless FOC using SVM inverter have been developed using MATLAB/Simulink. Aim of this section is to compare performance of FOC induction machine in FW region with and without sensors using SPWM and SVM in all the four quadrants.

2 Dynamic Model of Induction Machine

Induction motor drive system is becoming a more and more competitive system in many high performance motion drive applications. The Indirect Field Oriented Control (IFOC) can provide good dynamic torque response as obtained from DC motor drives [28]. Mathematical representation of induction motor is based on space vector notation of any three time varying quantities, its sum is always equal to zero, and are spatially separated by 120° can be expressed as space vector [29]. Figure 1 shows the Inverse Park transformation module and the stator voltage space vector and its component in stationary reference frame by using the transformation as:

$$\begin{bmatrix} V_{s\alpha} \\ V_{s\beta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} V_{sd} \\ V_{sq} \end{bmatrix} \tag{1}$$

Independent control of motor flux and torque can be achieved using FOC method by properly connecting coordinate system with rotor flux vector [30]. Figure 2 shows the phasor diagram in stationary and rotating reference frame. The reference frame d–q is rotating with the angular speed equal to rotor flux vector angular speed ω_e. Induction motor model equation is written as follows:

$$\vec{V}_s(t) = R_s \vec{i}_s + L_s \frac{d\vec{i}_s}{dt} + L_m \frac{d(\vec{i}_r e^{j\varepsilon})}{dt} \tag{2}$$

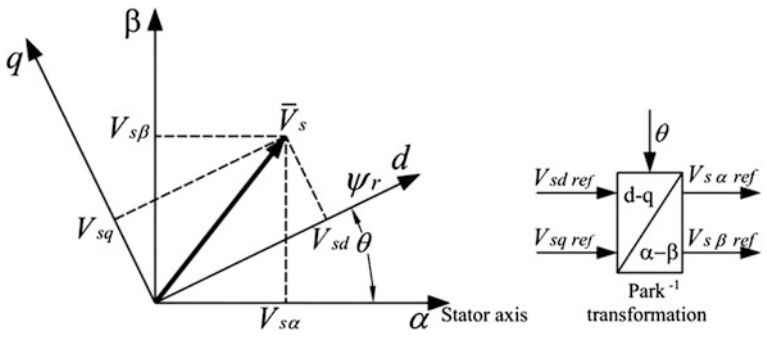


Fig. 1 *S*tator voltage space vector from *d–q* to *α–β*

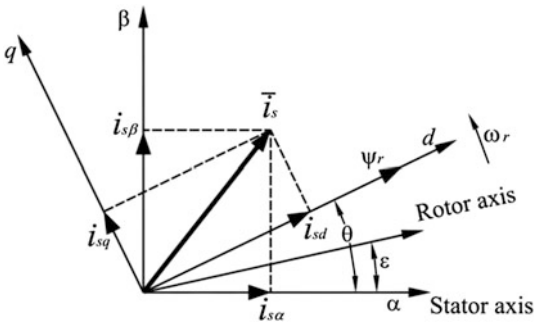


Fig. 2 Phasor diagram in stationary and rotating reference frame

$$0 = R_r \vec{i}_r e^{j\varepsilon} + L_r \frac{d\vec{i}_r}{dt} e^{j\varepsilon} + L_m \frac{d\vec{i}_s}{dt} \tag{3}$$

$$\frac{d\omega_r}{dt} = \frac{1}{J}\left(\frac{2}{3}\frac{P}{2}\frac{L_m}{L_r}\left(\psi_{rd} i_{sq} - \psi_{rq} i_{sd}\right) - T_L\right) \tag{4}$$

The two phase *d–q* model of an induction machine rotating at a speed will give the decoupled control concept. The complete motor dynamic equation is obtained by separating the real and imaginary components as:

$$\vec{V}_{sq} = R_s i_{sq} + L_s \frac{di_{sq}}{dt} + L_m \frac{d}{dt} i_{rq} + L_s \omega_e i_{sd} + L_m \omega_e i_{rd} \tag{5}$$

$$\vec{V}_{sd} = R_s i_{sd} + L_s \frac{di_{sd}}{dt} + L_m \frac{d}{dt} i_{rd} - L_s \omega_e i_{sq} - L_m \omega_e i_{rq} \tag{6}$$

$$0 = R_r i_{rd} + L_r \frac{di_{rd}}{dt} + L_m \frac{d}{dt} i_{sd} - L_r(\omega_e - \omega_r) i_{rq} + L_m(\omega_e - \omega_r) i_{sq} \tag{7}$$

$$0 = R_r i_{rq} + L_r \frac{di_{rq}}{dt} + L_m \frac{d}{dt} i_{sq} + L_r(\omega_e - \omega_r) i_{rd} + L_m(\omega_e - \omega_r) i_{sd} \tag{8}$$

The induction motor model is often used in vector control algorithms in which the reference frame may be aligned with the rotor flux linkage. In this reference frame, the torque can be instantaneously controlled by controlling the current i_{sq} after decoupling the rotor flux and torque producing component of the current components. The flux along the q axis must be zero, thus the field orientation concept in rotating reference frame is,

$$\frac{d\psi_{rq}}{dt} = 0 \tag{9}$$

The fundamental equations for vector control which allows the induction motor to act like a separately excited DC machine with decoupled control of torque and flux making the induction motor to operate as a high performance four quadrant servo drive [31]. From the voltage loop equation, the magnetizing current dependency on the d axis component of stator current is obtained as:

$$\frac{d\psi_{rd}}{dt} + R_r i_{rd} = 0 \tag{10}$$

$$T_r \frac{di_{mr}}{dt} + i_{mr} = i_{ds} \tag{11}$$

where

$$\tau_r = \frac{L_r}{R_r}$$

Slip speed and dynamic torque are calculated based on the following equations:

$$\omega_{slip} = \frac{1}{\tau_r} \frac{i_{sq}}{i_{mr}} \tag{12}$$

$$T_d = \frac{2}{3} \frac{P}{2} \frac{L_m}{(1 + \sigma_r)} i_{mr} i_{sq} \tag{13}$$

where,

$L_s = L_m(1+\sigma_s)$	Stator self inductance
L_m	Magnetizing inductance
$L_r = L_m(1+\sigma_r)$	Rotor self inductance
R_s	Stator resistance
R_r	Rotor resistance
T_d	Electromagnetic torque
ω	Angular speed
P	Number of poles
τ_r	Rotor time constant

i_{mr}	Rotor magnetizing current
ω_{slip}	Slip speed
ω_e	Rotor flux speed
θ_r	Rotor position
θ_{sl}	Slip angle

3 Field Weakening Control of Induction Machine

Field weakening is needed when motor operation above the rated speed is required. Field weakening operation consists of two steps (i) the choice of the proper flux reference to get maximum torque (ii) to produce the necessary current to meet the flux and torque reference. The operation of the induction motor can be divided into three speed ranges, (1) constant electromagnetic torque region (2) constant power region (3) constant slip frequency region. Figure 3 shows the typical capability curve of induction machine.

The limiting factor of maximum torque capability in field weakening mode of operation is the DC bus voltage. The performance of AC machine driven by a three phase PWM inverter in the high speed range is limited by the voltage and current rating of electric machine and inverter and the machine thermal rating. The stator voltage equation by considering stator resistance effect at higher operating speed as:

$$V_{sd} = R_s i_{sd} + \sigma L_s \frac{di_{sd}}{dt} + \frac{L_m}{L_r}\frac{d\psi_{rd}}{dt} - \omega_e \sigma L_s i_{sq} \tag{14}$$

$$V_{sq} = R_s i_{sq} + \sigma L_s \frac{di_{sq}}{dt} + \frac{L_m}{L_r}\omega_e \psi_{rd} + \omega_e \sigma L_s i_{sd} \tag{15}$$

The voltage limit boundary is an ellipse and the shape of the ellipse is determined by its eccentricity which depends on the leakage factor of the machine. The maximum phase voltage is decided by the PWM strategy and the d–q axis stator voltage should satisfy the following inequality condition, which sensibly influence the motor behaviour and the voltage limit ellipse equation as follows:

$$\frac{i_{sd}^2}{\left(\frac{V_{sm}}{\omega_e L_s}\right)^2} + \frac{i_{sq}^2}{\left(\frac{V_{sm}}{\omega_e L_s'}\right)^2} \leq 1 \tag{16}$$

where,

$$L_s' = \sigma L_s$$

The maximum current to the machine is also limited, and the current limit boundary is a circle, whose radius depends only on the current rating, as follows:

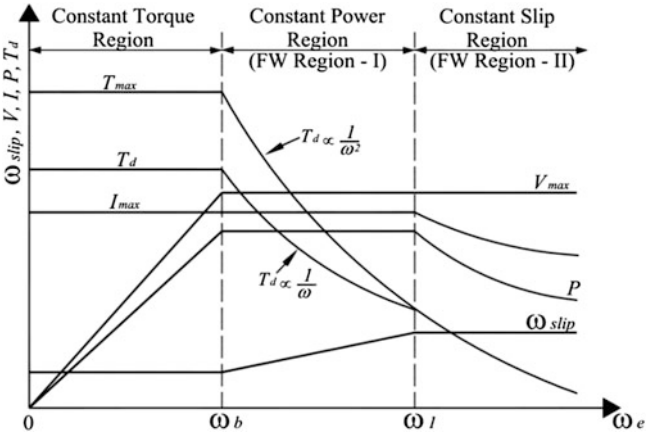


Fig. 3 Typical capability curve of induction machine

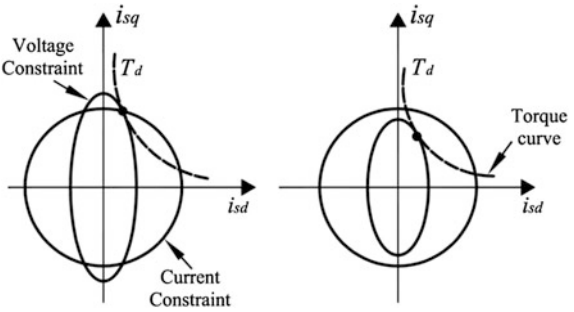


Fig. 4 Voltage constraint ellipse, current constraint circle and torque locus

$$i_{sd}^2 + i_{sq}^2 \leq I_{\text{sm}}^2 \tag{17}$$

At a given operating frequency, to satisfy voltage and current constraints, the command current must be inside the common area of a circle and an ellipse. Figure 4 shows voltage limit boundary, current constraint circle and constant torque locus. In steady state and rotor flux orientation condition, the torque of the field oriented induction machine is expressed as:

$$T_d = \frac{2}{3}\frac{P}{2}\frac{L_m^2}{L_r}\left(i_{sd} i_{sq}\right) \tag{18}$$

In the constant torque region, the speed is less than the rated speed and the d axis current is kept constant. In this current limit and the rated flux level determine the operating point to the maximum torque. The base frequency is the angular frequency where the constant torque operation region terminates at ω_b and is given as:

$$\omega_b = \frac{V_{sm}}{L_s}\sqrt{\frac{1}{[\sigma^2(I_{sm}^2 - i_{sd}^2) + i_{sd}^2]}} \tag{19}$$

In field weakening region I, the ellipse shrink as the speed increases and goes beyond its rated speed, it is necessary to reduce the rotor flux magnitude, the d axis current is no longer a constant. In this region, the power delivered to the load is nearly constant because the maximum torque is inverse proportion to the mechanical speed, the operation region starts from the base speed and ends at ω_1.

$$\omega_1 = \frac{V_{sm}}{\sigma L_s I_{sm}}\sqrt{\frac{(1+\sigma^2)}{2}} \tag{20}$$

The characteristic region of the induction machine referred to as field weakening region II, the operating frequency of the induction machine further increases from ω_1 and the ellipse is encircled by the current constraint circle, the power delivered to the load decreases proportionally with the rotor speed. In this constant slip frequency region, the maximum torque is inverse proportion to square of the mechanical speed. The maximum q axis current is:

$$i_{sq} = \frac{i_{sd}}{\sigma} \tag{21}$$

4 SPWM and SVPWM Based Three Phase Inverters

The three phase voltage source inverter is used for driving the motor by accepting the control signals generated by the controller and the modulation technique used here are sine triangle and space vector pulse width modulation. The basic concept of SPWM is to achieve symmetrical 3-phase sine voltage waveforms of adjustable voltage and frequency, while in SVM, both the inverter and motor are taken as whole using the eight fundamental voltage vector to realize variable frequency of voltage and speed adjustment [32]. To obtain PWM pulses for the three phases in SPWM, the controlled sine waves for the three phases are compared with a carrier triangular waveform and in SVM, the modulating functions are compared with the triangular waveform. These PWM pulses are given to the gate of the inverter switches to get a controlled three phase output voltage which can be given to the motor input.

In SVM, a space phasor can be considered as a rotating phasor and no separate modulators are needed for each of the three phases in comparison with SPWM. Six non-zero vectors called the active vectors, shape the axis of hexagonal and the angle between any adjacent two non-zero vectors is 60°. Switching logic signals for the second sector is shown in Fig. 5. For each switching period T_s, the

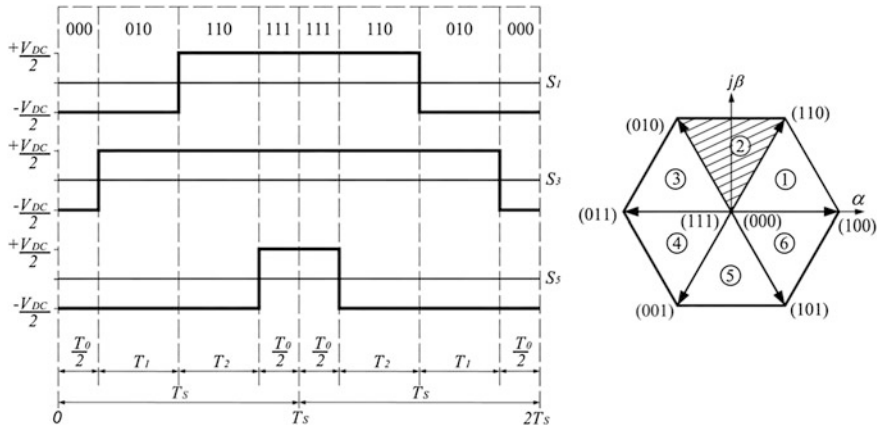


Fig. 5 Switching logic signals

reference vector as a geometric summation of two nearest space vectors is expressed mathematically by applying volt-second balance equation as:

$$\overrightarrow{V}_{ref} = \frac{T_1}{T_s} V_1 + \frac{T_2}{T_s} V_2 + \frac{T_0}{T_s} V_0 \tag{22}$$

5 Sliding Mode MRAS Speed Observer

The standard non linear time varying feedback system, which is said to be asymptotically stable if it posses the following two conditions; (1) the transfer function of the feed-forward linear time invariant block must be strictly positive real and (2) the error should converge asymptotically. In MRAS scheme, with two independent machine models, speed estimation is done continuously by comparing the output of the reference model with the output of the adaptive model until the error between the two models disappear. Figure 6 describes the modified MRAS with both references and adaptive models for rotor speed estimation. The reference model does not contain the speed to be computed, which represents stator equation and is usually known as voltage model. The reference value of the rotor flux components in the stationary frame are generated from the monitored stator voltage and current components, which are given by:

$$\frac{d}{dt}\begin{bmatrix}\psi_{r\alpha}\\ \psi_{r\beta}\end{bmatrix} = \frac{L_r}{L_m}\begin{bmatrix}V_{s\alpha}\\ V_{s\beta}\end{bmatrix} - \begin{bmatrix}R_s + \sigma L_s & 0\\ 0 & R_s + \sigma L_s\end{bmatrix}\begin{bmatrix}pi_{s\alpha}\\ pi_{s\beta}\end{bmatrix} \tag{23}$$

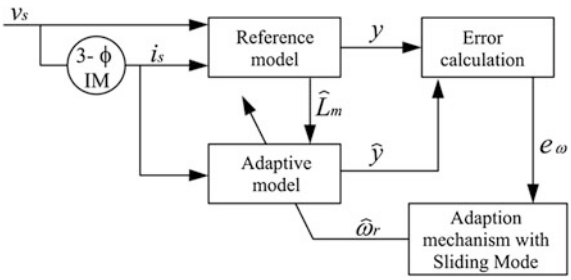


Fig. 6 Block diagram of modified MRAS-SM speed estimator for FW

where,

$$p = \frac{d}{dt} \quad \sigma = 1 - \frac{L_m^2}{L_s L_r} \tag{24}$$

The adaptive model contains the estimated rotor speed, which represents the rotor equation and is usually known as the current model flux equations. The adaptive values of rotor flux components are given by:

$$\frac{d}{dt}\begin{bmatrix}\hat{\psi}_{r\alpha}\\ \hat{\psi}_{r\beta}\end{bmatrix} = \frac{L_m}{\tau_r}\begin{bmatrix}i_{s\alpha}\\ i_{s\beta}\end{bmatrix} - \begin{bmatrix}\frac{1}{\tau_r} & \omega_e\\ -\omega_e & \frac{1}{\tau_r}\end{bmatrix}\begin{bmatrix}\hat{\psi}_{r\alpha}\\ \hat{\psi}_{r\beta}\end{bmatrix} \tag{25}$$

$$e_\omega = \psi_{r\beta}\hat{\psi}_{r\alpha} - \psi_{r\alpha}\hat{\psi}_{r\beta} \tag{26}$$

The adaptive scheme for the MRAS estimator can be designed based on Popov's criteria for hyper stability concept. The difference between the two estimated vectors is fed to an adaption mechanism to generate estimated value of rotor speed which is used to tune the adaptive model. The tuning signal, e_ω actuates the rotor speed, which makes the error signal zero. The major problems associated with the classical MRAS speed estimator described above are those related to initial condition and integrator drift which is solved by substituting pure integration with low pass filtering. At high frequencies (FW region) variation of stator resistance has practically no impact on the accuracy of the speed estimation.

SM control with variable structure system is an adaptive control where the structure of the control is deliberately varied to alleviate the control and to make its response robust. The sliding mode control should be chosen such that the candidate Lyapunov function, V which is a scalar function of S and its derivative satisfies the Lyapunov stability criteria:

$$V(S) = \frac{1}{2}S(x)^2 \tag{27}$$

In Lyapunov theory, if the time derivative of $V(S)$ along a system trajectory is negative definite, this will ensure that it constrains the state trajectories to a point towards the sliding surface $S(x)$ and once on the surface, the system trajectories

remain on the surface until the origin is reached asymptotically. Thus the sliding condition is achieved by the following condition (28) makes the surface an invariant set.

$$\dot{V}(x) \leq -\eta|S(x)| \tag{28}$$

where,

η strictly positive constant on outside of $S(x)$

The control signal is written as:

$$u(t) = u_{eq}(t) + u_{sw}(t) \tag{29}$$

The sliding mode control should be chosen such that the candidate Lyapunov function, V which is a scalar function of S and its derivative satisfies the Lyapunov stability criteria:

$$\dot{V}(S) = S(x)\dot{S}(x) \tag{30}$$

where, $u(t)$ is the control vector, $u_{eq}(t)$ is the equivalent control vector and $u_{sw}(t)$ is the switching vector and must be calculated so that stability condition as per (30) for the selected control is satisfied.

$$u_{sw}(t) = \eta sign(S(x,t)) \tag{31}$$

where,

$$sign(S) = \begin{cases} -1 & for \quad S < 0 \\ = 0 & for \quad S = 0 \\ +1 & for \quad S > 0 \end{cases}$$

The sliding mode control theory is now applied to the rotor flux MRAS scheme for speed estimation by replacing the conventional constant gain PI controller. With reference to dynamic model of induction machine and the speed tuning signal, the time varying sliding surface $S(x)$ is formulated and is given in (32):

$$S(x) = e_\omega + \int Ke_\omega dt = 0 \tag{32}$$

where, K is the switching gain which is strictly positive constant. When the system reaches the sliding surface, the error dynamics at the sliding surface, $S(x) = 0$ will be forced to exponentially decay to zero. Thus,

$$\dot{S} = \dot{e}_\omega + Ke_\omega = 0 \tag{33}$$

The time derivative of $V(S)$ is negative definite for the following conditions:

$$(a + Ke_{\omega} - \hat{\omega}_r b)\begin{cases} <0 & for \quad S > 0 \\ = 0 & for \quad S = 0 \\ > 0 & for \quad S < 0 \end{cases} \tag{34}$$

This can be attained when:

$$\hat{\omega}_r = u_{eq} + u_{sw} \tag{35}$$

where,

$$u_{eq} = \frac{a + Ke_{\omega}}{b} \tag{36}$$

$$u_{sw} = \eta.sign(S) \tag{37}$$

The equivalent control defines the control action which keeps the state trajectory on the sliding surface and the switching control depends on the sign of the switching surface and η is the hitting control gain which makes (30) negative definite, whose main purpose is to make the sliding condition viable and the value of η should be large enough to overcome the effect of external disturbance. The controller given will have chattering near sliding surface due to the presence of sign function. This drastic change is avoided by introducing a boundary layer with width, ϕ. By replacing $sign(s)$ with $sat(S/\phi)$, then (37) becomes:

$$u_{sw} = \eta sat(S/\phi) \tag{38}$$

A natural solution to reduce the chattering in the estimated speed is by means of a Low-Pass Filter (LPF) as in (39).

$$u_{sw} = \frac{1}{\mu s + 1} u_{sw} \tag{39}$$

6 Simulation

Simulation models for (1) FOC with sensor using SPWM inverter, (2) FOC with sensor using SVM inverter, (3) sensorless FOC using SPWM inverter and (4) sensorless FOC using SVM inverter have been developed using MATLAB/Simulink. The block diagrams of the drive system for FOC induction machine with and without sensor are shown in Figs. 7 and 8 respectively. The switching frequency selected for both the inverters is 5 kHz. A 10 kW squirrel cage induction machine used for the simulation having the motor parameters is given in Table 1. For comparing the performance of the developed drive system, simulations are carried out for both models, with and without sensor as Case-1 and Case-2 respectively.

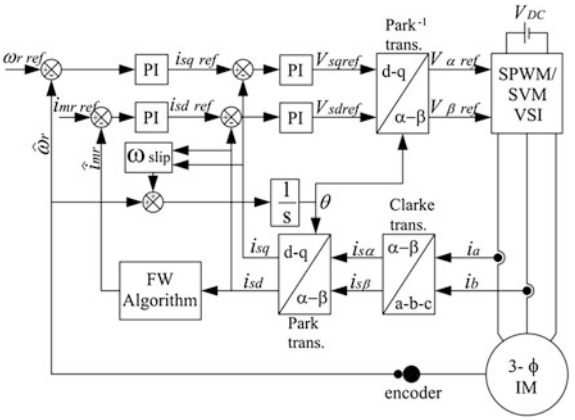


Fig. 7 Block diagram of FOC induction machine with sensor

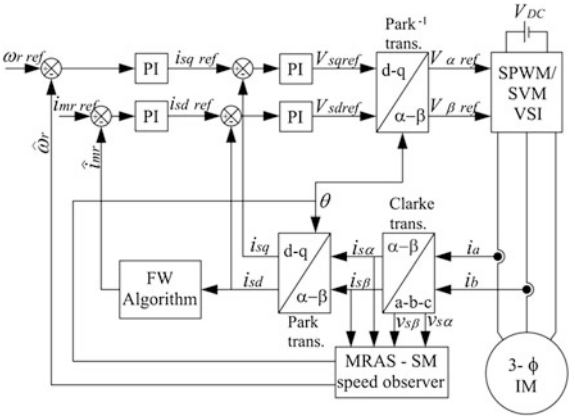


Fig. 8 Block diagram of FOC induction machine without sensor using MRAS-SM

Table 1 Parameters of induction motor

	Parameter	Rated values
Stator resistance	R_s	0.74 Ω
Rotor resistance	R_r	1.00 Ω
Stator inductance	L_s	0.0963 H
Rotor inductance	L_r	0.0963 H
Magnetizing inductance	L_m	0.0854 H
Line current	I_L	28 A
Number of poles	P	4
Moment of inertia	J	0.03 Nm
Viscous friction coefficient	B	0.00334 Nm/rad/sec

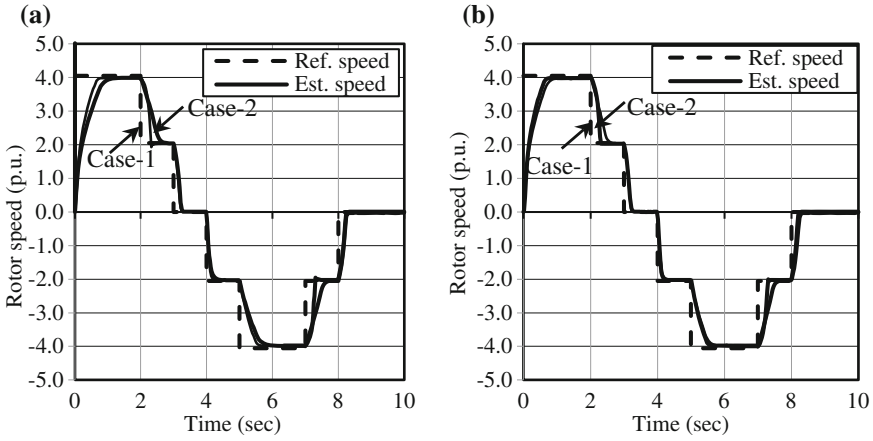


Fig. 9 Rotor speed versus Time. **a** With SPWM inverter. **b** With SVM inverter

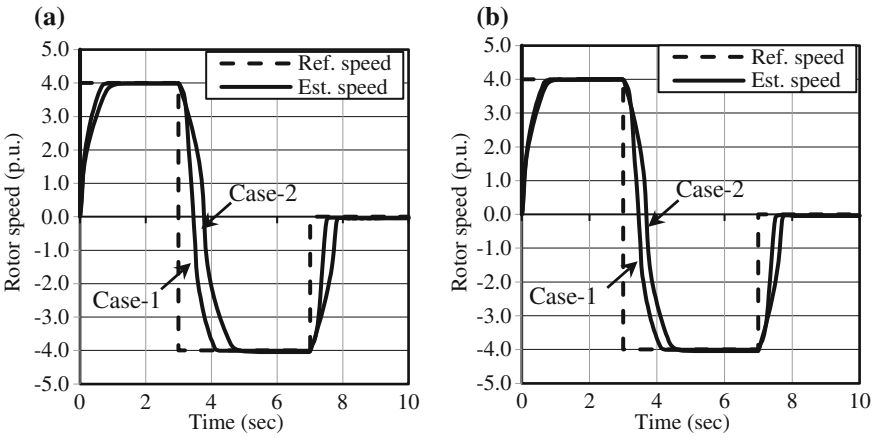


Fig. 10 Rotor speed versus Time. **a** With SPWM inverter. **b** With SVM inverter

7 Results and Discussion

The dynamic behavior of field weakening algorithm in the models is evaluated by applying a sequence of multiple and single step changes of the speed reference signal between 0 and 4 p.u. (4 times rated value) in four quadrant are shown in Figs. 9 and 10 respectively. The results show that MRAS-SM observer estimates the rotor speed well and close to that of with sensors in all ranges of speed and SVM inverter catches better performance in tracking the speed command compared to that of SPWM inverter.

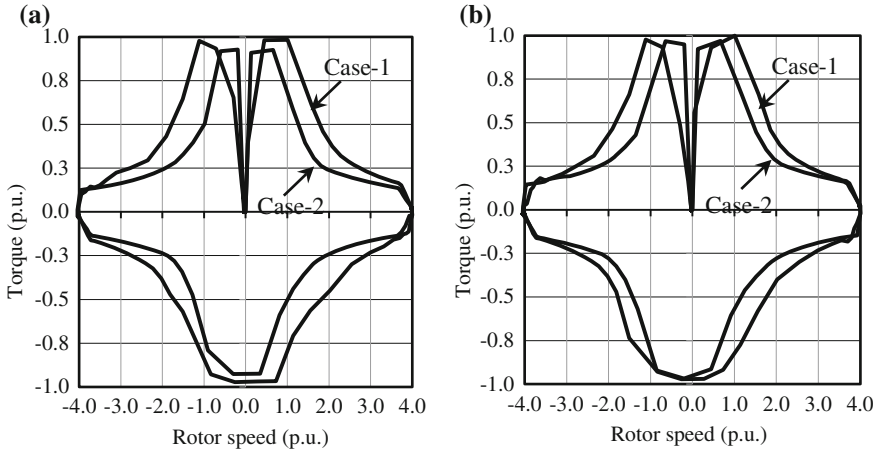


Fig. 11 Torque versus rotor speed. **a** With SPWM inverter. **b** With SVM inverter

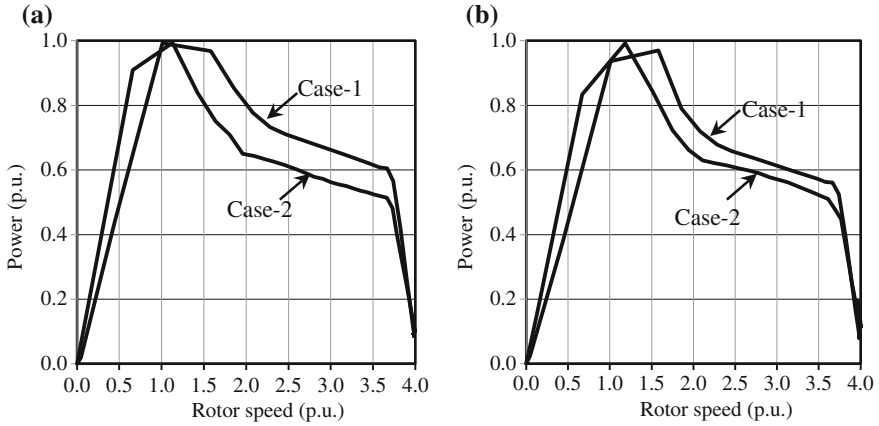


Fig. 12 Power versus rotor speed. **a** With SPWM inverter. **b** With SVM inverter

The Electromagnetic Torque (p.u.) versus rotor speed (p.u.) characteristics of the four models is presented in Fig. 11. The torque attainability rate is slightly less for Case-2 compared to Case-1 for both type of inverters and the characteristics is more or less same in all quadrants. But, SVM inverter improves the torque capability in all quadrants compared to SPWM inverter.

Figure 12 shows the variation of power (p.u.) with respect to rotor speed (p.u.) in first quadrant operation. Power maintains constant for rotor speed ranges between 1.0 and 1.5 in Case-1 while the range is less for Case-2. However, the power for Case-1 and Case-2 are very close when using SVM inverter compared to SPWM inverter.

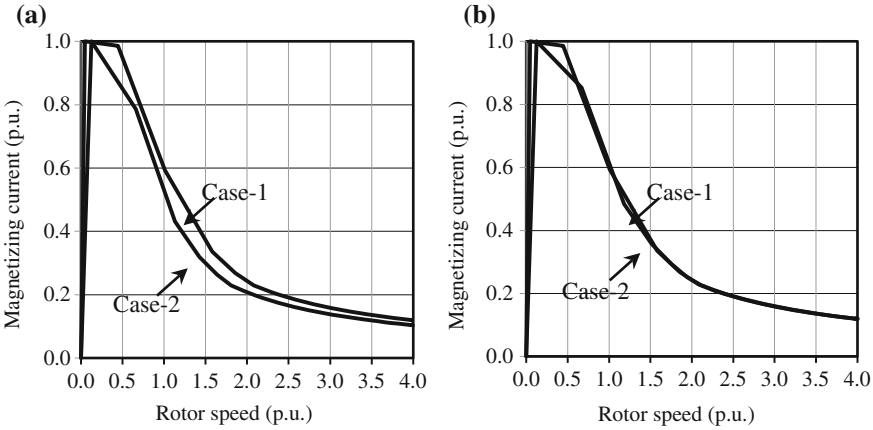


Fig. 13 Magnetizing current versus rotor speed. **a** With SPWM inverter. **b** With SVM inverter

Variation of magnetizing current (p.u.) with respect to rotor speed (p.u.) in first quadrant operation is shown in Fig. 13. The magnetizing current in Case-1 and Case-2 are very close in FW region by using SVM inverter.

By observing all the simulation results of the four models, it ensures that FOC induction motor drives with MRAS-SM observer using SVM inverter can respond quickly and provide accurate speed estimation in both base speed and FW regions of all the four quadrants which works well when the parameters are precisely measured and do not change during operation.

8 Conclusions and Future Work

In this paper, four simulation models of FOC induction motor drives with and without sensors using SPWM and SVM inverters are developed for wide ranges of speed including FW region and the performance of the models are compared in four quadrant operation. For sensorless model, a novel adaption mechanism using MRAS with SM control in FW region is proposed. The adaption mechanism is based on Lyapunov theory to ensure stability with fast error dynamics. The speed estimation by the rotor flux MRAS with SM observer using SVM inverter, instead of conventional SPWM inverter, has ensured very good accuracy in all ranges of speed control. In this method, the transition speed between base speed region and FW region is smooth, depending upon the voltage and current limits.

In future, the models can be implemented in a digital platform and hardware implementation of the above models has to be done.

Acknowledgment The first author acknowledges support from SPEED-IT Research Fellowship from IT Department of the Government of Kerala, India.

References

1. H. Abu Rub, A. Iqbal, J. Guzinski, *High Performance Control of AC Drives* (Wiley, NewYork, 2012), pp. 375–388
2. M.S. Huang, C.M. Liaw, Speed control for field weakened induction motor drive. IEE Elect. Power Appl. **152**(3), 565–576 (2005)
3. B.K. Bose, *Modern Power Electronics and AC Drives* (Prentice Hall of India Pvt. Limited, New Delhi, 2009), pp. 390–392
4. Y.D. Landau, *Adaptive Control: The Model Reference Approach* (Marcel Dekker, New York, 1979). 1979
5. C. Schauder, Adaptive speed identification for vector control of induction motors without rotational transducers. IEEE Trans. Ind. Appl. **28**(5), 1054–1061 (1992)
6. M. Comanescu, L. Xu, Sliding mode MRAS speed estimators for sensorless vector control of induction machine. IEEE Trans. Ind. Electron. **53**(1), 146–153 (2006)
7. V.I. Utkin, Sliding mode control design principles and applications to electric drives. IEEE Trans. Ind. Electron. **40**(1), 23–36 (1993)
8. J.J.E. Slotine, W. Li, *Applied Non Linear Control* (Prentice Hall, New Jersey, 1998), pp. 276–286
9. S.H. Kim, S.K. Sul, Maximum torque control of an induction machine in the field weakening region. IEEE Trans. Ind. Appl. **31**(4), 787–794 (1995)
10. F. Briz, A. Diez, M.W. Degner, R.D. Lorenz, Current and flux regulation in field weakening operation of induction motors. IEEE Trans. Ind. Appl. **37**(1), 42–50 (2001)
11. M.H. Shin, D.S. Hyun, S.B. Cho, Maximum torque control of stator flux oriented induction machine drive in the field weakening region. IEEE Trans. Ind. Appl. **38**(1), 117–122 (2002)
12. D. Casadei, G. Serra, A. Tani, L. Zarri, in *A Robust Method for Field Weakening Operation of Induction Motor Drive with Maximum Torque Capability*, Conference Record of IEEE Industry Applications, IAS 2006, Tampa, FL, pp. 111–117, 8–12 Oct 2006
13. H. Abu-Rub, J. Holtz, in Maximum torque production in rotor field oriented control of an induction motor at filed weakening. *Proceedings of IEEE International Symposium on Industrial Electronics, ISIE 2007*, Vigo, pp. 1159–1164, 4-7 June 2007
14. M. Wlas, H. Abu-Rub, J. Holtz, in Speed sensorless nonlinear control of induction motor in the field weakening region. *Proceedings of IEEE International Conference on Power Electronics and Motion Control, EPE-PEMC 2008*, Poznan, pp. 1084–1089, 1–3 Sept 2008
15. M. Mengoni, L. Zari, A. Tani, G. Serra, D. Casadei, Stator flux vector control of induction motor drive in the field weakening region. IEEE Trans. Power Electron. **23**(2), 941–949 (2008)
16. P.Y. Lin, Y.S. Lai, Novel voltage trajectory control for field-weakening operation of induction motor drives. IEEE Trans. Ind. Appl. **47**(1), 122–127 (2011)
17. M.-H. Shin, D.-S. Hyun, Speed sensorless stator flux oriented control induction machine in the field weakening region. IEEE Trans. Power Electron. **18**(2), 580–586 (2003)
18. E. Levi, M. Wang, A speed estimator for high performance sensorless control of induction motors in the field weakening region. IEEE Trans. Power Electron. **17**(3), 365–378 (2002)
19. A.V. Stancovie, E.L. Benedict, J. Vinod, T.A. Lipo, A novel method for measuring induction machine magnetizing inductance. IEEE Trans. Ind. Appl. **39**(5), 1257–1263 (2003)
20. G.K. Nisha, S. Ushakumari, Z.V. Lakaparampil, in Harmonic elimination of space vector modulated three phase inverter. *Lecture Notes in Engineering and Computer Science: Proceedings of International Multi Conference of Engineers and Computer Scientist, IMECS 2012*, HongKong, pp. 1109–1115, 14–16 March 2012
21. G.K. Nisha, S. Ushakumari, Z.V. Lakaparampil, in CFT based optimal PWM strategy for three phase inverter. *Proceedings of IEEE International Conference on Power, Control and Embedded Systems, ICPCES 2012*, Allahabad, India, pp. 1–6, 17–19 Dec 2012

22. G.K. Nisha, S. Ushakumari, Z.V. Lakaparampil, Online harmonic elimination of SVPWM for three phase inverter and a systematic method for practical implementation. IAENG Int. J. Comput. Sci. **39**(2), 220–230 (2012)
23. G.K. Nisha, Z.V. Lakaparampil, S. Ushakumari, FFT analysis for field oriented control of SPWM and SVPWM inverter fed induction machine with and without sensor. Int. J. Adv. Elect. Eng. **2**(4), 151–160 (2013)
24. G.K. Nisha, Z.V. Lakaparampil, S. Ushakumari, in Sensorless vector control of SVPWM fed induction machine using MRAS—sliding mode. *Proceedings of IEEE International Conference on Green Technologies, ICGT 2012*, Trivandrum, India, pp. 29–36, 18–20 Dec 2012
25. G.K. Nisha, Z.V. Lakaparampil, S. Ushakumari, in Sensorless field oriented control of SVM inverter fed induction machine in field weakening region using sliding mode observer. *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013, WCE 2013*, London, U.K., pp. 1174–1181, 3–5 July 2013
26. G.K. Nisha, Z.V. Lakaparampil, S. Ushakumari, Performance study of field oriented controlled induction machine in field weakening using SPWM and SVM fed inverters. Int. Rev. Model. Simul. **6**(3), 741–752 (2013)
27. G.K. Nisha, S. Ushakumari, Z.V. Lakaparampil, in Method to eliminate harmonics in PWM: a study for single phase and three phase. *Proceedings of International conference on Emerging Technology Trends on Advanced Engineering Research, ICETT 2012*, Kollam, India, pp. 598–604, 20–21 Feb 2012
28. W. Leonhard, *Control of Electrical Drives* (Springer, New York, 1996)
29. Z.V. Lakaparampil, K.A Fathima, V.T. Ranganathan, in Design modeling simulation and implementation of vector controlled induction motor drive. *Proceedings of the International Conference on Power Electronics, Drives and Energy Systems, PEDES 1996*, New Delhi, India, pp. 862–868, 8–11 Jan 1996
30. S.-K. Sul, *Control of Electric Machine Drive Systems* (Wiley, New Jersey, 2011), pp. 255–267
31. P. Vas, *Electrical Machines and Drives: A Space vector Theory Approach* (Clarendon Press, Oxford, 1992). 1992
32. D.G. Holmes, T.A. Lipo, *Pulse Width Modulation for Power Converters: Principles and Practice* (Wiley IEEE Press, New Jersey, 2003). 2003

The Investigation of the Optical and Electrochemical Characteristics for the Pani Thin Film by Cyclic Voltammetry and Potentiostatic Methods

Chia-Yu Liu, Jung-Chuan Chou, Yi-Hung Liao, Cheng Jung Yang and Hsueh-Tao Chou

Abstract The objective of this study is to investigate optical and electrochemical characteristics of Polyaniline (PANI)/indium tin oxide/glass (ITO/Glass) by cyclic voltammetry and potentiostatic method. The electrochromic behaviors of the PANI/ITO/Glass were performed in 0.1 M lithium perchlorate ($LiClO_4$)/propylene carbonate (PC) electrolyte. The coloration efficiency (η) of the PANI/ITO/Glass was 9.35 cm^2/C. Furthermore, the experimental results observed that the color of PANI/ITO/Glass was changed from Green to Light Green.

Keywords Chromaticity coordinate · Coloration efficiency · Cyclic voltammetry · Electrochromism · Polyaniline · Potentiostatic method

C.-Y. Liu · J.-C. Chou (✉) · C. J. Yang · H.-T. Chou
Graduate School of Electronic and Optoelectronic Engineering, National Yunlin University of Science and Technology, Douliou 64002, Taiwan, People's Republic of China
e-mail: choujc@yuntech.edu.tw

C.-Y. Liu
e-mail: u9813335@yuntech.edu.tw

C. J. Yang
e-mail: m10113321@yuntech.edu.tw

H.-T. Chou
e-mail: chouht@yuntech.edu.tw

Y.-H. Liao
Department of Information Management, TransWorld University, Douliou 64002, Taiwan, People's Republic of China
e-mail: liaoih@twu.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_29,

1 Introduction

The conjugated polymers are organic macromolecules, which consist of one or more backbone chains of alternating double and single bonds. Conjugated polymers have been widely used in variety of applications, such as polymer conductors [1], electronic components [2, 3], light-emitting diodes [4], batteries [5], and polymer electrochromic devices (PECDs) [6], because they exhibit several advantages, such as thermal stability, low cost and easy preparation.

Various methods have been proposed to deposit the conjugated polymers thin film, such as chemical and electrochemical polymerizations. The conjugated polymers prepared by chemical or electrochemical polymerization have received significant attention due to the wide range of electrical, electrochemical, and optical properties [7]. In this study, the thin film of the conjugated polymers was deposited on the indium tin oxide/glass (ITO/Glass) substrate by electrochemical polymerization. The advantages of electrochemical polymerization comparing with other methods include rapidity, simplicity, generation of the polymer directly on the electrode in the doped or undoped states, and easy controlled synthesis of these compounds [8]. According to these results, the preparation, characterization and application of electrochemically active and electronically conjugated polymeric systems are still investigated in electrochemistry [9].

Polyaniline (PANI) shows yellow color in the reduction state and green color in the oxidation state [10–13]. It can be applied to electrochromic displays [14]. According to the mentioned above, the electrochromic PANI thin film can be applied in display for green pixel.

In this study, the PANI thin film has been electrodeposited on the indium tin oxide/glass (ITO/Glass) substrate with the various deposition charges to optimize electrochromic property of the PANI thin film. Furthermore, the electrochromic property of the PANI thin film has been studied in a 0.1 M lithium perchlorate ($LiClO_4$)/propylene carbonate (PC) electrolyte solution, and color of the PANI thin film was switched between green (1.0 V (PANI vs. Platinum (Pt))) and yellow (−0.5 V (PANI vs. Pt)).

2 Experimental

2.1 Materials

Indium tin oxide/glass (ITO/Glass) substrate was manufactured by Sinonar Corp., Taiwan, and its sheet resistivity is 7 Ω/unit square area. Aniline (ANI) solution, hydrogen chloride (HCl) solution, lithium perchlorate ($LiClO_4$) powders and propylene carbonate (PC) solution were all purchased from Acros Organics Corp., USA.

2.2 Instrumentation

2.2.1 Ultraviolet-Visible (UV-Vis) Spectroscopy

In this study, all optical parameters of thin film, e.g. transmittance, absorbance and energy band gap were observed by Ultraviolet-visible (UV-vis) spectroscopy (LABOMED UVD-3500, USA), and these parameters were utilized to calculate the optical modulation, optical density and coloration efficiency of electrochromic thin film. The software of UV-vis measuring system was UVwin 5 (v5.1.0), and the scanning wavelength was set from 900 to 190 nm. The scanning speed and the interval were set at medium and 1 nm, respectively.

2.2.2 Micro Spectroscopy

In this study, chromaticity coordinates parameters of thin film was observed by micro spectroscopy (SD1200-LS-HA, Tiawan). The software of chromaticity coordinates measuring system was sprctrasmart 1.0.4.0. For the simulation of mid-morning to mid-afternoon natural light, the relative spectral power distribution of a D65 constant temperature (6,500 K black body radiation) standard illuminant was used in the calculations.

2.2.3 Cyclic Vlotammetric (CV) Measurement System

In this study, the cyclic voltammetric (CV) measurement system (BioLogic SP-150, France) was utilized to observe the redox reactions and coloration efficiency of the PANI thin film. The technique of CV is composed of [20]:

1. An initial rest potential sequence,
2. A starting potential setting block,
3. The 1st potential sweep with a final limit E_1,
4. The 2nd potential sweep in the opposite direction with a final limit E_2,
5. The possibility to repeat n_c cycles for the 1st and the 2nd potential sweeps.

[Note that all the different sweeps have the same scan rate (absolute value)]

The measurement was performed in electrolyte with the three electrodes arrangement comprising the PANI thin film as the working electrode, a platinum counter electrode and Ag/AgCl reference electrode, and the schematic diagram of CV measurement system.

2.3 Substrate Cleaning

The characteristics of ITO/Glass substrate were listed on Table 1. The processes of the substrate cleaning were described as follows:

Table 1 Characteristics of ITO/glass substrate

ITO/glass substrate	
Thickness of substrate	0.7 mm
Resistance of substrate	<15 ohm/square
Transmittance of substrate	>86 %

The purchased ITO/Glass substrate was cut into 1.4 cm × 5 cm.

The ITO/Glass substrate was immersed quickly in ethanol to remove acetone, and utilizing D. I. water to remove ethanol and dirt with ultrasonic cleaner for 10 min.

The nitrogen gas was utilized to remove the water spots on ITO/Glass substrate and the substrate was put into the oven at 120 °C for 20 min to evaporate the water molecular on ITO/Glass substrate.

2.4 Preparation of the PANI Thin Film

In this study, the electrochromic PANI thin film have been deposited on the indium tin oxide/glass (ITO/Glass) substrate by cyclic voltammetry and potentiostatic method. The PANI thin film was deposited on the ITO/Glass by cyclic voltammetry in 1 M ANI and 2 M HCl with deionized (D. I.) water, and the potentials were set at −0.5 V to +1.5 V, the scan rate is 100 mV s^{-1}. The deposited cycles were controlled for 10, 20, 30, 40, 50 and 60 cycles, respectively. On the other hand, the PANI thin film was deposited on the ITO/Glass by potentiostatic polymerization in 1 M ANI and 2 M HCl with deionized (D. I.) water, and the potential was set at +1.5 V.

The deposited time were controlled for 400, 800, 1,200, 1,600, 2,000 and 2,400 s, respectively. The PANI thin film electrodes were removed from the monomer/electrolyte solution after electrochemical polymerization and rinsed with 0.1 M HCl to produce cleaned surface without monomer. Figure 1 was the schematic diagram of the electrochromic PANI device in this study.

3 Results and Discussion

The PANI thin film was obtained by cyclic voltammetry and potentiostatic method. And the PANI thin film was obtained by electrochemical polymerization at potential slightly higher than the monomer oxidation onset potential, and could obtain a homogeneous thin film on the ITO/Glass. The redox behavior of the ANI monomer was obtained in 1 M ANI and 2 M HCl with D. I. water by cyclic voltammetry, as shown in Fig. 2. An oxidation onset potential was obtained about

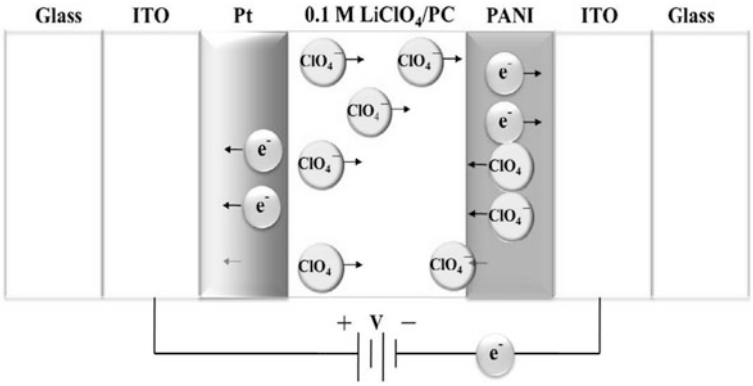


Fig. 1 Structure of electrochromic PANI device [15]

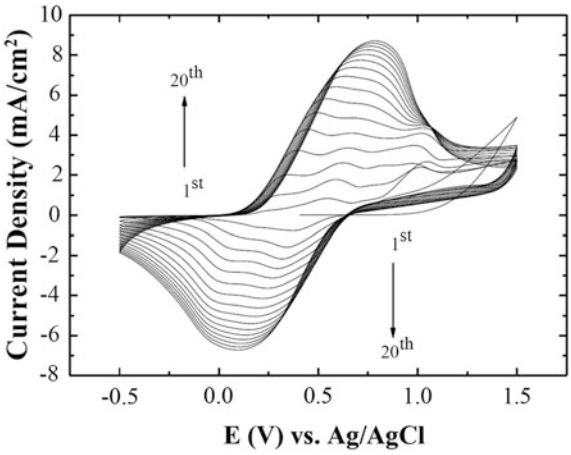


Fig. 2 Cyclic voltammograms of PANI thin film during electrodeposition (scan rate = 100 mV/s)

1.25 V. Because lower potential could not synthesize the ANI, the potential of polymerization of the PANI thin film was set at 1.5 V. The CV curves of the PANI thin film exhibited that the onset potential of oxidation of ANI was at +1.25 V in 1 M ANI and 2 M HCl with D. I. water, as well as, the oxidation peak of the PANI thin film at +0.75 V and reduction peak of the PANI thin film was at +0.15 V during the 20 cycles.

As shown in Fig. 3, the cyclic voltammetry of the PANI/ITO/Glass and the maximum current density were observed when the PANI thin film (the PANI/ITO/Glass with 10 cycles) was fabricated by cyclic voltammetry. During the oxidation (anodic peak at 0.75 V), the ClO_4^- ions of $LiClO_4$/PC electrolyte solution were injected from the PANI thin film. During the reduction (cathodic peak at −0.25 V), the ClO_4^- ions of $LiClO_4$/PC electrolyte solution were excluded the PANI thin film.

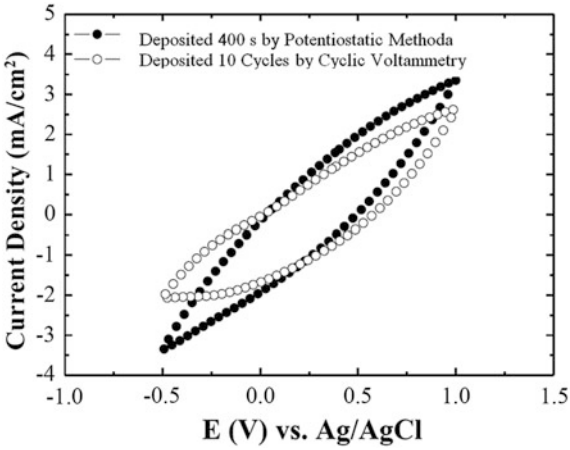


Fig. 3 Cyclic voltammograms of PANI/ITO/glass in 0.1 M $LiClO_4$/PC electrolyte solution with scan rate of 100 mV s − 1 by different methods: *black circle* deposited 400 s by constant potential +1.5 V; *circle* deposited 10 cycles by deposited voltage −0.5 V to +1.5 V

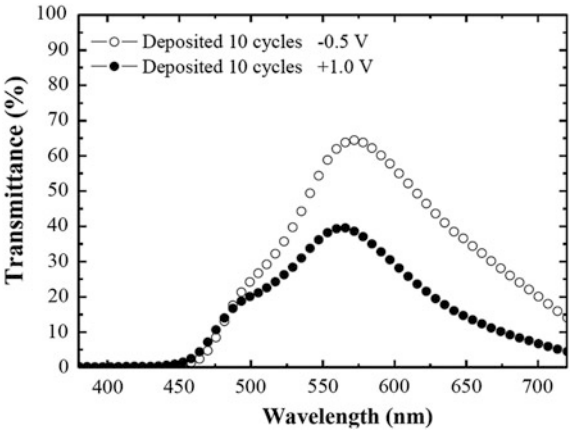


Fig. 4 Transmittances of PANI/ITO/glass (the PANI/ITO/glass with 10 cycles) was immersed in 0.1 M $LiClO_4$/PC electrolyte solution, and the oxidizing potential and reducing potential were set +1.0 V and −0.5 V, respectively

The overall reaction, involving ion diffusion -in and -out of the polymer matrix to balance the charge, can be represented as in Eq. (1) for PANI [13],

$$\underset{\text{(Light Green)}}{(\mathrm{ANI})_n} + ny(\mathrm{ClO_4^-}) \longleftrightarrow \underset{\text{(Green)}}{[(\mathrm{ANI+}) + y(\mathrm{ClO_4^-})]_n} + nye^- \quad (1)$$

where n is the number of repeated units and y is the stoichiometric number of the counter ion [13].

As shown in Fig. 4, the colored and bleached transmittances of the PANI/ITO/Glass were studied with 10 cycles and −0.5 V to +1.5 V by cyclic voltammetry. The PANI/ITO/Glass was performed in 0.1 M $LiClO_4$/PC electrolyte solution. The potentials were set −0.5 V and +1.0 V. The solid shape and hollow shape were coloring and bleaching, respectively. The PANI/ITO/Glass has the maximum optical transmittance variation (ΔT (%)) which was 20 % at 550 nm.

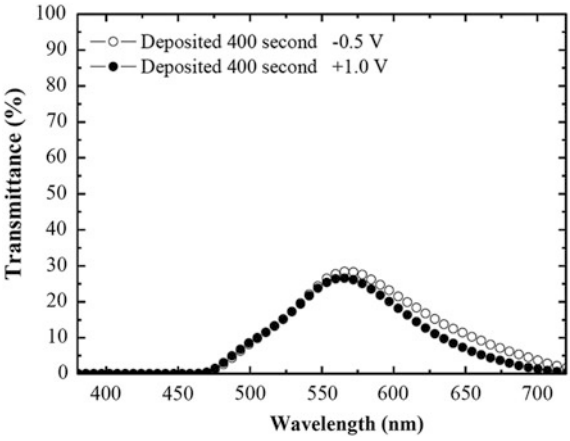


Fig. 5 Transmittances of PANI/ITO/glass (the PANI/ITO/glass with 400 s) was immersed in 0.1 M $LiClO_4$/PC electrolyte solution, and the oxidizing potential and reducing potential were set +1.0 V and −0.5 V, respectively

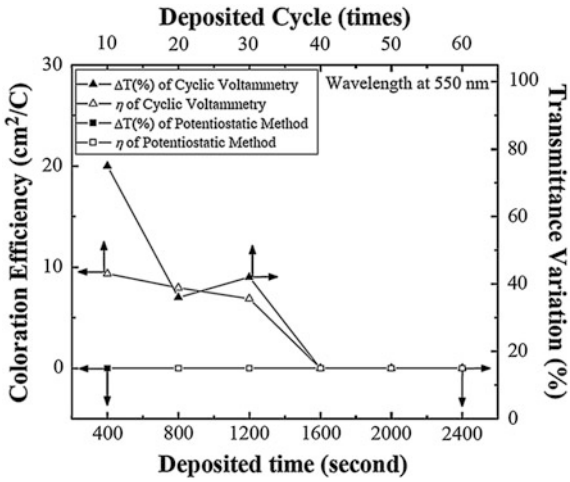


Fig. 6 Transmittance variation and coloration efficiency of the PANI/ITO/Glass in 0.1 M $LiClO_4$/PC electrolyte solution between oxidizing potential (+1.0 V) and reducing potential (−0.5 V)

As shown in Fig. 5, the colored and bleached transmittances of the PANI/ITO/Glass were studied with 400 s and +1.5 V by potentiostatic method. The PANI/ITO/Glass was performed in 0.1 M $LiClO_4$/PC electrolyte solution. The potentials were −0.5 V and +1.0 V. The solid shape and hollow shape were coloring and bleaching, respectively. The PANI/ITO/Glass has the maximum optical transmittance variation (ΔT (%)) which was 1 % at 550 nm.

As shown in Fig. 6, the coloration efficiency (η) of the PANI/ITO/Glass with different fabricated parameters. The PANI/ITO/Glass had the best transmittance variation (20 %) when the deposited cycles and voltage of the PANI thin film were set for 10 cycles and −0.5 V to +1.5 V, respectively. The coloration efficiency of the PANI/ITO/Glass was 9.35 cm^2/C. Compared with the above results, the coloration efficiency of the PANI/ITO/Glass was higher than other literatures [16–18], as summarized in Table 2.

Table 2 Comparison of the coloration efficiency of PANI/ITO/glass in this study and previous literatures for electrochromic devices

Electro chromic device	Method	Coloration efficiency (cm^2/C)	Wave length (nm)	Tran smittance variation (ΔT (%))	Time of manu facture	Cost of manu facture	References
PANI/ITO/ glass	Electro chemistry	9.35	550	20	Short	Low	In this study
PANI/ITO/ glass	Chemistry	42.80	550	41	Long	High	[16] (2009)
PANI/G/glass	Chemistry	39.60	550	45	Long	High	[16] (2009)
PANI/ITO/ glass	Screen-printed	52.00	550	N/A	Long	High	[17] (2009)
POSS-PANI/ ITO/glass	Screen-printed	69.00	550	N/A	Long	High	[17] (2009)
PANI/ITO/ glass	Electrochemistry	6.20	550	N/A	Short	Low	[18] (2009)
PANI/PASA/ ITO/glass	Solution bath	N/A	633	40	Long	High	[19] (2010)

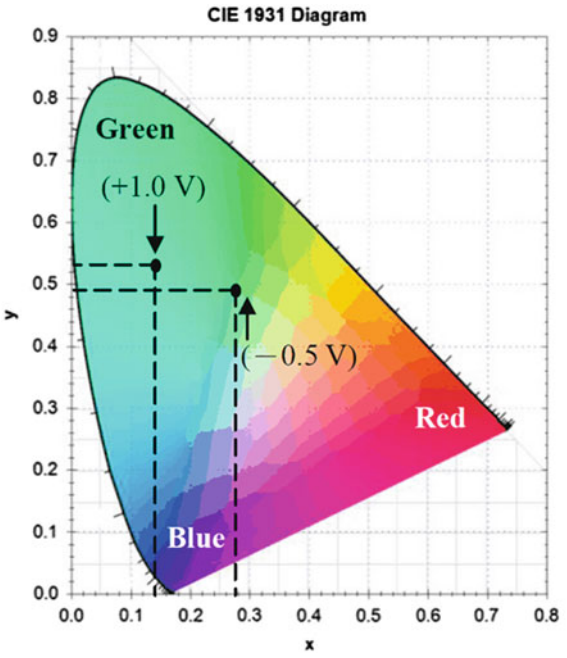


Fig. 7 Chromaticity coordinates of CIE of PANI/ITO/glass (the colored and bleached transmittances of the PANI/ITO/glass were studied with 10 cycles and −0.5 V to +1.5 V by cyclic voltammetry) were performed in 0.1 M $LiClO_4$/PC electrolyte solution

The variant colors of the electrochromic thin film were observed by the chromaticity coordinates of the commission internationale de l'Eclairage 1931 (CIE 1931) [21]. Figure 7 shows the chromaticity coordinates of CIE of PANI/ITO/Glass (the colored and bleached transmittances of the PANI/ITO/Glass were

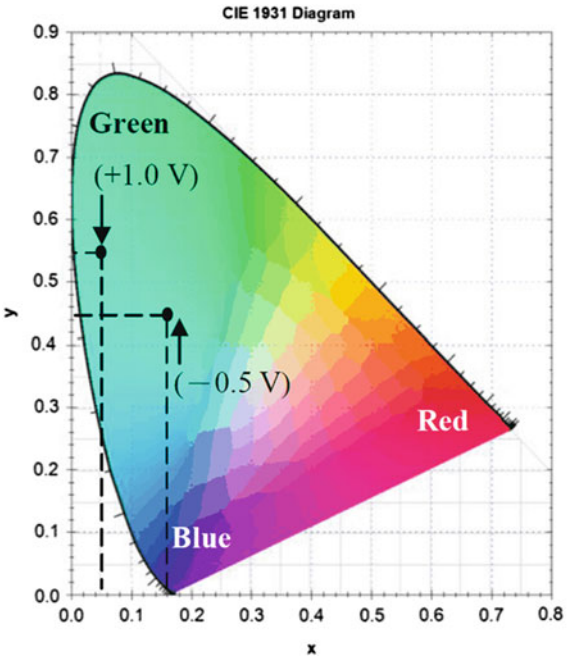


Fig. 8 Chromaticity coordinates of CIE of PANI/ITO/Glass (the PANI/ITO/Glass was studied with 400 s and +1.5 V by potentiostatic method) were performed in 0.1 M $LiClO_4$/PC electrolyte solution

studied with 10 cycles and −0.5 V to +1.5 V by cyclic voltammetry) were performed in 0.1 M $LiClO_4$/PC electrolyte solution. The PANI/ITO/Glass in yellow color state (reduction state) was at −0.5 V (PANI vs. Pt), and the chromaticity coordinate of the PANI/ITO/Glass was (0.28, 0.48). Green color state (oxidation state) of PANI thin film was at +1.0 V (PANI vs. Pt), and the chromaticity coordinate of the PANI/ITO/Glass was (0.14, 0.54).

Figure 8 shows the chromaticity coordinates of CIE of PANI/ITO/Glass (the PANI/ITO/Glass was studied with 400 s and +1.5 V by potentiostatic method) was performed in 0.1 M $LiClO_4$/PC electrolyte solution. The PANI/ITO/Glass in light green color state (reduction state) was at −0.5 V (PANI vs. Pt), and the chromaticity coordinate of the PANI/ITO/Glass was (0.16, 0.42). The PANI/ITO/Glass in dark green color state (oxidation state) was at +1.0 V (PANI vs. Pt), and the chromaticity coordinate of the PANI/ITO/Glass was (0.04, 0.44). Above descriptions of the Chap. 3 is part of description from literature [22].

4 Conclusion and Outlook

The electrochromic PANI thin film has been successfully deposited on ITO/Glass by cyclic voltammetry and potentiostatic method. The PANI/ITO/Glass had the best transmittance variation (20 %) when the deposited cycle and voltages of the PANI thin film were set for 10 cycles and −0.5 V to +1.5 V, respectively.

The coloration efficiency of the PANI/ITO/Glass was 9.35 cm^2/C. According to the experimental results, the electrochromic PANI thin film was prepared by cyclic voltammetry, which has the best transmittance variation and current density. Furthermore, the experimental results obtained the colors of the PANI/ITO/Glass were from Green (oxidation state) to Light Green (reduction state). According to the above results, the electrochromic PANI/ITO/Glass can be applied in display for green pixel. The electrochromic performance of the electrochromic PANI thin film can be improved by changing other parameters such as potential range, temperature of deposition and deposited cycle in cyclic vlotammetric system.

Because large area and array-type can increase the applications, to design large area and array-type flexible electrochromic device, colloidal electrolyte can be applied to flexible electrochromic device, which will improve the overflow of liquid electrolyte.

Acknowledgment This study has been supported National Science Council, Republic of China, under the contracts NSC 100-2221-E-224-017, NSC 101-2221-E-224-046, NSC 101-2221-E-265-001 and NSC 102-2221-E-224-075.

References

1. X. Crispin, F.L.E. Jakobsson, A. Crispin, P.C.M. Grim, P. Andersson, A. Volodin, C. van Haesendonck, M. van der Auweraer, W.R. Salaneck, M. Berggren, The origin of the high conductivity of poly (3, 4-ethylenedioxy-thiophene)-poly (styrenesulfonate) (PEDOT-PSS) plastic electrodes. Chem. Mater. **18**, 4354–4360 (2006)
2. H. Sirringhaus, T. Kawase, R.H. Friend, T. Shimoda, M. Inbasekaran, W. Wu, E.P. Woo, High-resolution inkjet printing of all-polymer transistor circuits. Science **290**, 2123–2126 (2000)
3. H.E. Katz, J. Huang, Thin-film organic electronic devices. Annu. Rev. Mater. Sci. **39**, 71–92 (2009)
4. I.D. Parker, Carrier tunneling and device characteristics in polymer light-emitting diodes. J. Appl. Phys. **75**, 1656–1666 (1994)
5. S.R. Sivakkumar, J.Y. Nerkar, A.G. Pandolfo, Rate capability of graphite materials as negative electrodes in lithium-ion capacitors. Electrochim. Acta **54**, 6844–6849 (2009)
6. A. Yildirim, S. Tarkuc, M. Ak, L. Toppare, Syntheses of electroactive layers based on functionalized anthracene for electrochromic applications. Electrochim. Acta **53**, 4875–4882 (2008)
7. G.A. Sotzing, J.R. Reynolds, Electrochromic conducting polymers via electrochemical polymerization of bis (2-(3, 4-ethylenedioxy) thienyl) monomers. Chem. Mater. **8,** 882–889 (1996)
8. Y. Pang, X. Li, H. Ding, G. Shi, L. Jin, Electro polymerization of high quality electrochromic poly (3-alkyl-thiophene)s via a room temperature ionic liquid. Electrochim. Acta **52**, 6172–6177 (2007)
9. G. Inzelt, M. Pineri, J.W. Schultze, M.A. Vorotyntsev, Electron and proton conducting polymers: recent developments and prospects, Electrochim. Acta **45**, 2403–2421 (2000)
10. K. Sheng, H. Bai, Y. Sun, C. Li, G Shi, Layer-by-layer assembly of graphene/polyaniline multilayer films and their application for electrochromic devices. Polymer **52**, 5567–5572 (2011)

11. S. Xionga, F. Yangb, G. Ding, K.Y. Mya, J. Ma, X. Lu, Covalent bonding of polyaniline on fullerene: enhanced electrical, ionic conductivities and electrochromic performances. Electrochim. Acta **67**, 194–200 (2012)
12. H. Karami, M.G. Asadic, M. Mansoori, Pulse electro polymerization and the characterization of polyaniline nano fibers. Electrochim. Acta **61**, 154–164 (2012)
13. K.Y. Shen, C.W. Hu, L.C. Chang, K.C. Ho, A complementary electrochromic device based on carbon nanotubes/conducting polymers. Solar Energy Mater Solar Cells **98**, 294–299 (2012)
14. P. Somani, A.B. Mandale, S. Radhakrishnan, Study and development of conducting polymer-based electrochromic display devices. Acta Mater. **48**, 2859–2871 (2000)
15. C.G. Granqvist, A. Azens, J. Isdorsson, M. Kharrazi, L. Kullman, T. Lindstrom, G.A. Niklasson, C.G. Ribbing, D. Ronnow, M. Stromme Mattson, M. Veszelei, Towards the smart windows: progress in electromics. J. Non-Crys. Solids **218**, 273–279 (1997)
16. L. Zhao, L. Zhao, Y. Xu, T. Qiu, L. Zhi, G. Shi, Polyaniline electrochromic devices with transparent graphene electrodes. Electrochim. Acta **55**, 491–497 (2009)
17. L. Zhang, S. Xiong, J. Ma, X. Lu, A complementary electrochromic device based on polyaniline-tethered polyhedral oligomeric silsesquioxane and tungsten oxide. Solar Energy Mater. Solar Cells **93**, 625–629 (2009)
18. J.H. Kang, Y.J. Oh, S.M. Peak, S.J. Hwang, J.H. Choy, Electrochromic device of PEDOT-PANI hybrid system for fast response and high optical contrast. Solar Energy Mater. Solar Cells **93**, 2040–2044 (2009)
19. R. Montazami, V. Jain, J.R. Heflin, High contrast asymmetric solid state 161 electrochromic devices based on layer-by-layer deposition of polyaniline and poly (aniline sulfonic acid). Electrochim. Acta **56**, 990–994 (2010)
20. EC-Lab Software: Techniques and Applications, Instrument Manual, Bio-Logic SAS, version 9.5 (2008)
21. T. Smith, J. Guild, Transactions of the Optical Society **33**, 73 (1931)
22. C.Y. Liu, J.C. Chou, Y.H. Liao, C.J. Yang, C.J. Huang, T.Y. Cheng, J.E. Hu, H.T. Chou, The investigation of the optical and electrochemical characteristics for the PANI thin film by cyclic voltammetry and potentiostatic method, in *Lecture Note in Engineering and Computer Science: Proceedings of the World Congress Engineering 2013*, London, UK, 3–5 July 2013 pp. 999–1003

Influence of Titanium Dioxide Layer Thicknesses and Electrolyte Thicknesses Applied in Dye-Sensitized Solar Cells

Jui-En Hu, Jung-Chuan Chou, Yi-Hung Liao, Shen-Wei Chuang and Hsueh-Tao Chou

Abstract In this study, the different TiO_2 layer thicknesses and electrolyte thicknesses were investigated, which were applied in the dye-sensitized solar cells (DSSC). The amount of dye adsorption was decided by thickness of TiO_2 layer, the appropriated TiO_2 thickness could increase the short-circuit current density of DSSC and decreased the resistance of TiO_2 layer, effectively. The oxidation-reduction reaction of inner electrochemical of DSSC was decided by thickness of electrolyte. The appropriate electrolyte thickness could increase the redox rate of DSSC and decreased the distance of electron transmission, effectively.

Keywords Dye adsorption · Dye-sensitized solar cells · Electrolyte thickness · Electron transmission · Oxidation-reduction reaction · TiO_2 layer thicknesses

J.-E. Hu · S.-W. Chuang
Graduate School of Electronic and Optoelectronic Engineering, National Yunlin University of Science and Technology, Douliou, 64002 Taiwan, R.O.C
e-mail: M10013323@yuntech.edu.tw

J.-C. Chou (✉) · H.-T. Chou
Department of Electronic Engineering and Graduate School of Electronic and Optoelectronic Engineering, National Yunlin University of Science and Technology, Douliou, 64002 Taiwan, R.O.C
e-mail: choujc@yuntech.edu.tw

H.-T. Chou
e-mail: chouht@yuntech.edu.tw

Y.-H. Liao
Department of Information Management, TransWorld University, Yunlin, 64063 Taiwan, R.O.C
e-mail: liaoih@twu.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_30,

1 Introduction

In recent years, some of petrochemical energies were dried up day by day, such as petroleum, coal, fuel and natural gas, which produced serious pollution to influence the environment and humanity. Solar cells are very effective solution for solving the problems of the depletion of fossil fuels and the emission of greenhouse gases [1]. Various solar cells, such as dye-sensitized solar cells (DSSCs), a-Si thin film solar cells [2], organic solar cells [3], and quantum dot solar cells [4] have been researched. Compared with silicon solar cell, glass-based DSSC has been extensively studied due to its low cost, easy fabrication and high transmittance [5].

Dye-sensitized solar cell (DSSC) was one of photoelectrochemical solar cells, which was composed of a dye-modified wide band working electrode, a platinum counter electrode, N3 dye and an electrolyte layer containing a redox couple (I^-/I_3^-), as shown in Fig. 1. The working electrode consisted of titanium dioxide (TiO_2) film which was fabricated on the substrate by screen printing technique. The counter electrode consisted of platinum (Pt) film which was fabricated on substrate by R. F. magnetic sputtering. The sensitized dye solution was adsorbed on the TiO_2 film of working electrode and the electrolyte containing I^-/I_3^- redox couple [6].

TiO_2 film of working electrode was porous nanostructure, which was fabricated on substrate by screen printed technique to be an adsorption layer. The TiO_2 film provided the high surface area which could be absorbed by sensitized dye. The most important thing was the energy level had to match each other between adsorption layer and sensitized dye. The TiO_2 adsorption layer could produce electron and hole pairs after illumination, the common materials were such as TiO_2 [7], $BaSnO_3$ [8], ZnO [9], $SrTiO_3$ [10] and CdS [11], the TiO_2 material had higher photocatalytic activity than others. And the TiO_2 material had some advantages, such as low cost, chemical stability, and no poison.

Electrolyte of dye-sensitized solar cell was utilized to produce oxidation-reduction reaction, which could reduce the dye molecule. DSSC almost utilized the liquid state of electrolyte, the advantages were more types and easy to control, but it had some disadvantage as follows: (1) Boiling point of organic solvent was lower and easy to evaporate, it was not good for stability of DSSC to be utilized for a long time. (2) Liquid state electrolyte was difficult to be packaged, which produced leakage of electrolyte for a long time. (3) Electrolyte was organic solvent. (4) Some of moisture in liquid state electrolyte caused the dye molecule to be adsorbed.

2 Experimental

2.1 Chemicals and Materials

Titanium dioxide (TiO_2) powder and Ruthenium-535 (N3) were purchased from UniRegion Bio-Tech, Taiwan. The ethanol was purchased from Katayama Chemical, Japan. The Triton X-100 was purchased from PRS Panreac, Spain. The

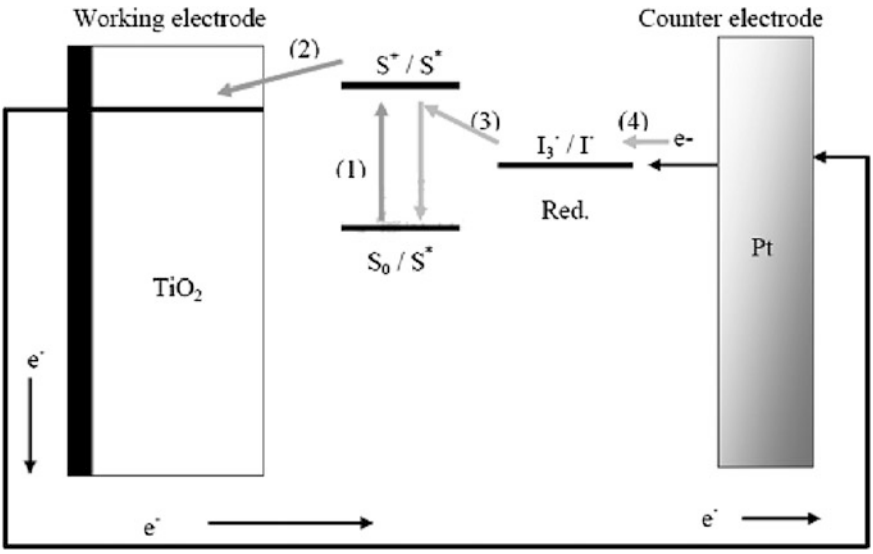


Fig. 1 Working principle scheme of the dye-sensitized solar cell [6]

Acetylacetone (AcAc), lithium iodide (LiI) and 4-Tert-Butylpyridine (TBP) were purchased from Sigma-Aldrich, United States. The iodine (I_2) was purchased from Riedel-deHaen, United States. The 1-propyl-2,3-dimethylimidazolium iodide (DMPII) was purchased from Tokyo Chemical, Japan. The Surlyn layer (Meltonix 1170-60 Series) was purchased from Solaronix, Switzerland.

2.2 Preparation of Solvent, Paste and Fabrication of Dye-Sensitized Solar Cell

The TiO_2 paste consists of 3 g TiO_2 powder (P25), 3.5 mL deionized (D. I.) water, 0.1 mL acetylacetone and 0.3 mL Triton X-100 [12, 13]. The TiO_2 working electrodes with an active area of 0.64 cm^2, which were fabricated on FTO and ITO-PET substrates [14] by screen-printing technique, as shown in Fig. 2. The working electrode was baked at 100 °C for 10 min, and then was immersed in an absolute ethanol solution of 3×10^{-4} M N3 dye at 75 °C for 1 h. Platinum was fabricated on ITO-PET substrate by sputtering for 90 s and was generally regarded as the counter electrode. The liquid-state electrolyte consists of 0.6 M DMPII, 0.5 M LiI, 0.05 M I_2, and 0.5 M TBP in 15 mL MPN [15]. Finally, Dye-sensitized solar cells (DSSCs) were sealed by Surlyn.

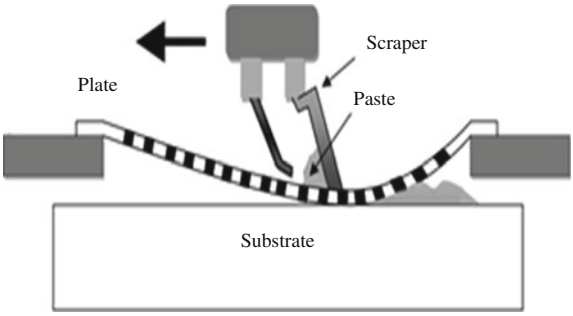


Fig. 2 Fabricated TiO_2 thin film of working electrode by screen printing technique

2.3 Design and Measurement of Dye-Sensitized Solar Cell

First, the 0.8 × 0.8 cm active area of TiO_2 thin films were fabricated on substrate by screen printing technique. The DSSC is based on a sandwich structure [16], which consists of working electrode, electrolyte and counter electrode, as shown in Fig. 3. The electrolyte thicknesses and TiO_2 layer thicknesses were changed by different materials and space layers quantities.

The short-circuit current density (J_{sc}), open circuit voltage (V_{oc}), fill factor (FF) and conversion efficiency (η) of DSSC were measured by Keithley 2,400 digital source meter under one sun illumination (AM 1.5 G, 100 mW/cm^2). And the thickness of Surlyn, Teflon tape, and TiO_2 thin film were measured by Stylus Surface Profiling System.

3 Results and Discussion

3.1 Analysis of Thicknesses of TiO_2 Layers by Different Fabrication Pressures

In this study, the pressure of screen printing technique was utilized to determine the optimal thickness of TiO_2 layer. The common pressure as the stander (0 circle) in this experiment, and the different pressures were investigated from large to small (−2 circle, −1 circle, 0 circle, +1 circle and +2 circle), The current density-voltage (J-V) curves of different pressures were shown in Fig. 4 and the characteristic parameters were listed in Table 1. The FTO surface was very smooth and low adhesion, which needed an appropriate pressure to fabricate TiO_2 layer. When the pressure was −1 circle, the adhesion between TiO_2 layer and FTO glass substrate was better than other pressures, which enhanced the amount of dye adsorption. According to the experiment result, when the fabrication pressure was −1 circle had better characteristics, where the short-circuit current density was 5.28 mA/cm^2 and the conversion efficiency was 2.15 %, respectively.

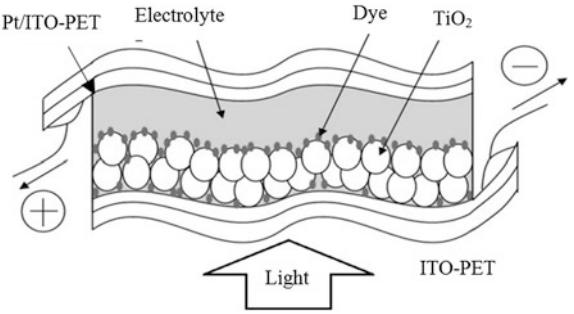


Fig. 3 Schematic diagram of flexible dye-sensitized solar cell

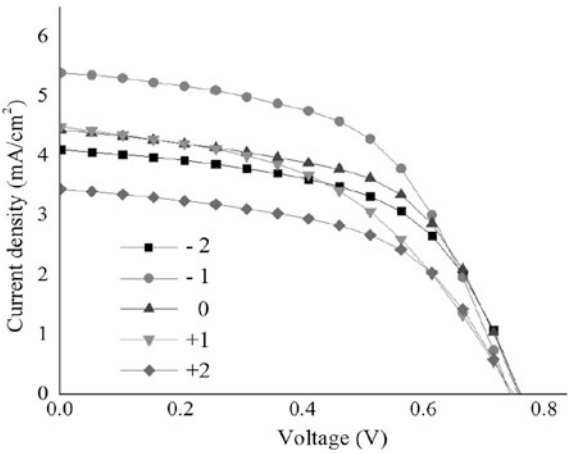


Fig. 4 J-V curves of different pressures of screen printing technique

Table 1 Different fabrication pressures of screen printing technique

Pressure (circle)	Thickness (μm)	V_{oc} (V)	J_{sc} (mA/cm^2)	FF (%)	η (%)
−2	17.26	0.76	4.11	55.38	1.73
−1	19.63	**0.74**	**5.28**	**55.07**	**2.15**
0	21.56	0.76	4.44	56.12	1.89
+1	23.87	0.75	4.49	46.03	1.58
+2	25.19	0.74	3.44	53.36	1.37

In literature [17], the TiO_2 layer was fabricated on FTO glass substrate by screen printing technique. The thicknesses parameters of TiO_2 layer were 3, 6, 8, 12, 15, 20 and 26 μm, the optimal thickness of TiO_2 layer was 20 μm, where the short-circuit current density was 14.42 mA/cm^2 and conversion efficiency was 7.85 %. The increasing percentages (%) of literature [17] and our research group were listed in Table 2. Compared with our experimental result, the increasing percentages of short-circuit current density and conversion were higher than literature [17].

Table 2 Increasing percentage of J_{sc}, V_{oc} and η with different TiO_2 thicknesses in our experimental result and previous literature

Thickness (μm)	Increasing percentage (%)		
	V_{oc} (V)	J_{sc} (mA/cm^2)	η (%)
17.26–19.63	−3.6	28.5	24.3
15.00–20.00 [17] (2010)	−1.4	16.5	18.8

3.2 Analysis of Different Thicknesses of Electrolyte

The thickness of space layer of flexible dye-sensitized solar cell (FDSSC) was an important part for reduction-oxidation reaction of electrolyte. Figure 5 showed the schematic diagram of space layer of DSSC. The thicknesses of electrolyte were changed by different materials and space layers quantities. When the space layer was too thin, the distance was short for electron to transmit, but the TiO_2 thin film touched the counter electrode and produced the leakage of electrolyte. When the space layer was too thick, the excessive amount of electrolyte produced higher resistance and longer electron transmission. Figure 5 showed the schematic diagram of different thicknesses of electrolyte. Figure 6a showed the regular DSSC which had one space layer [18], and Fig. 6b showed the DSSC which had thicker space [19], and the extra space region was called "region 2".

Figure 7 showed the short circuit current density-open circuit voltage (J-V) curves of flexible dye-sensitized solar cells (FDSSCs) with different space layers for Teflon tape. The thickness of space layer subtracts the thickness of TiO_2 thin film could get the real thickness of electrolyte to investigate the different thicknesses of space layers. The thicknesses of one, two and three space layers of electrolyte applied in Teflon tape were 38.47, 85.19 and 131.91 μm, respectively. According to Fig. 7, the one layer of Teflon tape touched the counter electrode which had no trend of fill factor, and the 3 layers of Teflon tape had better open circuit voltage and short circuit current density. Figure 8 showed the J-V curves of FDSSCs with different space layers for Surlyn. The thicknesses of one, two and three space layers of electrolyte applied in Surlyn were 51.75, 111.75 and 171.75 μm, respectively. According to Fig. 8, the one layer of Surlyn had less reduction-oxidation reaction, because the electrolyte was too less to be reacted, and three layers of Surlyn had enough electrolyte to enhance reduction-oxidation reaction, which could enhance open circuit voltage and short circuit current density, obviously.

The conversion efficiency of FDSSC with Surlyn 2 layers had more stable characteristics after measured 6 times. The values of characteristic parameters and thicknesses of space layers were listed in Table 3. According to Table 3, the experimental result of the thickness of thicker space had larger amount of electrolyte in FDSSC. Although the distance of electron transmission become longer, the FDSSC had better reduction-oxidation reaction to enhance short circuit current

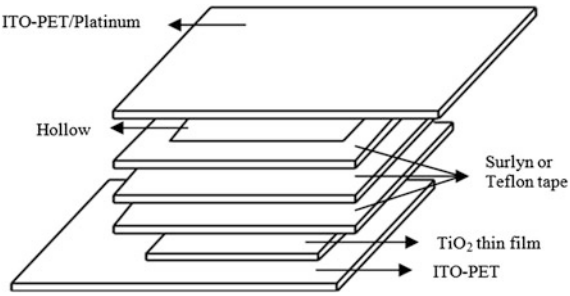


Fig. 5 Schematic diagram of space layer of flexible dye sensitized solar cell

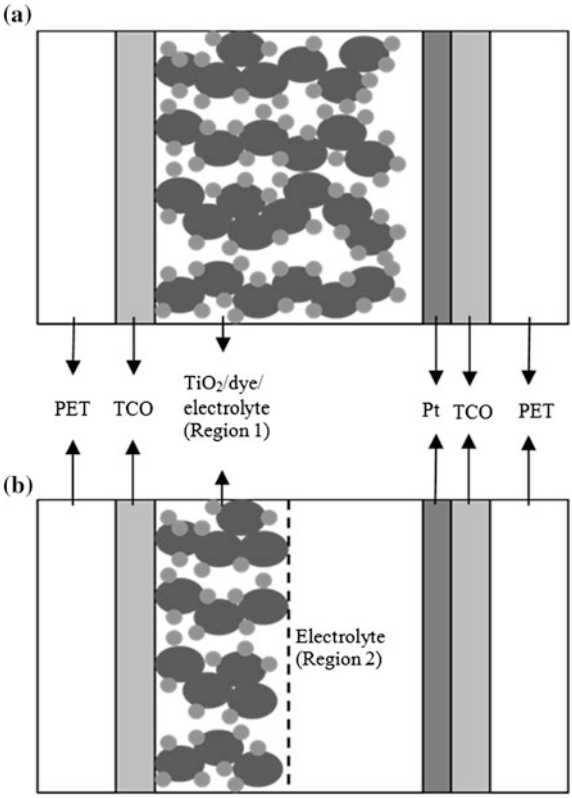


Fig. 6 Simulation models for DSSCs. **a** Cell a: model employed by Ferber et al. [18]. **b** Cell b: model proposed in literature [19]

density and conversion efficiency for FDSSC. These results proved that the reduction-oxidation reaction was more important than the distance of electron transmission between working electrode and counter electrode. Above descriptions of the Sect. 3.2 is part of description from literature [20].

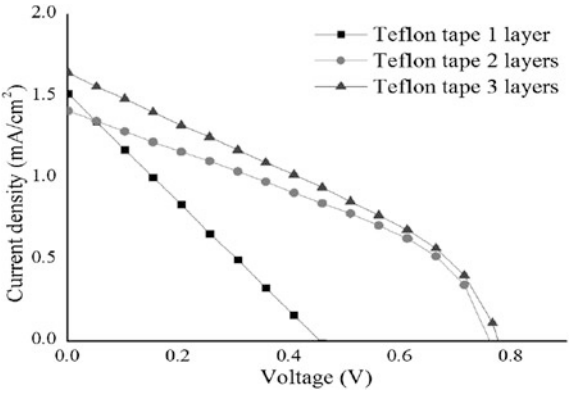


Fig. 7 Dependence of J-V curves to the spacer thickness of Teflon tape

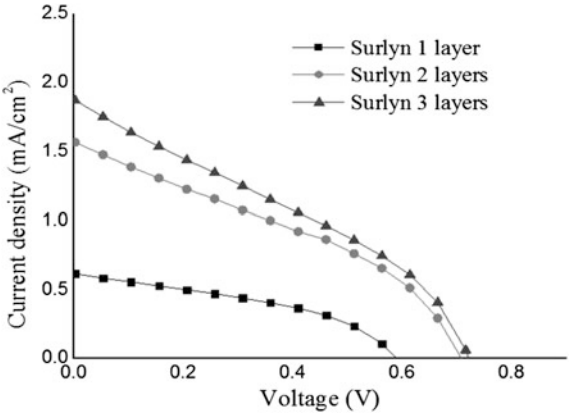


Fig. 8 Dependence of J-V curves to the spacer thickness of Surlyn

Table 3 Different space thicknesses of Teflon tape and surlyn applied in FDSSC

Materials	Electrolyte thickness (μm)	V_{oc} (V)	J_{sc} (mA/cm^2)	FF (%)	η (%)
Teflon tape 1 layer	38.47	0.46	1.52	24.82	0.17
Teflon tape 2 layers	85.19	0.76	1.41	37.24	0.40
Teflon tape 3 layers	**131.91**	**0.78**	**1.65**	**34.19**	**0.44**
Surlyn 1 layer	51.75	0.62	0.79	39.54	0.19
Surlyn 2 layers	111.75	0.71	1.57	35.83	0.40
Surlyn 3 layers	**171.75**	**0.72**	**1.88**	**32.86**	**0.45**

4 Conclusion and Outlook

In this study, the different thicknesses of space layers and thickness of TiO_2 layer of dye-sensitized solar cells were investigated to provide an improved method to enhance the reduction-oxidation reaction, dye adsorption and electron transmission,

which can enhance the short circuit current density (J_{sc}) and conversion efficiency (η), although the cell has a smaller inner resistance.

According to experiment results, the optimal conditions are Surlyn of 2 layers (111.75 μm), and the open circuit voltage, short circuit current density and conversion efficiency can reach to 0.71 V, 1.57 mA/cm^2 and 0.4 %, respectively. The optimal thickness of TiO_2 layer was 19.63 μm, where the short-circuit current density was 5.28 mA/cm^2 and conversion efficiency was 2.15 %.

At future prospects, the liquid electrolyte should change the gel electrolyte to avoid evaporating and leakage, which can increase the stability of DSSC. The TiO_2 layer is fabricated by screen printing technique, which can be manufactured numerously.

Acknowledgment This study has been supported by National Science Council, Republic of China, under the contracts NSC 100-2221-E-224-017, NSC 101-2221-E-224-046, NSC 101-2221-E-265-001 and NSC 102-2221-E-224-075.

References

1. B.O'Regan, M. Grätzel, A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. Nature **353**, 737–740 (1991)
2. F. Smole, M. Topic, J. Furlan, Analysis of TCO/p(a-Si:C:H) heterojunction and its influence on P-I-N A-Si:H solar cell performance. J. Non-Cryst. Solids **194,** 312–318 (1996)
3. E. Ahmed, A. Zegadi, A.E. Hill, R.D. Pilkington, R.D. Tomlinson, A.A. Dost, W. Ahmed, S. Lepphvuori, J. Levoska, O. Kusmartseva, Impact of annealing processes on the properties of $Culn_{0.75}Ga_{0.25}Se_2$ thin films. Solar Energy Mater. Solar Cells **36** 227–239 (1995)
4. P.K. Sharma, R.K. Dutta, M. Kumar, P.K. Singh, A.C. Pandey, Luminescence studies and formation mechanism of symmetrically dispersed ZnO quantum dots embedded in SiO_2 matrix. J. Luminescence **129**(6), 605–610 (2009)
5. Y. Chiba, A. Islam, Y. Watanabe, R. Komiya, N. Koide, L. Han, Dye-sensitized solar cells with conversion efficiency of 11.1 %. Jpn. J. Appl. Phys. **45**, L638–L640 (2006)
6. T. Yamaguchi, N. Tobe, D. Matsumoto, T. Nagai, H. Arakawa, Highly efficient plastic-substrate dye-sensitized solar cells with validated conversion efficiency of 7.6 %. Solar Energy Mater. Solar Cells **94**, 812–816 (2010)
7. B.L. Chen, H. Hu, Q.D. Tai, N.G. Zhang, F. Guo, B. Sebo, W. Liu, J.K. Yuan, J.B Wang, X.Z. Zhao, An inverted fabrication method towards a flexible dye sensitized solar cell based on a free-standing TiO_2 nanowires membrane. Electrochim. Acta **59**, 581–586 (2012)
8. F.A. Guo, L.I. GQ, W.F. Zhang, Barium staminate as semiconductor working electrodes for dye-sensitized solar cells. Int. J. Photoenergy, No. 105878 (2010)
9. M.C. Kao, H.Z. Chen, S.L. Young, C.Y. Kung, C.C. Lin, J.Z. Lai, The microstructure and ferroelectric properties of Sm and Ta-doped Bismuth titanate ferroelectric thin films. J. Supercond. Novel Mag. **23**, 897–900 (2010)
10. P. Jayabal, V. Sasirekha, J. Mayandi, K. Jeganathan, V. Ramakrishnan, A facile hydrothermal synthesis of $SrTiO_3$ for dye sensitized solar cell application. J. Alloys Compd. **586**, 456–461 (2014)
11. K. Meng, P.K. Surolia, O. Byrne, K.R. Thampi, Efficient CdS quantum dot sensitized solar cells made using novel Cu_2S counter electrode. J. Power Sources **248**, 218–223 (2013)
12. H. Chang, T.L. Chen, K.D. Huang, S.H. Chien, K.C. Hung, Fabrication of highly efficient flexible dye-sensitized solar cells. J. Alloys Compd. **504**, S435–S438 (2010)

13. Y.L. Lee, C.L. Chen, L.W. Chong, C.H. Chen, Y.F. Liu, C.F. Chi, A patinum counter electrode with high electrochemical activity and high transparency for dye-sensitized solar cells. Electrochem. Commun. **12**(11), 1662–1665 (2010)
14. Y.T. Cheng, J.J. Ho, C.K. Wang, W. Lee, C.C. Lu, B.S. Yau, J.L. Nain, S.H. Chang, C.C. Chang, K.L. Wang, Improvement of organic solar cells by flexible substrate and ITO surface treatments. Appl. Surf. Sci. **256**(24), 7606–7611 (2010)
15. J.C. Chou, Y.Y. Chiu, Y.M. Yu, S.Y. Yang, P.H. Shih, C.C. Chen, Research of titanium dioxide compact layer applied to dye-sensitized solar cell with different substrates. J. Electrochem. Soc. **159**(2), A145–A151 (2012)
16. S.R. Scully, M.T. Lloyd, R. Herrera, E.P. Giannelis, G.G. Malliaras, Dye-sensitized solar cells employing a highly conductive and mechanically robust nanocomposite gel electrolyte. Synth. Met. **144**(3), 291–296 (2004)
17. I. Shin, H. Seo, M.K. Son, J.K. Kim, K. Prabakar, H.J. Kim, Analysis of TiO_2 thickness effect on characteristic of a dye-sensitized solar cell by using electrochemical impedance spectroscopy. Curr. Appl. Phys. **10**(3), S422–S424 (2010)
18. J. Ferber, R. Stangl, J. Luther, An electrical model of the dye-sensitized solar cell. Solar Energy Mater. Solar Cells **53**(1–2), 29–54 (1998)
19. T. Oda, S. Tanaka, S. Hayase, Differences in characteristics of dye-sensitized solar cells containing acetonitrile and ionic liquid-based electrolytes studied using a novel model. Solar Energy Mater. Solar Cells **90**(16), 2696–2709 (2006)
20. J.E. Hu, J.C. Chou, Y.H. Liao, S.W. Chuang, S.H. Huang, X.Z. Lin, C.Y. Liu, T.Y. Cheng, H.T. Chou, Analysis of different thicknesses of electrolyte applied in flexible dye-sensitized solar cells, in *Lecture Note in Engineering and Computer Science: Proceedings of the World Congress Engineering 2013*, London, UK, 3–5 July 2013, pp. 1014–1018

Fabrication of Real-Time Wireless Sensing System for Flexible Glucose Biosensor

Jie-Ting Chen, Jung-Chuan Chou, Yi-Hung Liao, Hsueh-Tao Chou, Chin-Yi Lin and Jia-Liang Chen

Abstract In this study, the wireless sensor network (WSN) with Zigbee technique is integrated with the flexible glucose biosensor. The wireless sensing system is accomplished by the graphical language laboratory virtual instrumentation engineering workbench (LabVIEW). The wireless sensing system can be classified into two parts, which are the glucose detection system of front end and the transmission platform of back end. According to the experiment results, wireless sensing system can operate successfully on potentiometric sensor.

Keywords Glucose biosensor · LabVIEW · Potentiometric sensor · Wireless sensing system · Wireless sensor network · Zigbee

J.-T. Chen · J.-C. Chou (✉) · H.-T. Chou · C.-Y. Lin
Graduate School of Electronic and Optoelectronic Engineering, National Yunlin University of Science and Technology, Douliou 64002, Taiwan, R.O.C
e-mail: choujc@yuntech.edu.tw

J.-T. Chen
e-mail: m10013337@yuntech.edu.tw

H.-T. Chou
e-mail: chouht@yuntech.edu.tw

C.-Y. Lin
e-mail: m10113316@yuntech.edu.tw

Y.-H. Liao
Department of Information Management, TransWorld University, Douliou 64002, Taiwan, R.O.C
e-mail: liaoih@twu.edu.tw

J.-L. Chen
Department of Electronic Engineering, National Yunlin University of Science and Technology, Douliou 64002, Taiwan, R.O.C
e-mail: u9913022@yuntech.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_31,

1 Introduction

The wireless sensor network (WSN) is consisted of network nodes with sensor that designed to communicate via wireless radio. The recent development of WSN provides the advantages of low cost, low power consumption, small size, flexibility and distributed intelligence that compared with wired ones [1]. The sensors are combined with the WSN that has been widely used in various applications. The applications of WSN techniques have been proposed in the healthcare of patient monitoring [2, 3]. The monitored signals include heart rate (HR) electrocardiogram (ECG), blood glucose, and activity for ambulatory health monitoring. Environment monitoring [4, 5] has become an important area of management and protection that is provided real-time system and control communication from WSN. In literature [6], a wearable healthcare system is integrated with WSN for detecting falls of an elder person, that the healthcare system can reduce the cost of medical care, and improves primary care services. About the application of intelligent life is proposed in literature [7], this literature develops a smart medication system which utilizes the WSN techniques. The functions of medication system are medication reminding, pill-dispensing assisting and medication recording that help the patients with chronic diseases.

In 1970, Bergveld [8] presented a chemical sensor, ion sensitive field effect transistor (ISFET), that was fabricated by semiconducting process and electrochemistry technique. The physical difference in the ISFET structure replaced the metal gate of the metal-oxide-semiconductor field effect transistor (MOSFET) by the series combination of the reference electrode, electrolyte and chemical sensitive insulator or membrane [9]. Afterward, Spiegel et al. [10] proposed the extended gate ions sensitive field effect transistor (EGISFET) in 1983. The EGISFET was improved to become separative extended gate field effect transistor (SEGFET) [11]. The SEGFET structure only needs to change sensing electrode, and the MOSFET device of that can be used repeatedly. The structure is shown in Fig. 1 [11]. The SEGFET holds the advantages of small size and fast response time. An extended metal wire is used as the connection between metal gate and field effect transistor (FET), and the sensing film is deposited on the metal gate area to measure the various detections of environment.

Some advantages of the WSN are presented a low cost technique for collecting detection signals, and increase the space expansion. Consequently, the WSN for homecare, healthcare, and environmental monitoring in our life is required. The proposed system includes the glucose detection system and transmission platform. We achieve the WSN that is integrated with the flexible glucose biosensor, and detect in different concentrations of glucose solutions. Subsequently, the detection signals are communicated via Zigbee module to nearby computer. According to the experimental results, wireless sensing system can be operated successfully and replaces the wired devices.

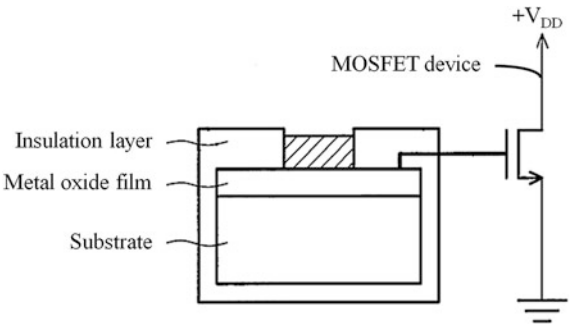


Fig. 1 Structure of the SEGFET [11]

2 Experimental

2.1 Glucose Detection System

Potentiometric electrochemical method is used to measure the output signal of the potential difference between reference electrode and the flexible glucose biosensor. This glucose detection system comprises: test solutions in the container; a silver/silver chloride (Ag/AgCl) reference electrode is providing stable reference potential in test solutions; a flexible glucose biosensor; a readout circuit device is amplifying the detection signals.

The flexible glucose biosensor is imitated the structure of SEGFET and a 99.99 % purity ruthenium metal target via radio frequency (R.F.) sputtering is used, which deposits ruthenium dioxide (RuO_2) thin film on PET substrate. The screen printing technique produces conductive wire and insulation layer. The insulating layer has an aperture for exposing a sensing window on the flexible glucose biosensor and the area of aperture is 3 × 3 mm. First, the 0.1 M potassium phosphate dibasic (K_2HPO_4) and 0.1 M potassium phosphate monobasic (KH_2PO_4) are mixed in distilled water to obtain 0.1 M phosphate buffer saline at pH 7. The glucose oxidase (GOX) powder of 3 mg is premixed with phosphate buffer saline of 5 ml as glucose oxidase solution, and then the 5 wt% Nafion solution and glucose oxidase solution are mixed by the chemical solution method, after compositing the solution of 3 μl is dropped on sensing window. The glucose sensing membrane is prepared by an optimal mixed ratio of 3:4 (vol%) with Nafion and glucose oxidase solution [12]. After the enzyme immobilization, the flexible glucose biosensor is stored at 4 °C in a refrigerator for 12 h. The cross-sectional of flexible glucose biosensor is shown in Fig. 2 [13].

The commercial instrumentation amplifier (LT1167 CN8, Linear Technology Corp., U.S.A.) is regarded to the appropriate component of readout circuit device. The LT1167 is a low power, precision instrumentation amplifier that requires only one external resistor to set gains of 1–10,000. The block diagram of instrumentation amplifier is shown in Fig. 3 that compares with Fig. 4, the descriptions of input and output pins are as follows. The second pin (−IN) of LT1167 is

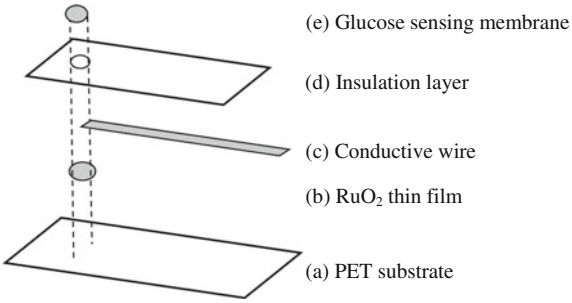


Fig. 2 Flexible glucose biosensor cross-sectional view [13]

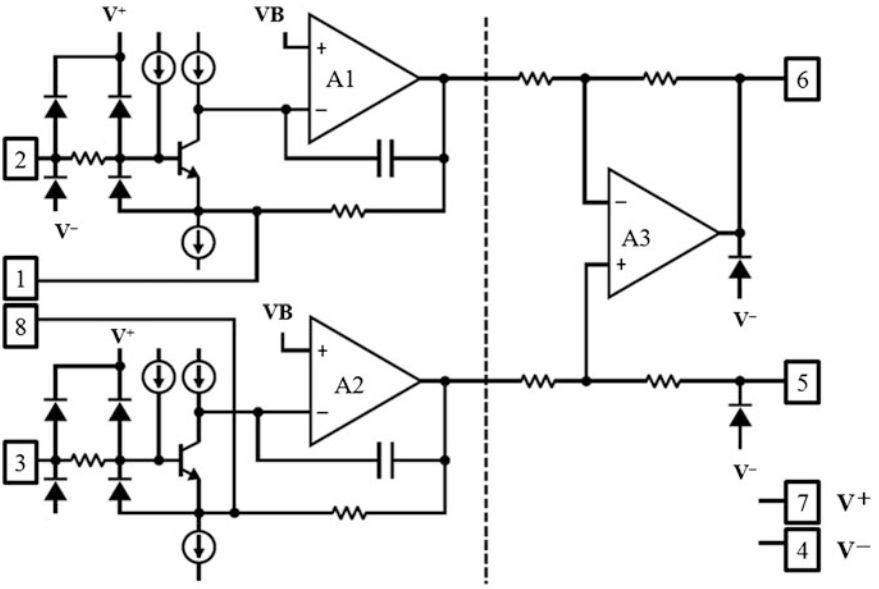


Fig. 3 Block diagram of instrumentation amplifier

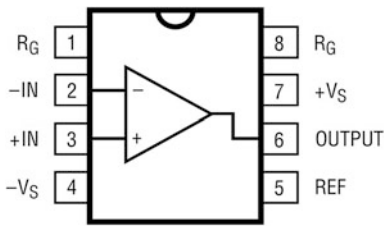


Fig. 4 Top view of the instrumentation amplifier

connected with flexible glucose biosensor. The reference electrode is connected to ground at the third pin (+IN). The fifth pin (REF) is grounded. Furthermore, the detection signal is calculated by instrumentation amplifier and transmitted to the measurement node at the sixth pin (OUTPUT). The power supply provides ±5 V at the seventh pin ($+V_S$) and the fourth pin ($-V_S$), respectively. The others are not connected.

2.2 Transmission Platform

The popular near field communications are such as Zigbee, Bluetooth, and Wi-Fi. Zigbee is best suited for periodic measurement, intermittent data or signal transmission from a sensor or a device. This aspect makes Zigbee very useful for monitoring and management. The transmission platform is consisted of wireless measurement devices and graphical language laboratory virtual instrumentation engineering workbench (LabVIEW). In this study, the wireless measurement devices use National Instruments (NI) WSN system of Zigbee module to transmit the detection signals. Wireless measurement devices consist of measurement node (Model: NI WSN-3202, National Instruments Corp., U.S.A.), and a gateway (Model: NI WSN-9791, National Instruments Corp., U.S.A.). The measurement node is directly connected via 2.4 GHz radio transmitted signals to the gateway, and it installed with four 1.5 V AA alkaline battery cells. The measurement node offers four analog input channels and four digital Input/Output channels. The gateway must be connected to a computer that running graphical language LabVIEW (Model: LabVIEW 2011, National Instruments Corp., U.S.A.). The front panel of graphical language LabVIEW can display detection signals and then process, analyze, and store. The schematic diagram of wireless sensing system is shown in Fig. 5. The diagram of actual installation of wireless sensing system is shown in Fig. 6. The indoor transmission distance of a single measurement node is about 10–15 m.

2.3 Description of Graphical Language

The graphical language LabVIEW is a kind of program language that is a method of graphic design to replace the traditional text program function. This study utilizes the graphical language LabVIEW to implement the real-time wireless sensing system, and the main functions are described below. Before running graphical language LabVIEW, the wireless measurement devices should be installed.

At first, we describe the gateway operation in program. As shown in Fig. 7, the 'WSN Open Gateway' creates a reference to the gateway. User can confirm the gateway internet protocol (IP) address is correct, and then the 'WSN Discover All

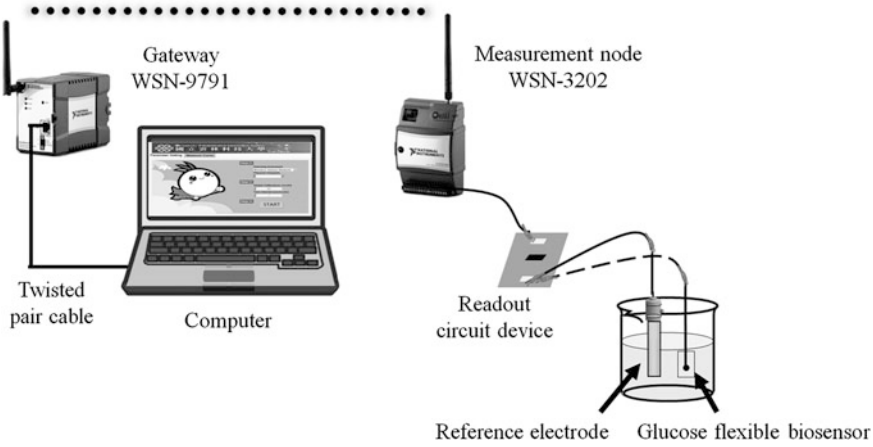


Fig. 5 Schematic diagram of the wireless sensing system

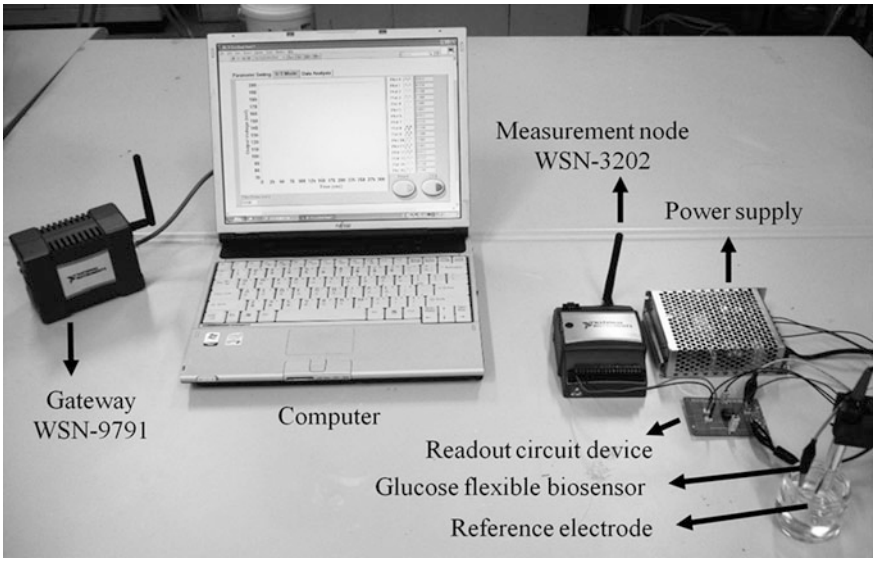


Fig. 6 Actual installation of the wireless sensing system

Nodes' is started to search the measurement nodes via the Zigbee module from the gateway. 'WSN Get Node Info.' is returned information about the specified measurement node. The information of specified measurement node are serial number, wireless identification (ID), measurement node type, status of the firmware update, current version of the firmware, state of the battery, network link quality, power supply type, and network mode. The state of the battery is shown no

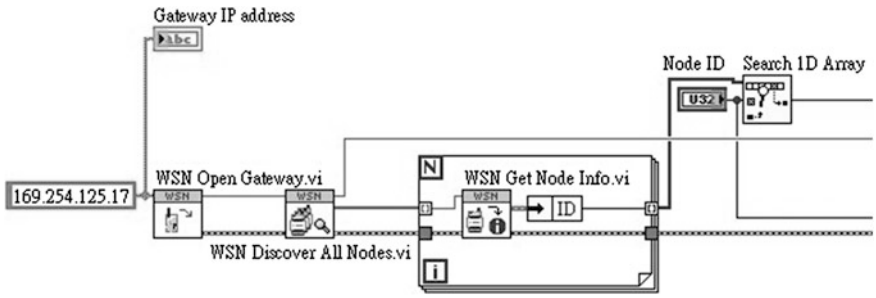


Fig. 7 Program framework of the gateway interface

signal, poor, fair, good, or excellent. Finally, the program will set parameter and scan the amount of measurement node.

If the measurement node is working normally that should create the shared variable node. The shared variable node represents to transmit information between different virtual instruments. As shown in Fig. 8, the analog inputs (AI0–AI3) are the shared variable nodes relative to the four analog channels of measurement node. The 'WSN Open Node Reference' creates a reference to the specified measurement node, and then 'WSN Close Node' closes the reference to the specified measurement node. At the same time, input signal of the potential difference through shared variable node can be read. The detection signals of the potential difference will output to the next step.

Next all functions are processed the detection signals. The descriptive programs of storage and display framework are shown in Fig. 9. In Fig. 9a, the function is referred to select saving path. In addition, after terminating the measurement, the program will automatically save two files: an excel files of measured data (*.xls) and a graph (*.png). In Fig. 9b, the function of 'Waveform Chart' can display the result of detection signals with real-time which will be written into selected saving path file. The saving path is retained in the beginning, which provides a function to auto-save the completed measurement curve for the '*.png' formation. Chapter Prediction of Thermal Deformation for a Ball Screw System Under Composite Operating Conditions is parts of extended description of literatures [14].

3 Results and Discussion

In this study, we build the glucose detection system to detect glucose values, and integrated with the wireless sensor network, transmitted signals with real-time and displayed the results. The flexible glucose biosensor is based on RuO_2/PET. The glucose detection system is measured in glucose solutions (100, 200, 300, and 400 mg/dL). The environment temperature is controlled at room temperature (25 °C). The immovable measurement time is set 180 s. If the measurement time needs to be changed, the parameter settings can be changed in the user interface.

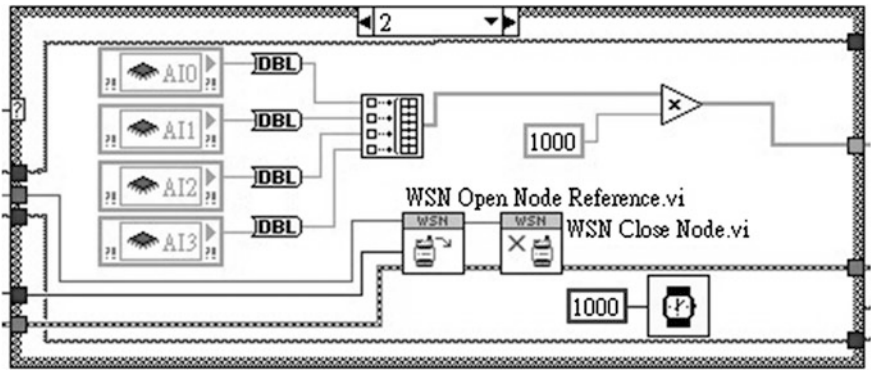


Fig. 8 Program framework of the measurement node interface

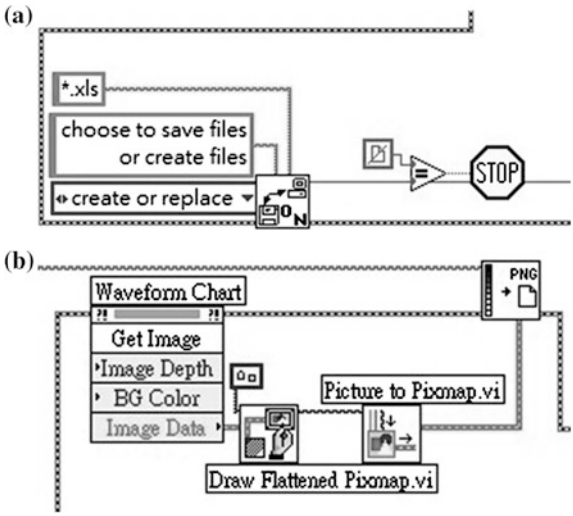


Fig. 9 **a** Program framework of the storage interface. **b** Program framework of the display interface

After operating the wireless sensing system, we can analyze the characteristic of flexible glucose biosensor. Figure 10 shows that the average sensitivity of flexible glucose biosensor is 0.192 mV(mg/dL)$^{-1}$ and the linearity is 0.975. The linear range is between 100 and 400 mg/dL. The average response voltage with 100, 200, 300 and 400 mg/dL are about −273.9, −248.8, −219.4, and −204.0 mV, respectively. Then the error bars of response voltage with 100, 200, 300 and 400 mg/dL are 6.65, 0.57, 5.81, and 4.20 mV, respectively. The measurement results show that the proposed wireless sensing system is used successfully to measure glucose solutions and stable measurement. The measured results of flexible glucose biosensor are compared with other literatures as shown in Table 1. Furthermore, we list the different applications of the biosensor and compare the WSN technique and analytical apparatus, as shown in Table 2.

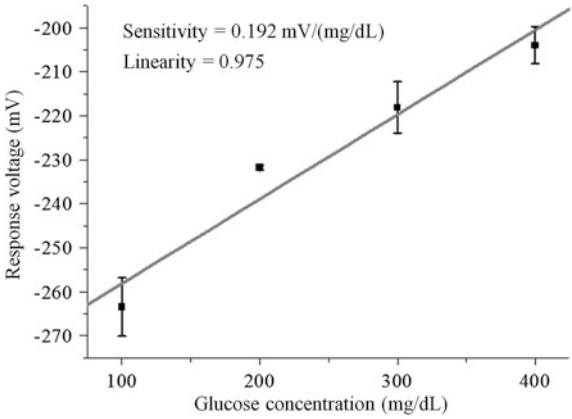


Fig. 10 The measurement results of the flexible glucose biosensor

Table 1 Glucose biosensor in this study is compared with other literatures [15, 16]

Sensing membrane/Substrate	Sensitivity $(mV(mg/dL)^{-1})$	Linearity	Linear range (mg/dL)	Reference
RuO_2/PET	0.192	0.975	100–400	In this study
RuO_2/Silicon	0.018	0.964	100–500	Chou and Yang [15]
Si_3N_4/Silicon nitride	0.050	0.989	0–1,800	Lee et al. [16]

Table 2 WSN system in this study is compared with other literatures [17, 18]

Method of transmission	A/D resolution	Analytical apparatus	Application	Reference
Zigbee, Twisted pair	16 bits	LabVIEW	pH, glucose	In this study
Bluetooth, RS-232	10 bits	LabVIEW	pH, potassium, sodium, chloride	Cheng et al. [17]
2.4 GHz wireless transceiver, RS-232	8 bits	ASCII	pH, temperature, chlorine	Chung et al. [18]

Different sensing membrane and characteristics of glucose biosensors are compared with other literatures [15, 16]. Chou et al. [15] reported that the average sensitivity and linearity of ruthenium oxide based glucose biosensor are 0.018 $mV(mg/dL)^{-1}$ and 0.964, respectively, this literature has bad sensing characteristics because the ruthenium metal thin film has smooth surface, which is not easy to adhere enzyme. Lee et al. [16] propose that the glucose biosensor is fabricated by utilizing silicon nitride (Si_3N_4)-based on ISFET, although the glucose biosensor has a wide detecting range (0–1,800 mg/dL), but it has lower sensitivity (0.05 $mV(mg/dL)^{-1}$). Besides, the structure of ISFET has the disadvantages such

as high cost, complicated process, not easy to package, etc. that compared with the structure of SEGFET.

Different WSN techniques and sensor analytical apparatus are presented in literatures [17, 18]. We compare the properties of literatures [17, 18] and are shown in Table 2. The WSN system of literature [17] is based on graphical language LabVIEW with the Bluetooth wireless technique for handheld devices. But the measurement system is limited to extend more sensor devices. In this study, the measurement node has four channels, and the 16 bits A/D resolution of NI WSN system can process more signal transfer quickly. However, a measurement node offers four analog input channels. If we expect to extend more channels to transmit analog signals, we just needed to add another measurement node to install in NI WSN system. In addition, the analytical apparatus of graphical language LabVIEW is the simpler and more flexible than literature [18]. The graphical language LabVIEW is easily customized and controlled with different sensor devices.

The reproducibility and storage stability of the proposed flexible glucose biosensor have been studied [19]. The lifetime of flexible glucose biosensor is affected by the number of use, acid or alkaline environment in test solution, and temperature in test solution or the preservation of environment when flexible glucose biosensor is not in use. The research of lifetime of flexible glucose biosensor is measured in different concentrations of glucose solution from 100 to 400 mg/dL per 7 days. After using, the flexible glucose biosensor is preserved in 4 °C and the measurement time is 30 days. Figure 11 is shown that the sensing characteristic of sensitivity of flexible glucose biosensor is kept about 78.0 % after 30 days. The analysis of measured results of flexible glucose biosensor is shown in Table 3.

The main objective of reproducibility and stability is the glucose biosensor to be used repeatedly over a long period. The proposed glucose biosensor has been studied in literature [20]. The glucose biosensor is stored dry at 4 °C and measured at intervals of 1 week, and it remained about 75 % of the original sensitivity after 5 weeks. The literature [21] has evaluated the stability of the glucose biosensor, which is stored at 4 °C. The storage stability of the glucose biosensor is tested by monitoring the response currents in 0.2 mM glucose concentration over 20 days. The optimum experimental result is the activity of the electrode remained about 83 % that contrast the initial current response, after the storage periods of 10 days.

In literature [22], the long-term stability of the glucose biosensor is studied by amperometric detection of 5 mM glucose solution every 2 or 3 days over a month. The glucose biosensor retains about 72.4 % of its initial response after 30 days. Although the maintainable proportion of glucose biosensor of literature [21] has been retained 83.0 % for 10 days, further the characteristic of glucose biosensor should be shown over 10 days whether its keeps superior proportion. We compare other literatures to demonstrate the proposed flexible glucose biosensor has excellent stability for long-term use.

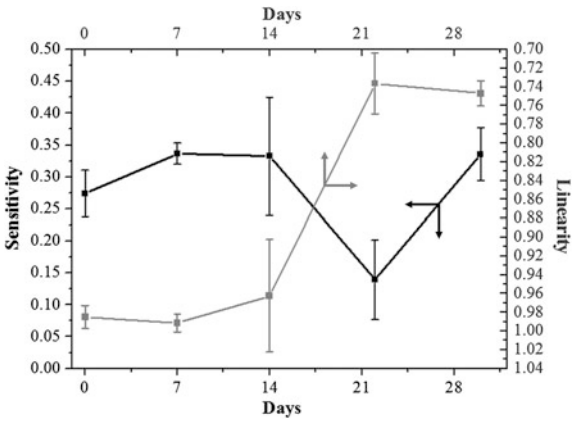


Fig. 11 Lifetime measurement of the flexible glucose biosensor

Table 3 Lifetime of flexible glucose biosensor in this study is compared with other literatures

Sensing membrane	Measurement method	Measurement range	Lifetime (day)/ maintainable proportion (%)	Reference
RuO_2	Potentiometric	100–400 mg/dL	30/78.0	In this study
Gold nanoparticles biocomposite	Amperometric	5.0 μM–2.4 mM	35/75.0	Luo et al. [20]
Silver nanowire	Amperometric	10.0 μM–0.8 mM	10/83.0	Wang et al. [21]
Wood ceramics	Amperometric	0.5–7.0 mM	30/72.4	Qian et al. [22]

4 Conclusion and Outlook

In this study, the wireless sensing system has been presented successfully for detecting glucose values. The system is integrated with glucose detection system and transmission platform. The wireless sensing system provides a real-time monitoring and rapid detection, and the detection range is from 100 to 400 mg/dL with average sensitivity is 0.192 mV(mg/dL)$^{-1}$, and linearity is 0.975. Moreover, the wireless sensing system is designed by using graphical language LabVIEW which can design different functions according to user's needs. The wireless sensing system receives the detection signals and analyzes the characteristics of flexible glucose biosensor.

The various potentiometric sensors or biosensors such as calcium ion, chlorine ion, potassium ion, sodium ion etc., are suitable for combining with the wireless sensing system in the application of future, which widely use in various applications and construct a healthcare system, monitoring function and offers the convenience of living. Wireless sensing system is a promising direction to develop for applications of portable mobile measurement, healthcare and homecare.

Acknowledgments This study has been supported by National Science Council, Republic of China, under the contracts NSC 100-2221-E-224-017, NSC 101-2221-E-224-046, NSC 101-2221-E-265-001, and NSC102-2221-E-224-075.

References

1. L.R Garcia, P. Barreiro, J.I. Robla, Performance of ZigBee-based wireless sensor nodes for real-time monitoring of fruit logistics. Int. J. Food Eng. **87**(3), 405–415 (2008)
2. H.J. Lee, S.H. Lee, K.S. Ha, H.C. Jang, W.Y. Chung, J.Y. Kim, Y.S. Chang, D.H. Yoo, Ubiquitous healthcare service using Zigbee and mobile phone for elderly patients. Int. J. Med. Inform. **78**(3), 193–198 (2009)
3. A. Milenkovic, C. Otto, E. Jovanov, Wireless sensor networks for personal health monitoring: issues and an implementation. Comput. Commun. **29**(13–14), 2521–2533 (2006)
4. M.F. Othman, K. Shazali, Wireless sensor network applications: a study in environment monitoring system, in *International Symposium on Robotics and Intelligent Sensors*, Sarawak, Malaysia, pp. 1204–1210 (2012)
5. W.S. Jang, W.M. Healy, M.J. Skibniewski, Wireless sensor networks as part of a web-based building environmental monitoring system. Autom. Constr. **17**(6), 729–736 (2008)
6. R. Paoli, F.J. Fernández-Luque, G. Doménech, F. Martínez, J. Zapata, R. Ruiz, A system for ubiquitous fall monitoring at home via a wireless sensor network and a wearable mote. Expert Syst. Appl. **39**(5), 5566–5575 (2012)
7. W.W. Chang, T.J. Sung, H.W. Huang, W.C. Hsu, C.W. Kuo, J.J. Chang, Y.T. Hou, Y.C. Lan, W.C. Kuo, Y.Y. Lin, Y.J. Yang, A smart medication system using wireless sensor network technologies. Sens. Actuators, A **172**(1), 315–321 (2011)
8. P. Bergveld, Development of an ion-sensitive solid-state device for neurophysiological measurements. IEEE Trans. Biomed. Eng. BME **17**(1), 70–71 (1970)
9. S. Swaminathan, S.M. Krishnan, L.W. Khiang, Z. Ahamed, G. Chiang, Microsensor characterization in an integrated blood gas measurement system, in *Proceedings of IEEE Asia Pacific Conference on Circuits and Systems*, Singapore, pp. 15–20 (2002)
10. J. Van Der Spiegel, I. Lauks, P. Chan, D. Babic, The extended gate chemical sensitive field effect transistor as multi-species microprobe. Sens. Actuators B Chem. **4**, 291–298 (1983)
11. J.C. Chou, J.M. Chen, An equivalent circuit model for simulating the separative extended gate field effect transistor. Sens. Lett. **6**(6), 924–928 (2008)
12. J.C. Chou, T.Y. Cheng, G.C. Ye, Y.H. Liao, S.Y. Yang, H.T. Chou, Fabrication and investigation of arrayed glucose biosensor based on microfluidic framework. IEEE Sens. J. **13**(11), 4180–4187 (2013)
13. J.C. Chou, W.C. Chen, C.C. Chen, Flexible sensor array with programmable measurement system, in *Proceedings of International Conference on Chemical and Biomolecular Engineering*, Tokyo, Japan, pp. 340–344 (2009)
14. J.T. Chen, J.C. Chou, Y.H. Liao, H.T. Chou, C.Y. Lin, J.L. Chen, Integration of the real-time remote wireless sensing system for glucose flexible biosensor, in *Proceedings of The World Congress on Engineering 2013, WCE 2013*, 3–5 July 2013. Lecture Notes in Engineering and Computer Science, London, UK (2013), pp. 1097–1101
15. J.C. Chou, H.Y. Yang, Potentiometric glucose biosensor based on ruthenium-modified RuO_2/Si sensing electrod, in *The 8th Asian Conference on Chemical Sensors*, Daegu, Korea, pp. 11–14 (2009)
16. S.R. Lee, K. Sawada, H. Takao, M. Ishida, An enhanced glucose biosensor using charge transfer techniques. Biosens. Bioelectron. **24**(4), 650–656 (2008)
17. J.F. Cheng, J.C. Chou, T.P. Sun, S.K. Hsiung, H.L. Kao, Study on a multi-ions sensing system for monitoring of blood electrolytes with wireless home-care system. IEEE Sens. J. **12**(5), 967–977 (2012)

18. W.Y. Chung, C.L. Chen, J. B. Chen, Design and implementation of low power wireless sensor system for water quality monitoring, in *Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering*, Wuhan, China, pp. 1–4 (2011)
19. Y.L. Tasi, Fabrication of arrayed flexible screen-printed glucose biosensor based on microfluidic framework, Report for Practical Project, Department of Electronic Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan, pp. 51–52 (2013)
20. X.L. Luo, J.J. Xu, Y. Du, H.Y. Chen, A glucose biosensor based on chitosan–glucose oxidase–gold nanoparticles biocomposite formed by one-step electrodeposition. Anal. Biochem. **344**(2), 284–289 (2004)
21. L. Wang, X. Gao, L. Jin, Qi Wu, Z. Chen, X. Lin, Amperometric glucose biosensor based on silver nanowires and glucose oxidase. Sens. Actuators B Chem. **176**, 9–14 (2013)
22. J.M. Qian, A.L. Suo, Y. Yao, Z.H. Jin, Polyelectrolyte-stabilized glucose biosensor based on wood ceramics as electrode. Clin. Biochem. **37**(2), 155–161 (2004)

Gate-Passing Detection Method Using WiFi and Accelerometer

Katsuhiko Kaji and Nobuo Kawaguchi

Abstract Gate-passing information is useful for daily activity recording. We propose a gate-passing detection method using WiFi and accelerometer. Since doors divide such physical areas as rooms and hallways, the WiFi environments tend to greatly vary. A gate should exist when the points in the WiFi environments are significantly different. We define such points as WiFi significant points and propose a detection method based on a WiFi propagation model and estimated moving distance according to an accelerometer. We evaluated our proposed method and found out that most door passings can be detected. We also found that we can estimate the existence of doors that have identical door passings with a high degree of accuracy. Furthermore, we propose a cumulative error correction method of pedestrian dead-reckoning based on our proposed method as an application.

Keywords Accelerometer · Activity recognition · Cumulative error correction of personal dead-reckoning · Gate passing detection · Signal propagation model · WiFi significant point

K. Kaji (✉)
Graduate School of Engineering, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
e-mail: kaji@nuee.nagoya-u.ac.jp

N. Kawaguchi
Graduate School of Engineering, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
e-mail: kawaguti@nagoya-u.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_32,

1 Introduction

Gate-passings, which refer to the entrances and exits to a building or a room and going by a corridor, are crucial information for indoor location-based services, especially for monitoring user activities, recognizing user migration pathways, and lifelogs.

Traditional gate-passing detection methods suffer from the following problems. The most general gate-passing detection method is IC card readers or RF tag readers attached to gates. In such situations, users touch the readers with their cards. The vision-based approach detects a gate [1]. Its door is extracted from the images captured by the camera attached to a robot or a user. The restrictions of camera locations burden general users. Another method uses proximity sensors [2], although general mobile terminals don't have them.

In this paper, we propose a gate-passing detection method [3]. We assume that users have general smartphones. In our method, we use WiFi signal information for gate detection and estimate the moving distance by accelerometers with which most smartphones are equipped. WiFi access points (APs) must be placed in the environment, even though many APs have already been placed in public buildings, universities, and offices.

The following is the outline of our proposed method. Since WiFi signal strength tends to be cut off or reduced by such gates as doors, we assume gates in a location where the WiFi environment greatly varies. To acquire the degree of variation of WiFi environments, we introduce and compare two kinds of moving distances that are based on WiFi and accelerometers. If the WiFi-based distance deviate from the accelerometer-based distance, we assume that the user is passing a gate.

2 Proposed Method

Many objects divide spaces, such as doors, elevators, and walls. Such objects tend to cut off or weaken WiFi signal strength. The degree of decay depends on the object's material and the physical relationship between the object and the AP. However, in many cases, WiFi environments separated by objects tend to be very different.

Figure 1 shows an example where a WiFi environment is different because it is separated by a door. If the user passes it, the WiFi environment changes. We assume that if the WiFi environment greatly varies, the user is passing a gate such as a door.

In this paper, we define a location where WiFi environments are separated by significantly different locations as a WiFi significant point. We assume a situation where users have standard smartphones and walk around indoors. Our method requires two kinds of moving distances. One is accelerometer-based step estimation, and the other is the distance based on the variation of WiFi signal strengths

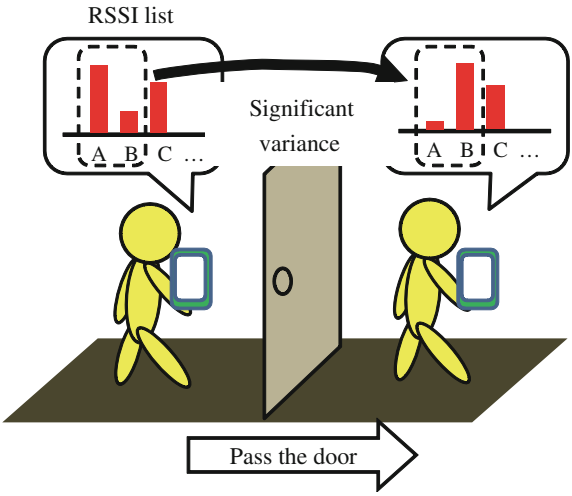


Fig. 1 WiFi environment variation by passing a door

and the signal propagation model. If the latter distance deviate from the former distance, the method judges that the user has passed a WiFi significant point.

One typical signal propagation model is the Seidel model [4], which represents the relationship between the distance to an AP and its received signal strength indication (RSSI). With the model, we can estimate the distance to AP using RSSI.

2.1 Fundamental Algorithm for Extracting WiFi Significant Points

First, we formulate the WiFi significant point extraction algorithm by defining a simple environment. It has only one AP whose location is unknown. Users walk around it freely, thus the trajectory is not necessarily linear. From WiFi and accelerometer information, two kinds of distances are estimated.

One distance is user minimum moving distance d_{min} that is estimated by the WiFi information. When RSSI r_{t1} at time t_1 changes to r_{t2} at time t_2, the minimum moving distance is represented as the following formula using WiFi propagation model f:

$$d_{min} = |f(r_{t1}) - f(r_{t2})|. \tag{1}$$

Figure 2 shows several possible trajectory examples where RSSI is changed from −30 to −40 dBm. When the user linearly moves away from the AP, the lengths of the distance of trajectories must be the shortest. The length is calculated as $|f(-30\ \text{dBm}) - f(-40\ \text{dBm})|$. If the user passed the WiFi significant point, estimated minimum distance d_{min} should be larger than the actual walking distance.

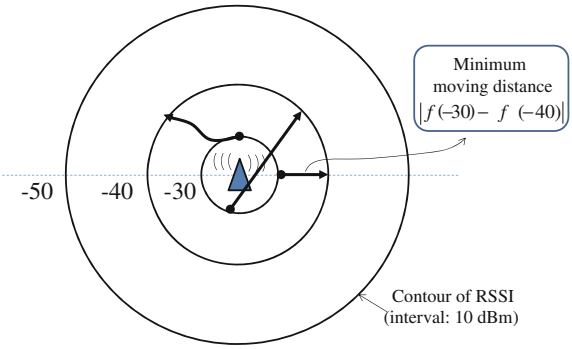


Fig. 2 Possible variation of user trajectory when RSSI varies from −30 to −40 dBm

The other distance, which is maximum moving distance d_{max} between times t_1 and t_2, is estimated by an accelerometer. Walking steps can be extracted by capturing the periodical local maximum and local minimum values of an accelerometer. Each step's distance is estimated using the user's height and the local maximum and local minimum values. Here, if the user is walking linearly, the distance is the sum of each step's distance. If the user isn't walking linearly, the distance between the user's positions at t_1 and t_2 must be shorter than the distance of the linear walking. The sum of the walking distance must be the maximum distance.

Based on d_{min} and d_{max}, we estimate whether the user passed the WiFi significant point during t_1 and t_2. The algorithm is shown in Fig. 3. If d_{max} exceeds d_{min}, the actual distance range can be estimated (Fig. 3, top). On the other hand, if the WiFi environment varies significantly during t_1 and t_2, d_{min} should be larger than the actual walking distance, and d_{min} is probably larger than d_{max} (Fig. 3, bottom). In such situations, we judge that value d_{min} is not reasonable. Consequently, we consider that the user passed the WiFi significant point during t_1 to t_2.

2.2 Extending Our Proposed Method for Real Environments

We introduce the effect of the fluctuation of RSSI and multiple WiFi information and extend our proposed method for real environments. In the real world, WiFi signals influence multipath fading so that RSSI is not constant. Using the average or median RSSI values that are observed multiple times, the effect of fluctuation can be reduced. We imagine a situation where users aren't standing, so WiFi RSSI cannot be observed multiple times. The effect of fluctuation cannot be ignored. At the same time, we must consider the multiple WiFi information transmitted by multiple APs. Recently, since many APs have been placed in various buildings, we can receive multiple AP signals at a number of locations.

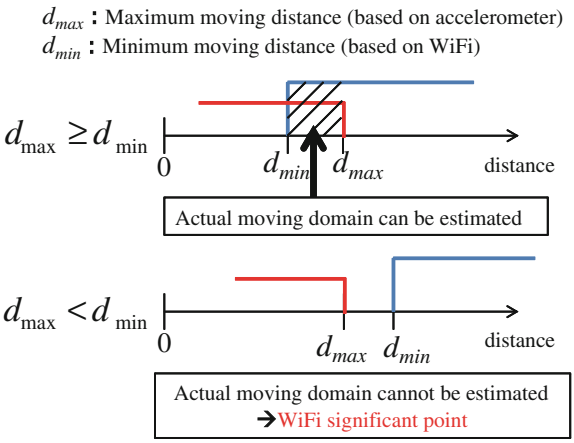


Fig. 3 Fundamental basis of WiFi significant point extraction

2.2.1 Effect of Fluctuation of RSSI

First, we introduce the effect of the fluctuation of RSSI and reconstruct our above scheme as a stochastic model. In this paper, we approximate the fluctuation as a Gaussian distribution. Several researches adopt Gaussian distribution to approximate RSSI fluctuation [5, 6]. We also regard the level of fluctuation as constant. In ideal environments, the distance can be calculated using function f and RSSI r_μ, and the distance is expressed as $f(r_\mu)$. The fluctuation is expressed as a Gaussian whose average is r_μ and the standard deviation is r_σ (Fig. 4, top). At the time, in the ideal environment, when RSSI is observed, distance $r_\mu - r_\sigma$ can be calculated as $f(r_\mu) - f(r_\mu - r_\sigma)$. Using the value, we approximate the distance fluctuation to AP as a Gaussian distribution where the average is $w_\mu = f(r_\mu)$ and the standard deviation is $w_\sigma = f(r_\mu) - f(r_\mu - r_\sigma)$ (Fig. 4, bottom).

Minimum distance d_{min}, which we introduced above, is expressed as a subtraction of Gaussian distributions. Consequently, minimum distance d_{min} is expressed as a Gaussian whose average is $d_{min\_\mu} = w_{\mu 1} - w_{\mu 2}$, and the standard deviation is $d_{min\_\sigma} = \sqrt{w_{\sigma 1} + w_{\sigma 2}}$.

In the Sect. 2.2, the existence probability of WiFi significant points is expressed as binary. On the other hand, by introducing fluctuation, the likelihood based on two kinds of distances d_{max} and d_{min} are expressed as cumulative probability (5 shaded area). The likelihood is calculated as Eq. 2. Here, $erf(x)$ is an error function.

$$p = \frac{1}{2}\left(1 + erf\left(\frac{d_{ref} - d_\mu}{\sqrt{2d_\sigma^2}}\right)\right) \tag{2}$$

The top of Fig. 5 is an example where cumulative probability p is high. In short, the observed RSSI should probably be fluctuated. On the other hand, if p is under threshold $p_{threshold}$ (Fig. 5, bottom), the observed RSSI is unlikely even where the

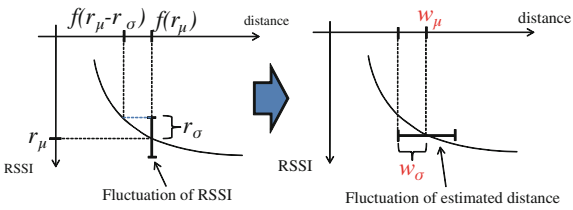


Fig. 4 Conversion from RSSI fluctuation to distance fluctuation. (*Top* Gaussian distribution of RSSI, *Bottom* Gaussian distribution of distance)

fluctuation is concerned. We assume that a WiFi significant point is passed between observation times t_1, t_2.

Based on the fluctuation, a weak RSSI value should not be used to extract WiFi significant points. If the RSSI is weak, the estimated distance to the AP is significantly different if the RSSI value is fluctuated. For example, using the WiFi propagation model from the evaluation section, the distance where the RSSI is -80 dBm is 83 m, and the distance where it is -81 dBm is 91 m. The variance is only 1 dBm, but the difference of the estimated distances is 8 m. Therefore, we use RSSI values that exceed threshold $r_{threshold}$ for WiFi significant point extraction.

2.2.2 Multiple APs' WiFi Information

Next, we introduce multiple AP RSSI information. When the user passes a point where the WiFi environment changes significantly, RSSIs don't always change simultaneously due to the mobile device's sensitivity and the device driver. Thus, the time instants that WiFi significant points are observed don't always match.

To reduce the problem, we aggregate WiFi significant points that come from each AP's RSSI as one WiFi significant point.

Based on the previous section, the existence of WiFi significant points from each RSSI is judged in each observation interval between t and $t + 1$. The WiFi significant points receive votes for their respective intervals. Then the interval that receives the most votes in a window, whose size is w, is deemed to be one WiFi significant point. Figure 6 shows an example of the voting and the aggregation of WiFi significant points. The window size is 4. In the example, four zones are voted as WiFi significant point at first (Fig. 6, top). Then, according to the voting count and window size, they are aggregated as two zones (Fig. 6, bottom). Finally, these two zones are considered as WiFi significant point.

2.3 *Identical Gate-Passing Detection and Passing Direction Estimation*

The aggregated WiFi significant points consist of multiple WiFi significant points from multiple APs' WiFi information. We believe that identical gate-passing detection can be realized using the pattern of the AP's information. The pattern of

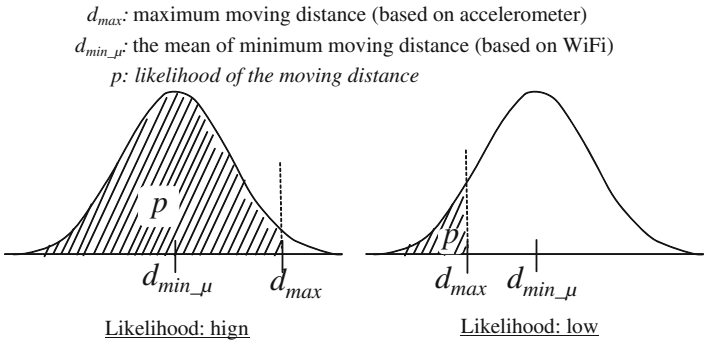


Fig. 5 Distance likelihood

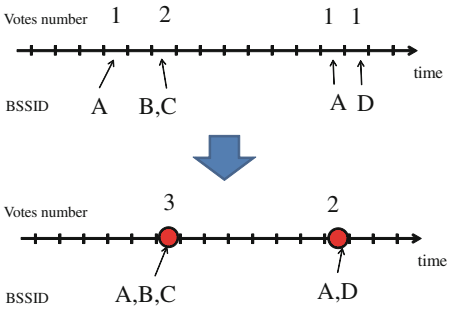


Fig. 6 Voting and aggregation of WiFi significant points. *Top* voting, *bottom* aggregation

*i*th WiFi significant point S_i is expressed as a vector using the number of votes and voted BSSIDs b.

$$S_i = [b_{i,0}, b_{i,1}, \ldots, b_{i,n}] \tag{3}$$

The similarity of two arbitrary WiFi significant points S_i, S_j is calculated using Tanimoto coefficient T [7]:

$$T = \frac{N(S_i \cap S_j)}{N(S_i) + N(S_j) - N(S_i \cap S_j)}. \tag{4}$$

The Tanimoto coefficient is a similarity metric to evaluate two sets. If they are completely identical, T is 1. They don't have a common element, and T is 0. Here, $N(x)$ is the number of elements in x.

When similarity T exceeds similarity threshold $t_{threshold}$, WiFi significant points S_i, S_j are estimated to be the same point, and the user is passing the gate again.

Furthermore, we estimated the passing direction using the pattern of the variance of the RSSIs. For each common BSSID b in S_i and S_j, we checked the variance direction to determine whether RSSI increased or decreased. If the

variance direction is the same, N_{same} is incremented. If the variance direction is different, N_{diff} is incremented. If N_{same} is larger than N_{diff}, the user passed the gate from the same direction, and if N_{diff} are larger than N_{same}, the user passed the gate from a different direction.

2.4 Correction of WiFi Significant Points Using Accelerometers

As above, RSSIs don't always change at the gate-passing moment. Based on our pilot study, the difference of the RSSI change timing and actual gate-passing timing is not zero, and the difference may be about 10 s.

Next we corrected the WiFi significant point with an accelerometer. Generally, when a person passes a gate, the step interval is long, and each step length is short, even though the continuing time of the state is not so long. Based on the heuristics, we developed simple gate-passing timing estimation using an accelerometer. In our method, when the accelerometer's local maximum and minimal are lower than threshold $g_{threshold}$ and the continuing time is lower than $w_{threshold}$, we assume the time zone is a gate passing. Here, $g_{threshold}$ means threshold of gate-acc and $w_{threshold}$ means threshold of gate-passing time.

If the time distance between a WiFi significant point and a gate-passing time is under window size w, the time of the WiFi significant point is corrected to the gate-passing time. When multiple WiFi significant points exist within the window, the nearest WiFi significant point is corrected as the gate-passing time.

Note that our door passing estimation is not very robust. Various situations probably exist where the estimation is not correct. For example, when the environment is crowded, people stand or walk slowly for a short time. The method is probably inaccurate when a person slow down to passes a corridor's corner.

2.5 Restrictions

Our proposed method is very dependent on the physical relationships between gates and APs. Thus, not all gate passings can be detected by our method. If there are no APs around a gate, gate passings cannot be detected. Even if an AP exists around a gate, there are patterns of physical relationships between the AP and the gate where our method cannot extract gate passings. Figure 7 shows two of the examples. In such a situation as the top of Fig. 7, the RSSIs at points A and B are almost the same, so the gate passings cannot be extracted by the RSSI variance. In such a situation as the bottom of Fig. 7, the pattern of the RSSI variance of passing rooms C and D is almost the same. Thus, using our proposed identical gate-passing detection, the doors of the two rooms should be detected as the identical door.

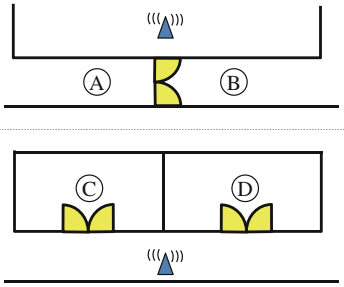


Fig. 7 Examples of situations where it is impossible to apply proposed method

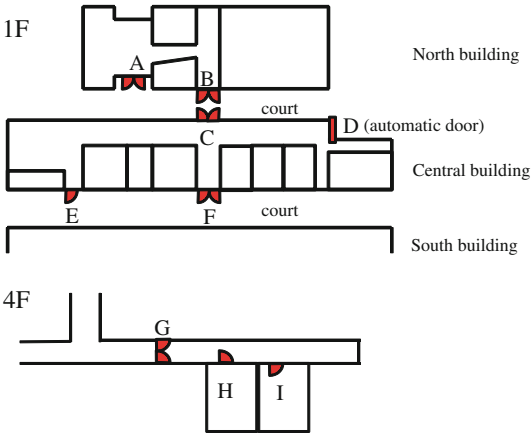


Fig. 8 Door alignment. *Top* 1F, *bottom* 4F

Additionally, there are several restrictions to apply our proposed method. First, the gate must physically divide the environment like doors and elevators. Second, the person himself should open a gate to pass. If the door is already open, the RSSI variance cannot be captured.

3 Experiments

We experimentally evaluated the accuracy of our method using the gate-passing detection method and the identical gate-passing estimation method.

3.1 Experimental Environment

We conducted our experiment on the 1st and 4th floors of the IB Information Buildings on Nagoya University. The door alignment and types are shown in Figs. 8 and 9. There were nine doors in the environment including one automatic door.

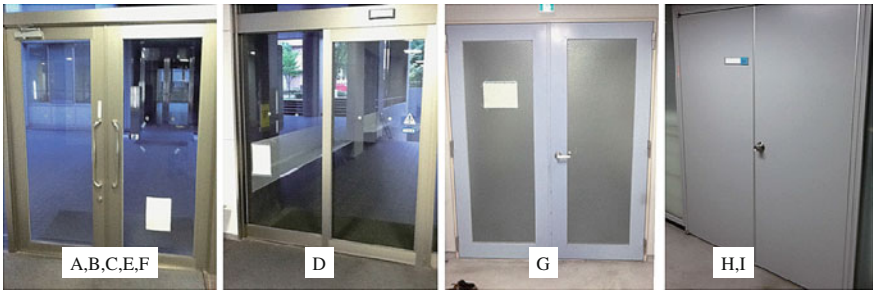


Fig. 9 Door types

Table 1 Overview of experimental data

Sampling rate of WiFi observation	1 Hz
Sampling rate of accelerometer	100 Hz
Number of doors	9
Number of door passings	A–F: 10 times, BG–I: 20 times
Total experimental time	5,300 s

Doors A F are the entrance doors of the buildings, and doors G I are inside the buildings.

Table 1 overviews the observation data. The subject is one of the authors of this paper who used an iPhone3G smartphone. He put it in his waist holder and walked around the experimental environment. His walking speed was not constant; standing and slow walking were included except for door passings. Our proposed method is applicable when users themselves open and close doors, so he opened and closed doors when passing them.

3.2 Settings

We adopted LaMarca's parameter of the Seidel model [8] (Eq. 5).

$$f(r) = -32 - 25\log_{10} r \tag{5}$$

Step length s is calculated by the following formula [9].

$$s = 0.26 \cdot height + (peakdiff - peakavg) \cdot 5.0. \tag{6}$$

Here, *peakdiff* is the difference between the value of the local maximum and the local minimum in each step and *peakavg* means the average value of *peakdiff*. The user's height is *height*. In this experiment, we set the values as $height = 1.80$ m, $peakavg = 1.11$ g (Table 2).

Table 2 Experimental parameters

Fluctuation of RSSI r_σ	2.5 dBm
Threshold of RSSI $r_{threshold}$	−60 dBm
Threshold of likelihood $p_{threshold}$	0.1 %
Threshold of similarity $t_{threshold}$	0.4
Window size w	10 s
Threshold of gate-acc $g_{threshold}$	0.15 G
Threshold of gate-passing time $w_{threshold}$	2.0 s

Table 3 Accuracy of gate-passing detection

Gate-passing detected points	157
Actual gate passings	120
Successful gate-passing detection	92
Precision (%)	59
Recall (%)	76
F-measure (%)	66

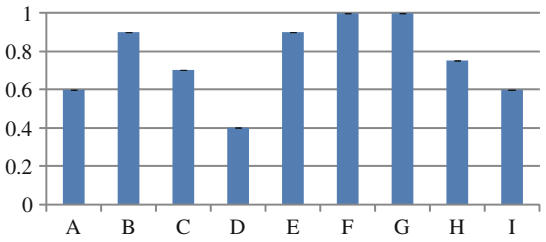


Fig. 10 Gate-passing detection accuracy for individual doors

3.3 Results

3.3.1 Gate-Passing Detection Method

Table 3 shows the result of gate-passing detection. We define correct answers to be when a detected gate passing is within 10 s of the actual door passing. The precision of the gate-passing detection was about 58 %, and the recall was about 76 %. Consequently, our proposed method detected about half of the door passings, but it doesn't always detect them.

Figure 10 shows the accuracy of the gate-passing detection for individual doors. The maximum accuracy is 100 %, and the minimum accuracy is 40.0 %. Based on the results, the accuracy of the gate-passing detection significantly differs by door, even though gate-passing detection is possible when the user passes the door many times.

Automatic doors provide minimum accuracy. When passing automatic doors, the step length around the door isn't shorter than manual doors. This explains why the accuracy of automatic door passing detection is low. Of course, our method is

Table 4 Accuracy of identical gate-passing estimation

WiFi significant points related to door passings	92
Pair of WiFi significant points that hline have identical gates	348 pairs
Pairs of WiFi significant points where hline identical gate detection was correct	245 pairs
Pairs of WiFi significant points where hline they should be estimated as same gate	508 pairs
Precision (%)	70
Recall (%)	48
F-measure (%)	57

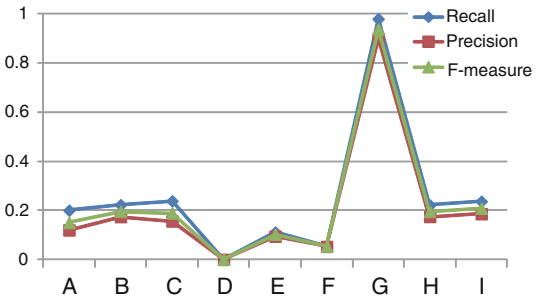


Fig. 11 Accuracy of identical gate-passing detection for each door

influenced by the door's material and the distribution of APs. This is one reason that the accuracy of gate-passing detection widely differs by door.

On the other hand, WiFi significant points were detected except for around the gate. One reason is the existence of WiFi hotspots caused by reflections and multipaths. For example, corridor's corner tends to be WiFi hotspot.

3.3.2 Identical Gate-Passing Estimation

Using successfully detected points (92 points), we evaluated the identical gate-passing estimation. Precision, recall, and F-measure are shown in Table 4. Figure 11 shows the individual door results of the identical gate estimation. The accuracy of door G is obviously higher than the other doors. Door G is the thickest, and one AP is placed near it. Such an ideal environment enhances the accuracy of identical gate-passing estimation.

The number of errors relevant to doors H and I is 43, 19 of which were mistaken for other doors. Doors H and I are located within 3 m of each other, so the pattern of their WiFi environments is similar.

Consequently, the accuracy of identical gate-passing estimation is not as high as gate-passing detection, even though we found doors on which the identical gate-passing estimation method was successfully performed. Therefore, we believe that our method is useful for restrictive situations.

For 245 pairs that were correctly estimated as the same gate, we applied the gate-passing direction estimation method, and the accuracy was 92 %.

Additionally, for door G whose accuracy of identical gate-passing estimation was high, the accuracy of the gate-passing direction estimation was 100 %. Consequently, the gate-passing direction estimation method is generally useful.

4 Related Works

There are several researches on gate detection. Patel proposed a person movement detection method based on air pressure sensors attached to HVAC units [10]. The method captures door openings and closings, door-passing based on the variation patterns of air pressure, although a case might exist where air pressure sensors are difficult to attach to HVAC units due to a building's structure. Moreover, if multiple persons exist, this method cannot track an individual.

GPS-based building entrance/exit detection methods have also been proposed [11]. Generally, GPS signal strength tends to be weak Inside buildings. In our method, with training data that were observed beforehand, we generated a detection model. Therefore, a labor cost problem exists for prior observation. On the other hand, our proposed method needs no preparation.

Hotta proposed a robust room-level location estimation method [12]. When generating WiFi fingerprints, the distance to the nearest door is input. Additionally, they introduced a room transition probability, which is generated using the distance to the nearest door; the probability will be high when the location is near a certain door. Our method doesn't just detect actual door passings; it also enhances the transition probability.

5 Application

We are currently trying to correct cumulative error of personal dead-reckoning (PDR) [13]. By using the proposed identical gate-passing detection method, PDR could be more accurate like Fig. 12.

PDR is relative position tracking method using multiple sensors such as accelerometer and magnetometer equipped in mobile devices. Sensor values contain noise, so that error of estimated position tends to be large when the tracking time duration is long. To solve the problem, a kind of absolute position estimation method should be combined. There are several methods to estimate absolute position by using GPS, RF tags and WiFi, and so on. Though, most of the methods are not so practical indoors. GPS signal cannot arrive indoors well. RF tag reader should be placed for each doors for RF tag based positioning. WiFi fingerprints should be corrected previously in WiFi positioning, so that labor-cost is high [14]. On the other hand, cumulative error correction based on our proposed identified gate-passing detection can be realized at a lower cost. The method don't need to place some kinds of special devices such as RF tag reader for each doors, and don't need any operation previously such as observing WiFi environment.

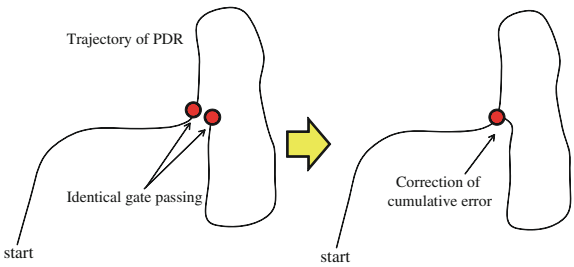


Fig. 12 Cumulative error correction of personal dead-reckoning based on gate identification

We think that the application is very practical for recording person's daily indoor activities such as working in office and live at home in detail. In daily life, person tends to pass same gate frequently. For example, office worker passes same door at the working room twice to go to and return from restroom. Therefore, our method could detect identical gate-passing and can correct trajectory of PDR a number of times on the same date.

6 Conclusion

We proposed a gate-passing detection method based on WiFi significant points. Our method is based on the assumption that WiFi environments, which are divided by gates, tend to be very different. Only WiFi and accelerometer information are used to detect gate passings. We conducted several experiments and found that our proposed method has the ability to detect more than half of the gate passings. Identical gate-passing detection has very low accuracy. However, we found gates whose accuracy of identical gate-passing methods is high.

Currently, we are developing an indoor pedestrian sensing corpus with a balance of gender and age for indoor positioning and floor-plan generation researches (HASC-IPSC) [15]. The corpus contains over one hundred subjects' pedestrian sensing data in certain buildings. As future work, we are going to refine the proposed method to achieve high accuracy by using the corpus. The corpus is almost ready to publish. We consider to publish the corpus for free.

References

1. E. Jauregi, E. Lazkano, B. Sierra, Approaches to door identification for robot navigation, in *Mobile Robots Navigation,* ed. by A. Barrera. InTech, pp. 241–262. 2010 ISBN: 978-953-307-076-6
2. G. Schindler, C. Metzger, T. Starner, A wearable interface for topological mapping and localization in indoor environments, in *Proceedings of Location- and Context-Awareness Second International Workshop* (2006), pp. 64–73

3. K. Kaji, N. Kawaguchi, Gate-passing detection method based on WiFi significant points, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE 2013*, London, pp. 1409–1414, 3–5 July 2013
4. S. Seidel, T. Papport, 914 Mhz path loss prediction model for indoor wireless communications in multifloored buildings, in *Proceedings of IEEE Transactions on Antennas and Propagation* (1992), pp. 207–217
5. B. Ferris, D. Fox, N. Lawrence, WiFi-SLAM using Gaussian process latent variable models, in *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (*IJCAI'07*) (2007), pp. 2480–2485
6. A. Goswami, L.E. Ortiz, S.R. Das, WiGEM: a learning-based approach for indoor localization, in *Proceedings of the Seventh Conference on Emerging Networking Experiments and Technologies (CoNEXT'11)* (2011)
7. T. Segaran, *Programming Collective Intelligence: Building Smart Web 2.0 Applications* (O'Reilly Media, Sebastopol, 2008)
8. A. LaMarca, J. Hightower, I. Smith, S. Consolvo, Self-mapping in 802.11 location systems, in *Proceedings of the Seventh International Conference on Ubiquitous Computing (Ubicomp2005)* (2005), pp. 87–104
9. K. Anzai, S. Okajima, H. Tsubokawa, The estimate of the indoor position that used a smartphone and the suggestion of the walk navigation systems, in *Multimedia, Distributed, Coorporative, and Mobile Symposium (DICOMO2011)* (2011), pp. 921–927
10. S.N. Patel, M.S. Reynolds, G.D. Abowd, Detecting human movement by differential air pressure sensing, in *Pervasive Proceedings of the 6th International Conference on Pervasive Computing* (2008), pp. 1–18
11. E. Katsuda, A. Uchiyama, H. Yamaguchi, T. Higashino, Distinguishing indoor/outdoor using training data of GPS status, in *IPSJ SIG Technical Report 2011-MBL-60(18)* (2011), pp. 1–8
12. S. Hotta, Y. Hada, Y. Yaginuma, A robust room-level localization method based on transition probability for indoor environments, in *Proceedings of International Conference on Indoor Positioning and Indoor Navigation, 2012* (2012)
13. D. Kamisaka, S. Muramatsu, T. Iwamoto, H. Yokoyama, Design and implementation of pedestrian dead reckoning system on a mobile phone. IEICE Trans. Info. Syst. **94**(6), 1137–1146 (2011)
14. K. Kaji, N. Kawaguchi, Design and implementation of WiFi indoor localization based on gaussian mixture model and particle filter, in *The 3rd International Conference on Indoor Positioning and Indoor Navigation (IPIN2012)* (2011), pp. 1–9
15. K. Kaji, H. Watanabe, R. Ban, N. Kawaguchi, HASC-IPSC: indoor pedestrian sensing corpus with a balance of gender and age for indoor positioning and floor-plan generation researches, in *International Workshop on Human Activity Sensing Corpus and its Application (HASCA2013)* (2013), pp. 605–610

Extended Performance Studies of Wi-Fi IEEE 802.11a, b, g Laboratory WPA Point-to-Multipoint and Point-to-Point Links

J. A. R. Pacheco de Carvalho, H. Veiga, C. F. Ribeiro Pacheco and A. D. Reis

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) WPA 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 WPA 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 · WPA point-to-multipoint links

J. A. R. P. de Carvalho (✉) · C. F. R. Pacheco · A. D. Reis
Unidade de Detecção Remota, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
e-mail: pacheco@ubi.pt

C. F. R. Pacheco
e-mail: a17597@ubi.pt

A. D. Reis
e-mail: adreis@ubi.pt

H. Veiga
Centro de Informática, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
e-mail: hveiga@ubi.pt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_33,

1 Introduction

Contactless communication techniques have been developed using mainly electromagnetic waves in several frequency ranges, propagating in the air. Examples of wireless communications technologies are Wi-Fi and FSO, whose importance and utilization have been growing.

Wi-Fi is a microwave based technology providing for versatility, mobility and favourable prices. The importance and utilization of Wi-Fi have been growing for complementing traditional wired networks. It has been used both in ad hoc mode and in infrastructure mode. In this case an access point, AP, permits communications of Wi-Fi devices 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 and point-to-multipoint 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. Wireless communications, wave propagation [2, 3] and practical implementations of WLANs [4] have been studied. Detailed information has been given about the 802.11 architecture, including performance analysis of the effective transfer rate. An optimum factor of 0.42 was presented for 11 Mbps point-to-point links [5]. Wi-Fi (802.11b) performance measurements are available for 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. 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. WPA implements the majority of the IEEE 802.11i standard [1]. It includes a MIC, message integrity check, 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], WPA [11] and WPA2 [12] links, as well as very high speed FSO [13]. In the present work new Wi-Fi (IEEE 802.11a, b, g) results arise, using personal mode WPA, through OSI levels 4 and 7. Performance is evaluated in laboratory measurements of WPA point-to-multipoint links using new available equipments. Comparisons are made to corresponding results obtained for WPA point-to-point (PTP) and Open point-to-multipoint (PTMP) links.

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 WPA point-to-multipoint links using available equipments. Comparisons are made to corresponding results obtained for WPA 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 [14], 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 [15]. 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 [16], to enable PTMP links to the access point. In every type of experiment, interference free communication channels were used (chapter "Identification of Multistorey Building's Thermal Performance Based on Exponential Filtering" for 802.11a; chapter "Performance Evaluation of the Valveless Micropump with Piezoelectric Actuator" for 802.11b, g). This was checked through a portable computer, equipped with a Wi-Fi 802.11a/b/g/n adapter, running NetStumbler

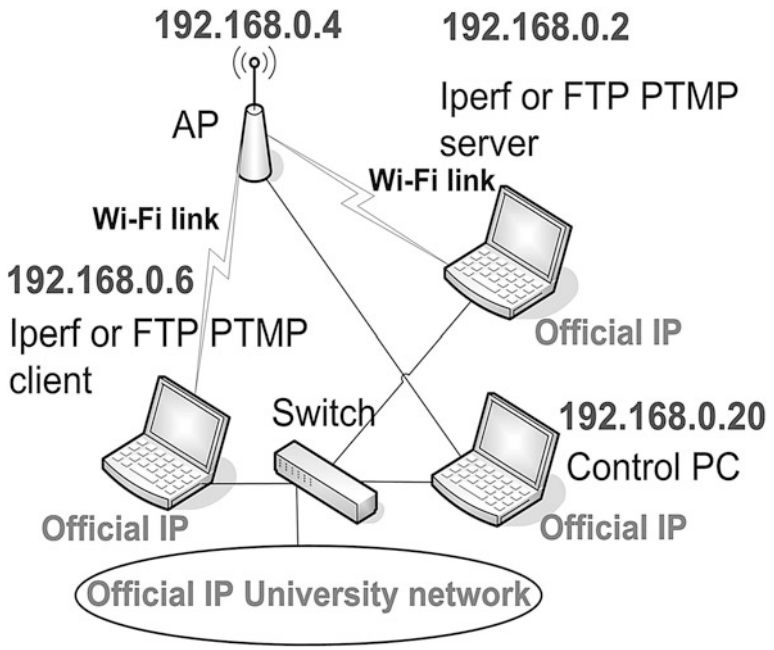


Fig. 1 Experimental laboratory setup scheme

software [17]. WPA encryption was activated in the AP and the wireless adapters of the PCs, using AES and 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 new laboratory setup has been planned and implemented for the PTMP measurements, as shown in Fig. 1. At OSI level 4, measurements were made for TCP connections and UDP communications using Iperf software [18]. 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 [12]. 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 RFC 1889 [19]. Another PC, with IP 192.168.0.20, was used to control the settings in the AP. The scheme of Fig. 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.

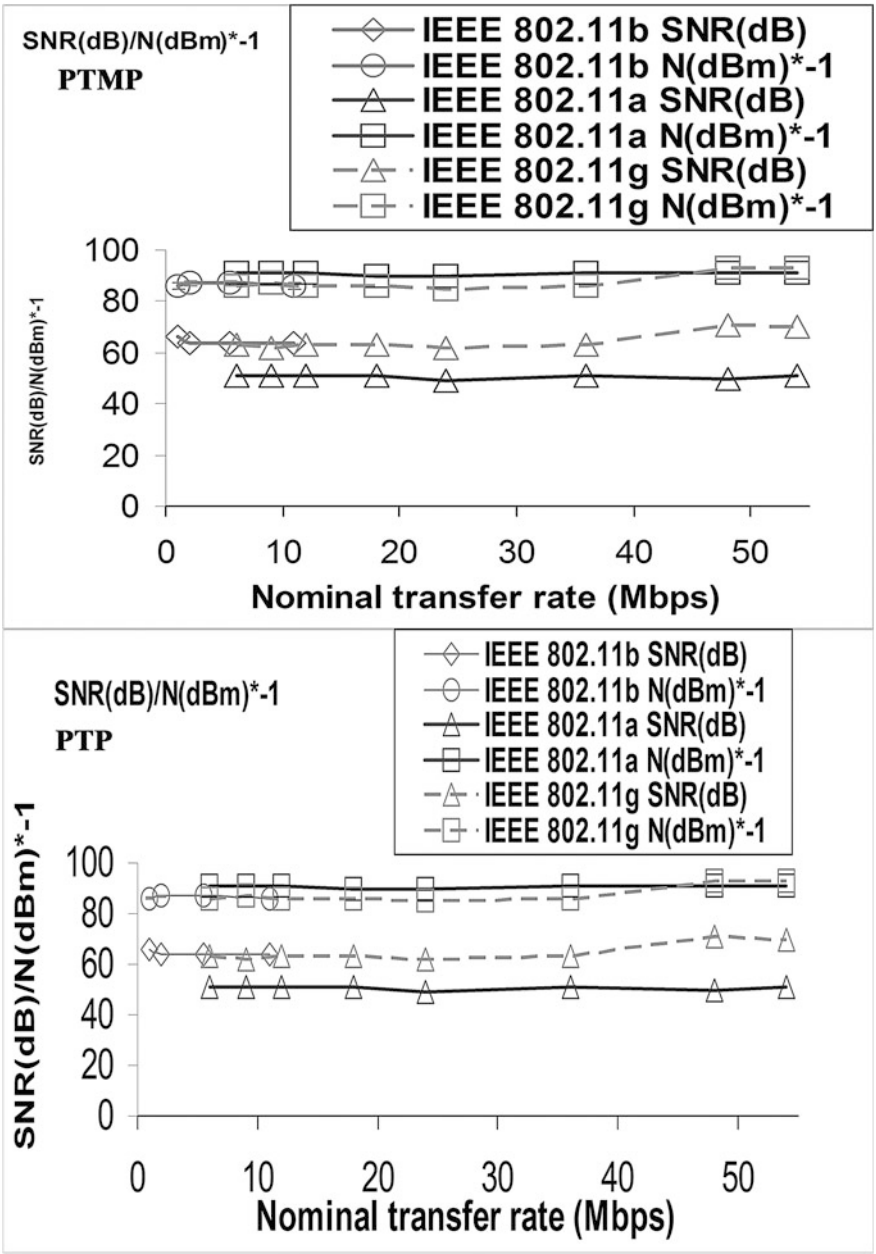


Fig. 2 Typical SNR (dB) and N (dBm); WPA PTMP and PTP links

Table 1 Average Wi-Fi (IEEE 802.11a, b, g) WPA 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	7.5 ± 0.2	6.3 + 0.2	2.9 + 0.1	15.9 + 0.5	13.4 ± 0.4
UDP-jitter (ms)	6.0 ± 0.9	3.3 ± 0.7	3.5 ± 0.5	5.5 + 0.2	2.5 + 0.5	2.3 + 0.1
UDP-% datagram loss	1.2 ± 0.2	2.2 ± 0.1	1.7 ± 0.1	1.2 + 0.2	1.2 + 0.2	1.8 + 0.2

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. 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. 2.

The main average TCP and UDP results are summarized in Table 1, both for WPA PTMP and PTP links. The statistical analysis, including calculations of confidence intervals, was carried out as in [20]. In Fig. 3 polynomial fits were made (shown as y vs. x), using the Excel worksheet, to the 802.11a, b, g TCP throughput data for PTMP and PTP links, respectively, where R^2 [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 (15.9 ± 0.5 Mbps, vs. 7.5 ± 0.2 Mbps for PTMP). On average TCP throughput for Open PTMP links (7.6 ± 0.2 Mbps) was found slightly better than for WPA PTMP links (7.5 ± 0.2 Mbps). In Figs. 4 and 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 802.11g and PTP links (2.3 ± 0.1 ms). On average, jitter performance was found similar for Open PMTP links (3.5 ± 0.4 ms) and WPA PTMP links (3.5 ± 0.5 ms). Concerning percentage datagram loss, the best performance was for 802.11a and PTP links (1.2 ± 0.2 %, vs. 2.2 ± 0.1 % for

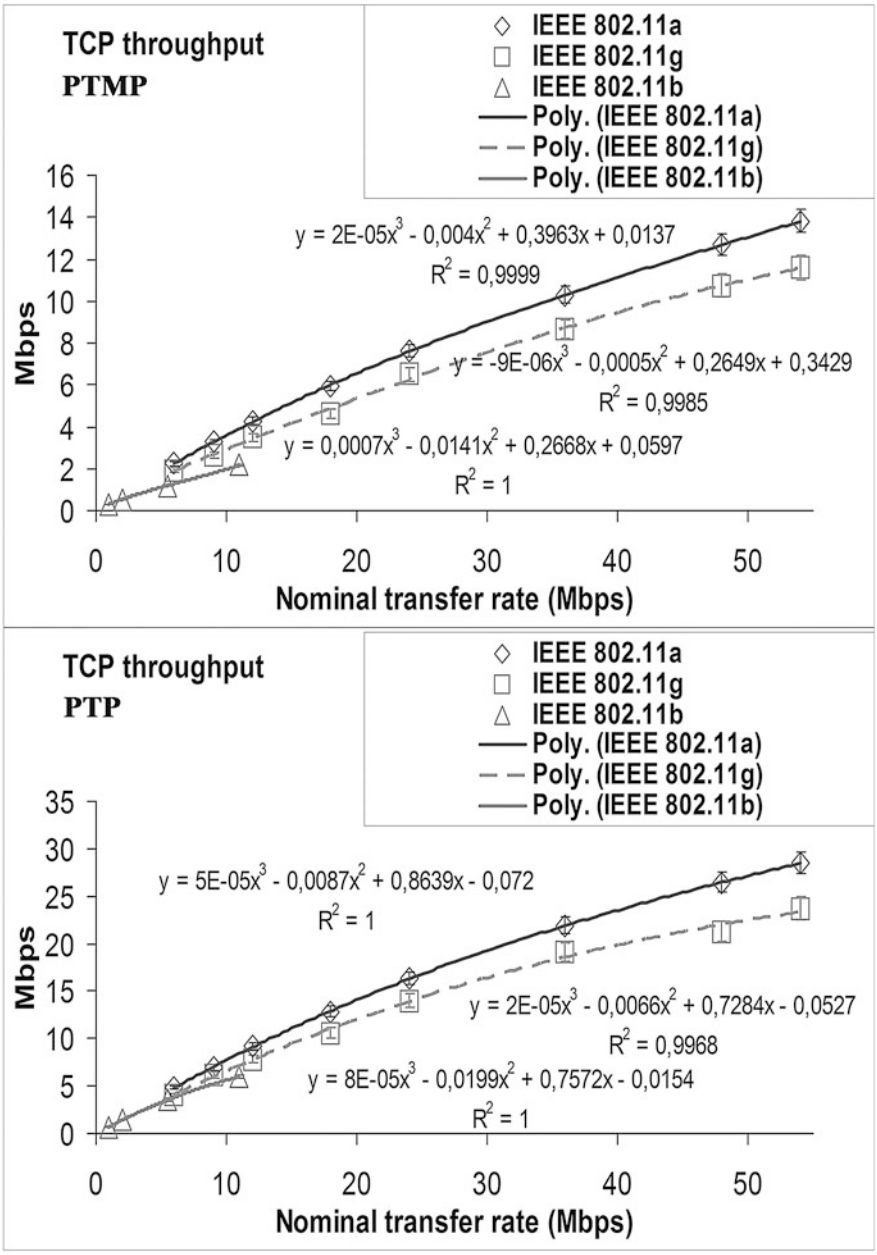


Fig. 3 TCP throughput results (y) versus technology and nominal transfer rate (x); WPA PTMP and PTP links

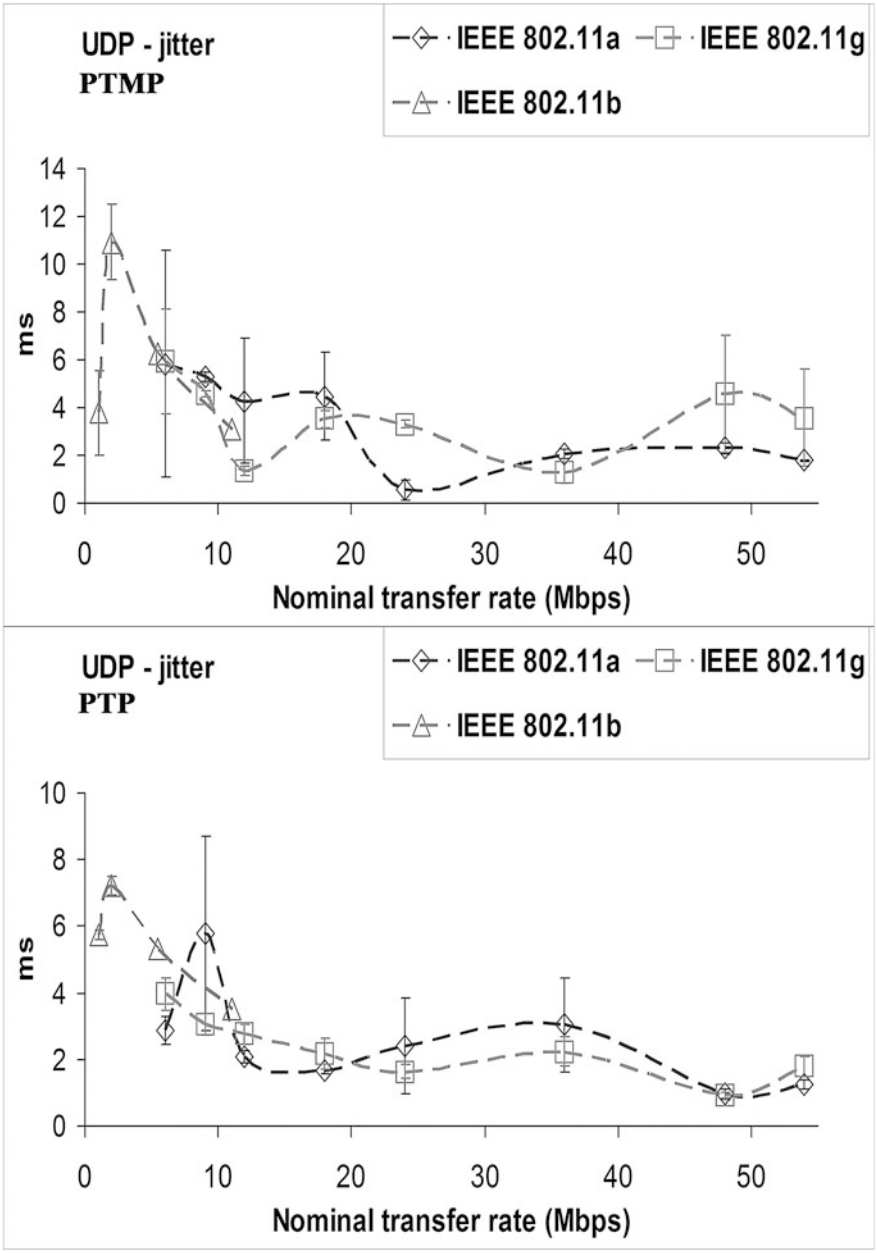


Fig. 4 UDP—jitter results versus technology and nominal transfer rate; WPA PTMP and PTP links

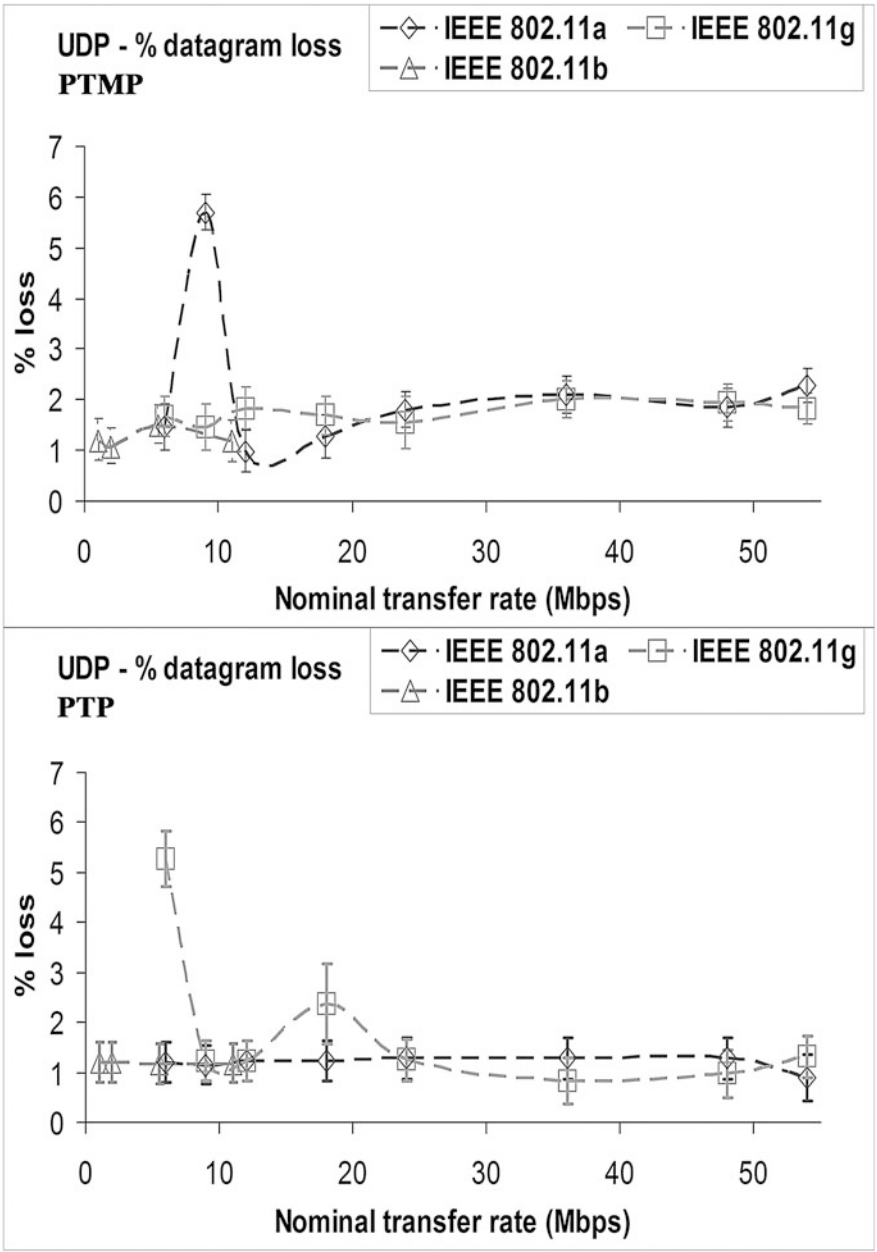


Fig. 5 UDP—percentage datagram loss results versus technology and nominal transfer rate; WPA PTMP and PTP links

PTMP). On average, percentage datagram loss performance was found better for Open PTMP links (1.0 ± 0.1 %) than for WPA PTMP Links (2.2 ± 0.1 %). Generally, in comparison to PTP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for PTMP links. In comparison to Open PTMP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for WPA 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 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 new 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 WPA point-to-multipoint links.

Through OSI layer 4, TCP throughput, jitter and percentage datagram loss were measured and compared for each standard and WPA PTMP and PTP links. It was found that, on average, the best TCP throughputs are for 802.11a and PTP links. On average, the best jitter performances were found for 802.11g and PTP links. Concerning percentage datagram loss, the best performance was for 802.11a and PTP links.

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. In comparison to Open PTMP links, TCP throughput, jitter and percentage datagram loss were found to show performance degradations for WPA 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 and security settings, 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/UI0524/2011(ProjectoEstratégico-UI524-2011-2012) 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. J.W. Mark, W. Zhuang, *Wireless Communications and Networking* (Prentice-Hall Inc, Upper Saddle River, 2003)
3. T.S. Rappaport, *Wireless Communications Principles and Practice*, 2nd edn. (Prentice-Hall Inc, Upper Saddle River, 2002)
4. W.R. Bruce III, R. Gilster, *Wireless LANs End to End* (Hungry Minds Inc, New York, 2002)
5. M. Schwartz, *Mobile Wireless Communications* (Cambridge University Press, Cambridge, 2005)
6. N.I. Sarkar, K.W. Sowerby, High performance measurements in the crowded office environment: a case study, in *Proceedings of ICCT'06-International Conference on Communication Technology*, Guilin, China, 27–30 Nov 2006
7. E. Monteiro, F. Boavida, *Engineering of Informatics Networks*, 4th edn. (FCA-Editor of Informatics Ltd, Lisbon, 2002)
8. J.A.R.P. de Carvalho, P.A.J. Gomes, H. Veiga, A.D. Reis, Development of a University Networking Project, in *Encyclopedia of Networked and Virtual Organizations*, ed. by G.D. Putnik, M.M. Cunha (IGI Global, Hershey, 2008), pp. 409–422
9. J.A.R.P. de Carvalho, H. Veiga, P.A.J. Gomes, C.F.R. Pacheco, N. Marques, A.D. Reis, Wi-Fi point-to-point links: performance aspects of IEEE 802.11a, b, g Laboratory links, in *Electronic Engineering and Computing Technology*. Lecture Notes in Electrical Engineering, vol. 60, ed. by S.-I. Ao, L. Gelman (Springer, Berlin, 2010), pp. 507–514
10. J.A.R. Pacheco de Carvalho, H. Veiga, N. Marques, C.F. Ribeiro Pacheco, A.D. Reis, Wi-Fi WEP point-to-point links: performance studies of IEEE 802.11a, b, g laboratory links, in *Electronic Engineering and Computing Technology*. Lecture Notes in Electrical Engineering, vol. 90, ed. by S.-I. Ao, L. Gelman. (Springer, Berlin, 2011), pp. 105–114
11. J.A.R.P. de Carvalho, H. Veiga, C.F.R. Pacheco, A.D. Reis, Performance studies of Wi-Fi IEEE 802.11A, G WPA point-to-multipoint links, in *Proceedings of WCE 2013—World Congress on Engineering 2013*. Lecture Notes on Computer Science (Imperial College London, London, 2013), pp. 1415–1419, 3–5 July 2013
12. J.A.R.P. de Carvalho, H. Veiga, N. Marques, C.F.R. Pacheco, A.D. Reis, Laboratory performance of Wi-Fi IEEE 802.11 B,G WPA2 point-to-point links: a case study, in *Proceedings of WCE 2011—World Congress on Engineering 2011* (Imperial College London, London), pp. 1770–1774, 6–8 July 2011
13. J.A.R.P. de Carvalho, N. Marques, H. Veiga, C.F.R. Pacheco, A.D. Reis, Experimental performance evaluation of a Gbps FSO link: a case study, in *Proceedigns of WINSYS 2010—International Conference on Wireless Information Networks and Systems* (Athens, Greece, 2010), pp. 123–128, 26–28 July 2010
14. DAP-1522 wireless bridge/access point technical manual (2010), http://www.dlink.com. Accessed 15 Jan 2012
15. AT-8000S/16 level 2 switch technical data (2009), http://www.alliedtelesis.com. Accessed 10 Dec 2010
16. WPC600N notebook adapter user guide (2008), http://www.linksys.com. Accessed 10 Jan 2012
17. NetStumbler software (2005), http://www.netstumbler.com. Accessed 21 Mar 2011
18. Iperf software (2003), http://dast.nlanr.net. Accessed 10 Jan 2008
19. Network Working Group, RFC 1889-RTP: A transport protocol for real time applications (1996), http://www.rfc-archive.org. Accessed 10 Feb 2008
20. P.R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences* (Mc Graw Hill Book Company, New York, 1969)

An Experimental Study of ZigBee for Body Sensor Networks

José Augusto Afonso, Diogo Miguel Ferreira Taveira Gomes and Rui Miguel Costa Rodrigues

Abstract We present an experimental performance evaluation of ZigBee networks in the context of data-intensive body sensor networks (BSNs). IEEE 802.15.4/ZigBee devices were mainly developed for use in wireless sensors network (WSN) applications; however, due to characteristics such as low power and small form factor, they are also being widely used in BSN applications, making it necessary to evaluate their suitability in this context. The delivery ratio and end-to-end delay were evaluated, under contention, for both star and tree topologies. The reliability of the ZigBee network in a star topology without hidden nodes was very good (delivery ratio close to 100 %), provided the acknowledgement mechanism was enabled. On the other hand, the performance in a tree topology was degraded due to router overload and the activation of the route maintenance protocol triggered by periods of high traffic load. The effect of the devices' clock drift and hidden nodes on the reliability of the star network was modeled and validated through experimental tests. In these tests, the worst-case delivery ratio when the acknowledgment is used decreased to 90 % with two sensor nodes, while for the non-acknowledged mode the result was of 13 %. These results show that a mechanism for distributing the nodes' traffic over the time is required to avoid BSN performance degradation caused by router overload, clock drift and hidden node issues.

Keywords Body sensor networks · Experimental study · IEEE 802.15.4 · Quality of service · Wireless sensor networks · ZigBee

J. A. Afonso (✉) · D. M. F. T. Gomes · R. M. C. Rodrigues
Centro Algoritmi, University of Minho, 4800-058 Guimarães, Portugal
e-mail: jose.afonso@dei.uminho.pt

D. M. F. T. Gomes
e-mail: a50035@alunos.uminho.pt

R. M. C. Rodrigues
e-mail: a50043@alunos.uminho.pt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_34,

1 Introduction

Recent advances in wireless communications, microelectronics and signal processing are enabling the development of body sensor networks (BSNs). These networks are mainly comprised by wearable or implantable sensor devices and a wireless network to transport the collected data from the users' bodies to a remote site [1]. BSNs can be used to monitor diverse physiological parameters and signals, such as temperature, heart rate, blood pressure, blood oxygen saturation, body posture, electroencephalogram (EEG), electrocardiogram (ECG) and electromyogram (EMG) [2].

BSNs can provide significant benefits in the long term diagnosis and treatments of patients, with minimum constrains to daily life activities. These networks allow the patients to move freely, inside or outside the hospital, while providing continuous monitoring, which can be particularly useful when long periods of monitoring are required. For example, many cardiac diseases are associated with episodic abnormalities, such as transient surges in blood pressure or arrhythmias [3], which cannot always be detected using conventional monitoring equipment. BSNs have the potential to provide early detection and prevention [4] of such pathologies, replacing expensive therapies later on.

IEEE 802.15.4 and ZigBee are widespread adopted network standards conceived primarily for wireless sensor networks (WSN) applications, which typically generate event based and low data rate traffic. Currently, these are also the most widely used network standards for BSNs [1, 2, 5]. However, unlike WSNs, BSNs usually generate periodic and, frequently, data-intensive traffic (e.g., ECG, EEG and body posture data). Therefore, the suitability of these standards to transport the traffic generated by this type of BSN sensors needs to be assessed.

Several works in the literature present performance evaluation results regarding IEEE 802.15.4 and/or ZigBee protocols, for different application scenarios. However, most of these results are based on analytical models [6–8] or simulations [9, 10]. On the other hand, this work, which presents a revised and extended version of our previous work [11], concerns the experimental performance evaluation of ZigBee and IEEE 802.15.4 using BSN traffic. This approach provides further insight on the performance of these systems, by taking into account variables present in real-world implementations that have impact in the performance but are overlooked in most theoretical models, such as the processing load in the network nodes.

In [12], the authors present a multihop ZigBee-based BSN system for patient monitoring in hospitals, where wearable patient units using MICAz motes are connected to a commercial blood pressure and heart rate monitor. Experimental tests in laboratory using three patient units resulted in no data loss. In [13], the authors present a multihop 802.15.4-based BSN system that measures the heart rate and blood oxygen levels of emergency room patients. The system was implemented using Telos motes and used the Collection Tree Protocol (CTP) provided by TinyOS to forward the measurements to a gateway. The measured delivery ratio was above 99.9 %.

Unlike these two works, which only use low data rate sensors, our work considers sensors that generate data-intensive traffic. Three relevant quality of service (QoS) metrics are studied: delivery ratio (DR), end-to-end delay and goodput. Clock drift and hidden nodes effects were also modeled and evaluated.

2 Network Standards and Platforms

2.1 IEEE 802.15.4 and ZigBee

The IEEE 802.15.4 standard [14] specifies the physical (PHY) and medium access control (MAC) layers for low power, low data rate and low cost wireless network devices. The PHY layer uses direct sequence spread spectrum (DSSS) and defines different transmission rates and bands: 250 kbps for the 2.4 GHz band and 20/40 kbps for 868/915 MHz band, among other possible optional configurations. The MAC layer defines two different operation modes: a non-beacon-enabled mode, which uses an unslotted CSMA-CA (Carrier Sense Multiple Access-Collision Avoidance) algorithm, and a beacon-enabled mode, which defines a superframe structure and uses a slotted CSMA-CA algorithm. The MAC layer provides also an optional guaranteed time slot (GTS) scheme, which allows the allocation of dedicated bandwidth for devices; however, this scheme is limited to a maximum of seven GTS allocations.

ZigBee [15, 16] is a standard designed for low power devices used on wireless monitoring and control systems. The protocol supports star, tree and mesh topologies. In star topology, all devices communicate directly with the coordinator. Tree and mesh topologies allow to increase the range of the network by introducing routers that relay the traffic from the end devices (EDs). The ZigBee stack is based on the Open Systems Interconnection (OSI) model. Each layer performs a specific set of services for the layer above. The stack is divided into four distinct layers: physical (PHY), medium access control (MAC), network (NWK) and application (APL). The IEEE 802.15.4 standard defines the two lower layers of ZigBee: PHY and MAC. The NWK layer enables multihop network communication and is responsible to create and maintain the network, discover new routes and assign the devices short addresses, among others tasks. The APL layer supports up to 240 applications on the same device.

2.2 Experimental Evaluation Platforms

The hardware platform used in the tests was the CC2530 development kit, which is provided by Texas Instruments, a leading supplier of ZigBee products. It is based on the CC2530 [17] SoC (System on Chip), which integrates a microcontroller and a

transceiver in the same chip. The microcontroller is based on the 8051 architecture, and the transceiver is compliant with the IEEE 802.15.4 standard in the 2.4 GHz frequency band.

The experimental tests presented in this work were developed using the ZigBee and IEEE 802.15.4 stack implementations provided by Texas Instruments: Z-Stack and TIMAC, respectively. The Z-Stack version that was used, Z-Stack-CC2530-2.4.0-1.4.0, supports the two stack profiles of the ZigBee 2007 specification: ZigBee and ZigBee Pro. This Z-Stack version is a combination of the ZigBee stack implementation version 2.4.0 and the IEEE 802.15.4 stack implementation version 1.4.0. Some of the experiments described in this work use only the IEEE 802.15.4 stack. In these cases, the standalone TIMAC version TIMAC-CC530-1.3.1 was used.

3 Evaluation Methods and Models

This section describes the experimental evaluation methods and models that were used to obtain the results presented in the next section. Channel 26 was used, due to the absence of interference from Wi-Fi networks and other sources, verified using a spectrum analyzer. Likewise, the transmission power and placement of the nodes was set to assure that there are no packet losses due to path loss or shadowing effects, since the purpose of this study is to evaluate only the losses due to collision and transmission attempt failures of the CSMA-CA protocol caused by contention, clock drift and hidden nodes. In the tests with hidden nodes, the signals of the sensor nodes were blocked from each other using metal plates and the nodes were placed inside an anechoic chamber to avoid multipath propagation.

The default parameters of the IEEE 802.15.4 unslotted CSMA-CA algorithm were used. The overhead introduced in the data packets by all ZigBee layers accounts for a total of 264 bits, in all evaluation scenarios. All tests finish after the coordinator has received 5,000 packets from the end devices. The tests presented in this work used the ZigBee Pro stack profile, but the same tests were performed using ZigBee stack and the results have shown no significant differences.

This work uses data-intensive traffic parameters extracted from a real implementation of a body posture monitoring system composed by multiple sensor modules, each one containing three accelerometers and three magnetometers [18], which are sampled at 30 Hz. Two different traffic configurations were used. In mode A, packets are transmitted at 200 ms intervals, and the data packet length, which includes six samples from each sensor plus the protocol overhead, is 89 bytes. In mode B, smaller packets of 62 bytes with half of the samples are transmitted every 100 ms. Similar data-intensive traffic can be found in other BSN applications, such as ECG monitoring, where the sampling rate can reach 250 Hz per electrode [19].

3.1 Delivery Ratio and Delay

In this evaluation scenario, the delivery ratio and end-to-end delay were measured in a contention environment where multiple EDs generate packets to the coordinator simultaneously. The delivery ratio (DR) represents the ratio of the number of successfully delivered packets to the number of packets generated by the source node application. The end-to-end delay is the time since the packet is delivered for transmission by the source node application layer until it reaches the destination node application layer.

Although star topologies are more common for BSNs, multihop topologies are considered in many works [12, 13]. Therefore, two topologies were evaluated: star and 2-hop tree. In the latter, a router forwards the packets from the EDs (sensor nodes) to the coordinator.

The same tests were performed with both Z-Stack and TIMAC, in order to observe the overall system behavior when supported by these two different stacks. Since the IEEE 802.15.4 standard does not define a network layer, for the tests using the TIMAC, the router of the 2-hop tree topology was simulated using a peer-to-peer network where all the EDs transmit the packets to a specific device, which relays the packets to the coordinator.

A wired trigger signal controlled by the coordinator was used to generate a periodic interrupt on the EDs according to the transmission period, which was set to 200 ms (mode A). The main objective of the trigger is to create a scenario of contention where all the EDs try to access the medium at same time, which represents a worst-case contention scenario. For the delay tests, an end device was chosen to be the reference device for the measured values.

3.2 Clock Drift

This section proposes a model that uses the differential clock drift between two ZigBee end devices to estimate the duration of two parameters: the interference period (T_{Int}), defined as the period during which the two end devices using the unslotted CSMA-CA algorithm will contend for the channel due to the clock drift, and the interference repetition interval (T_{IntRep}). This model uses the times associated to a packet transmission according to the unslotted CSMS-CA algorithm of the IEEE 802.15.4 standard, which are shown in Fig. 1.

The transmission period (T_{Tx}) is composed by the random backoff interval ($T_{Backoff}$), a transceiver turnaround time (T_{TA}) from RX to TX, the packet transmission time (T_{Packet}), a turnaround time from TX to RX and, finally, the ACK transmission time (T_{Ack}). The turnaround time is defined in IEEE 802.15.4 standard and corresponds to 192 μs. The ACK transmission time is 352 μs, while the packet transmission time depends on the payload length.

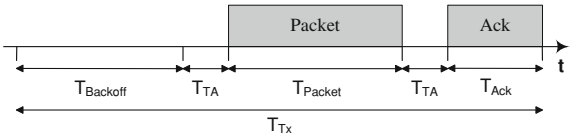


Fig. 1 Times associated to the IEEE 802.15.4 unslotted CSMA-CA algorithm

Each end device *EDn* was physically connected to the coordinator (base station), to measure the number of added or missing oscillations ($ticks_{drifted}$) within a period T, in comparison to the coordinator's clock. The differential clock drift between the base station (BS) and end device n can be calculated through Eq. 1, where f_{osc} is the nominal clock frequency of the CC2530 (32 MHz).

$$D_{BS,EDn} = \frac{ticks_{drifted}}{f_{osc} \times T} \tag{1}$$

The differential clock drift between *ED1* and *ED2* can be obtained, without the knowledge of the absolute clock drift of the end devices (D_{EDn}), from the respective differential clock drifts in relation to the BS:

$$D_{ED1,ED2} = D_{BS,ED1} - D_{BS,ED2} = D_{ED2} - D_{ED1} \tag{2}$$

Unsynchronized devices transmitting periodic traffic with the same nominal period will eventually contend for the wireless channel due to the clock drift effect. If the differential clock drift between *ED1* and *ED2* is $D_{ED1,ED2}$ and the nominal transmission period of the nodes is given by T_{ED}, then both nodes will start to contend for the wireless channel every T_{IntRep} seconds. The value of the interference repetition interval can be obtained through Eq. 3:

$$T_{IntRep} = \frac{T_{ED}}{D_{ED1,ED2}} \tag{3}$$

The interference period (T_{Int}) during which two devices will compete for the channel can be obtained through the Eq. 4, where T_{Vul} represents the vulnerability time window.

$$T_{Int} = \frac{T_{Vul}}{D_{ED1,ED2}} \tag{4}$$

Figure 2 shows this vulnerability time window under which the transmissions of two nodes may interfere with each other.

Equations 5 and 6 represent the instants of time when the interference period between devices *ED1* and *ED2* begins and ends, respectively.

$$t_{ED2} + T_{Backoff\_min} + T_{TA} = t_{ED1} + T_{Tx\_max} \tag{5}$$

$$t'_{ED1} + T_{Backoff\_min} + T_{TA} = t'_{ED2} + T_{Tx\_max} \tag{6}$$

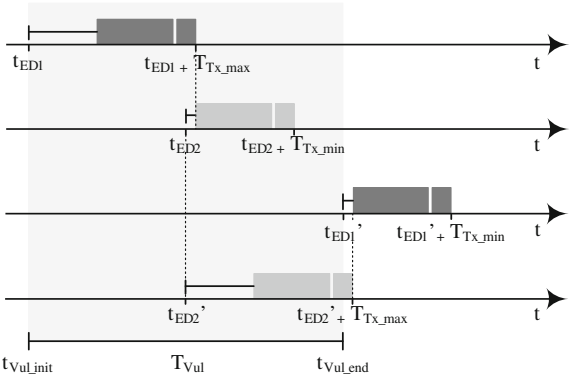


Fig. 2 Vulnerability window for the clock drift evaluation scenario

$T_{Backoff\_min}$ is the minimum backoff period, which is equal to zero. T_{Tx_max} represents the maximum period needed to transmit a packet and receive the respective acknowledgment (if required), which is calculated using the maximum backoff period ($T_{Backoff\_max} = 2.24$ ms). t_{EDn} and t'_{EDn} represent instants of time where device n starts the CSMA-CA algorithm. We obtain T_{Vul} from $t'_{ED1} - t_{ED1}$:

$$T_{Vul} = 2 \times (T_{Tx\_max} - T_{TA}) \tag{7}$$

To validate our model, we evaluated a ZigBee network formed by two end devices that transmit packets in mode B ($T_{ED} = 100$ ms) to the coordinator in a star topology. The packet transmission time in this case is 1.984 ms. In order to better observe the interference periods and interference repetition intervals, we forced a hidden node situation, so nodes are unable to backoff due to carrier sense, and the ACK mechanism was disabled. Therefore, in this case:

$$T_{Tx\_max} = T_{Backoff\_max} + T_{TA} + T_{packet} \tag{8}$$

3.3 Hidden Node Problem

In this test, two ZigBee end devices hidden from each other transmit packets in mode B in a star network topology. In order to analyze a worst-case scenario, the nodes generate packets at the same time, according to a trigger signal sent by the coordinator, and no acknowledgments are used.

The minimum transmission period (T_{Tx_min}) is associated to $T_{Backoff\_min}$ (zero), while the maximum transmission period (T_{Tx_max}) is achieved with $T_{Backoff\_max}$ (2.24 ms, which corresponds to 7 unit backoff periods). Given the packet transmission time in this test (1.984 ms), when the coordinator triggers a transmission in both EDs, the corresponding transmitted packets will not collide only if the transmission periods of *ED1* and *ED2* are equal to T_{Tx_min} and T_{Tx_max},

respectively, or vice versa. The probability for this specific case to occur (p_{TX}) can be obtained through the following equation:

$$p_{TX} = 2 \times p_{Backoff\_min} \times p_{Backoff\_max} \tag{9}$$

The two probabilities on the right side of this equation are equal to 1/8, since they come from a discrete uniform distribution with 8 possibilities (0 to 7). Therefore, p_{TX} is 3.125 %. This value corresponds to the expected DR of the network when the ACK mechanism is not used.

3.4 Maximum Goodput

In this scenario a single ED transmits packets continuously to the network coordinator in the star topology. The application layer waits for the indication that the ACK has arrived before sending the next packet. The theoretical maximum goodput is obtained using Eq. 10, where the average transmission period is the sum of the MAC times presented on Fig. 1, using the mean backoff interval (1.12 ms).

$$Goodput = \frac{Payload\ Length[bits]}{Average\ Transmission\ Period[s]} \tag{10}$$

4 Results and Discussion

4.1 Delivery Ratio and Delay

During the tests with the Z-Stack in the 2-hop tree topology and with the acknowledgment mechanism enabled, a router blocking problem was observed. Through the use of a packet sniffer, it was noticed that the router relays packets for just few seconds, then blocks for around 8 s, after what it becomes available again and the process repeats. Several other tests were performed in other conditions, and it was verified that this problem only occurred in tests where the router was subject to high traffic load. A possible explanation for this problem is that the router experiences an overload situation where it is not able to handle packet relaying at the NWK layer when new packets are constantly being received at the MAC layer (which is a higher priority task in the Z-Stack implementation). Therefore, in order to allow the evaluation of the delivery ratio and delay during the period where the router is not blocked, the number of packets received by the coordinator for this particular experiment was reduced from 5,000 to 1,000 packets.

Figure 3 presents the measured DR with Z-Stack as a function of the number of sensor nodes. For the star topology, the DR was close to 100 % when the ACK mechanism was used. However, the DR for the 2-hop tree topology with 3–5 end devices was lower (around 96 %). The explanation is that, due to the high traffic load generated by the end devices, the route maintenance protocol, triggered by the router's network layer, initiates the route discovery procedure frequently (each 5 s, on average). This procedure, which lasts for around 250 ms, forces the router to interrupt the packet relaying, causing packet drops due to buffer overflow. When the acknowledgments are disabled, the DR decreases significantly in both topologies as the number of sensor nodes increases, as expected.

In order to compare the TIMAC performance with Z-Stack, the length of the data packets has been equaled to the one used in the Z-Stack measurements through the introduction of dummy bytes, since the TIMAC has smaller protocol overhead. Figure 4 presents the DR with TIMAC as a function of the number of sensor nodes.

The results with the ACK mechanism enabled are worse than the ones obtained using the Z-Stack. This is explained by the fact that the Z-Stack network layer may retransmit a packet if the MAC layer has failed to transmit it (the default is one retransmission). The non-acknowledged experiments showed better results with the Z-Stack for the tree topology, due to the router's network layer capability for buffering the received packets and relaying them in lower contention periods, whereas the application that simulates the router in the TIMAC relays the received packets immediately.

Figures 5 and 6 show the measured average and maximum end-to-end delay, respectively, as a function of the number of sensor nodes, for both Z-Stack and TIMAC, and using the ACK mechanism. The delays measured with TIMAC are lower than those measured with the Z-Stack, due to the lower processing load introduced by the TIMAC stack. As expected, the delays increase with the number of nodes, because the contention, collisions and retransmissions also increase. The activation of the route maintenance protocol for the Z-Stack tree topology with 3–5 nodes causes the buffering of packets in the NWK layer, increasing significantly the maximum delay for the ZigBee network.

4.2 Clock Drift

Table 1 specifies the differential clock drifts between end device n and the BS ($D_{BS,EDn}$).

We have chosen end devices 0 and 1 for the experimental measurements and model validation; for these nodes, the differential clock drift is $D_{ED1,ED0} = 3.5$ ppm. Using these values in Eq. 4, we obtain a T_{Int} value of 40 min. The T_{IntRep} period, which can be obtained through Eq. 3, is 7 h and 56 min. Figure 7 shows the results obtained in this test, which uses a moving average window of 60 messages to compute the DR, corresponding to 6 s. The test

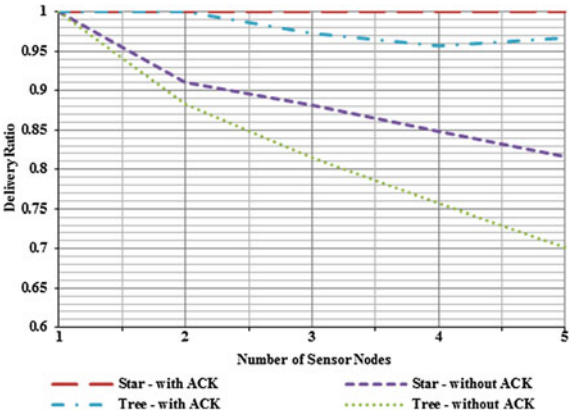


Fig. 3 DR measured with Z-Stack as a function of the number of nodes

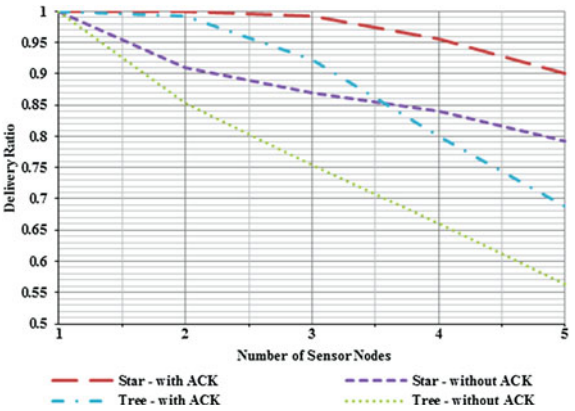


Fig. 4 DR measured with TIMAC as a function of the number of nodes

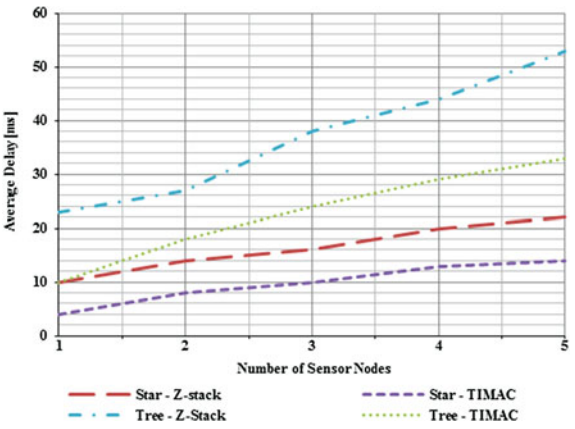


Fig. 5 Average delay as a function of the number of nodes for both Z-Stack and TIMAC

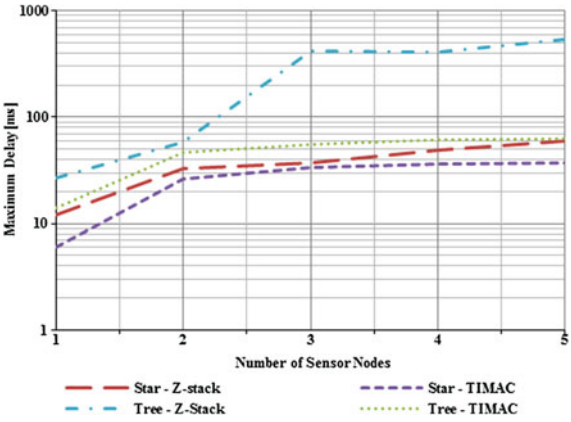


Fig. 6 Maximum delay as a function of the number of nodes for both Z-Stack and TIMAC

Table 1 Measured differential clock drifts to the BS in ppm

$D_{BS,ED0}$	$D_{BS,ED1}$	$D_{BS,ED2}$	$D_{BS,ED3}$	$D_{BS,ED4}$
3.6	0.1	−1.0	−0.5	0.2

started at 18:15:10 and ended at 13:02:44 the next day. The DR was 100 % most of the time, which corresponds to non-interference periods. The DR decreases when the interference period starts, reaches a minimum when both devices are generating packets simultaneously, and then increases again until the end of the interference period. Taking into account these boundaries, the interference period lasted for approximately 40 min. The interference repetition interval is approximately 7 h and 53 min. The measured T_{Int} matches the value predicted by the theoretical model, whereas T_{IntRep} presents an error of 0.6 %. These results validate the proposed model.

4.3 Hidden Node Problem

In this evaluation scenario, the measured DR when the acknowledgment is used was 90 %, whereas for the non-acknowledged mode the result was of 13 %, which is very close to the minimum DR verified in the clock drift experiment, shown in Fig. 7. Previous results showed DRs in the absence of hidden nodes of 100 % and 91 % for two end devices transmitting in acknowledged and non-acknowledged modes, respectively (Fig. 3). Therefore, when compared with the results without hidden nodes, the results with hidden nodes show accentuated decrease in the DR. These results show that, in a scenario of contention, the DR of a simple network constituted by only two hidden EDs decreases considerably. With more hidden nodes, the network performance would be even worse. This may seriously compromise the reliability of the network and, consequently, make it unable to satisfy the QoS requirements [20] of BSN applications.

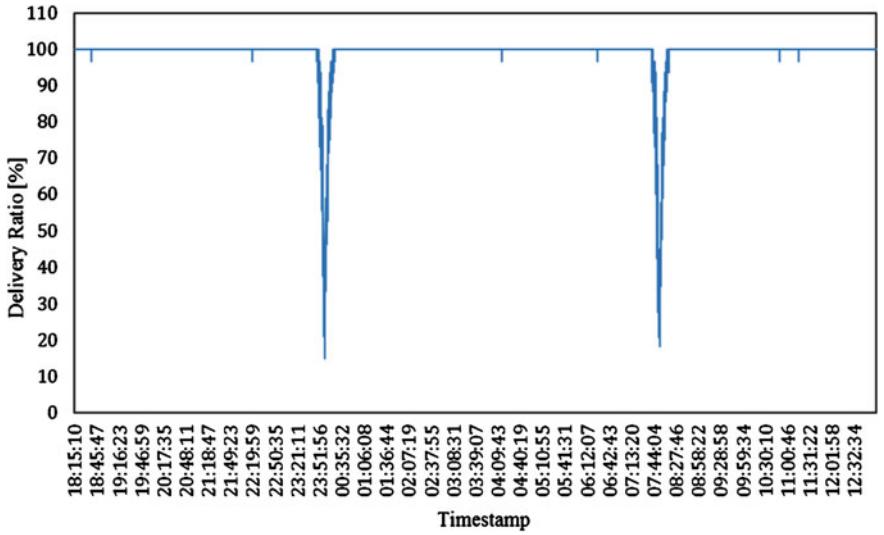


Fig. 7 DR with clock drift in a star topology with two hidden nodes

The DR measured in this experiment for the non-acknowledged mode (13 %) is higher that the value predicted by the theoretical analysis (3.125 %) performed on the previous section. In order to discover the origin of this discrepancy, we analyzed the log file of this specific experiment. The theoretical analysis assumes that the coordinator should only receive packets that were sent from the nodes in the absence of collision, which is only possible if node 1 selects $T_{Backoff\_min}$ and node 2 selects $T_{Backoff\_max}$ when the CSMA-CA is executed, or vice versa. Therefore, it should not be possible, in principle, to receive packets from only one of the nodes; however, this situation occurred, causing an increase on the DR. Using a packet sniffer, it was possible to observe that both nodes transmit their packets when triggered and, if one of the nodes was disabled, the coordinator receives all the packets from the other node. It was also observed that if the transmit power of the nodes were controlled in a way for the coordinator to receive equal power from both nodes, the DR decreased, while it increased if the packets were received with different power. Therefore, we conclude that the difference between theoretical and experimental results may be related with the capture effect, where, in the presence of collision, a packet may be successfully received if its power is sufficiently greater than the power of the interfering packet.

4.4 Maximum Goodput

Figure 8 presents the theoretical and measured maximum goodput for star topology, as a function of the payload length, using the Z-Stack. The measured goodput is significantly lower than the theoretical values given by Eq. 10 because the latter

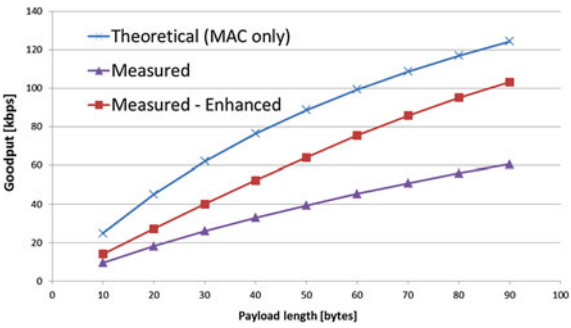


Fig. 8 Maximum goodput for star topology

was calculated using only the MAC times presented in Fig. 1. However, when the delays from the application to the MAC layer ($T_{APP \rightarrow MAC}$) and vice versa ($T_{MAC \rightarrow APP}$) are added to the average transmission period, the theoretical values become correct. For example, with 90-byte payload, the measured average $T_{APP \rightarrow MAC}$ and $T_{MAC \rightarrow APP}$ values were, respectively, 4.04 ms and 2.23 ms. Using these values, the theoretical maximum goodput becomes 59.7 kbps, which is very close to the measured value (60.5 kbps).

A simple enhancement, which consists in sending two packets from the application layer from the MAC layer at the beginning, was implemented and tested. As shown in Fig. 8, this enhancement provides a substantial increase in the measured maximum goodput (70.7 % with 90-byte payload). The rationale is that the MAC layer will always have a spare packet available on its buffer and therefore it can bypass most of the delay between the application and MAC layers.

5 Conclusion

This work presented an experimental performance analysis of ZigBee in the context of the BSNs, using the Texas Instruments implementations of ZigBee (Z-Stack) and IEEE 802.15.4 (TIMAC).

For 2-hop tree, tests have shown that successive periods of high traffic load can cause the ZigBee router to start the route discovery procedure, with negative impact on the delay and DR. A router blocking problem, which is also caused by high traffic loads and lasts several seconds, was also observed.

Results from the clock drift analysis showed that interference periods may last for a long time due to the small clock drifts between nodes. The experiments have also demonstrated the validity of the proposed clock drift model, where the theoretical and experimental results are close.

Other results have shown that the DR with hidden nodes is considerably worse. Although this experiment considered a worst-case contention scenario, due to the synchronization of packet generation instants, only two end devices were used.

Multiple hidden nodes combined with the clock drift effect may cause frequent network reliability problems during long periods.

Since BSN applications demand specific QoS requirements, these results suggest that it is necessary to provide a mechanism to distribute the traffic load generated by high traffic nodes along the time in ZigBee-based BSNs, in order to prevent the router overload, clock drift and hidden node issues.

Acknowledgments This work is funded by FEDER funds through "Programa Operacional Fatores de Competitividade—COMPETE" and by National Funds through FCT—Portuguese Foundation for Science and Technology in the scope of the Project FCOMP-01-0124-FEDER-022674 and Project PEst-OE/EEI/UI0319/2014.

References

1. M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, V.C.M. Leung, Body area networks: A survey. Mobile Netw. Appl. **16**(2), 171–193 (2011)
2. A. Pantelopoulos, N. Bourbakis, A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**(1), 1–12 (2010)
3. B. Lo, G.Z. Yang, Key Technical Challenges and Current Implementations of Body Sensor Networks. Proceedings of BSN 2005, London, UK, April 2005
4. A. Chen et al., HDPS: heart disease prediction system. Comput. Cardiol. **38**, 557–560 (2011)
5. H. Yan, H. Huo, Y. Xu, M. Gidlund, Wireless sensor network based e-health system—Implementation and experimental results. IEEE Trans. Consum. Electron. **56**(4), 2288–2295 (2010)
6. Z. Chen, C. Lin, H. Wen, H. Yin, An Analytical Model for Evaluating IEEE 802.15.4 CSMA/CA Protocol in Low-Rate Wireless Application. Proceedings of AINAW, Ontario, Canada 2007, pp. 899–904
7. X. Liang, I. Balasingham, Performance Analysis of the IEEE 802.15.4 based ECG Monitoring Network. Proceedings of Seventh IASTED International Conferences Wireless and Optical Communications, Montreal, Canada 2007, pp. 99–104
8. J.S. Choi, M.C. Zhou, Performance Analysis of ZigBee-Based Body Sensor Networks. Proceedings of IEEE SMC, Istanbul, Turkey 2010, pp. 2427–2433
9. J. Zheng, M.J. Lee, A comprehensive performance study of IEEE 802.15.4. Sensor Netw. Oper. **4**, 1–14 (2006)
10. G. Lu, B. Krishnamachari, C.S. Raghavendra, Performance Evaluation of the IEEE 802.15. 4 MAC for Low-Rate Low-Power Wireless Networks. Proceedings of IEEE IPCCC, Phoenix, USA 2004, pp. 701–706
11. D. Gomes, C. Gonçalves, J.A. Afonso, Performance Evaluation of ZigBee Protocol for High Data Rate Body Sensor Networks. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering, WCE 2013, London, UK, 3–5 July 2013, pp. 1468–1473
12. A. Hande, T. Polk, W. Walker, D. Bhatia, Self-powered wireless sensor networks for remote patient monitoring in hospitals. Sensors **6**(9), 1102–1117 (2006)
13. J. Ko, T. Gao, A. Terzis, Empirical Study of a Medical Sensor Application in an Urban Emergency Department. Proceedings of BodyNets '09, Los Angeles, USA 2009
14. IEEE Std 802.15.4-200, Part 15.4, Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs), September 2006

15. ZigBee Standards Organization, ZigBee Specification, Document 053474r17, January 2008
16. D. Gislason, *Zigbee wireless networking* (Newnes, UK, 2008)
17. Texas Instruments, A true system-on-chip solution for 2.4-GHz IEEE 802.15.4 and ZigBee applications. CC2530 datasheet, February 2011
18. J.A. Afonso, J.H. Correia, H.R. Silva, L.A. Rocha, Body kinetics monitoring system. International patent WO/2008/018810, February 2008
19. M. Paksuniemi, H. Sorvoja, E. Alasaarela, R. Myllylä, Wireless Sensor and Data Transmission Needs and Technologies for Patient Monitoring in the Operating Room and Intensive Care Unit. Proceedings of 27th IEEE EMBC, Shanghai, China 2005
20. H.F. López, J.A. Afonso, J.H. Correia, R. Simões, The Need for Standardized Tests to Evaluate the Reliability of Data Transport in Wireless Medical Systems. Lecture Notes of ICST, vol 102, 2012, pp. 137–145

Closed Form Solution and Statistical Performance Analyses for Regularized Least Squares Estimation of Retinal Oxygen Tension

Gokhan Gunay and Isa Yildirim

Abstract For improved estimation of oxygen tension in retinal blood vessels, regularization of least squares estimation method was proposed earlier and it was shown to be very effective. However, closed form solutions for the estimation, and bias and variance of the estimator were not provided and comprehensive statistical analyses were not done. In this chapter, we derive the closed form solution for the regularized least squares estimation, bias and variance of the regularized least squares estimator and with the help of the closed form solutions, statistical performance analyses of the estimator are realized for different values of estimation parameters.

Keywords Closed form solution · Least squares estimation · Phosphorescence lifetime imaging · Regularized estimation · Retinal oxygen tension · Statistical performance analysis

1 Introduction

Accurate estimation of oxygen tension (pO_2) in retinal vessels is of primary importance since abnormality of oxygenation in retinal tissue, in many cases, gives important clues regarding devastating common eye diseases such as diabetic retinopathy, glaucoma, and age related macular degeneration [1, 2].

G. Gunay (✉)
Department of Electrical and Electronics Engineering, Bozok University, 66900 Yozgat, Turkey
e-mail: ggunay@itu.edu.tr

G. Gunay · I. Yildirim
Department of Electrical and Electronics Engineering, Istanbul Technical University, 34469 Istanbul, Maslak, Turkey
e-mail: iyildirim@itu.edu.tr

I. Yildirim
Master of Engineering Department, University of Illinois, Chicago, IL 60607, USA

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_35,

Oxygen tension of retinal vessels can be estimated using phosphorescence lifetime imaging model (PLIM) [3, 4] whose mathematical model was developed by Lakowicz et al. [5] for fluorescence lifetime imaging model (FLIM). In [3, 4], the least squares (LS) estimation method was used to obtain estimate of oxygen tension in retinal vessels using PLIM.

While the LS estimation method is efficient in the computation sense, it produces high variance, and artificial peaks in the estimates, and therefore gives values outside of the physiological range. In order to overcome these shortcomings, regularization of the LS estimation method was proposed by Yildirim et al. [6].

Regularization has been extensively used in several problems such as image processing [7, 8], biomedical imaging [9, 10], and astronomical imaging [11] and its success in many problems was the main motivation of Yildirim et al. [6] to develop a regularized least squares (RLS) estimation method in the estimation of oxygen tension in retinal vessels.

In their study, after considering the physiology of retinal tissue in which oxygen tension of a retinal vessel does not vary rapidly in a small neighborhood [12], they assumed that mean value of a pixel value in an oxygen tension map of retinal blood vessel can be formulated as weighted mean of oxygen tension values of its neighboring pixels. It was shown that their RLS method is much better than the LS estimation approach in many senses such as robustness to noise, having much less variance, obtaining smoother pO_2 maps and therefore generating pO_2 values which are in the physiologically expected range. However, in their study, iterative procedures such as steepest descent algorithm were used to find minimum of the RLS cost function and a closed form solution was not provided. Bias and variance of the estimator were also estimated using some Monte-Carlo simulations.

Closed form solutions for the RLS estimations and its bias and variance was proposed in [13] and in this paper, we give derivation of them considering the RLS estimation method proposed in [6]. With the help of the closed form solutions RLS estimator, we examined and showed the effects of the regularization parameters, window size and weighting coefficients of neighboring pixels, which are used in the formation of regularization term in the model, and phosphorescence observation number on the statistical performance of the estimator.

2 Background

2.1 Phosphorescence Lifetime Imaging Model

In phosphorescence life time imaging model, a linear model developed earlier for the fluorescence life time imaging [5] is used. The intensity of the emitted phosphorescence, denoted by $I(t)$, is given as:

$$I(t) = I_o \exp(-t/\tau), \tag{1}$$

where I_o is the maximum intensity at time zero and τ is the lifetime of the phosphorescence.

The relationship between the phosphorescence lifetime and density of the quenching agent, which is oxygen, is given by the Stern-Volmer expression:

$$\frac{\tau_0}{\tau} = 1 + K_\varphi \tau_0 [pO_2], \tag{2}$$

where $[pO_2]$ $(mmHg)$ is the oxygen tension, K_φ is the quenching constant, and τ_0 is the lifetime in zero oxygen environments.

Considering (2) and for given constants K_φ and τ_0, τ first must be determined in order to find $[pO_2]$. The intensity depends on the phase angle between the modulated excitation laser light and the emitted phosphorescence, denoted by θ. Therefore, the lifetime can be calculated from the distribution of intensity values in different modulation phase images.

The relation between τ and θ, the phase angle of the phosphorescence, is given by;

$$\tan\theta = \omega\tau, \tag{3}$$

where ω is the modulation frequency. If the modulation frequency and modulation phase delay are respectively ω and θ_n then the integrated phosphorescence intensity at each pixel is:

$$I(\theta_n) = k[Pd]\left(1 + \frac{1}{2} m m_n \cos(\theta - \theta_n)\right) \quad n = 1, 2, \ldots S, \tag{4}$$

where k, $[Pd]$ and m are unknown and m_n is the known modulation profile. S is number of measurements at each pixel for different phase values of θ_n.

Using the trigonometric identity,

$$\cos(\theta - \theta_n) = \cos(\theta)\cos(\theta_n) + \sin(\theta)\sin(\theta_n), \tag{5}$$

the intensity values in (4) can be written as:

$$I(\theta_n) = k[Pd] + \frac{1}{2} k[Pd] m m_n (\cos(\theta)\cos(\theta_n) + \sin(\theta)\sin(\theta_n)) \tag{6}$$

Defining $a_0 = k[Pd]$, $a_1 = (1/2)k[Pd]mm_n \cos(\theta)$, and, $b_1 = (1/2)k[Pd]mm_n \sin(\theta)$, we can re-write (6) as:

$$I(\theta_n) = a_0 + a_1 \cos(\theta_n) + b_1 \sin(\theta_n) \tag{7}$$

from the equations above, the phase angle θ in (6) is obtained as:

$$\theta = \tan^{-1}(b_1/a_1) \tag{8}$$

Using Eqs. (2), (3) and (8), the oxygen tension values at each observation location can therefore be obtained from only a_1 and b_1.

In the absence of noise, the observed intensity vector for each pixel, by using (7), can be given as:

$$\begin{bmatrix} I_1 \\ \vdots \\ I_n \\ \vdots \\ I_S^i \end{bmatrix} = \begin{vmatrix} 1 \cos(\theta_1) \sin(\theta_1) \\ \vdots \\ 1 \cos(\theta_n) \sin(\theta_n) \\ \vdots \\ 1 \cos(\theta_S) \sin(\theta_S) \end{vmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \end{bmatrix}, \tag{9}$$

where $\theta_n = 2\pi(n-1)/S$, I_n is n-th phosphorescence intensity observation, and S denotes the number of observation per pixel.

In the presence of additive noise (9) becomes;

$$\begin{bmatrix} I_1 \\ \vdots \\ I_n \\ \vdots \\ I_S^i \end{bmatrix} = \begin{vmatrix} 1 \cos(\theta_1) \sin(\theta_1) \\ \vdots \\ 1 \cos(\theta_n) \sin(\theta_n) \\ \vdots \\ 1 \cos(\theta_S) \sin(\theta_S) \end{vmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \end{bmatrix} + \begin{bmatrix} N_1 \\ \vdots \\ N_n \\ \vdots \\ N_S \end{bmatrix} \tag{10}$$

The noise contaminated observation data in the phosphorescence lifetime images is modeled by (10). As can be seen explicitly, this estimation requires at least three observations for each location.

2.2 Estimation of Oxygen Tension from the LS Estimates of the Model Parameters

Considering the Eq. (2.10) for i-th pixel and defining $\boldsymbol{A}$, $\boldsymbol{y}^{\mathrm{i}}$, $\boldsymbol{n}^{\mathrm{i}}$ and $\boldsymbol{X}^{\mathrm{i}}$ as:

$$\begin{aligned} \boldsymbol{A} &= \begin{vmatrix} 1 \cos(\theta_1) \sin(\theta_1) \\ \vdots \\ 1 \cos(\theta_n) \sin(\theta_n) \\ \vdots \\ 1 \cos(\theta_S) \sin(\theta_S) \end{vmatrix}, \quad \boldsymbol{y^i} = \left[I_1^i \; \boldsymbol{I_2^i} \ldots I_S^i\right]^{\boldsymbol{T}}, \\ \boldsymbol{n^i} &= \left[N_1^i N_2^i \ldots N_S^i\right]^{\boldsymbol{T}}, \boldsymbol{x^i} = \left[a_0^i \; a_1^i \; b_1^i\right]^{\boldsymbol{T}}, \end{aligned} \tag{11}$$

we can re-write (10) as $\boldsymbol{y^i} = \boldsymbol{Ax^i} + \boldsymbol{n^i}$.

We model the additive noise as i.i.d. Gaussian and then the cost function can be given as:

$$C^i = \left\| \boldsymbol{y^i} - \boldsymbol{Ax}^{i2} \right\| \tag{12}$$

Using this cost function, the LS estimate of $\boldsymbol{x}$ can be given as:

$$\hat{\boldsymbol{x}}^i = \boldsymbol{Qy^i} = \boldsymbol{x^i} + \boldsymbol{Qn^i}, \tag{13}$$

where $\boldsymbol{Q}$ is the pseudo inverse matrix of $\boldsymbol{A}$, $\boldsymbol{Q} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$.

From the Eqs. (2), (3), (8) and (10), we see that assessment of oxygen tension values requires estimation of a_1 and b_1. $\hat{a}_1^i$ and $\hat{b}_1^i$ are respectively equal to $\hat{\boldsymbol{x}}^{\boldsymbol{i}}(2)$ and $\hat{\boldsymbol{x}}^{\boldsymbol{i}}(3)$ and using Eqs. (2), (3), (8) and (13) oxygen tension for the i-th pixel can be given as:

$$pO_2^i = \frac{1}{K_\varphi}\left(\frac{1}{\tau} - \frac{1}{\tau_0}\right) = \frac{1}{K_\varphi}\left(\frac{\omega\hat{a}_1^i}{\hat{b}_1^i} - \frac{1}{\tau_0}\right). \tag{14}$$

2.3 Regularization of the Least Squares Estimation

The RLS approach in [6] using phosphorescence lifetime imaging model will be briefly described here.

In [6], the RLS cost function for a_1 parameter was defined as follows:

$$f_{a_1}^i = (a_1^i - \hat{a}_1^i)^2 + \beta(a_1^i - a_1^{-i})^2, \tag{15}$$

where β is the regularization parameter, and $\overline{\mathrm{a}}_1^i$ denotes the mean value of the parameter to be estimated for the i-th pixel considering 3×3 neighborhood relation. $\hat{a}_1^i$ is the LS estimate of the parameter a_1 for i-th pixel defined by (6) and (7). Their assumption depends on that variation of oxygen tension values in a small neighborhood within the retinal arteries and veins should physiologically be minimal.

Additionally, it can be seen from the Eq. (15) that there is no pixel-wise solution. This is because for a pixel, regularization term in the cost function includes mean average of neighboring pixels' value. Therefore, the problem must be globally handled for all pixels in the oxygen tension map. The globalized cost functions for the parameters a_1 and b_1 are given as follows:

$$F_{a1} = \sum_{i=1}^{M} f_{a1}^i, \quad F_{b1} = \sum_{i=1}^{M} f_{b1}^i, \tag{16}$$

where M denotes number of pixels in the map.

In [6], a gradient-based iterative approach was used to find minimum of the global cost function.

3 Derivation of Closed Form Solutions for the Regularized Least Squares Estimation, its Bias and Variance

3.1 Closed Form Solution for the Regularized Least Squares Estimation

In the following expressions, phosphorescence intensity observations are shown in a vector form by reordering the matrix elements column wise. Traditionally, for the i-th pixel, the RLS cost function can be given as follows:

$$c_{RLS}^{i} = \left\|\boldsymbol{y}^{i} - \boldsymbol{A}\boldsymbol{x}^{i}\right\|_2^2 + \gamma\left\|\boldsymbol{x}^{i} - \bar{\boldsymbol{x}}^{i}\right\|_2^2, \tag{17}$$

where $\boldsymbol{y}^{i}$, γ and $\boldsymbol{A}$ stand for noise corrupted observation vector, regularization parameter and the system matrix, respectively. Additionally, the parameter to be estimated $\boldsymbol{x}^{i}$ and its mean are defined as follows:

$$\boldsymbol{x}^{i} = \left[a_0^{i}\ a_1^{i}\ b_1^{i}\right]^{T}, \quad \bar{\boldsymbol{x}}^{i} = \left[\boldsymbol{K}(i,:)\boldsymbol{a}_0\,\boldsymbol{K}(i,:)\boldsymbol{a}_1\,\boldsymbol{K}(i,:)\boldsymbol{b}_1\right]^{T}, \tag{18}$$

where $\boldsymbol{a}_0$, $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ are vectorized PLIM parameters and $\boldsymbol{K}$ is weighted mean matrix defining interrelations between $\boldsymbol{a}_0$, $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ parameters of the pixels (see Appendix). The global RLS cost function is written as:

$$C_{RLS} = \sum_{i=1}^{M} c_{RLS}^{i}, \tag{19}$$

where M is number of pixels in the oxygen tension map. Let $\boldsymbol{X} \in \Re^{IxJ}$, $\|\boldsymbol{X}\|_F$ is called as Frobenius norm in $\Re^{IxJ}$ space and defined as [13]:

$$\|\boldsymbol{X}\|_F = \sqrt{tr\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)} = \sqrt{tr\left(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}\right)} = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}}, \tag{20}$$

where tr denotes the trace of a matrix. The operation $tr\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)$ is called as Frobenius inner product [14]. Following that, the global cost function using Frobenius inner product can be given as follows:

$$C_{RLS} = \|\boldsymbol{Y} - \boldsymbol{AX}\|_F^2 + \gamma\|\boldsymbol{X} - \bar{\boldsymbol{X}}\|_F^2, \tag{21}$$

where $\boldsymbol{A}$ is the system matrix, γ is the regularization parameter and:

$$\boldsymbol{X} = \begin{vmatrix} a_0^1 \ldots a_0^i \ldots a_0^M \\ a_1^1 \ldots a_1^i \ldots a_1^M \\ b_1^1 \ldots b_1^i \ldots b_1^M \end{vmatrix} = [\boldsymbol{x}^1 \ldots \boldsymbol{x}^i \ldots \boldsymbol{x}^M] \text{ and } \boldsymbol{Y} = \begin{vmatrix} I_1^1 & \ldots & I_1^i & \ldots & I_1^M \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ I_p^1 & \ldots & I_p^i & \ldots & I_p^M \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ I_S^1 & \ldots & I_S^i & \ldots & I_S^M \end{vmatrix} \tag{22}$$

where I_p^i and S are respectively p-th phosphorescence intensity observation for i-th pixel and number of observation per pixel. Considering the definition of the $\boldsymbol{X}$, Eqs. (18) and (21) can be rewritten as follows:

$$\bar{\boldsymbol{x}}^i = (\boldsymbol{K}(i,:)\boldsymbol{X}^T)^T, \tag{23}$$

$$C_{RLS} = \|\boldsymbol{Y} - \boldsymbol{AX}\|_F^2 + \gamma\left\|\boldsymbol{X} - (\boldsymbol{KX}^T)^T\right\|_F^2. \tag{24}$$

For the PLIM parameters, the regularization term in the cost function (24) can be fragmented as follows:

$$\left\|\boldsymbol{X} - \left(\boldsymbol{KX}^T\right)^T\right\|_F^2 = \|\boldsymbol{a}_0 - \boldsymbol{Ka}_0\|_2^2 + \|\boldsymbol{a}_1 - \boldsymbol{Ka}_1\|_2^2 + \|\boldsymbol{b}_1 - \boldsymbol{Kb}_1\|_2^2 \tag{25}$$

The data fidelity term in Eq. (24) can be rewritten as:

$$\|\boldsymbol{Y} - \boldsymbol{AX}\|_F^2 = tr(\boldsymbol{Y}^T\boldsymbol{Y}) - 2tr(\boldsymbol{Y}^T\boldsymbol{AX}) + tr(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{AX}) \tag{26}$$

Since $\boldsymbol{A}^T\boldsymbol{A}$ is as follows:

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{vmatrix} S & 0 & 0 \\ 0 & S/2 & 0 \\ 0 & 0 & S/2 \end{vmatrix}, \tag{27}$$

where S is number of observation per pixel, $tr(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{AX})$ becomes:

$$tr\left(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{AX}\right) = S\boldsymbol{a}_0^T\boldsymbol{a}_0 + \frac{S}{2}\boldsymbol{a}_1^T\boldsymbol{a}_1 + \frac{S}{2}\boldsymbol{b}_1^T\boldsymbol{b}_1 \tag{28}$$

Additionally, from the definition of the Frobenius inner product, $tr(\boldsymbol{Y}^T\boldsymbol{AX})$ can be rewritten as follows:

$$tr\left(\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{X}\right) = tr\left(\boldsymbol{X}\boldsymbol{Y}^T\boldsymbol{A}\right) = \boldsymbol{a}_0^T\boldsymbol{Y}^T\boldsymbol{A}(:,1) + \boldsymbol{a}_1^T\boldsymbol{Y}^T\boldsymbol{A}(:,2) + \boldsymbol{b}_1^T\boldsymbol{Y}^T\boldsymbol{A}(:,3) \tag{29}$$

Considering the equations above, the global cost function can be given as:

$$\begin{aligned} C_{RLS} = tr\left(\boldsymbol{Y}^T\boldsymbol{Y}\right) + \boldsymbol{a}_0^T\left(S\boldsymbol{a}_0 - 2\boldsymbol{Y}^T\boldsymbol{A}(:,1)\right) + \boldsymbol{a}_1^T\left(\frac{S}{2}\boldsymbol{a}_1 - 2\boldsymbol{Y}^T\boldsymbol{A}(:,2)\right) \\ + \boldsymbol{b}_1^T\left(\frac{S}{2}\boldsymbol{b}_1 - 2\boldsymbol{Y}^T\boldsymbol{A}(:,3)\right) + \gamma\left(\|\boldsymbol{a}_0 - \boldsymbol{K}\boldsymbol{a}_0\|_2^2 + \|\boldsymbol{a}_1 - \boldsymbol{K}\boldsymbol{a}_1\|_2^2 + \|\boldsymbol{b}_1 - \boldsymbol{K}\boldsymbol{b}_1\|_2^2\right). \end{aligned} \tag{30}$$

After taking gradient of the cost function with respect to the parameter a_1 and equalizing the gradient to zero, we get the RLS estimate of the parameter a_1.

$$\nabla_{a_1} C_{RLS} = \left(S\boldsymbol{a}_1 - 2\boldsymbol{Y}^T\boldsymbol{A}(:,2)\right) + 2\gamma\left(\boldsymbol{I} - 2\boldsymbol{K} + \boldsymbol{K}^T\boldsymbol{K}\right)\boldsymbol{a}_1 = 0 \tag{31}$$

Assuming that $\gamma = S\beta/2$, we can rewrite Eq. (31) as follows:

$$\nabla_{a_1} C_{RLS} = \left(S\boldsymbol{a}_1 - 2\boldsymbol{Y}^T\boldsymbol{A}(:,2)\right) + S\beta\left(\boldsymbol{I} - 2\boldsymbol{K} + \boldsymbol{K}^2\right)\boldsymbol{a}_1 = 0,$$

and we get $\hat{\boldsymbol{a}}_{1-RLS}$ as:

$$\hat{\boldsymbol{a}}_{1.RLS} = \left(\boldsymbol{I} + \beta\left(\boldsymbol{I} - 2\boldsymbol{K} + \boldsymbol{K}^2\right)\right)^{-1}\left(\frac{2}{S}\boldsymbol{Y}^T\boldsymbol{A}(:,2)\right). \tag{32}$$

The same way can be followed for the parameter b_1:

$$\nabla_{b_1} C_{RLS} = \left(S\boldsymbol{b}_1 - 2\boldsymbol{Y}^T\boldsymbol{A}(:,3)\right) + S\beta\left(\boldsymbol{I} - 2\boldsymbol{K} + \boldsymbol{K}^2\right)\boldsymbol{b}_1 = 0 \tag{33}$$

Finally, we get $\hat{\boldsymbol{b}}_{1-RLS}$ as:

$$\hat{\boldsymbol{b}}_{1.RLS} = \left(\boldsymbol{I} + \beta\left(\boldsymbol{I} - 2\boldsymbol{K} + \boldsymbol{K}^2\right)\right)^{-1}\left(\frac{2}{S}\boldsymbol{Y}^T\boldsymbol{A}(:,3)\right) \tag{34}$$

Considering the definition of $\boldsymbol{A}^T\boldsymbol{A}$ and its inverse, it is explicit that $\left(\frac{2}{S}\boldsymbol{Y}^T\boldsymbol{A}(:,2)\right)$ and $\left(\frac{2}{S}\boldsymbol{Y}^T\boldsymbol{A}(:,3)\right)$ are the LS estimates of the a_1 and b_1 parameters, respectively. Therefore, we can rewrite the RLS estimates of a_1 and b_1 parameters as follows:

$$\hat{\boldsymbol{a}}_{1.RLS} = \left(\boldsymbol{I} + \beta\left(\boldsymbol{I} + \boldsymbol{K}^T\boldsymbol{K} - \boldsymbol{K} - \boldsymbol{K}^T\right)^{-1}\hat{\boldsymbol{a}}_{1-LS}, \tag{35}$$

$$\hat{\boldsymbol{b}}_{1.RLS} = \left(\boldsymbol{I} + \beta\left(\boldsymbol{I} + \boldsymbol{K}^T\boldsymbol{K} - \boldsymbol{K} - \boldsymbol{K}^T\right)^{-1}\hat{\boldsymbol{b}}_{1-LS}. \tag{36}$$

To abbreviate the notation, we define a new matrix $\boldsymbol{L}$ as:

$$\boldsymbol{L} = \boldsymbol{I} + \beta\left(\boldsymbol{I} + \boldsymbol{K}^T\boldsymbol{K} - \boldsymbol{K} - \boldsymbol{K}^T\right), \tag{37}$$

where, $\boldsymbol{I}$ stands for the identity matrix. Using the matrix $\boldsymbol{L}$, the RLS estimates of the parameters a_1 and b_1 can be rewritten in a simpler form as:

$$\hat{\boldsymbol{a}}_{1.RLS} = \boldsymbol{L}^{-1}\hat{\boldsymbol{a}}_{1-LS}, \tag{38}$$

$$\hat{\boldsymbol{b}}_{1.RLS} = \boldsymbol{L}^{-1}\hat{\boldsymbol{b}}_{1-LS}. \tag{39}$$

3.2 Bias of the Regularized Least Squares Estimation

It is well known that the LS estimator is unbiased. Following this fact and using (35), we can define the bias vector of the RLS estimator for a_1 as:

$$\boldsymbol{B}_{a_1} = \boldsymbol{a}_1 - E\{\hat{\boldsymbol{a}}_{1.RLS}\} = \boldsymbol{a}_1 - E\{\boldsymbol{L}\hat{\boldsymbol{a}}_{1.LS}\}, \tag{40}$$

$$\boldsymbol{B}_{a_1} = \boldsymbol{a}_1 - \boldsymbol{L}E\{\hat{\boldsymbol{a}}_{1.LS}\}. \tag{41}$$

Since $\mathrm{E}\{\hat{\boldsymbol{a}}_{1.LS}\} = \boldsymbol{a}_1$, the bias of the RLS estimator can be found as:

$$\boldsymbol{B}_{a_1} = \boldsymbol{a}_1 - \boldsymbol{L}\boldsymbol{a}_1\iota = (\boldsymbol{I} - \boldsymbol{L})\boldsymbol{a}_1. \tag{42}$$

3.3 Variances of the Least Squares and Regularized Least Squares Estimations

Oxygen tension depends on ratio of the model parameters b_1 and a_1. However, variance of this ratio cannot be found in our model due to the well-known fact that ratio of two Gaussian random variables does not have a defined variance value. In this regard, we only consider variances of $\mathrm{a_1}$ and $\mathrm{b_1}$ individually. Since we modeled the noise as i.i.d. Gaussian, covariance matrices of the a_0, a_1 and b_1 parameters are equal for each pixel. We define the LS estimate of the parameter vector as:

$$\boldsymbol{x}^i = \begin{bmatrix} \hat{a}_0^i \\ \hat{a}_1^i \\ \hat{b}_1^i \end{bmatrix} = \boldsymbol{Q}\boldsymbol{Y}(:, i), \tag{43}$$

where, i denotes pixel number under consideration. Covariance matrix of the observation vector is equal to the covariance matrix of the noise.

$$Cov\{\boldsymbol{Y}(:,i)\} = Var\left\{\begin{bmatrix} n_1^i \\ \vdots \\ n_S^i \end{bmatrix}\right\} = E\left\{\begin{bmatrix} n_1^i \\ \vdots \\ n_S^i \end{bmatrix}\begin{bmatrix} n_1^i & \dots & n_S^i \end{bmatrix}\right\}$$
$$= \begin{bmatrix} \sigma_n^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{bmatrix}_{SxS} \tag{44}$$

$$Cov\{\boldsymbol{x}^i\} = E\{\boldsymbol{QY}(:,i)(\boldsymbol{QY}(:,i))^T\} = \boldsymbol{Q}Var\{\boldsymbol{Y}(:,i)\}\boldsymbol{Q}^T \tag{45}$$

Since $\{\boldsymbol{Y}(:,i)\} = \sigma_n^2\boldsymbol{I}$,

$$Cov\{\boldsymbol{x}^i\} = \sigma_n^2\boldsymbol{Q}\boldsymbol{Q}^T = \sigma_n^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-T} = \sigma_n^2(\boldsymbol{A}^T\boldsymbol{A})^{-T} = \sigma_n^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}$$
$$= \sigma_n^2\begin{bmatrix} 1/S & 0 & 0 \\ 0 & 2/S & 0 \\ 0 & 0 & 2/S \end{bmatrix} \tag{46}$$

where S denotes number of phosphorescence intensity observation per pixel. Considering the parameters of i-th pixel, as can be seen from $Cov\{\boldsymbol{x}^i\}$, cross-covariances of $\hat{a}_0^i$, $\hat{a}_1^i$ and $\hat{b}_1^i$ are equal to zero and their auto-covariances are σ_n^2/S, $2\sigma_n^2/S$ and $2\sigma_n^2/S$ respectively. Since there is no relationship between pixels in the LS estimation, auto-covariance matrix of $\hat{\boldsymbol{a}}_{1-LS}$ can be given as:

$$Cov\{\hat{\boldsymbol{a}}_{1-LS}\} = (2\sigma_n^2/S)\boldsymbol{I}_{NxN}. \tag{47}$$

As shown above, $\hat{\boldsymbol{a}}_{1-RLS}$ is equal to $\boldsymbol{L}\hat{\boldsymbol{a}}_{1-LS}$. Using this, we can write variance of $\hat{\boldsymbol{a}}_{1-RLS}$ as:

$$Cov\{\hat{\boldsymbol{a}}_{1-RLS}\} = \boldsymbol{L}Var\{\hat{\boldsymbol{a}}_{1-LS}\}\boldsymbol{L}^T = (2\sigma_n^2/S)\boldsymbol{L}\boldsymbol{L}^T. \tag{48}$$

4 Results

4.1 Simulation Data

To generate the simulated data we used in this work, physiological features and topology of real retinal vessels were followed strictly based on the previous studies [12]. i.i.d. white Gaussian noise, for different SNR scenarios, was added to phosphorescence intensity observations. The performance comparison of the LS and RLS methods were made with bias, variance and mean absolute error (MAE)

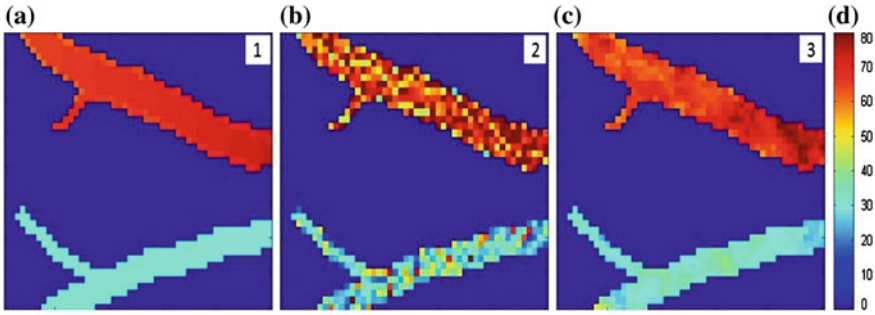


Fig. 1 Original simulated oxygen tension map (40 × 50 pixels) (**a**) and its estimates in the presence of noise with 20 dB signal to noise ratio using the LS (**b**), iterative regularized least squares (**c**) closed form regularized least squares. (**d**) *Color bar* represents oxygen tension values in millimeters of mercury

values. Performance of the RLS method is also examined for different values of the regularization parameters such as regularization coefficient, window size, and window weighting coefficients. Figure 1 shows the simulated oxygen tension map (a), and its estimates in the presence of 20 dB noise by the LS (b) and RLS (c) estimation methods, respectively. Unless otherwise is stated, pre-set values of the phosphorescence intensity observation number, window size and SNR are 10, 3 × 3 and 20 dB, respectively.

For 15 and 20 dB SNR values and for different regularization coefficient values, MAE performance of the LS and RLS estimation methods are given in Fig. 2. As regularization coefficient increases, MAE performance of the RLS estimators gets better as expected. However, if regularization coefficient is chosen too big, the RLS estimates suffer from over smoothing and local variation is suppressed extensively. In addition, increasing values of the regularization coefficient results in larger bias values in the estimates. Therefore, the regularization coefficient must be selected carefully.

As can be seen from the Fig. 2, there is no substantial difference in MAE performances of the iterative and closed form RLS methods as expected. Therefore, we from now on will not include results of the iterative RLS in the comparisons.

Monte-Carlo simulations had to be conducted before to compare variance and bias performances of the LS and RLS estimators. With the help of the closed form solution derived in this work, we are able to acquire variance and bias values of the RLS estimator analytically. It must be noted that relative variances shown in graphs are proportional to noise variance to facilitate visualization and have no dimension. Figure 3 shows relative variances and normalized biases of the LS, and closed form RLS estimators for different regularization parameter values.

Increasing values of the regularization parameter helps in decreasing variance of the RLS estimator as expected. Variance of the LS estimator relative to the

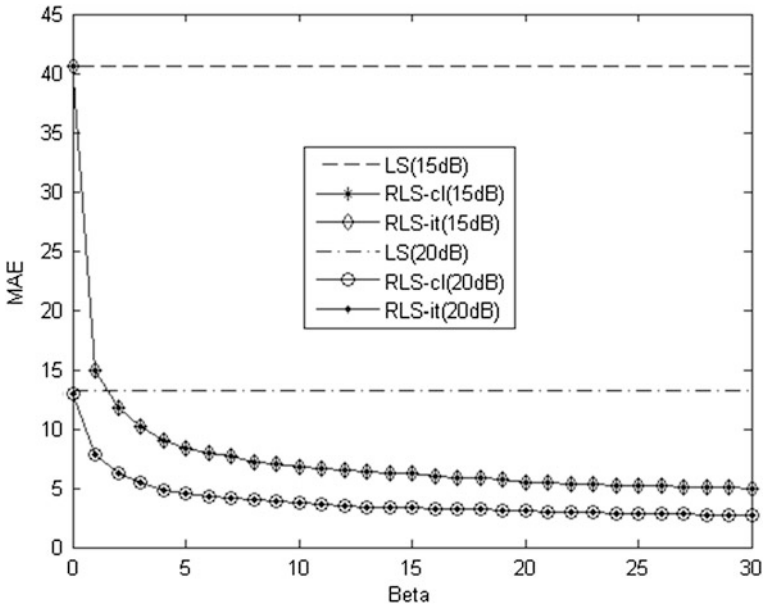


Fig. 2 Mean absolute errors (MAE) of the LS, iterative RLS (RLS-it) and closed form RLS (RLS-cl) methods for different regularization coefficient values and 15 and 20 dB SNR values. Mean absolute error values of oxygen tension values are in millimeters of mercury

noise variance is 0.2 whereas the relative variance of the RLS estimator gets lower than 0.025. As shown in the previous Section, bias of the RLS estimation method was given for every pixel in a vector form in Eq. (42). To compare bias performance of the RLS estimation method for different values of the regularization parameter, we used a normalized bias defined as follows:

$$Bias = \left(\sum_{i=1}^{N} |\boldsymbol{B}_{a_1}(i)| \right) \Big/ \left(\sum_{i=1}^{N} |\boldsymbol{a}_1(i)| \right). \tag{49}$$

Increasing values of the regularization parameter results in larger bias as expected whereas helps to obtain smaller variance values for the RLS estimates. Therefore, the regularization parameter cannot be selected arbitrarily large. As can be seen from Fig. 3, bias performance of the RLS method decreases rapidly for both regularization window sizes, whereas there is a relatively slow improvement in the variance performance when the regularization parameter takes values bigger than 5. Additionally, between the two different regularization window sizes, there is a slight superiority of 5×5 window size over 3×3 window size considering the variance performance of the RLS estimation method. However, the 3×3 window performs considerably better than 5×5 window size in the bias sense.

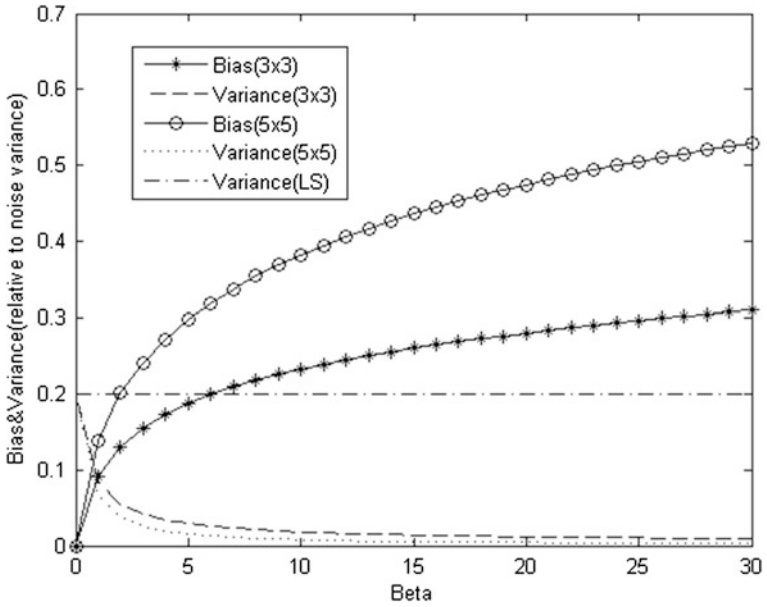


Fig. 3 Bias and variance of the LS and RLS estimation methods for 5 × 5 and 3 × 3 window sizes and different regularization parameter values

Therefore, in the sense of bias-variance performance of the RLS estimation method, 3 × 3 window size is more preferable than 5 × 5 window size.

As an experimental set up parameter, the phosphorescence intensity observation number is a key factor in the performance of the estimators. As mentioned in Sect. 2, at least three observations for three different gain modulation phases must be obtained to estimate the model parameters a_0, a_1 and b_1. As the observation number increases, performances of the LS and RLS estimation methods increase as shown in Figs. 4 and 5 with the expense of increased experimental cost. Therefore, it cannot be chosen arbitrarily large.

We also examined performance of the RLS estimation method for different values of weighting coefficients in the regularization window (see Appendix). In Figs. 6 and 7 (1) and (2) denote the RLS estimates for two different regularization windows of *l, p* and *q* coefficients: (1) $l = 2p = 2\sqrt{2}q$ (2) l = p = q. Between two types of windows (1) is more preferable over (2) when we compare their MAE, bias and variance results shown in Figs. 6 and 7.

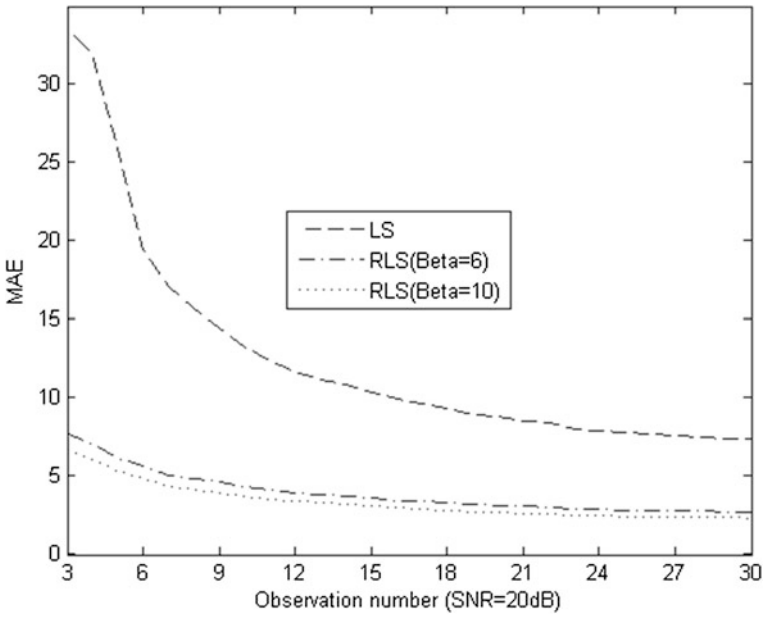


Fig. 4 MAE of the LS and RLS estimation methods for different phosphorescence intensity observation numbers. MAE values are of oxygen tension values in millimeters of mercury

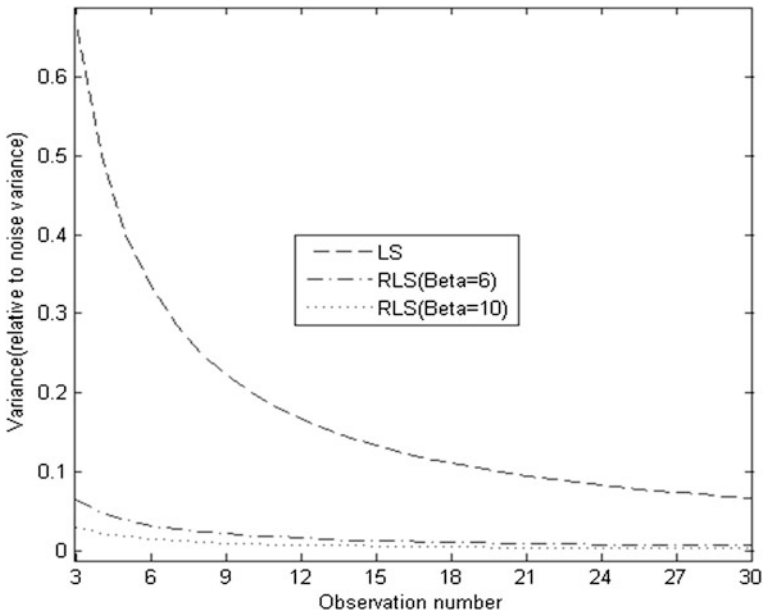


Fig. 5 Variances of the LS and RLS estimation methods considering different phosphorescence intensity observation numbers

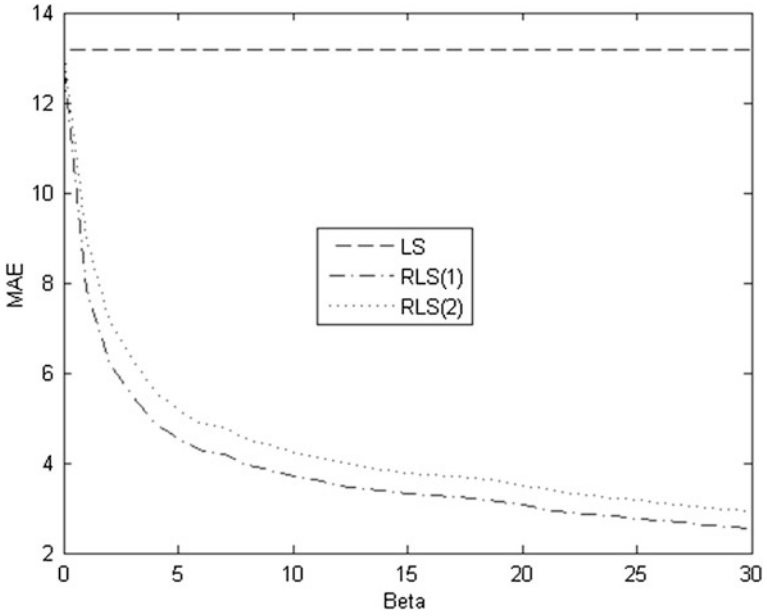


Fig. 6 MAE of the LS and RLS estimation methods for different regularization window weighting coefficients and regularization parameter values. MAE values are of oxygen tension values in millimeters of mercury

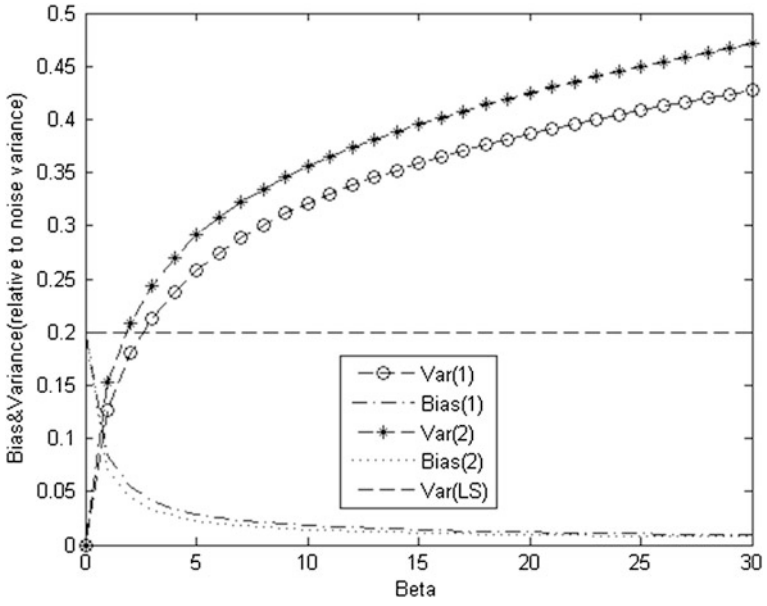


Fig. 7 Bias and variance performances of the LS and RLS estimation methods for different regularization window weighting coefficients and regularization parameter values

5 Conclusion and Discussions

In this study, performance analyses for different values of the RLS estimation parameters are carried out with the help of the closed form solution derived for the RLS estimation. These analyses are helpful for the further enhancements on the RLS estimation method by giving detailed information about estimation parameters' effects on the estimation performance. Moreover, the results of this study can be applied to other imaging problems, which use PLIM or FLIM; given the neighborhood information existing in retinal oxygenation problem is present.

Appendix

For 3×3 regularization window size, $\boldsymbol{K}$ is formed as follows:

First, we assume that the regularization window has coefficients as:

$$\begin{vmatrix} q & p & q \\ p & l & p \\ q & p & q \end{vmatrix}, = \frac{a}{4 * (b + c) + a}, \quad p = \frac{b}{4 * (b + c) + a}, \quad q = \frac{c}{4 * (b + c) + a} \tag{A.1}$$

where l, p and q denote weight of pixel to itself, to direct adjacent pixels and to cross adjacent pixels, respectively. In order to have mean of the regularization window coefficients be one, we normalize these coefficients.

After defining l, p and q, we form the $\boldsymbol{K}$ as follows:

$$\boldsymbol{K}(j,k) = \left\{ \begin{array}{l} l \;\; if \;\; j = k \\ p \;\; if \;\; j = k \pm 1 \\ p \;\; if \;\; j = k \pm M \\ q \;\; if \;\; j = k + M \pm 1 \\ q \;\; if \;\; j = k - M \pm 1 \\ 0 \;\; otherwise \end{array} \right\}, \tag{A.2}$$

where $\boldsymbol{K}(j, k)$ denotes weight coefficient of the k-th pixel on j-th pixel in the image, and M denotes the number of rows. For the regularization window 5×5 size, the same approach described above can be followed.

References

1. V.A. Alder, E.N. Su, D.Y. Yu, S.J. Cringle, P.K. Yu, Diabetic retinopathy: Early functional changes, Clin. Exp. Pharmacol. Physiol. **24**, 785–788 (1997)
2. B.A. Berkowitz, R.A. Kowluru, R.N. Frank, T.S. Kern, T.C. Hohman, M. Prakash, Subnormal retinal oxygenation response precedes diabetic-like retinopathy. Invest. Ophthalmol. Visual Sci. **40**, 2100–2105 (1999)

3. M. Shahidi, A. Shakoor, R.D. Shonat, M. Mori, N.P. Blair, A method formeasurement of chorio-retinal oxygen tension, Curr. Eye Res. **31**, 357–366 (2006)
4. R.D. Shonat, A.C. Kight, Oxygen tension imaging in the mouse retina. Ann. Biomed. Eng. **31**, 1084–1096 (2003)
5. J.R. Lakowicz, H. Szmacinski, K. Nowaczyk, K.W. Berndt, M. Johnson, Fluorescence lifetime imaging, Anal. Biochem. **202**, 316–330 (1992)
6. I. Yildirim, R. Ansari, J. Wanek, I.S. Yetik, M. Shahidi, Regularized estimation of retinal vascular oxygen tension from phosphorescence images. IEEE Trans. Biomed. Eng. **56**, 1989–1995 (2009)
7. J. Liu, P. Moulin, Complexity-regularized image restoration. Proc. IEEE Int. Conf. Image Process. **1**, 555–559 (1998)
8. R.C. Hardie, K.J. Barnard, E.E. Armstrong, Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. IEEE Trans. Image Process. **6**, 1621–1633 (1997)
9. W. Zhu, Y. Wang, Y. Deng, Y. Yao, R.L. Barbour, A wavelet-based multiresolution regularized least squares reconstruction approach for optical tomography. IEEE Trans. Med. Imaging **16**, 210–217 (1997)
10. X. He, L. Cheng, J.A. Fessler, E.C. Frey, Regularized image reconstruction algorithms for dual-isotope myocardial perfusion SPECT (MPS) imaging using a cross-tracer prior. IEEE Trans. Med. Imaging **30**, 1169–1183 (2011)
11. R.A. Frazin, Tomography of the solar corona, a robust, regularized, positive estimation method. Astrophys. J. **530**, 1026–1035 (2000)
12. A. Shakoor, N.P. Blair, M. Mori, M. Shahidi, Chorioretinal vascular oxygen tension changes in response to light flicker. Invest. Ophthalmol. Visual Sci. **47**, 4962–4965 (2006)
13. G. Gunay, I. Yildirim, Statistical Performance Analysis of the RLS Estimator for Retinal Oxygen Tension Estimation. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE, London, UK, 3–5 July 2013, pp. 2190–2194
14. R.A. Horn, C.R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 2012). ISBN 978-0-521-38632-6

Identification of Multistorey Building's Thermal Performance Based on Exponential Filtering

Vildan V. Abdullin, Dmitry A. Shnayder and Lev S. Kazarinov

Abstract This work examines the identification of thermal performance of a multistorey building, based on experimental data available for direct measurement. The authors suggest a new model structure with a reduced number of parameters. An identification method based on building an inverse dynamics model that uses exponential filtering is considered. The method makes it possible to estimate signals that cannot be measured directly: the signal of the general perturbation of the indoor air temperature and the signal of specific heat loss through the building envelope. Two examples are given of identifying the thermal performance of a building model: the one based on simulated test data and another one based on real measured data. The identification method proposed in the article puts in a strong performance in both the simulated data model as well as real-data model and may be used in engineering calculations for designing automatic control systems and in predictive control algorithms for heating buildings.

Keywords Building thermal conditions model · Case study · Dynamics model · Exponential filtering · Heating of buildings · Identification · Inverse dynamics operator

V. V. Abdullin (✉) · D. A. Shnayder · L. S. Kazarinov
Automatics and Control Department, South Ural State University, Lenina pr., 76, Chelyabinsk, Russia 454080
e-mail: vildan@ait.susu.ac.ru

D. A. Shnayder
e-mail: shnayder@ait.susu.ac.ru

L. S. Kazarinov
e-mail: kazarinov@ait.susu.ac.ru

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_36,

1 Introduction

One of the main objectives in the development of urban engineering infrastructure in countries with moderate climates is to improve the energy efficiency of building heating systems [1, 2]. The modern approach to saving thermal energy when heating buildings and increasing the comfort of building users assumes the introduction of automatic control systems that use model predictive control methods [3−6] in addition to simple arithmetic algorithms [7]. Another important issue is the development and identification of a mathematical model for building heating parameters [8–11].

The indoor air temperature T_{ind} of a building depends on its volume, building envelope type, the quantity of applied thermal energy Q_{source}, inner and external perturbing factors such as the outdoor air temperature T_{out}, solar radiation J_{rad}, wind V_{wind}, internal heat release Q_{int}, and the building's accumulated internal thermal energy Q_{acc} (Fig. 1).

However, the signals T_{ind}, Q_{source}, and T_{out} presented in Fig. 1 can be measured quite easily in practice, while direct measurement of J_{rad}, V_{wind}, Q_{int}, and Q_{acc}, which affect the temperature T_{ind}, is actually problematic.

Furthermore, it should be noted that the processes of heat transfer are distributed and generally described by partial differential equations [12]. However, these equations are not convenient for use in the identification process in their given form, because they contain a large number of parameters which are very difficult to determine in practice. The following is a method to identify the thermal characteristics of a building, based on a reduced set of experimental data, which makes it suitable for practical use.

2 A Method to Identify Model of a Building Heating System

The initial (empirical) data for modeling the thermal performance of a building include the heating power applied to the building, the outdoor air temperature, and the indoor air temperature. The indoor air temperature T_{ind} of a building, which is the average value of indoor temperatures in each room, accounting for differences in the area, is calculated as follows:

$$T_{\text{ind}}(t) = \frac{\sum\limits_i S_i \cdot T_{\text{ind}\,i}(t)}{\sum\limits_i S_i}, \tag{1}$$

where S_i, $T_{\text{ind}\,i}$ stand for the area and temperature of the i-th room, respectively, t is time. Using the average temperature T_{ind} permits us to estimate relatively fast perturbations—such as wind, solar radiation, or local heat sources, which affect the thermal performance of some rooms, for example, the rooms of one side of the

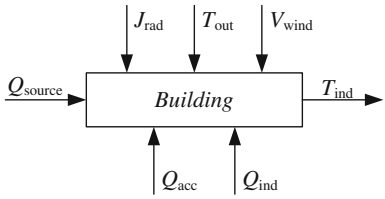


Fig. 1 Factors affecting the indoor air temperature

building—for the entire building. We can then assume that the response time of the model's output signal (indoor air temperature) to these perturbations is comparable to the time constants of the relatively slow processes of heat accumulation and heat loss through a building envelope with high heat capacity.

Consequently, the concept of a general temperature perturbation, T_z, may be introduced, characterizing the effect of the factors mentioned above on the indoor air temperature. Therefore, the heat balance equation takes the following form:

$$T_{ind}{}^{*}(t) = \frac{Q_0(t)}{q_0 \cdot V} + T_{ind}(t) - T_z(t), \tag{2}$$

where $T_{ind}{}^{*}(t)$ stands for the predicted value of the indoor air temperature (the prediction horizon is determined by the fluctuation of the indoor air temperature as a result of the perturbing factors (Fig. 1); $T^{*}_{out}(t)$ is the outdoor air temperature; $Q_0(t)$ stands for the heating power applied to the heating system; q_0 represents the specific heat loss (per cubic meter); and V is the external volume of the building.

Let us assume that the behavior of the indoor air temperature $T^{*}_{ind}(t)$ is described by a linear dynamic operator with a pure delay given by:

$$L_0(p) = \frac{\sum b_j p^j}{\sum a_i p^i} \cdot e^{-p\tau_d}, \tag{3}$$

where τ_d is the pure delay time.

A block diagram of a building thermal performance dynamics model composed in accordance with Eqs. (2) and (3) is presented in Fig. 2.

In the model presented in Fig. 1, we will consider the values $Q_0(t), T_{out}(t), T_{ind}(t)$ to be known, because they can be directly measured in practice. The unknown values are:

- polynomial coefficients a_i and b_j (3);
- delay time constant τ_d;
- building's specific heat loss q_0;
- general temperature perturbation $T_z(t)$;
- predicted value of the indoor air temperature $T_{ind}{}^{*}(t)$

The values of a_i, b_j, and τ_d can be determined from the building's response to a stepwise change in the heating power $Q_0(t)$ using well-known methods, for example, *Matlab's Ident* toolbox (*MathWorks, Inc.*, USA). However, to reduce the

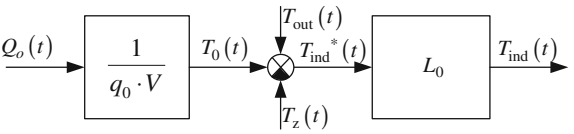


Fig. 2 Block diagram of the building's thermal performance dynamics model

effect of the perturbing factors $T_{out}^*(t)$ and $T_z(t)$ a series of experiments is conducted, and a_i, b_j, and τ_d are calculated according to the following equations:

$$\tau_d = \frac{1}{N}\sum_{k=1}^{N}\tau_{d\,k}, \quad a_i = \frac{1}{N}\sum_{k=1}^{N}a_{i,k}, \quad b_j = \frac{1}{N}\sum_{k=1}^{N}b_{j,k}, \tag{4}$$

where N is the number of experiments; k stands for a sequence number of an experiment; and $a_{i,k}$, $b_{j,k}$, and $\tau_{d\,k}$ represent the values obtained during the experiment.

Next, let us assume that $T_{ind}^*(t)$ and $T_{ind}(t)$ are statistically unbiased signals and that $T_z(t)$ is a signal with a mean of zero, then:

$$M_t\{T_{ind}^*(t)\} = M_t\{T_{ind}(t)\}, \tag{5}$$

$$M_t\{T_z(t)\} = 0, \tag{6}$$

where $M_t\{\bullet\}$ is the time-mean operator.

From Eqs. (2), (5), and (6), it follows that the specific heat loss through the building envelope can be calculated using:

$$q_0 = \frac{M_t\{Q_0(t)\}}{V(M_t\{T_{out}(t)\} - M_t\{T_{ind}(t)\})}. \tag{7}$$

It is evident from (2) that the general perturbation T_z can be determined by:

$$T_z(t) = \frac{Q_0(t)}{q_0 \cdot V} + T_{out}(t) - T_{ind}^*(t), \tag{8}$$

where $T_{ind}^*(t)$ is the predicted indoor air temperature.

According to the model presented in Fig. 1, the predicted indoor air temperature can be determined by the following equation:

$$T_{ind}^*(t) = L_0^{-1}\{T_{ind}(t)\}, \tag{9}$$

where $L_0^{-1}\{\bullet\}$ stands for the inverse dynamics operator.

From (3), the operator's formal inverse is:

$$L_0^{-1}(p) = \left(\left(\sum_{i=0}^{n} a_i p^i\right) \Big/ \left(\sum_{j=0}^{m} b_j p^j\right)\right) e^{p\tau_3}. \tag{10}$$

A block diagram of the operator's formal inverse is presented in Fig. 3.

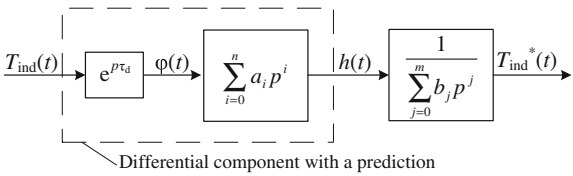


Fig. 3 Block diagram of the operator's formal inverse

Let us consider that constructing the dynamics operator $L_0^{-1}\{\bullet\}$ is based on the exponential filtering method [13, 14].

Let a signal decomposition in polynomial basis be given as follows:

$$T_{ind}(t-\lambda) \approx \sum_{i=0}^{n} g_i(t)\lambda^i, \tag{11}$$

where τ is the retrospective interval, and $g_i(t)$ stands for the decomposition's spectral components.

Considering a prediction at time τ_d (11) becomes:

$$T_{ind}(t-(\lambda-\tau_d)) \approx \sum_{i=0}^{n} g_i(t)(\lambda-\tau_d)^i. \tag{12}$$

According to the Newton binomial, we then obtain

$$(\lambda-\tau_d)^i = \sum_{k=0}^{i} c_k \lambda^k \tau_d^{i-k}(-1)^{i-k}, \tag{13}$$

where c_k stands for the binomial coefficients. Substituting (13) in (12), we get the relationship for a signal:

$$T_{ind}(t-(\lambda-\tau_d)) \approx \sum_{i=0}^{n} g_i(t) \sum_{k=0}^{i} c_k \lambda^k \tau_d^{i-k}(-1)^{i-k}, \tag{14}$$

which accounts for the prediction at time τ_d.

Then we decompose the signal $\varphi(t)$ in the polynomial basis:

$$\varphi(t-\lambda) \approx \sum_{i=0}^{n} g_i(t)\lambda^i. \tag{15}$$

Let us consider the decomposition of the signal $\varphi(t)$ into the Taylor series:

$$\varphi(t-\lambda) \approx \sum_{i=0}^{n} (-1)^i \frac{\varphi^{(i)}(t)}{i!}\lambda^i. \tag{16}$$

Comparing the expressions (15) and (16) yields the following equation:

$$\varphi^{(i)}(t) = (-1)^i i! g_i(t). \tag{17}$$

Equation (17) shows the relationship between the i-th derivative of the input signal $\varphi^{(i)}(t)$ and the corresponding spectral component $g_i(t)$.

Hence, the output of the filter's differential part, without accounting for the predictive component τ_d, will become:

$$h(t) = \sum_{i=0}^{n} a_i \varphi^{(i)}(t) = \sum_{i=0}^{n} (-1)^i i! a_i g_i(t). \tag{18}$$

Comparing the expressions (15) and (18), we conclude that obtaining an expression for the signal $h(t)$ requires the following substitution in (14):

$$\lambda^i \Rightarrow (-1)^i i! a_i. \tag{19}$$

By applying (19) to (14), we determine the output of the differential component of the predictive filter:

$$h(t) \approx \sum_{i=0}^{n} g_i(t) \sum_{k=0}^{i} c_k \tau_d^{i-k} (-1)^{i-k} (-1)^k k! a_k = \sum_{i=0}^{n} (-1)^i g_i(t) \sum_{k=0}^{i} k! c_k a_k \tau_d^{i-k}. \tag{20}$$

As a result, we obtain the inverse operator structure given in Fig. 4. Here ϕ_{inp} stands for the exponential filter of input signal moments; P^{-1} is the inverse correlation coefficient matrix; A is the coefficient matrix for the differential component of the inverse operator; $\mathbf{\mu}(t) = \{\mu_0(t), \mu_1(t), \ldots, \mu_n(t)\}^T$ represents the vector of input signals moments; and $\mathbf{g}(t) = \{g_0(t), g_1(t), \ldots, g_n(t)\}^T$ stands for a vector of the decomposition's coordinate functions.

Signal projections $\{g_i(t)\}$ are determined based on the criterion of minimum exponential mean error in the input signal (11):

$$E^2(t) = \frac{1}{T} \int_0^{\infty} \left[T_{ind}(t-\lambda) - \sum g_i(t) \lambda^i \right]^2 e^{-\frac{\lambda}{T}} d\lambda, \tag{21}$$

where T is the time constant of the averaging filter.

The solution is based on the minimum of function (21) along projections of $g_i(t)$:

$$\frac{\partial E^2(t)}{\partial g_i} = 0, \quad i = \overline{0, n}. \tag{22}$$

The solution to problem (22) is a system of recurrence relations [15]

$$\left.\begin{array}{c} \mu_{0,k} = \frac{1}{1+(\Delta t/T)} \left(\mu_{0,k-1} + \frac{\Delta t}{T} T_{\text{ind}\,k}\right); \\ \mu_{i,k} = \frac{1}{1+(\Delta t/T)} \left(\mu_{i,k-1} + i\Delta t \mu_{i-1,k}\right); \\ \mathbf{g}_k = P\mathbf{\mu}_k, \end{array}\right\} \tag{23}$$

where $T_{\text{ind},k}$ stands for the input signal at time t_k; Δt is the time sampling interval, and $P = Q^{-1}$ represents a matrix of constant coefficient determined from the following relationships:

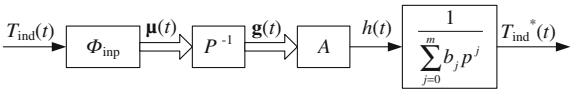


Fig. 4 Inverse operator structure

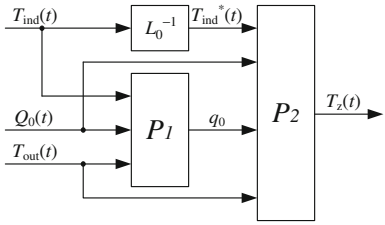


Fig. 5 Block diagram of a real-time identification system for signals q_0 and $T_z(t)$

$$Q = ||q_{i,j}||, \quad q_{i,j} = T^{(i+j)}(i+j)! \tag{24}$$

The block diagram of the identification system operating in real-time is presented in Fig. 5. Here P_1 is described by (7), and P_2 is described by (8).

Thus, the proposed method results in real-time identification of the unknown signals q_0 and $T_z(t)$ using the measured values of $Q_0(t)$, $T_{out}(t)$ and $T_{ind}(t)$.

3 A Simulated Example

Let us consider an example of identifying the thermal characteristics of a building based on the proposed method using *VisSim* visual simulation software (*Visual Solutions, Inc.*, USA).

Let us assume that operator $L_0(p)$ is as follows:

$$L_0(P) = \frac{1}{(1+pT_1)\cdot(1+pT_2)} \cdot e^{-p\tau_d}. \tag{25}$$

Let the parameters for the model presented in Fig. 1 take the following values: $V = 7{,}000$ m^3, $q_0 = 0.48$ W/(m^3 °C), $T_1 = 6$ h, $T_2 = 3$ h, $\tau_d = 2$ h.

Considering the cyclic nature of changes in outdoor air temperature, heat power, and the perturbing factors, let us use the harmonic test signals in Fig. 6 as the model's input signals $T_{out}(t)$ and $Q_0(t)$ as well as signal $T_z(t)$, which will be defined later in the identification process.

The graph of the corresponding variation in indoor air temperature $T_{ind}(t)$ for the model given in Fig. 2 with dynamics operator (25) is presented in Fig. 9 (solid line).

The dynamics operator (25) is inverted based on the structure of the inverse operator presented in Fig. 4. The target signals q_0 and $T_z(t)$ are calculated according to the identification system's block diagram, presented in Fig. 5.

Figures 7, 8, 9 and 10 present the modeling results. Figure 7 shows graphs of the source- and calculated signals of specific heat loss of the building. As you can see from the graph, the estimation error for signal q_0 does not exceed ±1 %.

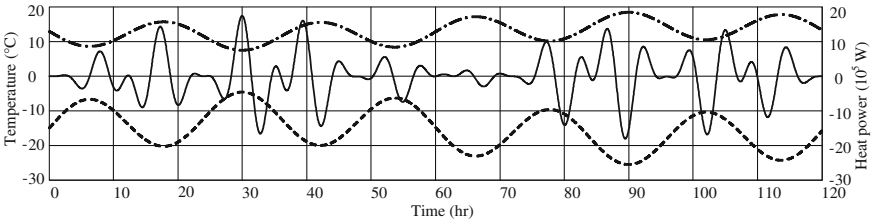


Fig. 6 Input signals. *Dash-dotted line* stands for $Q_0(t)$ [W]; *dashed line* stands for $T_{out}(t)$ [°C]; *solid line* stands for $T_z(t)$ [°C]

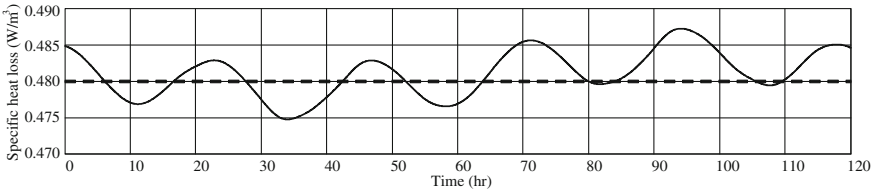


Fig. 7 Specific heat loss. *Dashed line* stands for actual value (average); *solid line* stands for predicted value

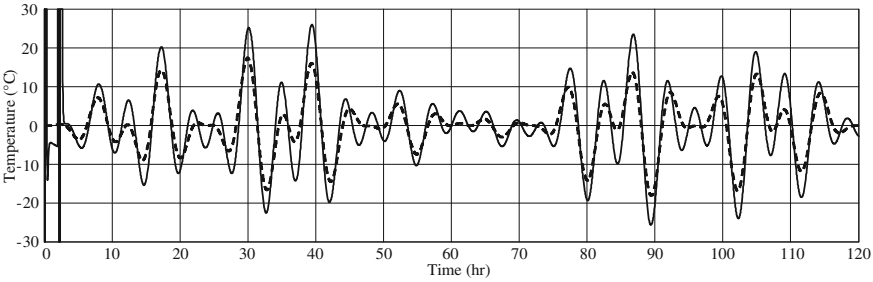


Fig. 8 General temperature perturbation. *Dashed line* stands for actual value; *solid line* stands for predicted value

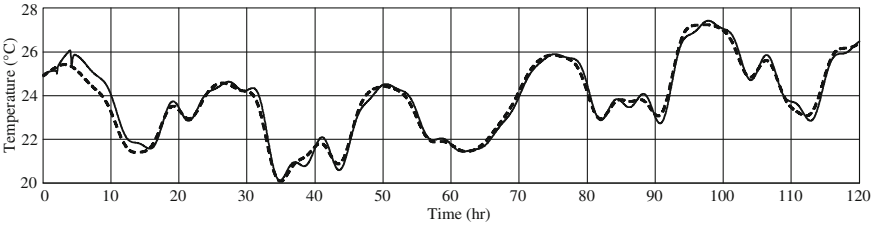


Fig. 9 Indoor air temperature. *Dashed line* stands for actual value; *solid line* stands for predicted value

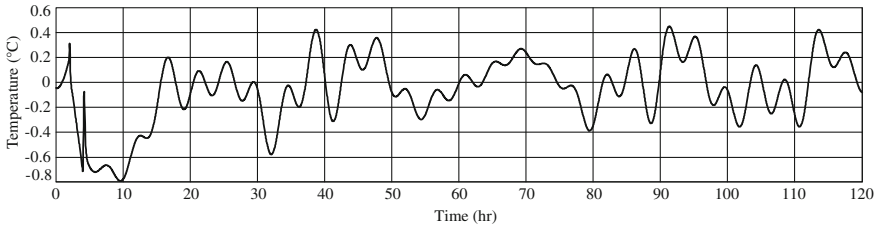


Fig. 10 Estimation error for indoor air temperature

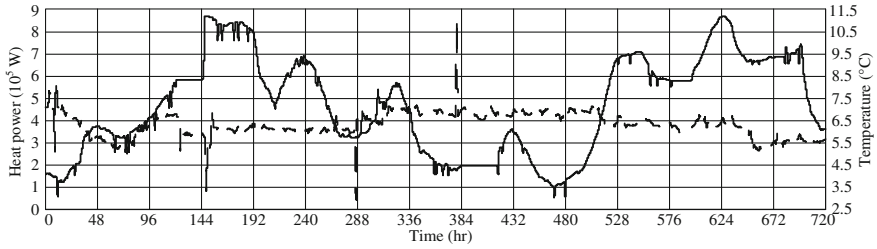


Fig. 11 Input signals. *Dashed line* stands for $Q_0(t)$ [W]; *solid line* stands for $T_{out}(t)$ [°C]

Figures 8 and 9 show similar graphs for the general temperature perturbation and the indoor air temperature. Figure 10 presents a graph of the estimation error for the indoor air temperature. As can be seen from the graph, the estimation error is about ±0.5 °C.

4 A Case Study

The next step to verify the obtained method is to perform the identification process using real data, taken from an existing building.

As a testbed we use a 10-storey academic building of the South Ural State University with outer walls made up of reinforced concrete large blocks. The building parameters are $V = 50{,}000$ m^3, $q_0 = 0.38$ W/(m^3 °C).

As for the model input parameters, we use measured values of $Q_0(t)$ and $T_{out}(t)$, taken from the building heat meter and the outdoor temperature sensor, respectively (Fig. 11). These metering devices are typically the parts of automatic heat supply units.

To obtain the third input parameter, i.e. indoor air temperature $T_{ind}(t)$ signal, we collect data from numerous temperature sensors deployed in different rooms of the building and calculate the average indoor temperature value, eliminating invalid data. The graph of the corresponding variation in average indoor air temperature

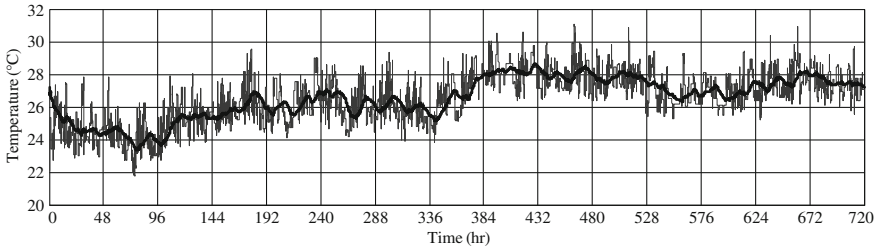


Fig. 12 Indoor air temperature average value. *Thin grey line* stands for measured value; *thick black line* stands for filtered data

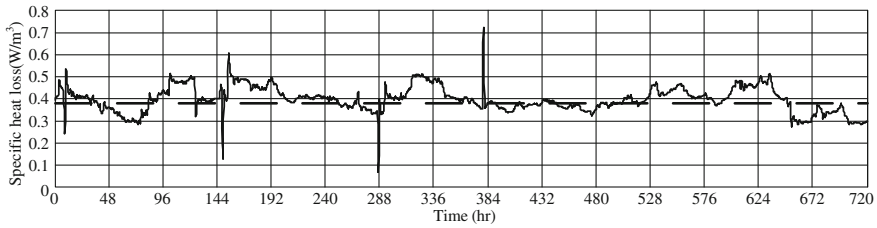


Fig. 13 Specific heat loss. *Dashed line* stands for actual value (average); *solid line* stands for predicted value

$T_{\text{ind}}(t)$ is presented in Fig. 12 (thin grey line). The average temperature is then filtered to distinguish the main trend of signal variation (Fig. 12, thick black line), which reflects the influence of $Q_0(t)$ and $T_{\text{out}}(t)$, at the same time eliminating signal noise. The filtering is performed using exponential filter in harmonic basis, though any appropriate filter algorithm with low delay value may be applied. We use *Dallas DS1921* temperature loggers as indoor temperature sensors but for commercial applications we recommend to use WSN-based sensors [16].

The target signals q_0 and $T_z(t)$ are calculated according to the identification system's block diagram as described in Sect. 3. The specific heat loss signal is presented in Fig. 13. As you can see from the graph, the fluctuations of the estimated value are more significant comparing to the same graph for test data. These fluctuations are mainly caused by numerous perturbing factors, specified in Sect. 1 and can be sufficiently reduced by implementing various filtering algorithms applied to $Q_0(t)$ and $T_{\text{out}}(t)$. Filtering of model input signals also reduces general temperature perturbation (Fig. 14).

Figure 15 presents a graph of the estimation error for the indoor air temperature. As you can see from the graph, the estimation error does not exceed ±0.5 °C that corresponds to the results obtained using simulated data.

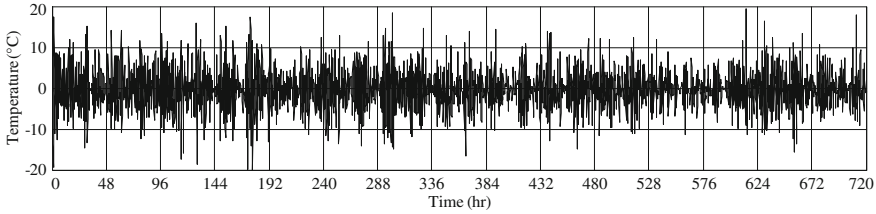


Fig. 14 General temperature perturbation

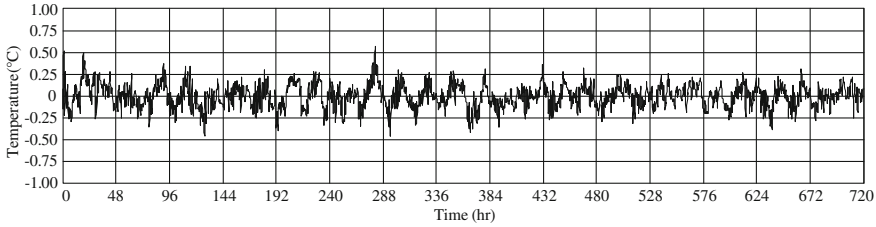


Fig. 15 Estimation error for indoor air temperature

5 Conclusion and Further Work

The results obtained demonstrate the overall viability of the proposed method to identify multistorey building's thermal characteristics based on experimental data and the possibility of its practical application in automatic heating control systems.

The academic building of the South Ural State University used in our test case is a typical representative of multistorey buildings. Hence, the results obtained are true for most multistorey office and apartment buildings.

The identification method examined in the article may be applied for designing automatic control systems and predictive control algorithms for heating buildings. This is the subject of our future research.

References

1. I.A. Bashmakov, An analysis of the main tendencies in the development of the heating systems in Russia and abroad, (Russian: Анализ основных тенденций развития систем теплоснабжения в России и за рубежом) Novosti teplosnabzhenia 2 (2008) [Online]. Available: http://www.cenef.ru/file/Heat.pdf
2. J.M. Salmerón, S. Álvarez, J.L. Molina, A. Ruiz, F.J. Sánchez, Tightening the energy consumptions of buildings depending on their typology and on climate severity indexes. Energ. Build. **58**, 372–377 (2013)
3. T. Salsbury, P. Mhaskar, S.J. Qin, Predictive control methods to improve energy efficiency and reduce demand in buildings. Comput. Chem. Eng. **51**, 77–85 (2013)

4. D. Zhou, S.H. Park, Simulation-assisted management and control over building energy efficiency—a case study. Energ. Procedia **14**, 592–600 (2012)
5. P.-D. Moroşan, A distributed MPC strategy based on Benders' decomposition applied to multi-source multi-zone temperature regulation. J. Process Control **21**, 729–737 (2011)
6. I. Jaffal, C. Inard, C. Ghiaus, Fast method to predict building heating demand based on the design of experiments. Energ. Build. **41**, 669–677 (2009)
7. D.A. Shnayder, V.V. Abdullin, A.A. Basalayev, in *Approach to Operations Analysis of Buildings Heat Supply*, (Russian: Подход к оперативному анализу эффективности теплоснабжения зданий), Bulletin of the South Ural State University, Series Computer Technologies, Automatic Control, Radio Electronics, vol. 13, 2 (219) 2011, pp. 70–73
8. A.P. Melo, D. Cóstola, R. Lamberts, J.L.M. Hensen, Assessing the accuracy of a simplified building energy simulation model using BESTEST: the case study of Brazilian regulation. Energ. Build. **45**, 219–228 (2012)
9. E. Žáčeková, S. Prívara, Z. Váňa, in *AUCC 2011*: *Model predictive control relevant identification using partial least squares for building modeling*, Proceedings of the 2011 Australian Control Conference, pp. 422–427 (Article number 6114301)
10. S. Ginestet, T. Bouache, K. Limam, G. Lindner, Thermal identification of building multilayer walls using reflective Newton algorithm applied to quadrupole modelling. Energ. Build. **60**, 139–145 (2013)
11. S. Prívara, J. Cigler, Z. Váňa, F. Oldewurtel, C. Sagerschnigc, E. Žáčeková, Building modeling as a crucial part for building predictive control. Energ. Build. **56**, 8–22 (2013)
12. Y.A. Tabunschikov, M.M. Brodach, Mathematical modeling and optimization of building thermal efficiency, (Russian: Математическое моделирование и оптимизация тепловой эффективности зданий.)—(Moscow: AVOK-PRESS, 2002)
13. V.V. Abdullin, D.A. Shnayder, L.S. Kazarinov, in *WCE 2013*: *Method of Building Thermal Performance Identification Based on Exponential Filtration*. Proceedings of the World Congress on Engineering 2013, London. Lecture Notes in Engineering and Computer Science, 3–5 July, 2013, pp. 2226–2230
14. D.A. Shnayder, L.S. Kazarinov, A method of proactive management of complicated engineering facilities using energy efficiency criteria (Russian: Метод упреждающего управления сложными технологическими комплексами по критериям энергетической эффективности). Manage. Large-Scale Syst. **32**, 221–240 (2011)
15. L.S. Kazarinov, S.I. Gorelik, Predicting random oscillatory processes by exponential smoothing (Russian: Прогнозирование случайных колебательных процессов на основе метода экспоненциального сглаживания). Avtomatika i telemekhanika **10**, 27–34 (1994)
16. D.A. Shnayder, V.V. Abdullin, in *TSP 2013*: *A WSN-based system for heat allocating in multiflat buildings*. Proceedings of 36th International Conference on Telecommunications and Signal Processing, Rome, Italy, 2–4 July 2013, pp. 181–185 (Article number 6613915)

DC-Image for Real Time Compressed Video Matching

Saddam Bekhet, Amr Ahmed and Andrew Hunter

Abstract This chapter presents a suggested framework for video matching based on local features extracted directly from the DC-image of MPEG compressed videos, without full decompression. In addition, the relevant arguments and supporting evidences are discussed. Several local feature detectors will be examined to select the best for matching using the DC-image. Two experiments are carried to support the above. The first is comparing between the DC-image and I-frame, in terms of matching performance and computation complexity. The second experiment compares between using local features and global features regarding compressed video matching with respect to the DC-image. The results confirmed that the use of DC-image, despite its highly reduced size, is promising as it produces higher matching precision, compared to the full I-frame. Also, SIFT, as a local feature, outperforms most of the standard global features. On the other hand, its computation complexity is relatively higher, but it is still within the real-time margin which leaves a space for further optimizations that could be done to improve this computation complexity.

Keywords Compressed domain · DC-image · Global features · Local features · MPEG · SIFT · Video matching

S. Bekhet (✉) · A. Ahmed · A. Hunter
School of Computer Science, University of Lincoln,
Brayford Pool, Lincoln LN6 7TS, UK
e-mail: sbekhet@lincoln.ac.uk

A. Ahmed
e-mail: aahmed@lincoln.ac.uk

A. Hunter
e-mail: ahunter@lincoln.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_37,

1 Introduction

The volume of video data is rapidly increasing, more than 72 hours of video are uploaded to YouTube every minute [1], and counters are still running fast. This is attributed to recent advance in multimedia technology. The majority of available video data exists in compressed format (e.g. MPEG), and the first step towards efficient video content retrieval is extraction of low level features, directly from compressed domain without full decompression to avoid the expensive computations and large memory requirement involved in decoding such compressed videos. Working on compressed videos is beneficial because of its richness of additional, pre-computed, features such as DCT coefficients, motion vectors and macro blocks types. DC coefficients specifically could be used to reconstruct a video frame with minimal cost [2]. However, most of the current techniques are still inefficient in directly handling compressed videos, without decompressing them first, which is a waste of valuable processing time and memory resources. All those advantages of detecting similarity from compressed videos are also expected to contribute to other higher-level layers of semantic analysis and annotation of videos, among other fields. An MPEG video consists of "**I**", "**P**" and "**B**" frames encoded using Discrete Cosine Transform (DCT) [3]. The DCT algorithm works by dividing an input image into 8×8 blocks (default block size). For each block, the DCT is computed and the result consists of one DC coefficient and 63 AC coefficients per block. A DC-image of an I-frame is the collection of all its DC coefficients, in their corresponding spatial arrangements. The DC image is 1/64 of its original I-frame size. Figure 1a shows an illustration of the DCT block structure. Figure 1b depicts samples of DC-images reconstructed from different I-frames.

The DC-image is usually an image of size around 40×30 pixels. However, the DC-image was found to retain most of the visual features of its corresponding full I-frame. It has also been found that human performance on scene recognition drops by only 7 % when using small images relative to full resolution images [4], as depicted in Fig. 2. This is very useful for computer vision algorithms, especially in relation to computation complexity of achieving the same complex tasks on the DC-image. Taking advantage of the this tiny size, fast reconstruction and richness of visual content, the DC-image could be employed effectively alone or in conjunction with other compressed domain features (AC coefficients, macro-block types and motion vectors) to detect similarity between videos for various purposes; as automated annotation [5] or copy detection or any other higher layer built upon similarity between videos.

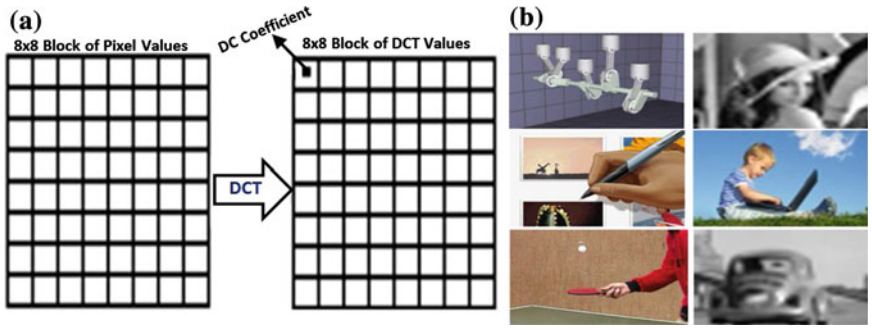


Fig. 1 **a** DCT block structure, showing the location of the DC coefficient, and **b** sample reconstructed DC-images, of size 30 × 40 (images are stretched for illustration)

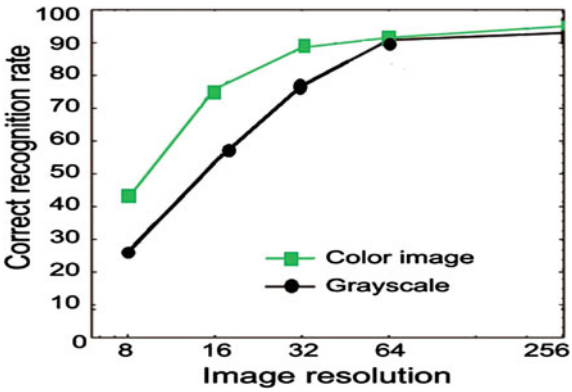


Fig. 2 Human eye performances on scene recognition as a function of image resolution [4]

2 Related Work

In this section, previous key work related to video matching in compressed domain is reviewed, focusing on the DC-image since it is a powerful feature compared to other MPEG features as depicted in Table 1. However, as the DC-image, is a small or lower-resolution image, the relevant work on low-resolution small images will also be reviewed. Initially the term "tiny image" was introduced in [4] during an attempt to construct a database of 80 million tiny images of size 32 × 32, labeled using non abstract English nouns, as listed in WordNet [6]. The aim of this work was to perform object and scene recognition by fusing semantic information extracted from WordNet with visual features extracted from the images, using nearest neighbor methods. Image similarity was computed using two measures; the first is the sum of squared differences (SSD), over the first 19 principal components of each image pixel values. The second similarity measure accounts for the potential small scaling (up to 10 pixels) and small translations (within a 5 × 5 window), by performing exhaustive evaluation of all possible image shifts. The concept of tiny

Table 1 MPEG compressed stream features

Feature	Type	Pros.	Cons.
DC coefficients	Spatial	• Partial decompression needed to extract from I frames [37]	• Cannot generate interest points easily due to its small size [37]
		• Used as a replacement of I frames [37]	• Full decompression needed to be extracted from P & B frames
		• Fast in applying complex operations	
		• Could be extracted either in grayscale or full color	
		• DC image of I frame could be used as a key frame of the entire GOP	
AC coefficients	Spatial	• Partial decompression needed to extract it	• Do not reveal any visual information unless reconstructed [3]
Motion vectors	Temporal		• Describe movement of a block
		• Partial decompression needed to extract	• Do not encode motion information across GOP's [38]
		• A pre-computed motion feature	• Only available for P & B frames
			• Do not encode visual information
Macroblock types	Spatial	• Partial decompression needed to extract	• Encodes only metadata about block compression information (eg. intra coded, skipped) [39]
		• Suitable for copy detection and fingerprinting [39]	• Do not encode visual information

image was then adopted and extended in [7–9] in an attempt to build a database of tiny videos. Approximately 50,000 tiny videos were used, in conjunction with the 80 million tiny images database, to enhance object retrieval and scene classification. Videos were collected with all their available metadata (e.g. title, description and tags), also all video frames were resized to 40 $\times$ 30 pixels stored as one dimensional vector that aggregates all the three color channels. Same similarity measures from tiny images [4] were adopted. Later the work was extended for the purpose of video retrieval using keyframes [7]. However, available video metadata were utilized, which is not always available neither accurate, in addition videos were treated as a set of unrelated images during the matching. Thus, our work is more focused on videos before they could have any tags or meta-data available which can be seen as a phase that can help in building such datasets for later use.

In the compressed domain, the DC-image has been used widely in shot-boundary detection and video segmentation due to its small size [10–14]. It was also utilized for keyframe extraction, instead of parsing the full frame [15–17], or even for video summarization purpose [18, 19]. For video retrieval, in [20] the DC-image was used to detect keyframes, then attention analysis is carried out on

full I-frames, to detect salient objects. SIFT [21] is applied to detect interest points and track them in successive spatial salient regions to build their corresponding trajectories. Color and texture were used to describe each salient region for matching purpose. But this method fails when either the visual features of the foreground object is not distinct or when video background contain rich details as it will produce meaningless salient regions which is not distinctive for a given video. An approach to match video shots and cluster them into scenes proposed in [22], the idea was taking into account variable number of frames to represent a shot (instead of only one keyframe). They utilized frame-to-frame matching based on color histograms computed for every DC or DC + 2AC depending on frame size, for frame of size 320×240 DC-image is selected and for frame size of 160×120DC + 2AC is selected, this makes representative frame images are always of size 40×30 and contains sufficient information for extraction, but with more full decompression for smaller size frames which affects the real-time processing. Regarding generating video signatures using the DC-image, in [23] matching between video clips was done using signatures built by extracting color values (Y-U-V) from DC-images sequence to form three different quantized histograms per each frame. The similarity between two videos is computed using sliding window technique, trying to find the best set of matching frames using histogram intersection. The approach of ordinal measures were applied on DC-images of each color component (Y-U-V) separately to generate fingerprint features for each frame which are accumulated to form video signature for later video matching [24]. Dimitrova et al. [25] demonstrated using DC coefficients of (Y-U-V) components separately and motion vectors to build signature for video retrieval. Signature was extracted from every frame and concatenated to form full video signature where hamming distance used to rank videos based on sliding window technique to determine the set of frames to compute signature from. A noticeable remark is that such approach used the DC-image as a set of numeric values leaving all the visual information behind, in addition to the slow operation of the sliding window technique as it applies an exhaustive search process to align two signatures together.

From a high level perspective, techniques that utilize the DC-image could be classified based on feature extraction level, feature types and applications as depicted in Fig. 3. For feature extraction level, there are two levels. The first; where every frame in video is being processed to extract low level features for later retrieval or signature building. The Second type is more compact and tries to reduce the amount of features being extracted by using keyframes only. Both approaches have disadvantages as they ignore the temporal dimension of a video and handles video as a bag of still images. Moreover, window alignment techniques will be needed in this case, which is based on exhaustive search among frames to find the best matching frames sequence. Regarding video signature built on those approaches it will be large and includes redundant information due to the concatenation of individual frames/keyframes signatures which violates the compactness of the signature. Furthermore for keyframe based schemes, there is

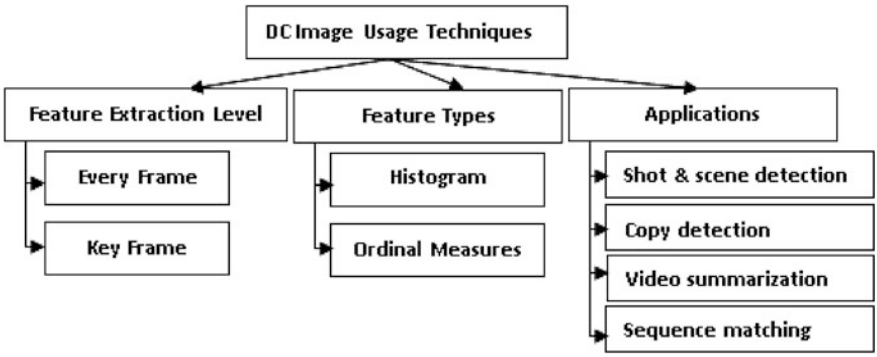


Fig. 3 DC images usage techniques

no fixed selection criteria for those keyframes which could be applied to all videos; some techniques uses the first and last frames within a shot as keyframes, while others uses the middle frame so, the resultant video signature may differ for same video with different keyframe selection criteria's.

For feature types that could be extracted from DC-images, exists: histogram [26] which is a global feature computed on frame level or video level (less common) where similarity between videos depends on the similarity of underlying histograms. Disadvantages of histograms are: (1) relatively high computational cost (pixel level processing) (2) high dependency on underlying color space as each one exhibit significant variations in color representation (3) histogram as a global feature don't capture neither spatial nor temporal information. The second common feature is ordinal measures [27], which also a global feature originally used for stereo matching and later adopted for video retrieval. The idea works by partitioning an image into equal-sized sub-images (blocks), then those sub-images are ranked based on their respective average color. Then, the finial ordering represents the ordinal matrix of the image. Ordinal measures are invariant to luminance change and histogram equalization effects within frame level only, but it is not invariant to geometric transformations, also it is based on color information only which is not robust against color format change. In addition, as a type of global feature it does not capture neither spatial nor temporal information. Recently it has been extended to capture the temporal dimension of videos, as the blocking process could be extended across video frames [28].

3 Proposed Approach

In this section, our proposed DC-image based system for video matching is introduced. The proposed idea is to utilize local features, such as SIFT [21] and SURF [29] on the small DC-images and track them across consecutive frames to compare

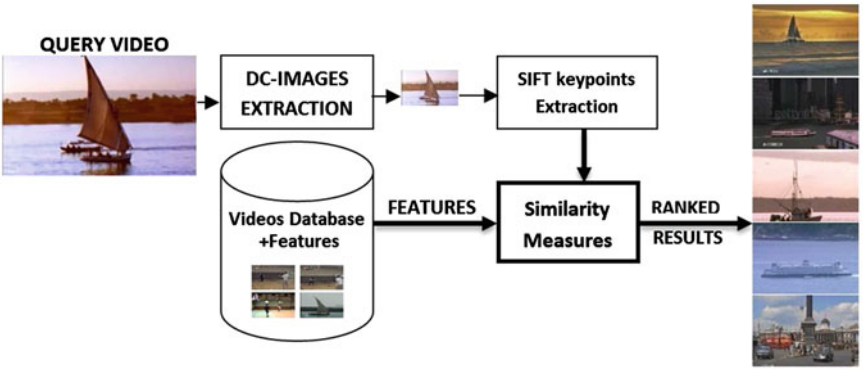


Fig. 4 Proposed system structure to measure videos similarity based on DC image

the similarity between videos. This idea introduces some challenges regarding local features extracting in such small size images, as discussed later. Figure 4 shows block diagram of the proposed system. The main stages of the system are:

1. Decoding video and extracting grayscale DC-image sequence.
2. Extracting SIFT keypoints and their descriptors, in each DC-image.
3. Video matching, using the extracted features.

The following sub-sections describe those stages, including challenges and our contribution to facilitate the video matching using small DC-images, without performing full decompression.

3.1 Extracting the DC Image Sequences

The process starts by decoding a video and extracting luminance DC-images sequence from I-frames only. Following Table 1, there are extra reasons for focusing on the DC-image, includes:

- I-frame's DC-image is the quickest part that could be extracted from a compressed video without performing full decompression of video stream.
- I-frames in GOPs (Group Of Pictures) are inserted by the encoder when there is large residual change (residual is the amount of motion estimation error accumulated at the end of GOP), this could be analogous to keyframes within a scene, in other words as a keyframe is reprehensive to a scene, DC-image of an I-frame could be used as representative of a GOP. In addition, GOPs could be merged to specific length to limit number of DC-images and map them to be key frames like.

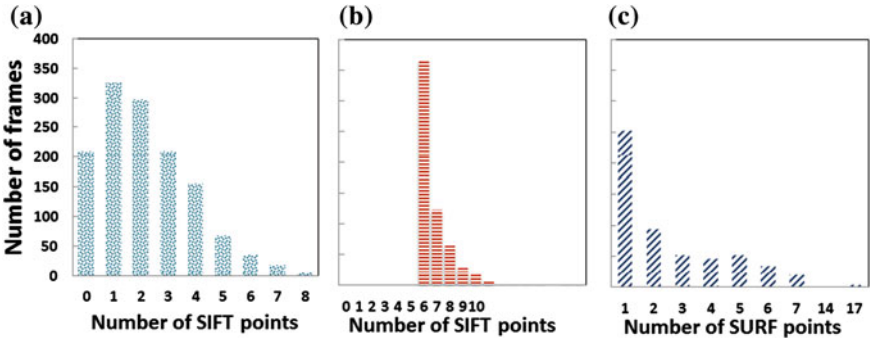


Fig. 5 **a** Number SIFT points per frame before adjusting sigma value. **b** Number SIFT points per frame after adjusting sigma value. **c** Number SURF points per frame

- I-frames will give about 10:1 initial compaction ratio assuming 10 frames per GOP [30] on average which means lower computations and faster results.
- Human eye is sensitive to small changes in luminance rather than chrominance [4]. Thus we can rely on luminance DC-image only.

3.2 Extracting Keypoints and Descriptors

The second stage in the proposed framework is extraction of keypoints and their respective descriptors. During our experiments we used SIFT and SURF for extracting keypoints, as they are the mostly reported effective feature detectors algorithms. While a typical full image of size 500 $\times$ 500 could generate more than 2,000 interest points [21]. However, most of the DC-images would generate less than three SIFT key points, which is not enough for matching [21]. We did an experiment using TRECVID BBC RUSHES [31] dataset to investigate the amount of local features a DC-image could generate. Figure 5a, c shows that $\sim$63 % of frames generate less than three keypoints for SIFT and SURF respectively. Since SIFT was reported for better keypoints localizing than SURF [32–34] we adapted SIFT by iteratively adjusting sigma value (the amount of gaussian blurring applied to an image) to generate a minimum of six keypoints in each DC-image. Figure 5b shows number of SIFT points per frame after our adjustment and enforcing the minimum of six SIFT keypoints per each DC-image. With this enforcement, we facilitated for the DC-image to be used for video matching. Regarding the precision of matching videos, using DC-images compared with the full I-frame, it will be presented later in Sect. 4.

video 2 \ Video1	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17
F1	66	83	50	33	83	66	50	50	50	0	33	83	66	66	66	83	66
F2	50	66	50	33	50	50	33	66	0	50	50	66	66	50	90	66	50
F3	50	66	50	33	50	50	33	50	16	33	50	50	50	50	66	50	33
F4	42	42	42	28	42	28	28	57	28	42	28	42	42	42	42	42	28
F5	66	66	50	33	50	50	50	50	50	50	50	50	66	33	66	50	33
F6	66	66	33	16	50	50	50	50	33	50	50	66	50	33	66	50	33
F7	28	57	71	57	28	57	42	57	57	57	71	71	57	4	57	75	75
F8	42	42	57	71	28	42	28	57	57	57	42	42	28	28	42	42	42
F9	33	50	66	50	33	50	50	50	50	50	50	33	33	50	50	50	50
F10	66	66	83	83	66	66	66	66	50	66	66	50	50	66	66	50	50
F11	66	50	66	83	50	50	66	66	33	50	50	33	33	50	50	33	50

Fig. 6 Finding matching similarity score between two videos

3.3 Video Matching Using DC-Images and SIFT

The third and final stage in our proposed framework is the actual matching between videos. For simplicity, we adopted the frame-to-frame matching, as for each video pair; we compute a similarity measure between SIFT keypoints taking into account the temporal order of video frames. This is done by searching for the longest optimal matching frames sequence between two videos using dynamic programming. Optimality in this case, means finding the best matching video frames that maximizes the overall similarity score with respect to the temporal order of frames. Figure 6 shows a sample confusion matrix of given two videos and the optimal matching values for their respective frames are highlighted in grey. Following is the pseudo-code of the dynamic programming algorithm used to compute the optimal matching cost:

```
SET M to video.1 frames number +1;
SET N to video.2 frames number +1;
CREATE_MATRIX OPT_MATCH[M][N];
INITALIZE DISTANCE to all frame-to-frame similarity based SIFT;
SET OPT_MATCH to 0;
FOR I=1 to M DO
    FOR J=1 to N DO
      SET OPT_MATCH[I][J] to MAX   of
      {
                    OPT_MATCH [I-1][J],
                    OPT_MATCH[I-1][J-1]+DISTANCE[I-1][J-1],
                    OPT_MATCH[I][J-1]
      }
RETURN OPT_MATCH[M-1][N-1] ;
```

Where **DISTANCE** is the confusion matrix between both video frames, computed based on the number of underlying matched SIFT keypoints, and **OPT_MATCH** is the matrix which will contain the finial matching score, this value will be located in location (**M − 1**, **N − 1**) and **MAX** is function returns the maximum value of a given group of numbers. The algorithm works by scanning the confusion matrix from left to right and from up to bottom trying to find the highest match for each frame taking into account the previous and next frames matching scores, in addition to the sequence of frames.

We can see that our proposed dynamic programming algorithm performs one to one mapping as each video frame will be matched to only one frame in the other video. Since the matching is one-to-one some frames may not be matched at all, for two reasons; the first that it might reduce the overall matching value between videos (e.g. frames 1, 4 in video 1). The second case happens if the currently matching videos are of different number of frames (e.g. frames 6, 7, 13 and 16 in video 1).

4 Experiments and Results

In this section we explain the experiments and present the results that support our work explained earlier. This section presents two experiments; the first is regarding comparing the DC-image to I-frame, in terms of matching performance and computation complexity. The second experiment compares between using local features and global features in compressed video matching, with respect to the small size images. We used the TRECVID BBC RUSHES [31] standard data set for video retrieval which contains diverse set of challenging videos; mainly man-made moving objects (cars, tanks, planes and boats). But, since the videos were only available in uncompressed format; all the videos were re-encoded to MPEG-2 format with frame size (352 × 240), so that all the DC-images are of equal size (44 × 30 pixels). The experiments ran on Intel Core i3-3.30 GHZ computer with 4 Gb of RAM.

4.1 DC-Image Versus I-Frame

The purpose of this experiment is to evaluate the performance of the DC-image, in terms of matching and computational complexity, compared to the corresponding I-frame. The experiment used the framework explained in Fig. 4. Regarding matching time based on the DC-image with SIFT [21] features; it took a total of 58.4 min for all videos in the dataset, while it took a total of 166.6 h for the same dataset using the full I-frame. The average time (per frame) is 0.017 s for the DC-image, compared to 1.05 s for the I-frame (time includes reconstruction, SIFT keypoints extraction and matching). This shows that the computation complexity using the DC-image is only 1.6 % of the corresponding I-frame, which means a

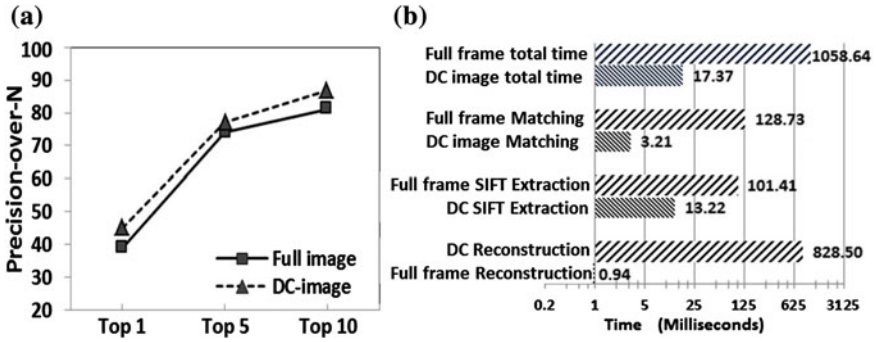


Fig. 7 **a** DC-image versus I-frame retrieval precision, and **b** DC-image versus I-frame timing performance

total reduction of 98.4 % in processing time. Figure 7b shows the timing details for the DC-image and the I-frame respectively. To compare the matching precision, we adopted the precision-over-N [35] standard measure over ranks 1, 5 and 10. The DC-image, despite its highly reduced size, was found to have a slightly higher precision than the I-frame at all ranks, as depicted in Fig. 7a.

4.2 Local Versus Global Features

The purpose of this experiment is to evaluate the performance of using local and global features, on the DC-image, in terms of matching precision and computational complexity. The experiment also used the framework described earlier in Fig. 5. For local features, we utilized SIFT [21] as a local feature descriptor, in addition to dense SIFT [36] to verify the results for a larger number of keypoints. For global features, we applied matching based on the luminance histogram, ordinal measures [27] and the pixel difference [8].

The results, presented in Fig. 8a, shows that SIFT as a local feature descriptor outperforms dense SIFT, in addition SIFT outperforms global feature descriptors by 15.4 % (compared to ordinal matching as the highest precision global feature method). However, SIFT's computation complexity was the highest, as depicted in Fig. 8b. SIFT took 16.43 ms to match two DC-image frames, compared to only 2 ms in pixel difference matching (maximum time in case of global features). But SIFT still works within real-time margin, while producing better matching performance. Knowing that all measures are being used in their generic form, using dynamic programming to incorporate the temporal dimension based on the finial fame-to-frame confusion matrix. We also developed results visualization software, a snapshot is depicted in Fig. 9 based on real example of SIFT matching using the BBC RUSHES dataset.

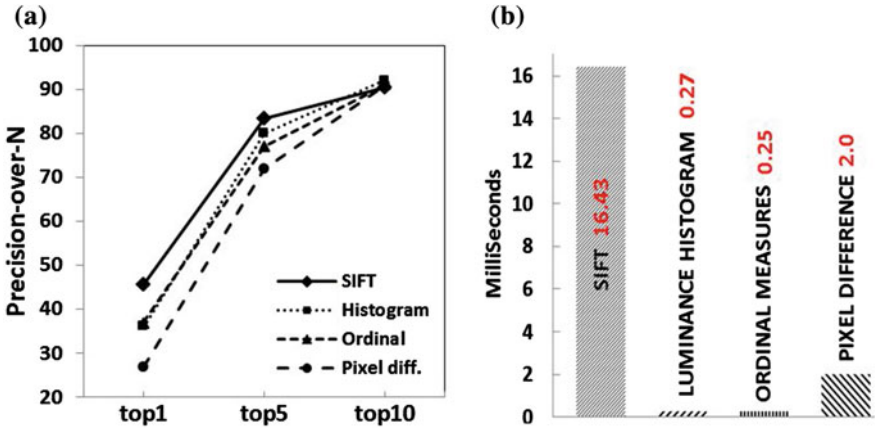


Fig. 8 **a** DC image retrieval precision-over-N curves using SIFT-luminance histogram-ordinal measures-pixel difference, and **b** DC timing analysis using SIFT-luminance histogram-ordinal measures-pixel difference

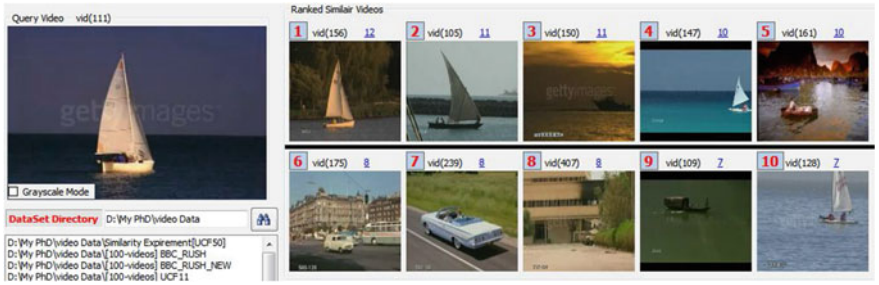


Fig. 9 Snapshot of visualization software based on SIFT matching

5 Conclusion and Future Work

In this paper, we presented a framework for video matching based on local features extracted only from the DC-images of MPEG compressed videos, without full decompression. Also, supporting experiments regarding DC-image precision and complexity versus the full I-frame were presented. But we had to address the issue of using SIFT on such small-size images, before it could be used. The results show that the DC-image, despite its small size, produces similar (if not better) similarity precision scores, compared to its corresponding I-frame. But using the DC-image has dramatically improved the computational performance ($\sim$62 times faster), which makes it a high candidate for more sophisticated use. Also, local features, such as SIFT, were compared to standard global features for the purpose of video similarity. The results shows that using SIFT, on DC-image only, slightly

outperformed the accuracy of the global features. On the other hand, the computational complexity of using SIFT is relatively higher than those for the global features. But SIFT extraction and matching is still within the real-time margins, and still we have a number of optimizations to be introduced to reduce this computation complexity. We also plan to introduce more complex matching, instead of the frame-to-frame approach, and better incorporate the temporal information actively.

Acknowledgments This work is funded by SouthValley University, Egypt.

References

1. YouTube Statistics [Online]. Available http://www.youtube.com/yt/press/statistics.html (2013)
2. S. Bekhet, A. Ahmed, A. Hunter, Video matching using DC-image and local features, in *Proceedings of the World Congress on Engineering*, UK (2013), pp. 2209–2214
3. A.B. Watson, Image compression using the discrete cosine transform. Math. J. **4**, 81 (1994)
4. A. Torralba, R. Fergus, W.T. Freeman, 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1958–1970 (2008)
5. A. Altadmri, A. Ahmed, A framework for automatic semantic video annotation. Multimedia Appl. Tools **64**(2), 1–25 (2013)
6. G. Miller, C. Fellbaum, *Wordnet: An Electronic Lexical Database* (1998)
7. A. Karpenko, P. Aarabi, Tiny videos: A large data set for nonparametric video retrieval and frame classification. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 618–630 (2011)
8. A. Karpenko, P. Aarabi, Tiny videos: A large dataset for image and video frame categorization, in *11th IEEE International Symposium on Multimedia (ISM '09)* (2009), pp. 281–289
9. A. Karpenko, P. Aarabi, Tiny videos: non-parametric content-based video retrieval and recognition, in *Tenth IEEE International Symposium on Multimedia (ISM 2008)* (2008), pp. 619–624
10. B.-L. Yeo, B. Liu, A unified approach to temporal segmentation of motion JPEG and MPEG compressed video, in *Proceedings of the International Conference on Multimedia Computing and Systems* (1995), pp. 81–88
11. J. Meng, Y. Juan, S.F. Chang, Scene change detection in a MPEG compressed video sequence, in *IS&T/SPIE Symposium Proceedings* (1995)
12. J.-L. Zheng, M.-J. Li, M.-X. Zhang, J. Zhou, Z.-D. Liu, An effective framework of shot segmentation based on I-frame in compressed-domain videos, in *International Conference on Wavelet Analysis and Pattern Recognition* (2012), pp. 84–90
13. P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, H. Sun, Algorithms and system for segmentation and structure analysis in soccer video, in *Proceedings of ICME* (2001), pp. 928–931
14. A. Divakaran, H. Ito, H. Sun, T. Poon, Scene change detection and feature extraction for MPEG-4 sequences, in *Electronic Imaging'99* (1998), pp. 545–551
15. G. Liu, J. Zhao, Key frame extraction from MPEG video stream, in *Third International Symposium on Information Processing* (2010), pp. 423–427
16. E.K. Kang, S.J. Kim, J.S. Choi, Video retrieval based on scene change detection in compressed streams. IEEE Trans. Consum. Electron. **45**, 932–936 (1999)

17. O.N. Gerek, Y. Altunbasak, Key frame selection from MPEG video data, in *Electronic Imaging* (1997), pp. 920–925
18. J.C.S. Yu, M.S. Kankanhalli, P. Mulhen, Semantic video summarization in compressed domain MPEG video, in *Proceedings International Conference on Multimedia and Expo*, vol. 3 (2003), pp. 329–332
19. J. Almeida, N.J. Leite, R.S. Torres, Online video summarization on compressed domain. J. Vis. Commun. Image Represent. (2011)
20. H.-P. Gao, Z.-Q. Yang, Content based video retrieval using spatiotemporal salient objects, *in 2010 International Symposium on Intelligence Information Processing and Trusted Computing (IPTC)* (2010), pp. 689–692
21. D.G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
22. M.M. Yeung, B. Liu, Efficient matching and clustering of video shots, in *Proceedings of the International Conference on Image Processing.* vol. 1 (1995), pp. 338–341
23. M.R. Naphade, M.M. Yeung, B.L. Yeo, A novel scheme for fast and efficient video sequence matching using compact signatures, in *Proceedings of SPIE, Storage and Retrieval for Media Databases* (2000), pp. 564–572
24. R. Mohan, Video sequence matching, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6 (1998), pp. 3697–3700
25. N. Dimitrova, M.S. Abdel-Mottaleb, Video retrieval of MPEG compressed sequences using dc and motion signatures, in *Video Retrieval of MPEG Compressed Sequences using DC and Motion Signatures*, 1999
26. R.C. Gonzalez, E.W. Richard, *Digital Image Processing*, 3rd edn edn. (Prentice Hall, Englewood Cliffs, 2008), pp. 142–143
27. D.N. Bhat, S.K. Nayar, Ordinal measures for visual correspondence, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR '96)* (1996), pp. 351–357
28. J. Almeida, N.J. Leite, R. S. Torres, Comparison of video sequences with histograms of motion patterns, in *18th IEEE International Conference on Image Processing (ICIP)* (2011), pp. 3673–3676
29. H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF). Comput. Vis. Image Underst. **6**(110), 346–359 (2008)
30. D. Fuentes, R. Bardeli, J.A. Ortega, L. Gonzalez-Abril, A similarity measure between videos using alignment, graphical and speech features. Expert Syst. Appl. **39**, 10278–10282 (2012). 9/1
31. A. Basharat, Y. Zhai, M. Shah, Content based video matching using spatiotemporal volumes. Comput. Vis. Image Underst. **110**, 360–377 (2008)
32. K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 2 (2003), pp. II-257–II-263
33. L. Juan, O. Gwun, A comparison of sift, pca-sift and surf. Int. J. Image Process. (IJIP) **3**, 143–152 (2009)
34. K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 1582–1596 (2010)
35. D.M. Powers, Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation, in *School of Informatics and Engineering*, Flinders University, Adelaide, Australia, Tech.Rep.SIE-07-001, 2007
36. T. Tuytelaars, Dense interest points, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2010), pp. 2281–2288
37. S. Bekhet, A. Ahmed, A. Hunter, Video matching using DC-image and local features, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering, WCE 2013*, London, 3–5 July 2013, pp. 2209–2214

38. N. Dimitrova, F. Golshani, Motion recovery for video content classification. ACM Trans. Inf. Syst. (TOIS) **13**, 408–439 (1995)
39. A.S. Abbass, A.A.A. Youssif, A.Z. Ghalwash, *Compressed Domain Video Fingerprinting Technique Using the Singular Value Decomposition* (2012)
40. P. Panchal, S. Merchant, Performance evaluation of fade and dissolve transition shot boundary detection in presence of motion in video, in *1st International Conference on Emerging Technology Trends in Electronics, Communication and Networking (ET2ECN)* (2012), pp. 1–6

Automated Diagnosis and Assessment of Dysarthric Speech Using Relevant Prosodic Features

Kamil Lahcene Kadi, Sid Ahmed Selouani, Bachir Boudraa and Malika Boudraa

Abstract In this paper, linear discriminant analysis (LDA) is combined with two automatic classification approaches, the Gaussian mixture model (GMM) and support vector machine (SVM), to automatically assess dysarthric speech. The front-end processing uses a set of prosodic features selected by LDA on the basis of their discriminative ability, with Wilks' lambda as the significant measure to show the discriminant power. More than eight hundred sentences produced by nine American dysarthric speakers of the Nemours database are used throughout the experiments. Results show a best classification rate of 93 % with the LDA/SVM system achieved over four severity levels of dysarthria, ranged from not affected to the more seriously ill. This tool can aid speech therapist and other clinicians to diagnose, assess, and monitor dysarthria. Furthermore, it may reduce some of the costs associated with subjective tests.

Keywords Dysarthria · GMM · LDA · Nemours database · Prosodic features · Severity-level assessment · SVM

K. L. Kadi (✉) · B. Boudraa · M. Boudraa
Faculty of Electronics and Computer Science, University of Sciences and Technology Houari Boumediene, 32 El Alia 16111 Bab Ezzouar Algiers, Algeria
e-mail: kkadi@usthb.dz

B. Boudraa
e-mail: bboudraa@usthb.dz

M. Boudraa
e-mail: mboudraa@usthb.dz

S. A. Selouani
Department of Information Management, University of Moncton, Campus of Shippagan, 218 boulevard J.-D.-Gauthier Shippagan NB E8S 1P6, Moncton, Canada
e-mail: selouani@umcs.ca

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_38,

1 Introduction

Dysarthria is a motor speech disorder resulting from disturbed muscular control of the speech mechanism, and is caused by damage to the central or peripheral nervous system. These disturbances to the brain and the nerve stimuli of muscles involved in the production of speech cause incoordination, paralysis, or weakness in the speech musculature. A few causes of dysarthria include stroke, head injury, Parkinson's disease, tumors, muscular dystrophy, and cerebral palsy. Both adults and children are affected by dysarthria. Millions of people across the world suffer from the condition, which induces perturbations in the timing and accuracy of the movements needed for normal prosody and intelligible speech [1–3].

Depending on the severity of the dysarthria, the intelligibility of speech can range from near-normal to unintelligible [4]. Usually, a large battery of tests assesses the degree of intelligibility by measuring the disease's severity or a treatment's progress. In practice, automatic methods of assessment can be helpful to clinicians in the diagnosis and monitoring of dysarthria.

Diverse methods have been developed for the automatic assessment of dysarthric speech. In [5], a combination of the statistical Gaussian mixture model (GMM) and soft-computing technique of artificial neural networks (ANNs) was applied along Mel-frequency cepstral coefficients (MFCCs) and speech rhythm metrics, and achieved 86.35 % accuracy over four severity levels of dysarthria. Feed-forward ANNs and support vector machines (SVMs) have been successfully used to design discriminative models for dysarthric speech with phonological features [6]. In [7], a Mahalanobis distance-based discriminant analysis classifier was proposed to classify the severity of dysarthria using a set of acoustic features. In this latter study, the classification achieved 95 % accuracy over two levels (mid-to-low and mid-to-high) by considering an improved objective intelligibility assessment of spastic dysarthric speech.

This paper presents an approach for assessing the severity level of dysarthria by combining linear discriminant analysis (LDA) with two classification methods: GMMs and SVMs. Discriminant analysis is used to select a pool of relevant prosodic features with a prominent discrimination capacity. We compare the performance of two combinations: LDA-GMM and LDA-SVM. The task consists of classifying four severity levels of dysarthria using the Nemours speech database [8].

The original contribution reported in this paper lies in the selection of the most relevant prosodic features for use in the front-end processing of the discriminant analysis to achieve better performance than existing dysarthria severity-level classification systems. Furthermore, the proposed approach reduces the processing time, as it represents each observation (sentence) by only one vector of eleven prosodic features, instead of using many acoustic vectors for each observation.

The remainder of the paper is structured as follows. Section 2 gives some definitions related to the prosodic features used by the proposed system. Section 3 presents the discriminant function analysis. In Sect. 4, the experiments and their outcome are presented and discussed. Section 5 contains our concluding comments.

2 Prosodic Features of Speech

Speech is primarily intended to transmit a message through a sequence of linguistic sound units. Prosody is defined as the branch of linguistics devoted to the description and representation of speaking elements. Prosodic cues include intonation, stress, and rhythm; each is a complex perceptual entity, fundamentally expressed using three acoustic parameters: pitch, duration, and energy [9]. The stress, timing, and intonation in speech are closely related to the speech prosody, enhancing the intelligibility of conveyed messages and allowing listeners to easily segment continuous speech into words and phrases [10].

According to [11], "speech is the most complex of innately acquired human motor skills, an activity characterized in normal adults by the production of about 14 distinguishable sounds per second through the coordinated actions of about 100 muscles innervated by multiple cranial and spinal nerves". In dysarthria, some neurological damage typically affects the nerves that control the articulatory muscle system involved in speech, causing weakness, slowness, and incoordination. Depending on the severity of the dysarthria, this disturbance has a variety of effects on prosody.

The extraction of a reasonably limited, informative, and meaningful set of features is an important step towards automatic dysarthria severity classification. In this work, we use a discriminant analysis with Wilks' lambda measure to select those prosodic features best adapted to dysarthria classification [12].

The proposed front-end processes the speech waveform at the sentence level; patients are able to repeat individual units (phonemes or words) of speech with a fairly normal consistency [13]. For each sentence completed by each speaker, eleven features are considered: jitter, shimmer, mean pitch, standard deviation of pitch, number of periods, standard deviation of period, proportion of the vocalic duration (%V), harmonics to noise ratio (dB), noise to harmonics ratio (%), articulation rate, and degree of voice breaks.

2.1 Mean Pitch

The physical correlate of pitch is the fundamental frequency (*F0*) estimated by the vibration rate of the vocal folds during the phonation of voiced sounds [10]. The ensemble of pitch variations during an utterance is defined as intonation [14]. The typical range of a male speaker is 80–200 Hz (for conventional speech), depending on the mass and length of the vocal chords [15]. In this work, mean pitch is calculated by averaging the fundamental frequency across one sentence using the autocorrelation method. The mean pitch value in dysarthric speech can help to detect a glottic signal abnormality.

2.2 Jitter

Jitter represents the variation of fundamental frequency within the time evolution of an utterance. It indicates the variability or perturbation of the time period *(T0)* across several cycles of oscillation. Jitter is mainly affected by a deficiency in the control of vocal fold vibration [16]. The threshold of comparison for normal/pathologic jitter is 1.04 %, according to the Multi-Dimensional Voice Processing Program (MDVP) designed by Kay Elemetrics Company [17]. The raw and normalized jitter are respectively defined as:

$$Jitter(\mathrm{s}) = \sum_{i=1}^{N-1} |T_i - T_{i+1}| / N - 1 \tag{1}$$

$$Jitter(\%) = Jitter(\mathrm{s}) / \frac{1}{N} \sum_{i=1}^{N} T_i \tag{2}$$

where T_i is the period and N represents the number of periods.

2.3 Shimmer

Shimmer indicates the perturbation or variability of the sound amplitude. It is related to variations in vocal emission intensity, and is partially affected by the reduction of glottic resistance [16]. MDVP gives a value of 3.81 % as a threshold for pathology. Shimmer is estimated in a similar manner to jitter, but uses amplitude as a parameter.

Both of intensity and pitch speech might be more difficult to control if the supply of air to the vocal folds is highly variable or extremely low [18].

2.4 Articulation Rate

The articulation rate is the number of syllables pronounced per second, excluding pauses [19]. In our study, the greater the severity level of dysarthria, the lower the articulation rate.

2.5 Proportion of Vocalic Duration

Vocalic duration describes the separation between the release and constriction when framing a vowel [20]. The proportion of vocalic duration (%V) is the fraction of the utterance duration that is composed of vocalic intervals [19].

Trouble maintaining the voice over a sustained vowel can be considered as a sign of pathology [21].

2.6 Harmonics to Noise Ratio

The harmonics to noise ratio (HNR) represents the degree of acoustic periodicity. Harmonicity is measured in dB, calculated as the ratio of the energy of the periodic part related to the noise energy. HNR can be used as a measure of voice quality. For example, a healthy speaker can produce a sustained "*a*" with an HNR of around 20 dB [21]. HNR is defined as:

$$HNR(dB) = 10\log\left(\frac{E_p}{E_n}\right) \tag{3}$$

where E_p is the energy of the periodic part and E_n is the energy of the noise.

2.7 Degree of Voice Breaks

The degree of voice breaks is the total duration of the breaks over the signal divided by the total duration, excluding the silence at the beginning and end of the sentence [19]. A voice break can occur with a sudden stoppage of the air stream due to a transient deficiency in the control of the phonation mechanism [22].

3 Discriminant Function Analysis

Feature selection can make a huge contribution to the classification task, as the selected features should avoid certain software mispredictions and improve the classification rate. In this work, we use discriminant analysis with Wilks' lambda as a tool to measure and select the effective features among a large number of computed speech characteristics.

Discriminant analysis is used to model a dependent categorical variable based on its relationship with one or more predictors. From a set of independent variables, discriminant analysis determines the linear combinations of those variables that best discriminate the classes. These combinations are called discriminant functions, and are defined by [23]:

$$d_{ik} = b_{0k} + b_{1k}\, x_{i1} + \cdots + b_{pk}\, x_{ip} \tag{4}$$

where d_{ik} is the value of the k^{th} discriminant function for the i^{th} class, p is the number of predictors (independent variables), b_{jk} is the value of the jth coefficient of the kth function, and x_{ij} is the value of the ith class of the jth predictor.

The number of functions is equal to *min(number of classes-1, number of predictor).*

The procedure automatically chooses a first function that will separate the classes as much as possible. It then selects a second function that is both uncorrelated with the first function and provides as much further discrimination as possible. The procedure continues adding functions in this way until the maximum number of functions is achieved, as determined by the number of predictors and categories in the dependent variable. To select the best variables for the model, a stepwise method can be used [23].

Wilks' lambda method of variable selection for stepwise discriminant analysis selects variables on the basis of their capacity to minimize Wilks' lambda. At each step, the variable that reduces the overall Wilks' lambda is entered [23]. The Wilks' lambda method needs some measure of discrimination capacity.

To measure the discriminant capacity of every variable X_p, we use the univariate ANOVA (Analysis of Variance). Its decomposition formula is [24]:

$$\underbrace{\sum_{i=1}^{I}\sum_{n=1}^{N_i}\left(X_{jin}-\bar{X}_j\right)^2}_{Total\ covariance}=\underbrace{\sum_{i=1}^{I}N_i\left(\bar{X}_{ji}-\bar{X}_j\right)^2}_{\substack{Separate-groups\\ covariance}}+\underbrace{\sum_{i=1}^{I}\sum_{n=1}^{N_i}\left(X_{jin}-\bar{X}_{ji}\right)^2}_{\substack{Within-groups\\ covariance}} \tag{5}$$

We consider a dataset with N observations constituted by $X_1,\ldots,X_p$ variables. These observations are partitioned by a qualitative variable into I classes of sizes $N_1,\ldots,N_I$, respectively.

X_{jin} is the value of X_j for the nth observation of class i
$\bar{X}_{ji}$ is the average of X_j over class i
$\bar{X}_j$ is the average of X_j.

For each variable X_p, Wilks' lambda is calculated as the ratio of the within-groups covariance to the total covariance. Smaller values of lambda indicate greater discrimination ability [24].

$$\Lambda_p=\frac{within\ group\ sum\ of\ squares}{total\ sum\ of\ squares} \tag{6}$$

In this work, we create a discriminant model that classifies dysarthric speakers into one of the four predefined "severity levels of dysarthria" groups. This model uses eleven prosodic features that have been selected by the Wilks' lambda method using discriminant analysis. To determine the relationship between a categorical dependent variable (severity level) and the independent variables (eleven features), we use a linear regression procedure [12].

4 Experiments and Results

4.1 Speech Material

Nemours is one of the few databases of recorded dysarthric speech. It contains 814 short, nonsensical sentences spoken by 11 American patients with varying degrees of dysarthria. Additionally, the database includes two connected-speech paragraphs: the "Grandfather" and "Rainbow" passages produced by each of the 11 speakers. Each sentence in the database is of the form "*The X is Ying the Z*", generated by randomly selecting *X* and *Z* from a set of 74 monosyllabic nouns without replacement, and selecting *Y* from a set of 37 disyllabic verbs without replacement. This process generated 37 sentences, from which 37 other sentences were produced by swapping *X* and *Y* [8]. Therefore, each noun and verb was spoken twice by each patient over the complete set of 74 sentences. The whole database has been marked at the word level; sentences for 10 of the 11 talkers have also been marked at the phoneme level. The entire speech corpus was spoken by one non-dysarthric speaker, a speech pathologist who conducted the recording session; this is considered as the healthy control (HC). All speech materials were recorded using a 16 kHz sampling rate and 16-bit sample resolution after low-pass filtering at 7,500 Hz cutoff frequency with a 90 dB/octave filter [8].

4.2 Subjects

The speakers are eleven young adult males suffering different types of dysarthria resulting from either cerebral palsy (CP) or head trauma (HT), and one male adult control speaker. Seven of the talkers had CP, among whom three had spastic CP with quadriplegia, two had athetoid CP (one quadriplegic), and two had a mixture of spastic and athetoid CP with quadriplegia. The remaining four subjects were victims of head trauma. The speech from one of the patients (head trauma, quadriplegic) was extremely unintelligible. This was considered so poor that it was not marked at the phoneme level, and perceptual data were not collected for this patient. A code of two letters was assigned to each patient: BB, BK, BV, FB, JF, KS, LL, MH, RK, RL, and SC. The patients can be divided into three subgroups based on their Frenchay Dysarthria Assessment scores (see Table 1): 'mild L1', including patients FB, BB, MH, and LL; 'severe L2' includes RK, RL, and JF, and 'more severe L3' includes KS, SC, BV, and BK. The speech assessment and perceptual data did not take into consideration the too mild case (subject FB) and the too severe case (KS) [2, 4].

Table 1 Frenchay dysarthria assessment scores of dysarthric speakers with the nemours database [8]

Patients	KS	SC	BV	BK	RK	RL	JF	LL	BB	MH	FB
Severity (%)	–	49.5	42.5	41.8	32.4	26.7	21.5	15.6	10.3	7.9	7.1

Table 2 Wilks' lambda of the acoustics features

Features	Wilks' lambda
Articulation rate	0.565
Number of period	0.595
Mean pitch	0.701
Voice breaks	0.835
%V	0.861
HNR	0.864
Jitter	0.925
Shimmer	0.962
Standard pitch	0.979
Standard period	0.984
NHR	0.989

4.3 LDA

In this section, we present discriminant analysis results using the stepwise method, the linear regression procedure, and Wilks' lambda.

The stepwise method is used in so far as we have many extracted features and don't have real expectations about which ones are important. The risk of this method is to choose a variable that have no practical significance as it's based only on statistical merit.

For the stepwise discriminant analysis, we consider Wilks' lambda as a feature selection method. The coefficients of the linear equations are estimated by the linear regression procedure, involving prosodic features that best predict the severity level of dysarthria.

Table 2 shows the eleven selected prosodic metrics and their ability to discriminate the four severity levels of dysarthria. Wilks' lambda varies from 0 to 1; smaller values of lambda reveal greater discrimination ability.

The discriminant analysis generated three discriminant functions to distinguish the four severity levels of dysarthria. The first two functions are more meaningful for the classification. Figure 1 represents the four classes discriminated by the first two discriminant functions.

The summary discriminant analysis, together with the rate of correct classification, is displayed in Table 3.

The four group's sizes are equals. We assume equal prior probabilities of group membership, for all severity level classes.

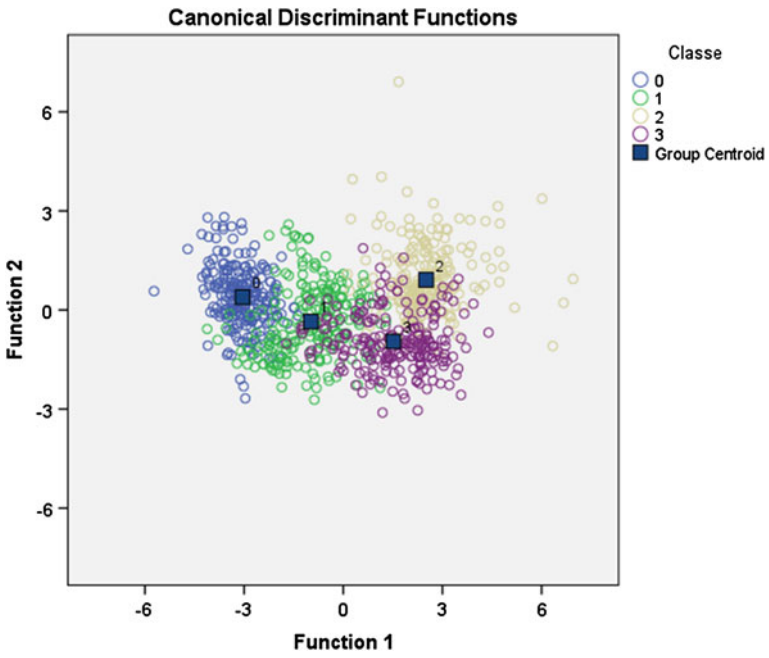


Fig. 1 Representation of combined groups

Table 3 Classification results

	Class	Predicted group membership				Total
		0	1	2	3	
Count	0	215	7	0	0	222
	1	27	185	1	9	222
	2	0	1	193	28	222
	3	0	31	29	162	222
%	0	96.8	3.2	0.0	0.0	100
	1	12.2	83.3	0.5	4.1	100
	2	0.0	0.5	86.9	12.6	100
	3	0.0	14.0	13.1	73.0	100

The largest number of misclassifications occurs for the 'severe L3' level (60/222). This can also be seen in Fig. 1, where the representation of L3 is scattered. The most severe dysarthria level is the hardest to characterize with prosodic features. Most other errors occur between two nearby classes. In this case, the values of the features can be close and confused. To reduce the misclassification rate, automatic classifiers with a high discrimination capacity are used [12].

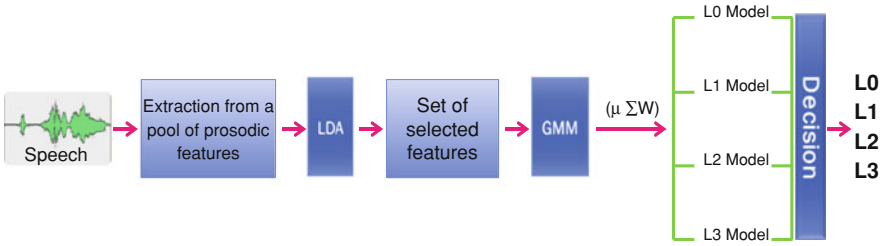


Fig. 2 LDA-GMM system

4.4 Automatic Classifiers of Severity Level

We compare two approaches for automatic classification as a front-end for the eleven prosodic features selected by LDA. The two methods, GMM and SVM, perform training and classification. We divide the entire set of sentences of the corpus into two subsets: a training subset that contains 70 % of the sentences with different severity levels of dysarthria, and a test subset that contains 30 % of the sentences. The training subset includes 459 sentences of dysarthric speech and 153 sentences of non-dysarthric speech (HC); the test subset contains 207 sentences of dysarthric speech and 69 sentences of non-dysarthric speech.

4.4.1 GMM

The automatic classification of observed vectors into one of I classes can be performed using GMM. The combined LDA-GMM system is represented in Fig. 2.

Training: For each class C_i in the corpus, the training is initiated to obtain a model containing the characteristics of each Gaussian distribution m of the class: the average vector $\mu_{i,m}$, the covariance matrix $\sum_{i,m}$, and the weight of the Gaussian $w_{i,m}$. These parameters are calculated after performing a certain number of iterations of the expectation-maximization (EM) algorithm [25]. One model is generated for each severity level of dysarthria.

Recognition: Each extracted signal X is represented by the acoustical vector x of p components. The size of the acoustical vector d is the number of acoustical parameters extracted from the signal. The likelihood of each acoustical vector for a given class C_i is estimated, and the likelihood is defined by [26]:

$$p(x\backslash C_i) = \sum_{m=1}^{M} w_{i,m} \cdot \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_{i,m}\right|}} . e^{A_{i,m}}$$
$$A_{i,m} = \left(-\frac{1}{2} \left(x - \mu_{i,m}\right)^T . \frac{1}{\sum_{i,m}} . \left(x - \mu_{i,m}\right) \right) \tag{7}$$

where M is the number of Gaussians.

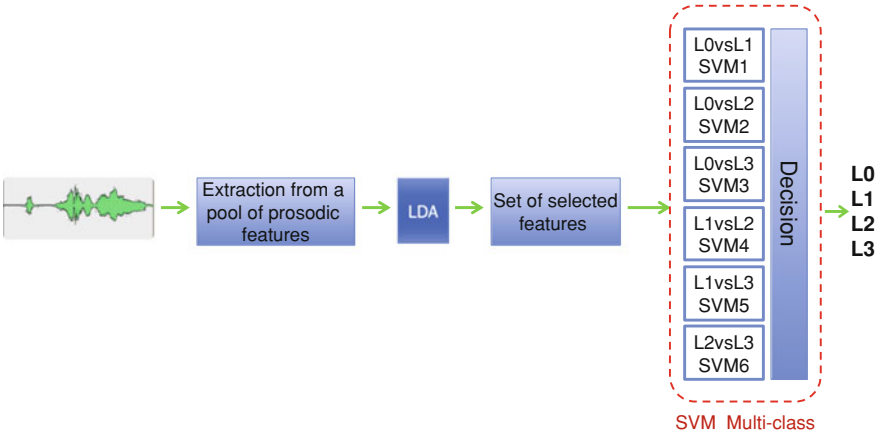


Fig. 3 LDA-SVM system

Each sentence is represented by one acoustical vector containing eleven prosodic features (not by one vector for each frame), and the likelihood of the signal is denoted as $p(x\backslash C_i)$. The algorithm estimates that signal X will belong to the group C_i in which $p(x\backslash C_i)$ is greater. The highest rate was achieved using eight Gaussians ($M = 8$), where 88.89 % of dysarthria severity levels were correctly classified.

4.4.2 SVM

SVMs are binary classifiers that use an optimal hyperplane to discriminate between two classes. Figure 3 illustrates the LDA-SVM system.

The SVM was proposed by Vapnik as a new method of machine learning, via introduction of the kernel function [27]. The kernel function projects the (non-linearly separable) data to a new high-dimensional space, where a linear separation is possible. SVMs determine the linear hyperplane separator that maximizes the margin between two classes of data.

The key component of an SVM is the kernel function. The choice of kernel function will affect the learning ability and generalization capacity of machine learning [28]. In our experiments, a radial basis function (RBF) is used as a kernel. The properties of the RBF depend on the Gaussian width σ and the error penalty parameter C. The RBF kernel is defined as:

$$k(x, y) = exp\left(\frac{\left\|x - y^2\right\|}{\sigma^2}\right) \tag{8}$$

Table 4 Performance comparison of proposed methods

System	LDA	LDA-GMM	LDA-SVM
% of correct classification	84.2	88.9	93

A multiclass-SVM using the 'one-against-one' method was set to classify the severity levels of dysarthria. Binary classifiers are built to differentiate classes C_i and C_j, $0 < i \leq I$ and $0 < j < i$, where I is the number of classes [29]. The number of binary classifis (SVMs) necessary to classify I classes is $\frac{I(I-1)}{2}$.

The multiclass-SVM includes six SVMs and a decision function based on majority voting (best candidate) using all classifiers. For each of the six SVMs, a cross-validation was carried out over four subsets of the corpus to determine the most relevant pair (C, σ) of the RBF kernel function. This method of automatic assessment achieved correct dysarthria severity level classification in **93 %** of cases.

Table 4 compares performance between LDA, LDA-GMM, and LDA-SVM systems.

5 Conclusion and Future Work

In this paper, we proposed and compared GMM and SVM discriminative approaches to perform an assessment of the dysarthria severity level. A reliable front-end processing technique using relevant prosodic features was developed, and these features were selected after performing LDA. We believe that the proposed system could constitute an appropriate objective test for the automatic evaluation of dysarthria severity. This tool might be useful for clinicians, and can be used to prevent incorrect subjective diagnosis and reduce the time required to monitor or diagnose dysarthric speech disorder.

Further works need to be done, to test this system on other dysarthric speech databases for more validation, and to develop an ergonomic tool interface adapted for clinicians. Moreover, this method can be used to characterize other speech pathologies.

Acknowledgments This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. C. Roth, in *Encyclopedia of Clinical Neuropsychology*, ed. by: B. Caplan, J. Deluca, J.S. Kreutzer (Springer, Heidelberg, 2011), pp. 905–908
2. S.-A. Selouani, H. Dahmani, R. Amami, H. Hamam, in *SOCO 2011*: *Dysarthric Speech Classification Using Hierarchical Multilayer Perceptrons and Posterior Rhythmic Features*. Proceedings of 6th International Conference. Soft Computing Models in Industrials and Environmental Applications

3. American Speech-Language-Hearing Association [Online]. Available: http://www.asha.org
4. J.B. Polikoff, H.T. Bunnel, in *ICPhS*: *The Nemours database of dysarthric speech: a perceptual analysis*. Proceedings of 14th International Congress of Phonetic Sciences, 1999, pp. 783–786
5. S.-A. Selouani, H. Dahmani, R. Amami, H. Hamam, Using speech rhythm knowledge to improve dysarthric speech recognition. Int. J. Speech Technol. **15**(1), 57–64 (2012)
6. F. Rudzicz, in *ICASSP 2009*: *Phonological Features in Discriminative Classification of Dysarthric Speech*
7. M.S. Paja, T.H. Falk, in *Automated Dysarthria Severity Classification for Improved Objective Intelligibility Assessment of Spastic Dysarthric Speech*. Interspeech, 2012
8. X. Menendez-Pidal, J.B. Polikoff, S.M. Peters, J.E. Leonzio, H.T. Bunnell, in *ICSLP*: *The Nemours Database of Dysarthric Speech*. Fourth International Conference on Spoken Language, vol. 3 (IEEE, New York, 1996) pp. 1962–1965
9. L. Mary, B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition. Speech Commun. **50**(10), 782–796 (2008)
10. E. Shriberg, A. Stolcke, D. Hakkani, Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication **32**(1–2), 127–154 (2000). (Special Issue on Accessing Information in Spoken Audio)
11. J.R. Duffy, Motor Speech Disorders: Clues to Neurologic Diagnosis, in *Parkinson's Disease and Movement Disorders*, ed. by C.H. Adler, J.E. Ahlskog (Springer, Heidelberg, 2000), pp. 35–53
12. K.L. Kadi, S.-A. Selouani, B. Boudraa, M. Boudraa, in *WCE 2013*: *Discriminative Prosodic Features to Assess the Dysarthria Severity Levels*. Proceedings of The World Congress on Engineering 2013, 3–5 July London. Lecture Notes in Engineering and Computer Science, pp. 2201–2205
13. R. Kent, H. Peters, P. Van-Lieshout, W. Hulstijn, *Speech Motor Control in Normal and Disordered Speech* (Oxford University Press, London, 2004)
14. J.T. Hart, R. Collier, A. Cohen, *A Perceptual Study of Intonation* (Cambridge University Press, Cambridge, 1990)
15. L. Mary, in *Extraction and Representation of Prosody for Speaker, Speech and Language Recognition* (Springer Briefs in Speech Technology, 2012), chap. 1
16. H.F. Westzner, S. Schreiber, L. Amaro, Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. Braz. J Orthinolaryngol. **71**(5), 582–588 (2005)
17. Multi-Dimensional Voice Processing Program (MDVP), Kay Elemetrics Company: http://www.kayelemetrics.com
18. L. Baghai-Ravary, S.W. Beet, in *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. Springer Briefs in Electrical and Computer Engineering, 2013
19. J.M. Liss, L. White, S.L. Mattys, K. Lansford, A.J. Lotto, S.M. Spitzer, J.N. Caviness, Quantifying speech rhythm abnormalities in the dysarthrias. J Speech, Lang. Hear. Res. **52**, 1334–1352 (2009)
20. Calliope, La parole et son traitement automatique, Dunod, 1989
21. P. Boersma, D. Weenink, Praat, a system for doing phonetics by computer. Glot Int **5**(9–10), 341–345 (2001)
22. C.E. Guerra, D.F. Lovey, in *EMBS 2003*: *A Modern Approach to Dysarthria Classification*. Proceedings of the 25th Annual International Conference of the IEEE, New York. Engineering in Medecine and Biology Society
23. Copyright IBM Corporation., 1989–2012. Available: http://www.ibm.com
24. A. El Ouardighi, A. El Akadi, D. Aboutadjine, in *ISCCIII*: *Feature Selection on Supervised Classification Using Wilk's Lambda Statistic*. International Symposium on Computational Intelligence and Intelligent Informatics, 2007, pp. 51–55
25. A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm. J. Acoust. Soc. Am. **39**(1), 1–38 (1977)

26. D. Istrate, E. Castelli, M. Vacher, L. Besacier, J. Serignat, Information extraction from sound for medical telemonitoring. IEEE Trans. Inf. Technol. Biomed. **10**(2), 264–274 (2006)
27. V.N. Vapnik, An overview of statistical learning theory. IEEE Trans. Neural Networks **10**(5), 988–999 (1999)
28. H. Gao, A. Guo, X. Yu, C. Li, in WiCOM'08: *Rbf-Svm and its Application on Network Security Risk Evaluation*. Proceedings of 4th International Conference on Wireless Communication, Networking and Mobile Computing, 2008
29. A. Fleury, M. Vacher, N. Noury, SVM-Based multimodal classification of Activities of daily living in health smart homes: sensors, algorithms, and first experimental results. IEEE Trans. Inf. Technol.Biomed. **14**(2), 274–283 (2010)

Parallelization of Minimum Spanning Tree Algorithms Using Distributed Memory Architectures

Vladimir Lončar, Srdjan Škrbić and Antun Balaž

Abstract Finding a minimum spanning tree of a graph is a well known problem in graph theory with many practical applications. We study serial variants of Prim's and Kruskal's algorithm and present their parallelization targeting message passing parallel machine with distributed memory. We consider large graphs that can not fit into memory of one process. Experimental results show that Prim's algorithm is a good choice for dense graphs while Kruskal's algorithm is better for sparse ones. Poor scalability of Prim's algorithm comes from its high communication cost while Kruskal's algorithm showed much better scaling to larger number of processes.

Keywords Distributed memory · Kruskal · MPI · MST · Paralellization · Prim

1 Introduction

A minimum spanning tree (MST) of a weighted graph $G = (V, E)$ is a subset of E that forms a spanning tree of G with minimum total weight. MST problem has many applications in computer and communication network design, as well as indirect applications in fields such as computer vision and cluster analysis [12].

V. Lončar · S. Škrbić (✉)
Faculty of Science, University of Novi Sad, Trg Dositeja Obradovica 4, Novi Sad, Serbia
e-mail: srdjan.skrbic@dmi.uns.ac.rs

V. Lončar
e-mail: vladimir.loncar@dmi.uns.ac.rs

A. Balaž
Scientific Computing Laboratory, Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, Belgrade, Serbia
e-mail: antun.balaz@scl.rs

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_39,

In this paper we implement two parallel algorithms for finding MST of a graph, based on classical algorithms of Prim [23] and Kruskal [18], building upon our previous work in [19]. Algorithms target message passing parallel machine with distributed memory. Primary characteristic of this architecture is that the cost of inter-process communication is high in comparison to cost of computation. Our goal was to develop algorithms which minimize communication, and to measure the impact of communication on the performance of algorithms. Our primary interest were graphs which have significantly larger number of vertices than processors involved in computation. Since graphs of this size cannot fit into the memory of a single process, we use a partitioning scheme to divide the input graph among processes. We consider both sparse and dense graphs.

First algorithm is a parallelization of Prim's serial algorithm. Each process is assigned a subset of vertices and in each step of computation, every process finds a candidate minimum-weight edge connecting one of its vertices to MST. The root process collects those candidates and selects one with minimum weight which it adds to MST and broadcasts result to other processes. This step is repeated until every vertex is in MST.

Second algorithm is based on Kruskal's approach. Processes get a subset of G in the same way as in first algorithm, and then find local minimum spanning tree (or forest). Next, processes merge their MST edges until only one process remains, which holds edges that form MST of G.

Implementations of these algorithms are done using C programming language and MPI (Message Passing Interface) and tested on a parallel cluster PARADOX using up to 256 cores and 256 GB of distributed memory.

Section 2 contains references to the most important related papers. In Sect. 3 we continue with the description and analysis of algorithms—both serial and parallel versions, and their implementation. In the last section we describe experimental results, analyze them and draw our conclusions.

2 Related Work

Algorithms for MST problem have mostly been based on one of three approaches, that of Boruvka [3], Prim [23] and Kruskal [18], however, a number of new algorithms has been developed. Gallager et al. [10] presented an algorithm where processor exists at each node of the graph (thus $n = p$), useful in computer network design. Katriel and Sanders designed an algorithm exploiting cycle property of a graph targeting dense graph, [17], while Ahrabian and Nowzari-Dalini's algorithm relies on depth first search of the graph [1].

Due to its parallel nature, Boruvka's algorithm (also known as Sollin's algorithm) has been the subject to most research related to parallel MST algorithms. Examples of algorithms based on Boruvka's approach include Chung and Condon [4], Wang and Gu [14] and Dehne and Götz [7].

Parallelization of Prim's algorithm has been presented by Deo and Yoo [8]. Their algorithm targets shared memory computers. Improved version of Prim's algorithm has been presented by Gonina and Kale [11]. Their algorithm adds multiple vertices per iteration, thus achieving significant speedups. Another approach targeting shared memory computers presented by Setia et al. [24] uses the cut property of a graph to grow multiple trees in parallel. Hybrid approach, combining both Boruvka's and Prim's approaches has been developed by Bader and Cong [2].

Examples of parallel implementation of Kruskal's algorithm can be found in work of Jin and Baker [16], and Osipov et al. [21]. Osipov et al. proposes a modification to Kruskal's algorithm to avoid edges which certainly are not in a graph. Their algorithm runs in near linear time if graph is not too sparse.

Bulk of the research into parallel MST algorithms has targeted shared memory computers like PRAM, i.e. computers where entire graph can fit into memory. Our algorithms target distributed memory computers and use partitioning scheme to divide the input graph evenly among processors. Because no process contains info about partition of other processes, we designed our algorithms to use predictable communication patterns, and not depend on the properties of input graph.

3 The Algorithms

Let us assume that graph $G = (V, E)$, with vertex set V and edge set E is connected and undirected. Without loss of generality, it can be assumed that each weight is distinct, thus G is guaranteed to have only one MST. This assumption simplifies implementation, otherwise a numbering scheme can be applied to edges with same weight, at the cost of additional implementation complexity.

Let n be the number of vertices, m the number of edges ($|V| = n$, $|E| = m$), and p the number of processes involved in computation of MST. Let $w(v, u)$ denote weight of edge connecting vertices v and u. Input graph G is represented as $n \times n$ adjacency matrix $A = (a_{i,j})$ defined as:

$$a_{i,j} = \begin{cases} w(v_i, v_j) & \text{if } (\mathrm{v_i}, \mathrm{v_j}) \in \mathrm{E} \\ 0 & \textit{otherwise} \end{cases} \tag{1}$$

3.1 Prim's Algorithm

Prim's algorithm starts from an arbitrary vertex and then grows the MST by choosing a new vertex and adding it to MST in each iteration. Vertex with an edge with lightest weight incident on the vertices already in MST is added in every iteration. The algorithm continues until all the vertices have been added to the MST. This algorithm requires $O(n^2)$ time. Implementations of Prim's algorithm commonly use auxiliary array d of length n to store distances (weight) from each

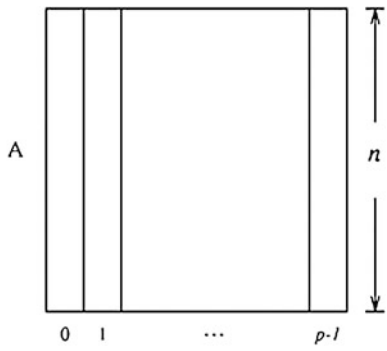


Fig. 1 Partitioning of adjacency matrix among p processes

vertex to MST. In every iteration a lightest weight edge in d is added to MST and d is updated to reflect changes.

Parallelizing the main loop of Prim's algorithm is difficult [13], since after adding a vertex to MST lightest edges incident on MST change. Only two steps can be parallelized: selection of the minimum-weight edge connecting a vertex not in MST to a vertex in MST, and updating array d after a vertex is added to MST. Thus, parallelization can be achieved in the following way:

1. Partition the input set V into p subsets, such that each subset contains n/p consecutive vertices and their edges, and assign each process a different subset. Each process also contains part of array d for vertices in its partition. Let V_i be the subset assigned to process p_i, and d_i part of array d which p_i maintains. Partitioning of adjacency matrix is illustrated in Fig. 1.
2. Every process p_i finds minimum-weight edge e_i (*candidate*) connecting MST with a vertex in V_i.
3. Every process p_i sends its e_i edge to the root process using all-to-one reduction.
4. From the received edges, the root process selects one with a minimum weight (called global minimum-weight edge e_{min}), adds it to MST and broadcasts it to all other processes.
5. Processes mark vertices connected by e_{min} as belonging to MST and update their part of array d.
6. Repeat steps 2–5 until every vertex is in MST.

Finding a minimum-weight edge and updating of d_i during each iteration costs $O(n/p)$. Each step also adds a communication cost of all-to-one reduction and all-to-one broadcast. These operations complete in $O(\log p)$. Combined, cost of one iteration is $O(n/p + \log p)$. Since there are n iterations, total parallel time this algorithm runs in is:

$$T_p = O\left(\frac{n^2}{p}\right) + O(n \log p) \tag{2}$$

Prim's algorithm is better suited for dense graphs and works best for complete graphs. This also applies to its parallel formulation presented here. Ineffectiveness of the algorithm on sparse graphs stems from the fact that Prim's algorithm runs in $O(n^2)$, regardless of the number of edges. A well-known modification [5] of Prim's algorithm is to use binary heap data structure and adjacency list representation of a graph to reduce the run time to $O(m \log n)$. Furthermore, using Fibonacci heap asymptotic running time of Prim's algorithm can be improved to $O(m + n \log n)$. Since we use adjacency matrix representation, investigating alternative approaches for Prim's algorithm was out of the scope of this paper.

3.2 Kruskal's Algorithm

Unlike Prim's algorithm which grows a single tree, Kruskal's algorithm grows multiple trees in parallel. Algorithm first creates a forest F, where each vertex in the graph is a separate tree. Next step is to sort all edges in E based on their weight. Algorithm then chooses minimum-weight edge e_{min} (i.e. first edge in sorted set). If e_{min} connects two different trees in F, it is added to the forest and two trees are combined into a single tree, otherwise e_{min} is discarded. Algorithm loops until either all edges have been selected, or F contains only one tree, which is the MST of G. This algorithm is commonly implemented using Union-Find algorithm [22]. *Find* operation is used to determine which tree a particular vertex is in, while *Union* operation is used to merge two trees. Kruskal's algorithm runs in $O(m \log n)$ time, but can be made even more efficient by using more sophisticated Union-Find data structure, which uses *union by rank* and *path compression* [9]. If the edges are already sorted, using improved Union-Find data structure Kruskal's algorithm runs in $O(m\alpha(n))$, where $\alpha(n)$ is the inverse of the Ackerman function.

Our parallel implementation of Kruskal's algorithm uses the same partitioning scheme of adjacency matrix as in Prim's approach and is thus bounded by $O(n^2)$ time to find all edges in matrix. Having that in mind, our parallel algorithm proceeds through the following steps:

1. Every process p_i first sorts edges contained in its partition V_i.
2. Every process p_i finds a local minimum spanning tree (or forest, MSF) F_i using edges in its partition V_i applying the Kruskal's algorithm.
3. Processes merge their local MST's (or MSF's). Merging is performed in the following manner. Let a and b denote two processes which are to merge their local trees (or forests), and let F_a and F_b denote their respective set of local MST edges. Process a sends set F_a to b, which forms a new local MST (or MSF) from $F_a \cup F_b$. After merging, process a is no longer involved in computation and can terminate.
4. Merging continues until only one process remains. Its MST is the end result.

Creating a new local MSF during merge step can be performed in a number of different ways. Our approach is to perform Kruskal's algorithm again on $F_a \cup F_b$.

Computing the local MST takes $O(n^2/p)$. There is a total of log p merging stages, each costing $O(n^2 \log p)$. During one merge step one process transmits maximum of $O(n)$ edges for a total parallel time of:

$$T_p = O(n^2/p) + O(n^2 \log p) \tag{3}$$

Based on speedup and efficiency metrics, it can be shown that this parallel formulation is efficient for $p = O(n/\log n)$, same as the first algorithm.

3.3 Implementation

Described algorithms were implemented using ANSI C and Message Passing Interface (MPI). Fixed communication patterns in parallel formulation of the algorithms map directly to MPI operations. Complete source code can be found in [25].

4 Experimental Results

Implementations of algorithms were tested on a cluster of up to 32 computing nodes. Each computer in the cluster had two Intel Xeon E5345 2.33 GHz quad-core CPUs and 8 GB of memory, with Scientific Linux 6 operating system installed. We used OpenMPI v1.6 implementation of the MPI standard. The cluster nodes are connected to the network with a throughput of 1 Gbit/s. Both implementations were compiled using GCC 4.4 compiler. This cluster has enabled testing algorithms with up to 256 processes as shown in Table 1.

We tested graphs with densities of 1, 5, 10, 15 and 20 % with number of vertices ranging from 10,000 to 100,000, and number of edges from 500,000 to 1,000,000,000. Distribution of edges in graphs was uniformly random, and all edge weights were unique. Due to the high memory requirements of large graphs, not every input graph could be partitioned in a small number of cluster nodes, as can be seen in Table 1.

4.1 Results

Due to the large amount of obtained test results, we only present the most important ones here. Complete set of results can be found in [25].

In the Table 2 we show the behavior of algorithms with increasing number of processes on input graph of 50,000 vertices and density of 10 %.

Results show poor scalability of Prim's algorithm, due to its high communication cost. Otherwise, computation phase of Prim's algorithm is faster than that of

Table 1 Testing parameters

Processes	Nodes	Processes per node	No. of vertices (k)
4	4	1	10–50
8	8	1	10–60
16	16	1	10–80
32	32	1	10–100
64	32	2	10–100
128	32	4	10–100
256	32	8	10–100

Table 2 CPU time (in seconds) for algorithms with increasing number of processes

	4	8	16	32	64	128	256
Kruskal	38.468	19.94	10.608	5.342	2.958	1.796	1.382
Prim	16.703	15.479	25.201	30.382	30.824	32.661	39.737

Table 3 CPU time (in seconds) for algorithms with increasing density

	1 %	5 %	10 %	15 %	20 %
Kruskal	0.607	2.603	5.342	8.164	10.663
Prim	30.189	30.007	30.382	30.518	30.589

Kruskal's. Due to the usage of adjacency matrix graph representation, Prim's algorithm performs almost the same regardless of the density of the input graph. This can be seen from the results of input graph with 50,000 vertices and 32 processes with varying density shown at Table 3.

On the other hand, Kruskal's algorithm shows degradation of performance with increasing density. Results of Kruskal's algorithm show that majority of local computation time is spent sorting the edges of input graph, which grows with larger density. Increasing the number of processes makes local partitions smaller and faster to process, thus allowing this algorithm to achieve good scalability. If the edges of input graph were already sorted, Kruskal's algorithm would be significantly faster than other MST algorithms.

4.2 Impact of Communication Overhead

Cost of communication is much greater than the cost of computation, so it is important to analyse the time spent in communication routines. During tests we measured the time spent waiting for the completion of the communication operations. In case of Prim's algorithm, we measured the time that the root process spends waiting for the completion of MPI_Reduce and MPI_Bcast operations. Communication in Kruskal's algorithm is measured as total time spent waiting for messages received over MPI_Recv operation in the last active process (which will

Table 4 Communication versus computation time (in seconds)

Processes	4	8	16	32	64	128	256
Prim's algorithm							
Total	16.703	15.479	25.201	30.382	30.824	32.661	39.737
Communication	8.188	11.183	23.009	29.248	30.237	32.322	39.467
Kruskal's algorithm							
Total	38.468	19.94	10.608	5.342	2.958	1.796	1.382
Communication	0.171	0.356	0.371	0.288	0.317	0.253	0.256

contain the MST after last iteration of the merge operation). This gives us a good insight into the duration of communication routines because the last active process will have to wait the most.

The Table 4 shows communication times of processing input graph of 50,000 vertices with 10 % density.

When comparing communication time with a total computation time it can be noted that the Prim's algorithm spends most of time in communication operations, and by increasing number of processes almost all the running time of the algorithm is spent on communication operations. A bottleneck in Prim's algorithm is the cost of MPI_Reduce and MPI_Bcast communication operations. These operations require communication between all processes, and are much more expensive than local computation within each process, because all processes must wait until the operation is completed, or until the data are transmitted over the network. This prevents Prim's algorithm from achieving substantial speedup of running time with increasing number of processes. Therefore, this algorithm is most efficient on the fewest number of processes that the partitioned input graph can fit.

On the other hand Kruskal algorithm spends much less time in communication operations, but instead spends most of the time in local computation. These differences are illustrated in Figs. 2 and 3. The diagrams show that communication in Prim's algorithm rises sharply with increasing number of processes, while execution time slowly reduces. In Kruskal's algorithm, the situation is reversed.

4.3 Analysis of Results

The experimental results confirmed some of the assumptions made during the development and analysis of algorithms, but also made a couple of unexpected results. Results of these experiments gave us directions for further improvement of the described algorithms.

Prim's algorithm has shown excellent performance in computational part of the algorithm, but a surprisingly high cost of communication operations spoils its final score. Finding candidate edges for inclusion in MST can be further improved by using techniques described in [5], but it will not significantly improve the total

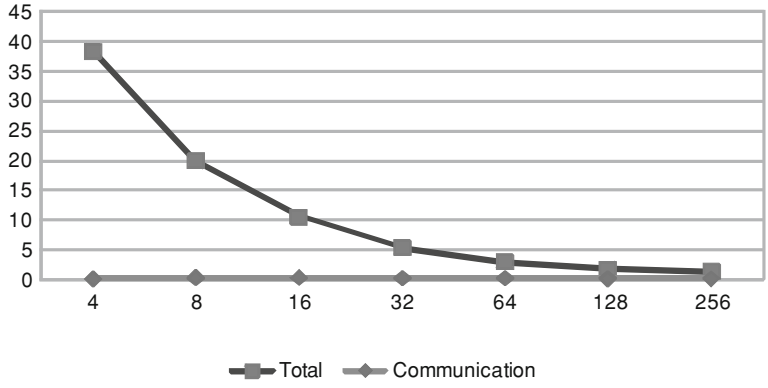


Fig. 2 Communication in Kruskal's algorithm

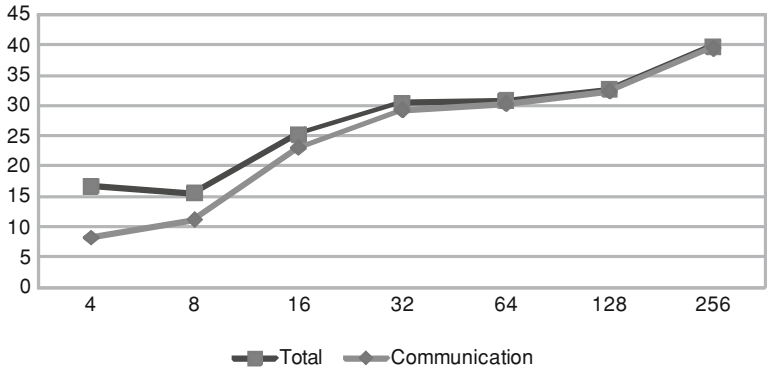


Fig. 3 Communication in Prim's algorithm

time of the algorithm, as communication routines will remain the same. Unfortunately, the communication can not be further improved by changing the algorithm. The only way to reduce the cost of communication is to use a cluster that has a better quality network, or to rely on the semantics of the implementation of the MPI operation MPI_Allreduce.

Kruskal's algorithm has shown good performance, especially for sparse graphs, while the performance degrades with increasing density. It is important to note that many real-world graphs have density much smaller than 1 % (for example, graph of roads as egdes and junctions as vertices has a density much smaller than 1 %). Also, this algorithm showed much better scaling to larger number of processes than Prim's algorithm. Cost of communication in Kruskal's algorithm is much smaller than in Prim's algorithm, but the local computation is slower. This can be improved by using more efficient Union-Find algorithms [9], or by improving merging of local trees between processes. Kruskal's algorithm does not use a lot of

slow messages like Prim's algorithm, but can send very large messages depending on the number of processes and the size of the graph. This can be improved by introducing techniques for compressing messages, or changing the structure of the message.

5 Alternate Parallelization Approaches

In this section we will give a brief overview of two other parallelization approaches we considered using for implementation of these algorithms. One approach would be using graphics processing unit (GPU) technologies like Nvidia CUDA or OpenCL. Another would be using shared-memory parallelization API like OpenMP to utilize multi-core processors on cluster nodes. We will go over advantages and disadvantages of both approaches.

With the introduction of CUDA and OpenCL programming models, using GPU for general-purpose computing (GPGPU) has become a powerful alternative to traditional CPU programming models. Nowadays GPUs can be found in most high-ranking supercomputers and even ordinary clusters. GPUs have their own RAM, which is separate from main RAM of a computer and was not accessible for distributed-memory technologies like MPI. This made writing multi-GPU programs more difficult, since it required expensive copy operations between GPU memory and host (CPU) memory which MPI could access. However, recent developments in MPI implementations have alleviated this problem, and newer versions of popular MPI implementations like OpenMPI and MVAPICH can access GPU memory directly. This unfortunately still doesn't make GPU the perfect platform for implementations of our algorithms. GPUs still have much smaller amount of RAM when compared to main memory (recently released models like Tesla K10 have up to 8 GB of memory [20]). This means that GPU solution could only be used on much smaller graphs. Alternatively, a different graph representation (like adjacency lists) would allow graphs with greater number of vertices, but would still be only useful for sparser graphs. Primary part of Prim's algorithm which could be accelerated by GPU is finding local (and then global) vertex with the smallest distance to the tree. This could be achieved by slightly modifying well-known parallel reduction algorithm for GPU [15]. Communication pattern between nodes would remain the same. Kruskal's algorithm is more complex to implement on GPU due to Union-Find data structure. Other important portions of Kruskal's algorithm, like sorting of input could be done using various GPU libraries.

Unlike the relatively new technology that is GPGPU, OpenMP has been successfully used to parallelize serial code since the late 90s. In some cases, OpenMP allows developers to parallelize their with programs with minimal effort, using compiler directives around loops, often with good performance [6]. This technique could be used in parallelization of Prim's algorithm for finding local (and later

global) vertex with the smallest distance to the tree. Graph would be partitioned in such a way that each node in cluster receives an equal part, then each node would use all it's processors and cores with OpenMP to find local minimum, and use MPI for communication between nodes. Kruskal's algorithm can be parallelized in similar way, although it would require a slightly greater effort for implementation of sorting and Union-Find data structure.

Acknowledgements Authors are partially supported by Ministry of Education, Science, and Technological Development of the Republic of Serbia, through projects no. ON174023: "Intelligent techniques and their integration into wide-spectrum decision support", and ON171017: "Modeling and numerical simulations of complex many-body systems", as well as European Commission through FP7 projects PRACE-2IP and PRACE-3IP.

References

1. H. Ahrabian, A. Nowzari-Dalini, Parallel algorithms for minimum spanning tree problem. Int. J. Comput. Math. **79**(4), 441–448 (2002)
2. D.A. Bader, G. Cong, Fast shared-memory algorithms for computing the minimum spanning forest of sparse graphs. J. Parallel Distrib. Comput. **66**(11), 1366–1378 (2006)
3. O. Boruvka, O Jistém Problému Minimálnm (about a certain minimal problem) (in Czech, German summary). *Práce Mor. Prrodoved. Spol. v Brne III*, vol. 3 (1926)
4. S. Chung, A. Condon, Parallel implementation of borvka's minimum spanning tree algorithm. in *Proceedings of the 10th International Parallel Processing Symposium, IPPS '96* (IEEE Computer Society, Washington, DC, 1996), pp. 302–308
5. T.H. Cormen, C. Stein, R.L. Rivest, C.E. Leiserson, *Introduction to Algorithms*, 2nd edn. (McGraw-Hill Higher Education, Boston, 2001)
6. M. Curtis-Maury, X. Ding, C.D. Antonopoulos, D.S. Nikolopoulos, *An Evaluation of Openmp on Current and Emerging Multithreaded/Multicore Processors*, ed. by M.S. Mueller, B.M. Chapman, B.R. Supinski, A.D. Malony, M. Voss. *OpenMP Shared Memory Parallel Programming*, vol 4315 (*Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2008), pp. 133–144
7. F. Dehne, S. Gtz, Practical Parallel Algorithms for Minimum Spanning Trees, *in Workshop on Advances in Parallel and Distributed Systems* (1998), pp. 366–371
8. N. Deo, Y.B. Yoo, Parallel algorithms for the minimum spanning tree problem, in *Proceedings of the International Conference on Parallel Processing* (1981), pp. 188–189
9. Z. Galil, G.F. Italiano, Data structures and algorithms for disjoint set union problems. ACM Comput. Surv. **23**(3), 319–344 (1991)
10. R.G. Gallager, P.A. Humblet, P.M. Spira, A distributed algorithm for minimum-weight spanning trees. ACM Trans. Program. Lang. Syst. **5**(1), 66–77 (1983)
11. E. Gonina, L.V. Kale, Parallel prim's algorithm on dense graphs with a novel extension, in *PPL Technical Report*, Oct 2007
12. R.L. Graham, P. Hell, On the history of the minimum spanning tree problem. IEEE Ann. Hist. Comput. **7**(1), 43–57 (1985)
13. A. Grama, G. Karypis, V. Kumar, A. Gupta, *Introduction to Parallel Computing, 2nd edn*. (Addison Wesley, Reading, 2003)
14. W. Guang-rong, G. Nai-jie, An efficient parallel minimum spanning tree algorithm on message passing parallel machine. J. Softw. **11**(7), 889–898 (2000)
15. M. Harris, Optimizing parallel reduction in CUDA. CUDA tips and tricks

16. M. Jin, J.W. Baker, Two graph algorithms on an associative computing model, in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 2007*, vol 1, Las Vegas, Nevada, 25–28 June 2007, pp. 271–277
17. I. Katriel, P. Sanders, J.L. Trff, J.L. Tra, A practical minimum spanning tree algorithm using the cycle property, in *11th European Symposium on Algorithms (ESA),* vol. *2832 in LNCS* (Springer, New York, 2003), pp. 679–690
18. J.B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**(1), 48–50 (1956)
19. V. Lončar, S. Škrbić, A. Balaž, Distributed memory parallel algorithms for minimum spanning trees, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE 2013*, London, 3–5 July 2013, pp. 1271–1275
20. Nvidia, Nvidia tesla GPU accelerators. Nvidia Tesla Product Datasheet (2012)
21. V. Osipov, P. Sanders, J. Singler, The filter-kruskal minimum spanning tree algorithm, in *ALENEX'09* (2009), pp. 52–61
22. D.-Z. Pan, Z.-B. Liu, X.-F. Ding, Q. Zheng, The application of union-find sets in kruskal algorithm, in *Proceedings of the 2009 International Conference on Artificial Intelligence and Computational Intelligence (AICI '09),* vol 2 (IEEE Computer Society, Washington, DC, 2009), pp. 159–162
23. R.C. Prim, Shortest connection networks and some generalizations. Bell Syst. Technol. J. **36**, 1389–1401 (1957)
24. R. Setia, A. Nedunchezhian, S. Balachandran, A new parallel algorithm for minimum spanning tree problem, in *Proceedings of the International Conference on High Performance Computing (HiPC)* (2009), pp. 1–5
25. S. Škrbić, Scientific Computing Seminar (2013)

Experiments with a Sparse Distributed Memory for Text Classification

Mateus Mendes, A. Paulo Coimbra, Manuel M. Crisóstomo and Jorge Rodrigues

Abstract The Sparse Distributed Memory (SDM) has been studied for decades as a theoretical model of an associative memory in many aspects similar to the human brain. It has been tested for different purposes. The present work describes its use as a quick text classifier, based on pattern similarity only. The results found with different datasets were superior to the performance of the dumb classifier or purely random choice, even without text preprocessing. Experiments were performed with a popular Reuters newsgroups dataset and also for real time web ad serving.

Keywords SDM · Sparse distributed memory · Text classification · Text comparison · Long range correlations · Vector space model

1 Introduction

Text classification and text similarity calculation are areas of increasing interest. They are important for modern search engines, information retrieval, targeted advertising and other applications that require retrieving the most appropriate, best matching texts, for specific purposes. As more and more data are stored in digital

M. Mendes (✉) · J. Rodrigues
ESTGOH, Polytechnic Institute of Coimbra, Coimbra, Portugal
e-mail: mmendes@estgoh.ipc.pt

J. Rodrigues
e-mail: jorge.rodrigues@estgoh.ipc.pt

M. Mendes · A. Paulo Coimbra · M. M. Crisóstomo · J. Rodrigues
Institute of Systems and Robotics, Pólo II, University of Coimbra, 3000 Coimbra, Portugal
e-mail: acoimbra@deec.uc.pt

M. M. Crisóstomo
e-mail: mcris@isr.uc.pt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_40,

databases, quick and accurate retrieval methods are necessary for a better user experience.

In many applications, retrieval does not need to be exact, and there may not even be an *exact* match. An example is choosing the right ads to exhibit in a website based on the website's contents: the ad server needs to output relevant ads as quickly as possible and in general all on-topic ads could be acceptable from the semantic point of view. Another example may be service robots. Service robots may execute tasks where accuracy is very important, and have other tasks where accuracy may be traded for speed, memory or other important tradeoff. A service robot which guides visitors inside a building must not take them to wrong destinations. But a service robot reading the latest news about artificial intelligence occasionally skipping one or two entries is not problematic at all.

On the other hand, it has long been known that non-random texts exhibit long-range correlations between lower level symbol representations and higher level semantic meaning [1]. Thus, in theory it may be possible to achieve some results by processing lower level symbols directly without getting to the upper structure levels. This possibility is worth exploring for real time applications, where speed is often more important than accuracy, as long as the expected results fall into acceptable limits.

The Sparse Distributed Memory (SDM) is a type of associative memory proposed by Pentti Kanerva in the 1980s [2]. It is based on the properties of high-dimensional spaces, where data are stored based on pattern similarity. Similar vectors are stored close to one another, while dissimilar vectors are stored farther apart in the memory. As for data retrieval, a very small clue is often enough to retrieve the correct datum. In theory, knowing only 20 % of the bits of a binary vector may be enough to retrieve it from the memory. To a great extent, the SDM exhibits characteristics very similar to the way the human brain works. The SDM has been successfully used in tasks such as predicting the weather [3] and navigating robots [4].

The present work is an extension of [5]. In the previous work the SDM was used for text classification, taking TF-IDF (Term Frequency-Inverse Document Frequency) vectors as input and raw text directly without any text processing. The results obtained with the SDM were inferior to the results obtained with other modern methods, but superior to the performance of a "dumb classifier." Interestingly, it was possible to obtain those results directly with raw input, skipping any text processing. Later, the SDM has been applied as a classifier to a web ad server, to choose which ad to show in a publisher website. The Click Through Rate (CTR) obtained was superior to the CTR obtained if the ads were just chosen randomly. Additionally, it has been shown that the way the information is encoded may influence the performance of the SDM [6]. Thus, it may be possible to achieve better results in the future by just changing the text encoding or by making other small adjustments to the input texts.

Section 2 briefly explains the process of text classification and the origins of long-range semantic correlations. Section 3 briefly describes the SDM. Section 4

describes the experiments performed using popular datasets. Section 5 describes the experiments performed in an ad server. Section 6 draws some conclusions and opens perspectives of future work.

2 Text Classification

Often, texts are processed as *bags of words* and methods such as k-nearest neighbour and support vector machines are applied. In *bag of words* methods, texts are processed in order to remove words which are considered irrelevant, such as *the*, *a*, etc. The remainder words are then reduced to their *invariant* forms (stemmed), so that different forms of the same word are counted as the same—e.g., *reserved* and *reserve* may be mapped to *reserv*. The remainder words are then counted and the text is finally represented by a vector of word frequencies. Text classification is then processed by means of applying different operations to the frequency vectors. But those methods invariably require pre-processing the texts. It is necessary to process the texts several times in order to extract the information necessary to create the sorted vectors from the *bags of words*. Pre-processing poses additional challenges for real time operation, specially if the method is applied to all words. Focusing operation on just keywords greatly reduces dimensionality of the vectors and processing time, at the cost of loosing the information conveyed by the overlooked words.

On the other hand, it is known that semantically meaningful texts (i.e., texts that carry some information, not "artificial" texts generated by randomly juxtaposing symbols) exhibit long-range correlations between lower level representations and higher semantic meaning. The correlations have been observed for the first time many years ago, and the topic has been subject to some research. Recently Altmann et al. [1] published an interesting analysis of those long-range correlations. In a text, a topic is linked to several words, which are then linked to letters, which are then linked to lower symbols, as represented in Fig. 1. Altmann claims that correlations between high-level semantic structures and lower level structures unfold in the form of a bursty signal, thus explaining the ubiquitous appearance of long-range correlations in texts.

3 Sparse Distributed Memory

The Sparse Distributed Memory is an associative memory model suitable to work with high-dimensional binary vectors. Thus, all information that can accurately be described by arbitrary sequences of bits may be stored into such a memory.

Kanerva shows that the SDM *naturally* exhibits the properties of large boolean spaces. Those properties can be derived mathematically and are, to a great extent, similar to that of the human cerebellum. The SDM implements behaviours such as

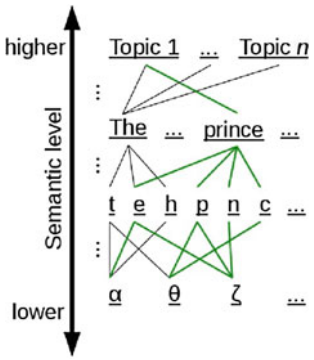


Fig. 1 Hierarchy of levels and links between different representation levels of a text. Correlations are preserved between different level structures

high tolerance to noise, operation with incomplete data, parallel processing and *knowing that one knows*.

3.1 Implementation of a SDM

The underlying idea behind the SDM is the mapping of a huge binary memory onto a smaller set of physical locations, so-called *Hard Locations* (HL). As a general guideline, those hard locations should be uniformly distributed in the virtual space, to *mimic* the existence of the larger virtual space as accurately as possible. Every datum is stored by distribution to a set of hard locations, and retrieved by *sampling* those locations.

Figure 2 shows a model of a SDM. "Address" is the reference address where the datum is to be stored or read from. It will activate all the hard locations in a given access radius, which is predefined. Kanerva proposes that the Hamming distance, that is the number of bits in which two binary vectors are different, be used as the measure of distance between the addresses. All the locations that differ less than a predefined number of bits from the input address are selected (activated) for the read or write operation.

3.1.1 Writing and Reading

Data are stored in arrays of counters, one counter for every bit of every location. Writing is done by incrementing or decrementing the bit counters at the selected addresses. To store 0 at a given position, the corresponding counter is decremented. To store 1, it is incremented. The counters may, therefore, store either a positive or a negative value, which should, in theory, most of the times fall into the interval $[-40, 40]$.

Reading is done by sampling the values of all the counters of all the active locations. A natural candidate to extract the correct value is the average. Summing

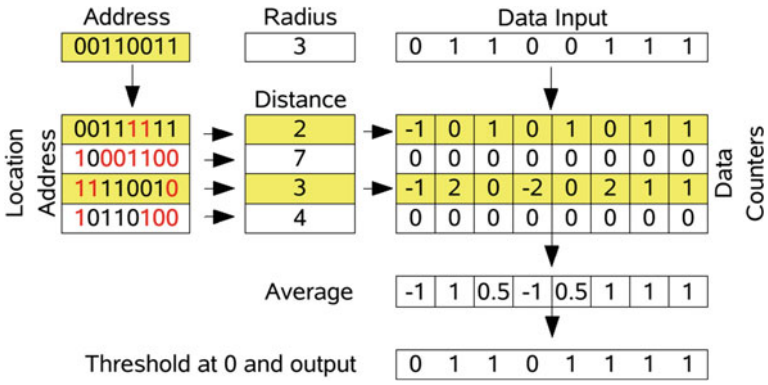


Fig. 2 Diagram of a SDM, according to the original model, showing an array of bit counters to store data and an array of addresses. Memory locations within a certain radius of the reference address are selected to read or write

all the counters columnwise and dividing by the number of active locations gives the average value of the counters. The average value can then be compared to a predefined threshold value. A threshold of 0 is appropriate for normal data—if the average value is below the threshold the bit to be read is zero, otherwise it is one. Figure 2 illustrates the method.

But the average value is only one possible sampling method. Another possible method is to pool the information by taking a *vote*—the value that is the most popular among the contributing counters is preferred. Yet another alternative is to weigh the contribution of each counter based on the distance between each hard location and the reference address.

3.1.2 Starting the Memory

Initially, all the bit counters must be set to zero, for the memory stores no data. The bits of the hard locations' addresses should be set randomly, so that those addresses would be uniformly distributed in the addressing space. However, many authors prefer to start with an empty memory, with neither data counters nor addresses, and then add more addresses where and when they are needed [7], in order to avoid processing unneeded locations and reduce startup time.

3.1.3 Characteristics of the SDM

Due to the nature of the model, there is no guarantee that the data retrieved is exactly the same that was written. However, it is provable that under normal circumstances the hard locations will be correctly distributed over the binary space and, if the memory has not reached saturation, the correct data will be retrieved

with high probability most of the times. The conclusion arises from the properties of high-dimensional boolean spaces. The mathematical details, due to their length, are out of the scope of the present work. They are elegantly described by Kanerva in [2]. A summary can also be found in [8].

Other important characteristics of the SDM are:

1. It is immune to noise up to a high threshold. Using coding schemes such as n-of-m codes, the immunity is even increased [6, 9], at the cost of reducing the addressable space.
2. SDMs are robust to failure of individual locations, just like neural networks.
3. SDMs degrade gracefully, when some locations fail or the memory approaches its maximum capacity.
4. One-shot learning is possible. If the memory is not close to saturation, it will learn in a single pass.
5. SDMs can be "open" and subject to analysis of individual locations. That is important namely for debugging purposes, or to track the learning process.
6. It is possible to change memory's structure without retraining all the memory [7]. For example, it is possible to add locations where they are needed as well as remove unused locations. That is an important characteristic to build modular or adaptive systems.

The main drawbacks of using Sparse Distributed Memories are:

1. Once a datum is written, it cannot be erased, only *forgotten* as time goes by. Under certain circumstances that may be an undesirable feature. If unnecessary memories cannot be deleted, they may interfere with more recent and important data.
2. If the SDM is simulated in a common computer, storage capacity may be as low as 0.1 bits per bit of traditional computer memory, although many authors reported techniques to improve storage performance [10].
3. If implemented in software, a lot of computer processing is required to run the memory alone.

3.2 Present Implementation

In the present approach, a variation of the original SDM was used. Many experiments were performed using an implementation which used 8-bit data counters (7 bits + sign), but differences for the results obtained without using data counters were negligible, when they existed at all. In the simplified model, without data counters, each memory location contains only one copy of the last datum. Any new datum will overwrite the previously stored information. This simplification greatly reduces memory and processing requirements, and previous experimental evidence reports no significant performance differences to the original model, in different fields [9, 11].

Additionally, in the model used, the hard locations were not placed randomly in the binary space as Kanerva proposes. The memory locations are managed using the Randomised Reallocation (RR) algorithm proposed by Ratitch et al. [7]. Using the RR, the system starts with an empty memory and allocates new locations when there is a new datum which cannot be stored into enough existing locations. The new locations are placed *randomly* in the neighbourhood of the new datum address.

4 Dataset Experiments

In a first stage of the present work, experiments were performed using a popular dataset, before a real application as described in Sect. 5.

4.1 Datasets Used

The datasets used in the experiments were pre-processed subsets of Reuters 21578 dataset, available from Cardoso-Cachopo's website[1] [12]. Those datasets were chosen because of their popularity and the fact that they were available in pre-processed form from Cardoso-Cachopo's website. The subset named R52 is a selection of documents which are classified into just one of the topics (single-labelled). R52 contains 9,100 documents distributed over 52 different topics. The subset named R8 is a selection of documents which are also single-labelled, but contains documents of just 8 of the 10 most frequent topics. Tables 1 and 2 show the number of documents per topic. As the tables show, data is very skewed, with the most popular class accounting for about half of the documents.

Cardoso-Cachopo makes available the datasets with different pre-processing applied. The subsets that are relevant for the present experiments are:

- **All-terms**—Obtained from the original datasets by applying the following transformations: Substitute TAB, NEWLINE and RETURN characters by SPACE; Keep only letters (that is, turn punctuation, numbers, etc. into SPACES); Turn all letters to lower-case; Substitute multiple SPACES by a single SPACE; The title/subject of each document is simply added in the beginning of the document's text.
- **Stemmed texts**—Obtained from the previous file, by removing all words that are less than 3 characters long; removing the 524 SMART system's stop-words[2] and applying Porter's Stemmer to the remaining words.[3]

[1] Datasets available at http://web.ist.utl.pt/acardoso/ (last checked 2013-02-10).

[2] ftp://ftp.cs.cornell.edu/pub/smart/english.stop (last checked 2013-02-13).

[3] http://tartarus.org/~martin/PorterStemmer/ (last checked 2013-02-13).

Table 1 Documents contained, in the training and test subsets, for each class of the dataset R8

Class	Train set	Test set	Total
Acq	1,596	696	2,292
Crude	253	121	374
Earn	2,840	1,083	3,923
Grain	41	10	51
Interest	190	81	271
Money-fx	206	87	293
Ship	108	36	144
Trade	251	75	326
Total	5,485	2,189	7,674

4.2 Experiments

To assess the performance of the SDM as a classifier, two types of vectors were used to represent the texts: Term Frequency-Inverse Document Frequency (TF-IDF) vectors and direct storage of the texts using ASCII characters.

The experiments were performed in two steps: learning and testing. In the learning stage, representations of all the documents in the training sets were stored into the SDM. In the testing stage, the SDM was then queried with documents of the test set. The category of the document retrieved was then used as the SDM's best match for the category of the test document.

The first experiment was done using TF-IDF vectors as representations of the texts. TF-IDF is a popular statistic used to represent documents in text categorisation. In general, the use of TF-IDF vectors produces good results for text categorisation with popular methods such as Support Vector Machines, K-Nearest Neighbours and similar. In the present work, TF-IDF vectors were calculated for each single document in the train set. Those vectors were then normalised and mapped in the interval [0, 127], so that all numbers could be represented as unsigned 7-bit integers. That conversion meant to loose a lot of precision, but considering the length of the vectors it was accepted as a good compromise between SDM simulation time and precision.

In a second experiment, the vectors stored into the SDM were chunks of up to 8 KB of the texts, coded in plain ASCII. When the length of the text was superior to 8 KB, the text was truncated. When the length was inferior to 8 KB, the remainder bytes were set at random.

In both experiments, the encoded text was used as address for the SDM. The data stored was the text with the category juxtaposed—i.e., the composition of the data vector was $\langle address, category \rangle$, where the category was always stored in plain ASCII text.

During the learning stage, the vectors were stored into the SDM with a radius of zero. Copies of all vectors were stored, since all vectors were different from each other in both datasets.

In the testing stage, the memory was queried with texts of the test set, encoded using the same method used during the learning stage. When plain texts were used,

Table 2 Documents contained, in the training and test subsets, for each class of the dataset R52

Class	Train set	Test set	Total
Acq	1,596	696	2,292
Alum	31	19	50
Bop	22	9	31
Carcass	6	5	11
Cocoa	46	15	61
Coffee	90	22	112
Copper	31	13	44
Cotton	15	9	24
Cpi	54	17	71
Cpu	3	1	4
Crude	253	121	374
Dlr	3	3	6
Earn	2,840	1,083	3,923
Fuel	4	7	11
Gas	10	8	18
Gnp	58	15	73
Gold	70	20	90
Grain	41	10	51
Heat	6	4	10
Housing	15	2	17
Income	7	4	11
Instal-debt	5	1	6
Interest	190	81	271
Ipi	33	11	44
Iron-steel	26	12	38
Jet	2	1	3
Jobs	37	12	49
Lead	4	4	8
Lei	11	3	14
Livestock	13	5	18
Lumber	7	4	11
Meal-feed	6	1	7
Money-fx	206	87	293
Money-supply	123	28	151
Nat-gas	24	12	36
Nickel	3	1	4
Orange	13	9	22
Pet-chem	13	6	19
Platinum	1	2	3
Potato	2	3	5
Reserves	37	12	49
Retail	19	1	20
Rubber	31	9	40
Ship	108	36	144
Strategic-metal	9	6	15
Sugar	97	25	122

(continued)

Table 2 (continued)

Class	Train set	Test set	Total
Tea	2	3	5
Tin	17	10	27
Trade	251	75	326
Veg-oil	19	11	30
Wpi	14	9	23
Zinc	8	5	13
Total	6,532	2,568	9,100

if the length of the test document was inferior to 8 KB, only part of the vector was used to compute the similarity measure in the SDM. For example, if the length of the text was just 1 KB, similarity was computed using only the first 1,024 coordinates (bytes) of the addresses of the hard locations. That should not affect the expected characteristics of the SDM, unless the text was too small.

4.3 Results

Table 3 summarises the results. The first column identifies the type of classifier. The second column shows the type of input used. The third column shows the results obtained for dataset R8 and the last one shows the results obtained for dataset R52. Experiments with other datasets and memory types were performed, but the performance of the SDM was very similar. Thus, for clarity, it was chosen to summarise the results into just Table 3. Table 3 summarises the percent of texts correctly classified using each method.

The "Dumb classifier" is shown just as a reference. It is a hypothetical classifier that always returns the most popular class. Thus, its performance is equal to the percentage of documents of the most popular class. In the datasets used, data is very skewed. Thus, the dumb classifier actually seems to have a good performance. In more homogeneous datasets the result is different. For example, in the "20 newsgroups dataset," which consists of approximately 1,000 messages of 20 different newsgroups (total close to 20,000 messages), the dumb classifier has a performance of 5.3 %, while the SDM achieves 20.9 % using "all-terms" and 22.2 % using the stemmed texts.

As for the SDM, two different methods were used for the prediction phase. First, the memory was configured in a way that for each prediction it always returned the *nearest* neighbour found in the neighbourhood of the input address. Second, the memory was configured in a way that it enlarged the access radius to encircle at least 10 data points, and then it returned the *most popular* class among the classes of data points found. Different numbers of data points were tested, between 2 and 10, and the differences were only negligible. In general, 10 seemed a good compromise. Using smaller numbers, the results tend to the results obtained

Table 3 Performance of the classification methods compared, for the dataset experiments

Method	Input	R8	R52
Dumb classifier	–	49.5	42.2
	TF-IDF vectors	55.9	51.0
SDM with shortest radius	Stemmed texts	60.7	50.4
	All terms	60.2	50.0
	TF-IDF vectors	52.3	51.0
SDM choosing the *most popular*	Stemmed texts	64.6	55.5
	All terms	64.6	55.2

in the first experiment. Enlarging the circle the results will tend to the performance of the dumb classifier.

As Table 3 shows, the results obtained are humble if compared to other modern classifiers. For example, [13] reports an accuracy of up to 96.98 % in dataset R8 and up to 93.8 % in R52 using SVMs. However, the best results obtained with the SDM can be achieved *naturally*, with almost no text processing. The stemmed datasets apparently have a marginal improvement in the results. But even if the results are the same, stemming and removing stop-words contributes to reducing dimensionality of the input vectors. High-dimensionality is not a problem for the SDM, but if the SDM is implemented in common serial processors more dimensions mean more processing time. Thus, removing data that carries no useful semantic information speeds up the process without compromising accuracy.

In summary, the results show that the SDM is able to grasp high-level semantic information from raw data input. There may be ways to improve the accuracy of the process, for example trying different methods of encoding the data. In some applications where real time processing is necessary, the SDM can still be a good option, even if the results are only humble compared to other modern methods. The results also open good perspectives for use of the SDM in other applications where text-matching is necessary, besides single-label text classification.

5 Ad Server Experiments

The second batch of experiments was carried in a real application (production) environment, in a web server hosting several different websites. The text classification application was used to choose the best matching ads to exhibit in publisher websites' pages—i.e., contextual advertising. The advertisers' database contained a total of 1,077 ads. Those ads advertised products from 4 different online shops. The number of ads available might have been less than 1,077 at some points, for the shop managers could hide ads while the experiments were running at any time. That could happen, for example, in case of stock shortage or other commercial reason. The ads were served to different publisher web pages. There were 27 different publisher websites, and the total number of pages of those

Table 4 Performance of the ad server system using different methods to choose the ads to serve

Method	Hits	Clicks	CTR (%)
Vector method, augmented frequencies	143,942	5,303	3.7
SDM, without data counters	317,484	4,399	1.4
Random	205,617	1,272	0.6

websites summed up to 77,964 different pages. It should also be mentioned that the majority of the products in the web shops, as well as the contents of the web sites, had something in common. They would have some relation, however faint, to the vegetarian lifestyle, animal rights, ecology and related topics.

The experiments were run in a way that different methods were used to choose the ads to exhibit on each webpage when it was served, during periods of 7 days. The system logged the number of ads served and whether each ad was clicked or not, so that it was possible to calculate the Click Through Rate (CTR) for each method. Table 4 shows the results.

First, for 7 days, the ads were selected randomly. In that period more than 200 thousand ads were served and the CTR was just about 0.6%.

For another period, the ads were chosen based on text similarity between the page being visited and the product being advertised. Similarity was computed using the vector space method. The vector space method was implemented using tf-idf "augmented frequencies." To get some "randomness" in case of repeated accesses to the same page by the same user, the ad selected was randomly picked among the 5 best matching ads. This method showed the best results, since it achieved a CTR of 3.7 %.

Since the texts of the ads themselves are very succinct, the application developed scanned the page being advertised and extracted the texts for comparison from the page itself. The texts used for computing the vector space model, for the vector space experiments, or to store in the SDM in the SDM experiments, included the page title, meta tags keywords and description, and also some headings when available.

Finally, for another 7-days period, the ads were chosen based on predictions from a Sparse Distributed Memory. The memory was loaded with chunks of up to 1 KB of the ads' texts. Another modification was made to the SDM: when reading, the access radius was enlarged in order to encircle at least 5 data points—i.e., 5 best matching ads. Then the return was not just one ad, but a list of the 5 best matching ads. The ad shown was then one of those 5 best matches. Using this method, the CTR achieved was 1.4 %. It is still far from the results obtained using the vector space model, but more than twice the CTR of the random server. And that was achieved using just a fraction of the requirements from the server. For example, the data structure in memory to implement the SDM was 5.5 MB, while the data structure necessary to implement the vector space model was 10.9 MB.

6 Conclusions

Text matching is a topic of increasing relevance, as it is important for information retrieval, text categorisation and other applications. It is known that language exhibits long-range correlations, from the lowest-level representations to the highest semantic meanings. The SDM is an associative memory model that works based on the properties of high-dimensional boolean spaces, exploring in part similarities between long binary vectors. The experimental results described in the present paper show that the original model of the SDM alone, without any text processing, is able to work as a surprisingly good text classifier, even taking plain ASCII text as input. In an ad server for the web, it has achieved a CTR of more than twice the CTR of a purely random method, though inferior to the CTR that can be achieved using a vector space model with approximately the same data. In future work different methods of encoding the information or tuning the SDM may be tried, seeking to improve the performance of the SDM as a text classifier.

References

1. E.G. Altmanna, G. Cristadorob, M.D. Esposti, On the origin of long-range correlations in texts. Proc. Nat. Acad. Sci. U.S.A. **109**(29) (2012)
2. P. Kanerva, *Sparse Distributed Memory* (MIT Press, Cambridge, 1988)
3. D. Rogers, Predicting weather using a genetic memory: a combination of Kanerva's sparse distributed memory with Holland's genetic algorithms, in *NIPS* (1989)
4. M. Mendes, M.M. Crisóstomo, A. Paulo Coimbra, Robot navigation using a sparse distributed memory, in *Proceedings of IEEE International Conference on Robotics and Automation*, Pasadena, California, USA, May 2008
5. M. Mendes, A. Paulo Coimbra, M.M. Crisóstomo, Exploring long-range correlations for text classification using a sparse distributed memory, in *Lecture Notes in Engineering and Computer Science, Proceedings of The World Congress on Engineering 2013, WCE 2013*, London, U.K., 3–5 July 2013. IAENG
6. M. Mendes, M.M. Crisóstomo, A. Paulo Coimbra, Assessing a sparse distributed memory using different encoding methods, in *Proceedings of the World Congress on Engineering* (*WCE*), London, UK, July 2009
7. B. Ratitch, D. Precup, Sparse distributed memories for on-line value-based reinforcement learning, in *ECML* (2004)
8. M. Mendes, A. Paulo Coimbra, M. Crisóstomo, AI and memory: studies towards equipping a robot with a sparse distributed memory, in *Proceedings of the IEEE International Conference on Robotics and Biomimetics*, Sanya, China, December 2007, pp. 1743–1750
9. S.B. Furber, J. Bainbridge, J.M. Cumpstey, S. Temple, Sparse distributed memory using *n*-of-*m* codes. Neural Netw. **17**(10), 1437–1451 (2004)
10. J.D. Keeler, Comparison between Kanerva's SDM and Hopfield-type neural networks. Cogn. Sci. **12**(3), 299–329 (1988)
11. A.P. Coimbra, M. Mendes, M.M. Crisóstomo, *Vision-Based Robot Navigation: Quest for Intelligent Approaches Using a Sparse Distributed Memory* (Universal Publishers, Boca Raton, 2012)

12. A. Cardoso-Cachopo, *Improving Methods for Single-label Text Categorization*. PhD thesis, Universidade Técnica de Lisboa, Oct 2007
13. A. Cardoso-Cachopo, A. Oliveira, Combining LSI with other classifiers to improve accuracy of single-label text categorization, in *First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning*. EWLSATEL (2007)

CO_2 Purify Effect on Improvement of Indoor Air Quality (IAQ) Through Indoor Vertical Greening

Ying-Ming Su

Abstract Modern people usually situate inside buildings, therefore indoor air quality (IAQ) is important to living environment. In this research, we measured the transpiration rate of 50 experimental plants first, and then we chose Bird's-Nest Fern (*Asplenium nidus*) as experimental subject to measure CO_2 adsorption, which was one of the common indoor plants with high transpiration rate in Taiwan. We constructed an indoor 189 pots green-wall of Bird's-Nest Fern. CO_2 adsorption was measured continuously before, during and after the experiment. The experiment consisted of three tests, A: Without plants (day time), B: With plants (day time), C: With plants (with light, night). Three measurements were compared between different tests from A to C. The result of the experiment indicated that the density of CO_2 was reduced from 2,000 to 600 ppm in 5.37 h of test C. In average, the efficiency of CO_2 adsorption was 1.38 ppm/h each plant. The indoor temperature decreased to 2.5 °C and relative humidity increased about 2–4 %. We concluded that Bird's-Nest Fern had superior properties to CO_2 adsorption, temperature and humidity conditioning to indoor air quality improvement.

Keywords Bird's-Nest Fern (*Asplenium nidus* Linn) · CO_2 · Herbaceous foliage plant · Indoor air quality (IAQ) · Indoor vertical greening · Leaf porometer · Transpiration rate

1 Introduction

Plants are the lungs of the earth: they produce the oxygen that makes life possible, add precious moisture, and purifying air such as toxins, dust fall, organ volatilize materials, and VOCs. Indoor plants can also have the function above, and restrain

Y.-M. Su (✉)
Department of Architecture, National Taipei University of Technology, 1, Sec. 3, Zhongxiao E. Rd, Taipei 10608, Taiwan R.O.C
e-mail: ymsu@ntut.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_41,

micro-organisms, maintain humidity and make our daily activity space more comfortable and healthy. Indoor plants can perform these essential functions in home or office with the same efficiency as a rainforest in our biosphere. In research designed to create a breathable environment for a NASA lunar habitat, noted scientist Dr. B.C. Wolverton discovered that houseplants are the best filters of common pollutants such as ammonia, formaldehyde, and benzene.

Hundreds of these poisonous chemicals can be released by furniture, carpets, and building material, and then trapped by closed ventilation systems, leading to the host of respiratory and allergic reactions now called Sick Building Syndrome [1–3]. Owing to modern people spend about 80 to 90 % of their time indoors [4, 5], the importance of indoor environment quality have been taken seriously. Although enhancing ventilation and using equipment could reduce SBS, the most natural ways to reducing SBS is to decorate the building with indoor plants. Lim [6] survey with the condition whether to have houseplants indoors and ventilated or not to see the SBS of the dweller, and the result shows that when the room is ventilated and decorated with houseplants, it can reduce the SBS by 35 %.

Plants would cause temperature difference between leaf surface and air flow when evaporate strongly. When the plants evaporate the water in the air, it will lead to the purification of the water into atmosphere and cool down the temperature, and reducing the usage of air-conditioner so that can save energy. Some research also point out that the larger the leaf is, the higher the transpiration rate is, and have better efficiency in indoor air purifying [2, 7, 8]. Indoor plants also become an important design method for energy saving, indoor air quality and work performance. Indoor plants can not only decorate the indoor spaces, some scientific researches also show that raising plants will contribute to the release of pressure and tiredness. The green environment will obviously increase the amplitude of the alpha wave inside human brain, and lower blood pressure, myoelectricity and skin conductivity. It will also release pressure and anxiousness and improve working attention [9–11]. Besides, indoor plants can reduce the employees' absence from work; increase their degree of satisfaction for work and feelings for life [12, 13]. In a word, decorating plants inside a room will indeed benefit to our physical and psychological health.

1.1 Harm of CO_2 on Human Body

Indoor CO_2 mainly comes from human breathing, smoking and open-fire heating etc. In concentrations between 0.2 and 0.3 % in a closed-air-circuit area shared by crowds, CO_2 will cause nausea, dizziness, and headache. With the rise of the concentration, temperature and humidity goes up, as well as dusts, bacteria and body odor, while oxygen and the numbers of ions in the air decreasing, causing uncomfortable [14]. Environment with low concentration of CO_2 is not classified as toxic; however, a high concentration will do great harm to human body, leading

to serious health condition such as suffocation. Nowadays, most offices have common problems that they are overused by too many people and are often not equipped with qualified ventilation systems [15, 16].

1.2 Transpiration Rate

Transpiration is part of water cycle. It is the loss of water vapor from parts of plants and through the process water is purified and plants are cooled. To be specific, the diffusion of water vapor in the atmosphere lead to the cross-ventilation, temperature decreased which help purify the air, moisturize the air, remove biological exhaled gas and chemical toxicant, and inhibit the micro-organisms in the air so that indoor air quality can be improved [17]. Personal breathing zones placed high evapotranspiration rates of plant; you can increase the humidity and removal of exhaled gases and chemical substances. Also can inhibit the microorganism in the air [18, 19].

2 Experimental Environment

This study was focused on IAQ effect of indoor plants and composed of three stages: (1) Optimum Setting (2) Numerical Simulation (3) Evaluation and Control. The experiment procedure proceeds were shown in Fig. 1.

We used Leaf Porometer as laboratory equipment to record the transpiration rate data, and chose the 50 experimental object plants in their proper sizes recommended by The Environmental Protection Administration Executive Yuan, 50 plants into 4 categories according to their characteristics: 19 kinds of herbaceous foliage plant, 12 of herbaceous flowering plant, 4 of trailing plant and 15 of woody plant.

This study had experimented 50 plants once every 12 h, 10 leaves at a time and lasted for 144 h continually. Finally we survey on the preferences for top 5 of herbaceous foliage plants, and select Bird's-Nest Fern (*Asplenium nidus* Linn) as the main indoor plant of this research, which is local plant of high transpiration rate, low requirement of optical activity, drought-and-shade-tolerance, and often of an appropriate size for indoor plating.

We set up the laboratory which is $4.8 \times 3.4 \times 3.1$ m^3 shown in Fig. 2. The experimental conditions used were summarized in Table 1. The average temperature of daytime was 28.15 °C and that of nighttime was 27.56 °C. The average moisture of daytime was 76.93 % and that of nighttime was 67.71 %. Results of the average transpiration rate and the standard deviation of the four categories measurements were analyzed through SPSS analysis. We had adopted the following instruments for CO_2 absorption experiment: (1) CO_2 monitor: KD Air-Box, (2) temperature and humidity monitor: iLog /Escort. All data values in the study were monitored in the seven-day-experiment.

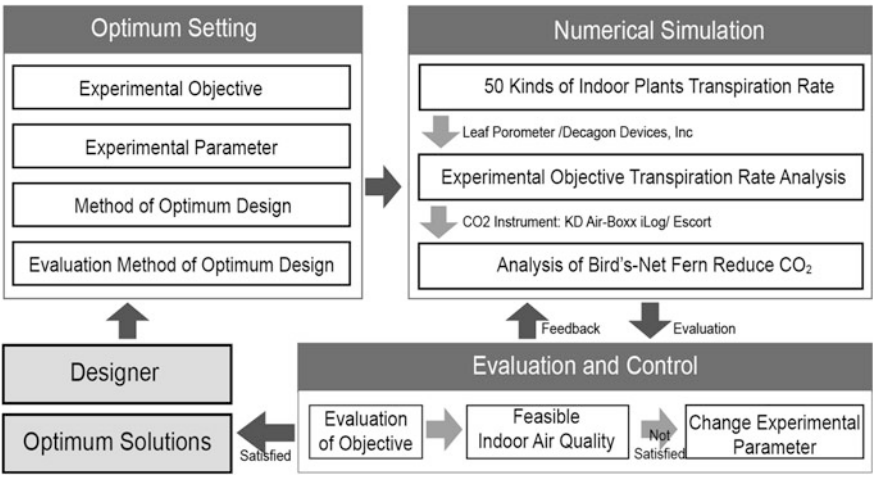


Fig. 1 Study procedure

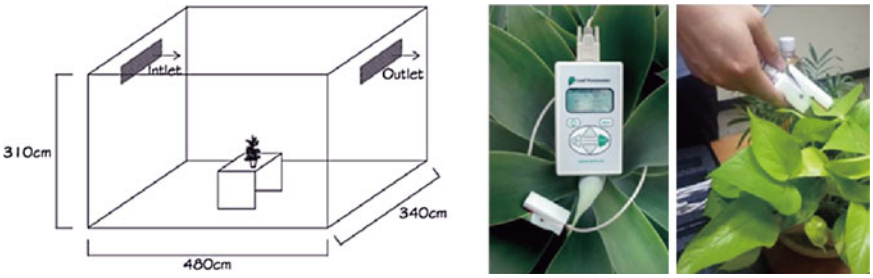


Fig. 2 The image of experimental lab for transpiration rate of Bird's-Nest Fern

Table 1 Experimental conditions (leaf porometer)

Instrument	Operating environments: 5–40 °C; 0–90 % relative humidity (non-condensing)
Air inlet	Accuracy: 10 %
	Sample chamber aperture: 6.35 mm (0.25 in.)
	Measurement range: 0–1,000 mmol/m^2 s
	Velocity Uin = 1.2 m/s, ventilation rate = 6 l/h
	kin = 3/2 (Uin × 0.05)2, εin = 0.09 × kin3/2/0.4
Temperature	28.15 °C (daytime average) 27.56 °C (nighttime average)
Moisture	76.93 % (daytime average) 67.71 % (nighttime average)

The experiment was conducted in a L360 × W360 × H300 cm laboratory (Fig. 3). A 1-cm thick Styrofoam on the windows and the surrounding walls and ceiling were covered with tinfoil. A galvanized steel sheet covered the ground and special cautions were taken to stuff the gaps around the door. Monitoring

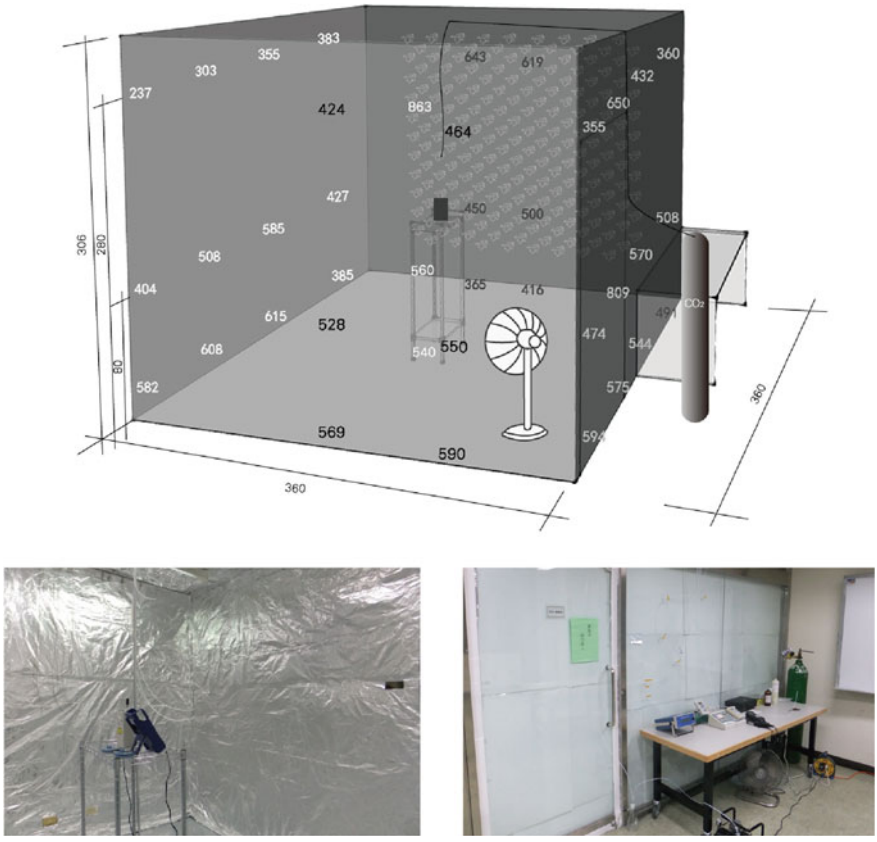


Fig. 3 Lab and monitor outside the lab for CO_2 absorption of Bird's-Nest Fern

equipment was placed in the center of the room and a monitoring spot was left outside the laboratory for another monitor. The laboratory was equipped with fluorescent of T5 lamps.

In accordance with the National Standards of Republic of China (CNS standard), the illumination range should be between 750 and 300 lux. After the installation of lighting, the measurements of laboratory's illumination were taken at the height of 80 and 280 cm, every 75 cm in width. The laboratory's average illumination of 13 sets altogether 39 data was 512.5 lux.

One of the laboratory's walls (L300 × W260 cm) (Fig. 4) was covered with a composite vertical greening by Bird's-Nest Fern in 3-in. pots, altogether 189 pots (Fig. 5). Each pot approximately included 8 large leaves (18 × 4 cm), six medium-sized leaves (15 × 3 cm) and eight small leaves (8 × 2 cm) altogether occupying an area of 15 m^2. Plastic flower pots and space between soil and plant as well as the composite wall were coated by aluminum foil to reduce other media's effect on the experimental data. A fan was placed to evenly spread the formaldehyde gas in the confined space.

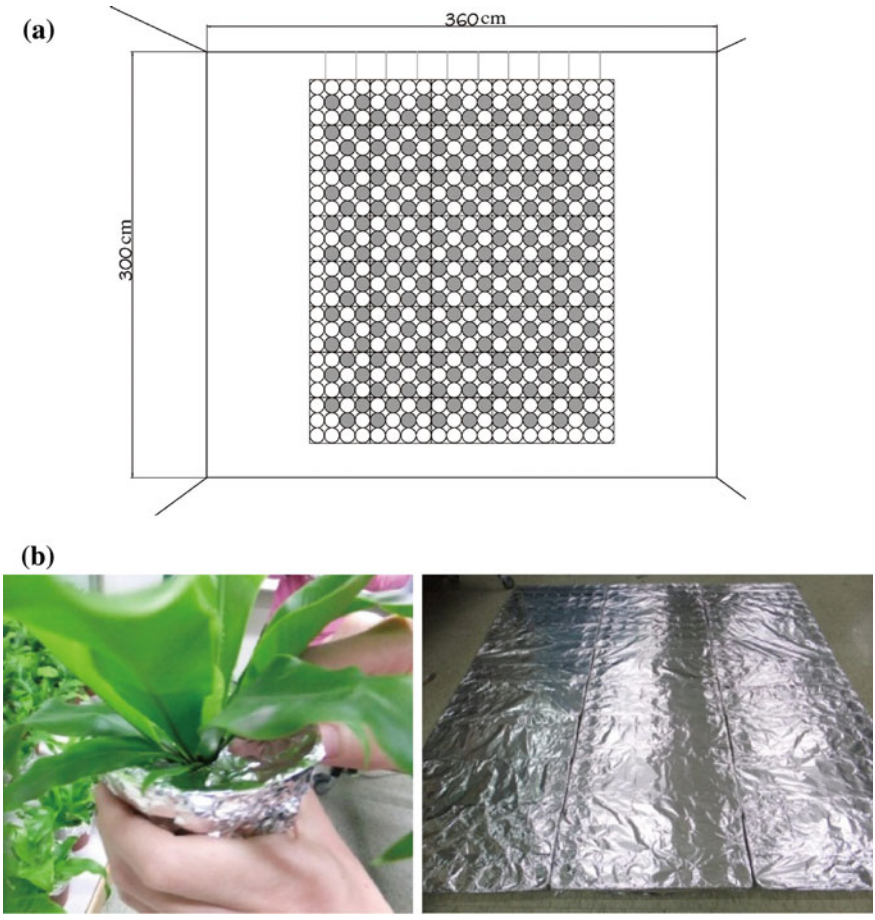


Fig. 4 Photos of experimental environment. **a** Photo of vertical greening. **b** Plastic flower pots and space between soil and plant coated by aluminum foil

3 Results and Discussion

Plants could turn CO_2 and H_2O into carbohydrates C (H_2O) and release oxygen through photosynthesis [20]. The rate of plant's CO_2 absorption could reach 10–30 mg per decimeter square per hour, and thus could be used an indicator of surface photosynthesis [21]. The indoor lamp lighting took the place of sunlight in the process. A fan was used instead of natural ventilation to evenly spread the formaldehyde in the space area. To ensure the accuracy of statistics, the laboratory was confined when statistics is monitored.

First, chose two fully expanded leaves and fix them twice a day respectively at 11:30 and 17:30. Then we recorded eight sets of data every day so that altogether 48 sets of data were recorded after six days (Fig. 6). A higher value represented a

Fig. 5 Vertically greening composite wall in the lab

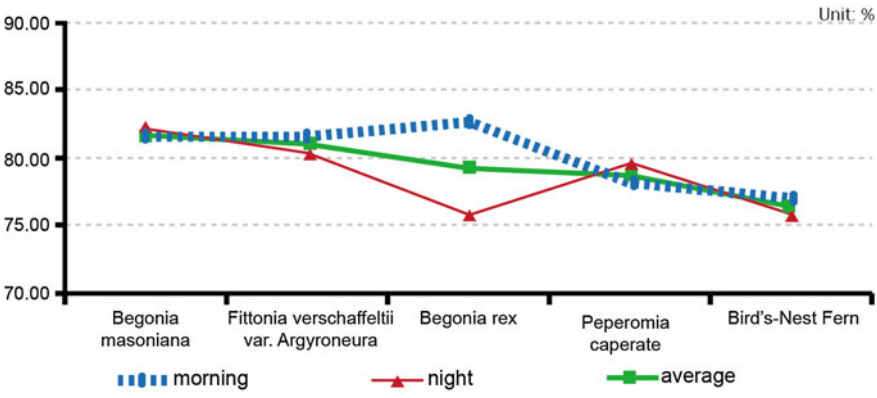


Fig. 6 The standard deviation and transpiration rate of top 5

higher rate of transpiration and indicated that the plant could effectively adjust the humidity as well as release more anion to clean air. The plants ranking on the top of the list for experiment were *Begonia masoniana*, *Fittonia verschaffeltii var. Argyroneura*, *Begonia res*, *Peperomia caperata* and Bird's-Nest Fern (*Asplenium nidus*).

Therefore, this study had chosen Bird's-Nest Fern (*Asplenium nidus*), a local plant of high transpiration rate, low requirement of optical activity, drought-and-shade-tolerant, and often of an appropriate size for indoor planting, as the detection objects.

The recorded data showed that Bird's-Nest Fern's transpiration rate was among the highest, of 78 mmol/(m^2 · s). According to the results, the interior designers could select suitable indoor plants not only to decorate house but also make it more efficiency and healthy. Choosing the indoor plants with high transpiration rate to decorate houses could have more possibility to improve the heat radiation, thus it would cool down the air.

3.1 Transpiration Rate of Bird's-Nest Fern

Three measurements were compared between different tests from A: Without plants (day time), B: With plants (day time) to C: With plants (with light, night) in average. The indoor temperature decreases to 2.5 °C and relative humidity increases about 2–4 %. The 48 average transpiration rate of 78 mmol/(m^2 · s) (Fig. 7, Table 2).

3.2 Reaction Rate with Release of CO_2

Results showed when CO_2, concentration of 2,000 ppm, was released, A: With plants (day time), it took 8 h without any plants to decrease the concentration to 600 ppm. B: With plants (day time), it took 6 h and 18 min at day. C: With plants (with light, night), it took 5 h and 37 min at night to decrease to the same concentration (Fig. 8). Time cost was least in the third experiment with plant at night, which was 2 h and 25 min less than the experiment without plants. This indicated that the longer acclimation time was, the better effect plants achieved. Also, though at night, the photosynthesis of plants continued as long as there was light.

3.3 Change of Temperature and Humidity of Experiment

In the experiment, the change of temperature and humidity was recorded both under the condition with and without plants. Result showed that the temperature of the laboratory with plants was about 2 °C lower than that without plants which was 22.5–23.5 °C. With plants, the temperature of the laboratory was 21.8–22.1 °C at daytime and 21.9–22.2 °C at night (Fig. 9).

The laboratory with plants in it could keep to a certain temperature. The humidity in the laboratory without plants was relatively low while it rises 10 %RH

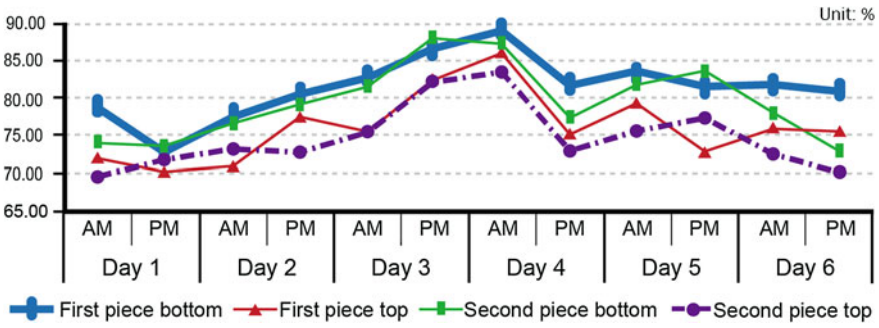


Fig. 7 The data of Bird's-Nest Fern transpiration rate and indicating velocity change of transpiration rate

Table 2 Transpiration rate experimental data of Bird's-Nest Fern

Day	Time	First leaf bottom	First leaf top	2nd leaf bottom	2nd leaf top
Day 1	Morning	78.8	72	74.1	69.7
	Afternoon	73	70.3	73.6	71.8
Day 2	Morning	77.7	71	76.7	73.2
	Afternoon	80.5	77.6	79.3	72.7
Day 3	Morning	82.7	75.6	81.6	75.5
	Afternoon	86.8	82.3	88.2	82.5
Day 4	Morning	88.9	86.2	87.4	83.7
	Afternoon	81.8	75	77.5	72.9
Day 5	Morning	83.5	79.3	81.8	75.5
	Afternoon	81.6	73	83.7	77.3
Day 6	Morning	82	75.9	78.2	72.7
	Afternoon	80.8	75.6	73	70.1

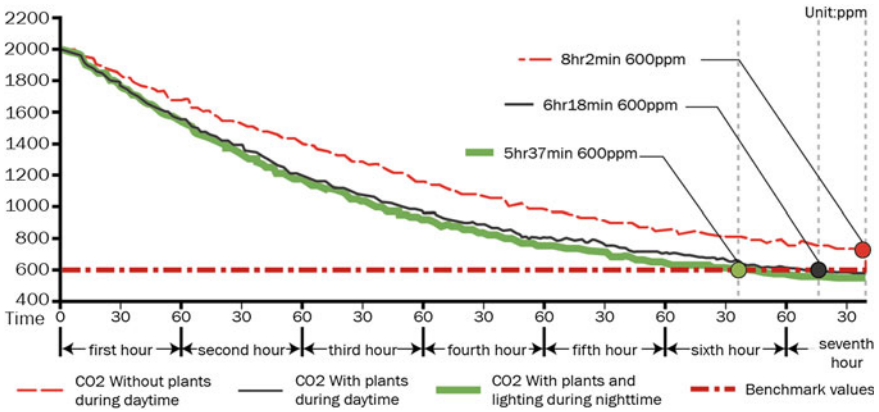


Fig. 8 CO_2 reaction time of experiments

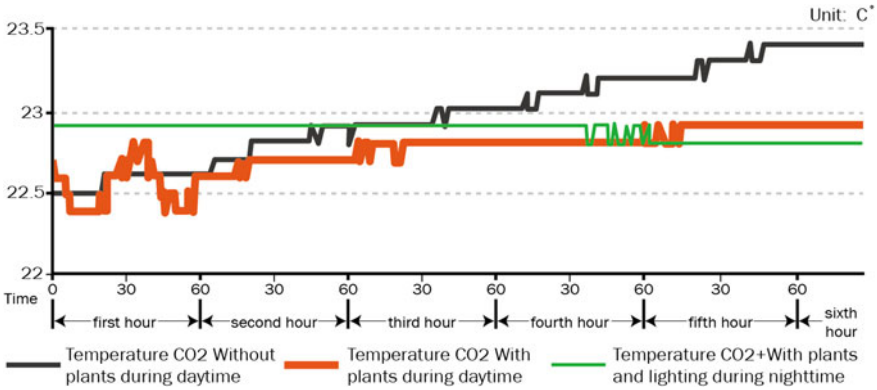


Fig. 9 Data overlapped results of temperature of experiment

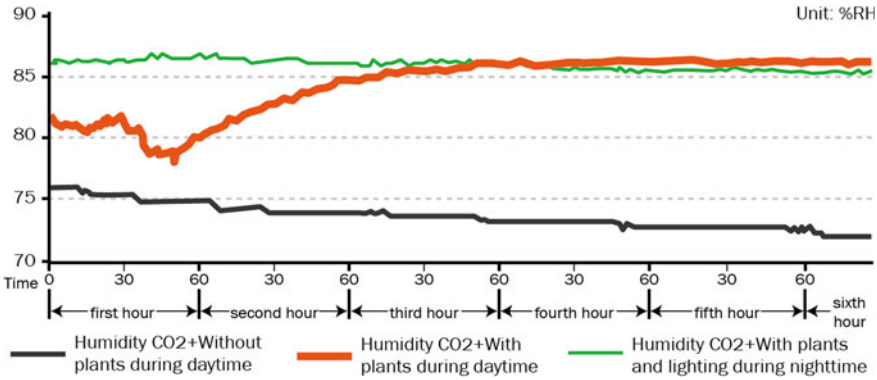


Fig. 10 Data overlapped results of humidity of experiments

when there were plants (Fig. 10). This indicated that plants could cool the temperature, kept the room to a certain temperature and meanwhile increased the humidity by 10 %RH.

4 Conclusion

The quality of the indoor environment significantly affected our health. The research indicated that 189 pots of Bird's-Nest Fern could reduce the concentration of CO_2 from 2,000 to 1,000 ppm in 2 h 6 min from 600 ppm to in 5 h 37 min of test C: With plants (with light, night). Bird's-Nest Fern could keep the room to a certain temperature and increase the humidity by 10 %RH. In average, the efficiency of CO_2 adsorption was 1.38 ppm/h each plant. The indoor temperature

decreased to 2.5 °C and relative humidity increased about 2–4 %. The 48 average transpiration rate of Bird's-Nest Fern was 78 mmol/($m^2 \cdot s$). This indicated that the longer acclimation time was, the better effect plants achieved. Also, though at night, the photosynthesis of plants continued as long as there was light.

Conversion of the total absorption by 189 pots of Bird's-Nest Fern: Absorption of CO_2:

2,000–600 ppm = 1,400 ppm

1,400/5.37 = 260.7 ppm/h

260.7/189 = 1.3794 ≒ 1.38 ppm/h.

Indoor plants could not only decorate the indoor spaces, release of pressure and tiredness, they could improve the quality of indoor air condition and reduce the pollutants in the air. This study showed that indoor plants can also contribute to lowering the temperature through better transpiration rate potentially.

Acknowledgments This paper represented part of the results obtained under the support of the National Science Council, Taiwan, ROC (No. NSC 101-2627-E-027-001-MY3).

References

1. B.C. Wolverton, (Penguin Books, 1996) p. 144
2. B.C. Wolverton, J.D. Wolverton, Interiorscape **11**, 17 (1993)
3. B.C. Wolverton, in *How to Grow Fresh Air* (Apple House, 2008)
4. M.C. Abbriti, G. Muzi, Indoor air quality and health effects in office buildings, in *International Conference on Healthy Buildings in a Mild Climate*, vol. **1**, Italy, pp. 185–195 (1995)
5. N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, J.V. Behar, S.C. Hern, W.H. Engelmann, J. Exposure Anal. Environ, Epidem. **11**, 231 (2001)
6. Y. Lim, H. Kim, J. Yang, K. Kim, J. Lee, D. Shin, J. Japan Soc. Hort. Sci. **78**(4), 456 (2009)
7. B.C. Wolverton, *How to Grow Fresh Air—50 Houseplants that Purify Your Home or Office* (Penguin Books, London, 1996), p. 144
8. Y. Su, C. Lin, Y. Chen, A preliminary study of plants transpiration effect of indoor environment, in *2011 Landscape Forum*, Taipei, Taiwan (2011)
9. M. Wang, Sci. Agri. **41**, 192 (1993)
10. C. Chang, P. Chen, HortScience **40**(5), 1354 (2005)
11. R.S. Ulrich, R.F. Simonds, B.D. Losito, E. Fiorito, M.A. Miles, M. Zelson, J. Environ. Psychol. **11**, 201 (1991)
12. T. Bringslimark, T. Hartig, G.G. Patil, HortScience **42**(3), 581 (2007)
13. A. Dravigne, T.M. Waliczek, R.D. Lineberger, J.M. Zajicek, HortScience **43**(1), 183 (2008)
14. Y. Zhang, *Sustainable Development and Green Indoor Environment* (Machinery Industry Press, Beijing, 2008)
15. L.T. Wong, K.W. Mui, P.S. Hui, A multivariate-logistic model for acceptance of indoor environmental quality (IEQ) in offices. Build. Environ. **43**, 1–6 (2008)
16. M. Frontczak, R.V. Andersen, P. Wargocki, Questionnaire survey on factors influencing comfort with indoor environmental quality in Danish housing. Build. Environ. **50**, 56–64 (2012)
17. M. Frontczak, P. Wargocki, Literature survey on how different factors influence human comfort in indoor environments. Build. Environ. **46**, 922–937 (2011)

18. J.-C. Min, *Research on the Ventilation of Technique of Building Regulation* (1997), pp. 61–62
19. E.-E. Lee, *Research on Strategy of Management of Quality of indoor Air in Office Space Settings*, Thesis of National Chen-Gon University (2004)
20. R.-T. Yeh, *The Physiological Basis of Plant Productivity* (National Translation Bureau, 1982)
21. S.-S. Liu, *Physiology of Plants* (Xu Foundation, 1990)

Eliciting Usability from Blind User Mental Model for Touch Screen Devices

Mohammed Fakrudeen, Maaruf Ali, Sufian Yousef and Abdelrahman H. Hussein

Abstract A novel mental model of born blind users for touch screen devices to aid software developers in designing and building better usability applications for blind users is presented. The work of Kurniawan was the basis to examine the mental model of blind users. A user case study was carried out to assess the subjective mental model of users in regards to application of usability. This revealed that two stages: listening and interaction are used to facilitate the development of the Kurniawan mental model. This paper also suggests twelve usability features that can be added to facilitate the design phase of the system.

Keywords Blind user · Heuristics · Mental model · Smartphone · Touch screen · Usability

1 Introduction

In spite of the authority of the human sensory system, the insight of the person is supplemented and enriched by past experiences developed by relationship with others. Without this ability, s/he will lack understanding of a particular situation.

M. Fakrudeen · M. Ali (✉) · A. H. Hussein
Department of Computer Science and Software Engineering, University of Ha'il, College of Computer Science and Engineering, PO Box 2440 Ha'il, Kingdom of Saudi Arabia
e-mail: maaruf@ieee.org

M. Fakrudeen
e-mail: m.fakrudeen@uoh.edu.sa

A. H. Hussein
e-mail: ar.hussein@uoh.edu.sa

M. Fakrudeen · S. Yousef
Science and Technology, Anglia Ruskin University, Cambridge, UK
e-mail: Sufian.Yousef@anglia.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_42,

In the field of cognitive science, these perceptual relations are referred to as "mental models" [1]. The field of Human-Computer Interaction (HCI) has adopted and adapted these concepts to further the study in the field of usability [2].

When software developers create new products or applications, they are articulating ideas that they deem will speed up specific end-user experiences. The idea is known as a "conceptual model" and when suitably utilized in the product construction, it will aid in complementing a mental model that typical end-users have and acknowledge [3]. Products that are not proficient to make this connection consistently and spontaneously are typically perceived by end-users as unwieldy or perplexing since the users have no means to associate or envisage the experience.

Usability is the extent to which specific users can use a product to their satisfaction in order to effectively achieve specific goals in a specific context of the user [4]. Usability is strongly tied to the extent to which a user's mental model matches and predicts the action of a system [2]. By and large, accessibility and usability are addressed at the end of the construction process of the software which often involves major amendments to it. To evade this, it has been recommended that it should be dealt with positively at the preliminary stage instead of retroactively after testing [5]. Thus the usability features should be incorporated at the requirement stage of the software development cycle. For the novice developer, it will be difficult to analyze the usability functionalities for a blind user unless s/he knows the mental model of blind users in interacting with the touch screen device. The designing of a mental model for blind users is a challenging task. Some research has explored the blind users' mental model and there is a necessity to give more focus on this area to ease the usability problems that blind users continue to face. As a result, we embraced a bottom-up approach based on fieldwork observation (for proposing a mental model) to depict a set of scenarios representing usability issues that have consequences on the final software architecture [6, 7]. Our research purpose is to derive possible usability features of the mental models of blind users of touch screen devices. Furthermore to explore and modify these models pertaining to the touch screen environment and their strategies in dealing with the device.

2 Mental Models

Kenneth Craik in 1943 was the first to invent the "Mental Model" theory [8]. He declared that the mind predicts the future action and reactions in advance by making a miniature dimensional model. The mental model can also be defined as how our mind represents an event/object to predict the future outcome [2, 9–11] affirmed that the mental model is indeed indispensable for designers since it reflects the user's insight of what the system is confined to, how it works and why it works in that way. In 2002, Puerta-Melguizo et al. [12] asserted that the content and structure of the mental model spur on how users interrelate with the system application. It avoids the interaction problem arising between the user and the computer by removing the unconstructive feelings towards the system.

Little research has been conducted on the mental model for blind users. The work on mental model for blind user by Murphy et al. [13] was focused on the challenge faced by the blind user using assistive technologies when browsing the web. The study found that the mental model of a blind user as a "vertical list" and that they perceive all information in a web page as single column like structure. In this case, the blind user will remember the sequence of interested items. The web page having huge information has to memorize a large set of items which is a burden for the blind user mental model. Furthermore, by converting 2-d information in a web page into a single columnar list, many navigational hints would be lost. Due to these factors, usability is achieved after spending more effort and time by a blind user as compared to the sighted user [14, 15]. Another study on the mental model by Takagi et al. [16] confirms that the blind user consider the information as a vertical list of online shopping web sites. They initiated the searching process based on their own scheme to speed up the process.

Some research has also been undertaken in the blind user model that required a more thorough understanding of the blind user's behaviour with the system. The standard model of Kurniawan and Sutcliffe [17] was studied because they investigated the blind user's mental model of the new window environment. He tested blind users in three processes, namely: exploration, task-action and configuration. In the exploration stage, the user first explored the desktop to see what applications are available. In the next stage, the user created another loop of planning how to interact with the system and executing the task. At the last stage, the user configured the system if needed when comfortable with the application.

Kurniawan and Sutcliffe's model [17] is modified by Saei et al. [18] to include more components such as: skill based, knowledge-based, domain user expert and system help to minimize the gap between the developer and the blind user. One of the important observations from our study is that neither Kurniawan nor Saei concern themselves about the action performed in each process of the mental model. Most of the traditional usability such as help, feedback, error prevention were based on what they think and how they handle each situation. No usability is elucidated for what action was performed for each stage of the mental model. Without understanding what action will be performed for the different stages, it will be difficult for developers to understand the exact functioning of the mental model. This is one of the major motivations for us to conduct this study.

3 Touch Screen Usability

The work of Kurniawan [17] gave an idea about how a blind user thinks and acts based on cues given by screen readers using desktop computers. Touch screen interaction differs from the interaction using a desktop screen reader in three ways. Firstly, it facilitates one to interact directly in a significant way than indirectly with a mouse or touchpad. Secondly, it can be held in the hand without requiring any intermediate devices. Finally, touch screen interaction can also be performed

through voice-activated search tools such as Siri or Vlingo. The distinctive features and characteristics of touch screen based Smartphones that makes usability evaluation a demanding process are: (1) they have a small screen size despite the fact that they still have to display large amounts of information, (2) the buttons of the device generally have more than one function, and (3) the devices have limited processing and memory capabilities [19].

Another reason is that the interaction in touch screen between the sighted to blind user varies using touch screen. Generally, the sighted user uses dynamic layout and identify the items through vision. But the blind user uses static layout where s/he has to flicker to identify the items. As a consequence, usability changes for the blind user and therefore the mental model should be adopted to derive their usability.

From the literature, we found that the Kurniawan mental model for blind users [17] is based on a study carried out for screen readers only and not for touch based devices. Not only that, the processes studied concerned: exploration, task action and the configuration processes. However, our user study for blind touch screen users reveals that each process is not executed in a single step. It consists of multiple stages. Therefore, this study will extend the previous works by studying the stages available in each process to propose an updated, more universal and thus an enhanced mental model [20]. Subsequently, factors affecting these stages will be studied to elicit usability features to facilitate the design of the new system.

4 Experiment

The user study was carried out to elicit the mental models of the blind user when interacting with the touch screen based smart phone through audio and haptic feedback. We recruited around seven blind participants with an average age of 35 years. All participants have enough experience of using mobiles with screen readers. None of the participants, however, have experience of using a touch screen mobile. Since the cohort size is small, we conducted on average about 8.5 trials per participants. *All* users have English as a second language (L2). A Smartphone running our prototype generates the speech according to Android development code. The prototype was deployed on the Samsung Galaxy S2 running the Android Ice Cream Sandwich OS touch based Smartphone and tested with the blind users.

The mental model literature suggests that if the system is too complex such as web and touch screen devices without prior training, a mental model is elicited [21]. We adopted the same strategy.

We conducted Verbal and Hands-on Scenario in which the user is required to perform a series of tasks to achieve the target [22]. The user is prompted to think aloud during the experiment. The user has to answer several questions pertaining to hands on the task using the system. It helps the investigator to extract the usability problem they faced.

The blind users were given the target name. The audio cues were given to reach the target. The blind users used these audio cues in order to reach this target. On pressing each target chosen by the blind user, the audio cue informs the name of the target. The blind user has to repeat this task until they reach the desired target. The mental model and usability are elucidated based on observation and discussion with the blind user. Some of our findings are supported by the literature which is mentioned appropriately.

5 Extended Mental Model

The study observed that a blind user for every exploration, task action and configuration process, adopted two stages of strategy to acquire the target: listening and interaction. In the listening stage, a blind user listens to the audio cues to navigate. Based on the listening comprehension, interaction took place. This activity is iterated until the target is reached. Each stage of the strategy based on our understanding, in a developer point of view is explained.

The process of listening is divided into: listening, hold in memory, build the images, search in their database, retrieved, compared, test the image and execute the task. The interaction technique will be faster if the image was already stored. Hence it is imperative for developers to use a common vocabulary for effective interaction between the blind user and the device [20].

The perception of the image may be different between the blind user to the sighted user. But it does not affect the quality of interaction. For instance, the image form for the word 'tiger' will be the same for all sighted users but vary for each blind user. Although it varied widely, based on their own individual perception, the blind user would proceed to the next level.

According to physics, interaction is a transfer of our energy from human to any device. In the exploration process and task action process, interaction is the next stage after listening. This interaction occurs through gestures such as touch and flickering in touch screen devices. When a blind user has performed the interaction, he waits for feedback. If feedback is provided, the user proceeds to the next task. The unexpected feedback cause the user to be stuck from proceeding any further. If the user did not receive the feedback, the user will repeat the task. The application will be terminated if the user incorrectly presses the close button [20].

6 Usability Elucidation

The user study reveals that the listening and interaction stage is either strengthened or weakened by many usability features. In this extended paper, more components are accommodated for listening in support of listening comprehension (LC). Thus, the listening stage is dependent on audio features, listener characteristics, speech

synthesizer and text characteristics. The interaction stage is influenced by gesture, orientation, content and sub-content features. The other features such as user characteristics and environmental factors overlap both the stages if stimulated. The usability features are discussed with their metrics.

6.1 Audio

Audio is the main component for blind users to listen to. The source of audio may be from the environment or the device itself. While the ambient sound from the environment may hinder the interaction, the audio from the device, however, facilitates the interaction. Controlling the audio such as the volume [23] stop, pause and repeat what they listen to determines the effect of the listening.

6.2 Speech Synthesizer

The blind users receive the information aurally. While receiving the information, there is a chance of passing the erroneous message to the user through a speech synthesizer. Thus, choosing an intelligent speech synthesizer is a daunting task. This section deals with the usability problem that the blind user will face while listening to the audio due to the speech synthesizer.

6.2.1 Type of Synthesizer

The type of synthesizer has substantial impact on the quality of the output speech. Natural speech is more intelligent than synthesized speech [24]. The compressed synthesized speech is more intelligible than uncompressed natural speech.

6.2.2 Speech Rate

The L2 listener has more comprehension with a low speech rate compared to a higher speech rate.

6.2.3 Intonation

Accent: It identifies the person is from regionally or socially by the features of pronunciation [25]. Familiar accents are easier to understand than an unfamiliar accent. For instance, it may be difficult for an Arab user to comprehend a British or American English accent than an accustomed Indian English accent.

Table 1 Syntactic phrases

Punctuation	Syntactic parse
(A+B) * C	A plus B star C
We saw Peter, and Mary saw Scott	We saw (Peter and Mary) saw Scott

Phrase: The natural speakers often break up sentences into several phrases, which can be articulated with pausing. Sometimes punctuation can be misleading (Table 1) and sometimes can serve as a guide.

Melody: It refers to the patterns of tones with which a phrase, sentence or paragraph is spoken. The speech synthesizer has a small taxonomy of melodic patterns. It has the limitation of uttering only assigned patterns.

6.3 Listener Characteristics

While considerable individual difference factors may affect both native language (L1) and L2 listening comprehension, this review covers only the subset of factors deemed by our participant (L2) as relevant to the question of difficulty of listening passages on our prototype. The factors discussed here include working memory capacity, the use of metacognitive strategies, language proficiency and experience with the L2 and anxiety.

6.3.1 Working Memory Capacity

It refers to those who are most competent to the presence of mind, attentive and understand easily what they have listened to and have strong provisional storage. The comprehension correlates with greater working capacity [26].

6.3.2 Metacognitive Strategies

The L2 listener having good metacognitive strategies—those who are aware of and use effective strategies shows better listening comprehension [27].

6.3.3 Proficiency

It involves familiarity with non-native language's vocabulary size and phonology; amount of exposure to the language and background information about, scheme, structure, text and culture.

The listener's ability to correctly decipher the phonology and vocabulary improves with an increase in proficiency and experience. Prior experience in the relevant study compensates for mishearing or encountering unfamiliar words, which can increase comprehension.

6.3.4 Anxiety

When a listener feels the message is too complex or difficult to understand, concentration falters and comprehension declines. Anxiety mostly occurs when the listener is trying to sort conflicting information, listening to illogical passage(s) and also to new information. It causes a negative impact on comprehension.

6.4 Text Characteristics

Although there are many factors which influence the L2 listener, our study is based on the factors which influence the usability of the system. The factors discussed here include the passage: length, complexity, type, organization, authenticity and readability grade.

6.4.1 Passage Length

The passage length has been defined by the researchers with a number of measures such as syllabus/second, duration, number of words or sentences. This section is classified into passage length, and redundancy.

Passage length: The research reveals that the longer length passage has some difficulty in LC, but the effect is weak. Since the user has to listen to the keywords of the sentences and then build-up the sentences on their own.

Redundancy: Repetition of information increases the LC consistently but it depends on the type of redundancy such as whether the sentence is paraphrased or an exact repetition. It is supported by Sasaki [28].

6.4.2 Complexity

Passage complexity is referred directly to dissimilar properties such as syntactic structure, pragmatic information and directness.

Syntactic feature: Newly listened to vocabulary have a detrimental impact on LC. Negatives (negative prefixes like '–un' and negative markers like 'not') have higher impact than positive keywords.

Pragmatic information: Inclusion of pragmatic information such as idioms and culturally specific vocabulary decreases the LC among L2 listeners which was proved by Sasaki [28].

Directness: The sentence with implied meaning is more difficult for the L2 low proficiency listener.

6.4.3 Organization

Passage type: The passages about familiar topics are easier for L2 listeners than the passages about unfamiliar topics.

Orality: It is the extent to which a passage contains more spoken words (non-academic) than written languages (academic) words [29]. The passages that are more oral are less difficult to understand for L2 listeners. Such passage have more disfluencies, greater redundancy and simpler syntax (not grammatical).

Coherence: It involves the appearance of logicality and relevance in a passage. The less coherence between the passage the more they seem off-topic or tangential.

Discourse Markers: Discourse markers such as *but*, *however*, increases the coherence of the sentences. Thus, it improves LC among L2 listeners as mentioned in [30].

Readability grade: The reading scale is the ease in which the text can be interpreted and comprehended. This scale is used when the listening user are kids [31].

6.5 Handedness

Hand movements are very crucial for gesture based interaction such as touch and flick. Mostly a finger is used for interaction. Often, the index finger is used for interaction. Sometimes if the proximity of a finger to a device is narrow, there is a chance for other fingers to hit the surface of touch screen leading to execute unexpected events. Generally, a blind user seldom uses multiple hands for interaction.

The study reveals the size and shape of the finger also play a vital role in the exploration process. The bigger size finger will hit many targets simultaneously which will lead to mayhem during navigation. If the shape of the finger was not normal then it may hit the wrong target on the touch screen. As a result, developers should take care of the target size, avoiding the target to be placed in crowded areas and keeping the padding size normal to facilitate easy exploration.

6.6 Contents and Sub-contents

The contents can be classified into: text, images, audio/video and widgets. The structure of the text was already discussed in Sect. 6.4. The image information was delivered to the blind user through alternative text. It needs the advice of a domain expert to deliver precisely [18].

Delivering video is identical to audio delivery which was discussed in Sect. 6.1. Widgets are the interaction point in the touch screen to operate the given kind of data or application. The interaction varies with the types of widget such as buttons, text input, list and menus. The orientation of the widget also causes an impediment for interaction for blind users. At present, blind users are able to interact only with the textbox and a button. Therefore, more study is needed possibly in the future for direct exploration of different classes of widgets.

6.7 User Characteristics

User characteristics determine the effectiveness of listening and interaction. User characteristic such as age affects the exploration process. The rate of hand movement and the preciseness of hitting the target were decreasing with an increase in age.

Mental state affects the rate of the exploration process. While positive mood enhances the process, negative mood retards it. At some posture, the body may assume a great variety of shapes and positions. If the body and mind are not stressed, comfortability can be achieved and hence the exploration phase will take place faster.

Physiological factors such as illness, stress and fatigue will cause discomfort during the interaction. A high level of exposure (familiarity) enhances the interaction level.

6.8 Environmental Factors

Environmental factors such as: noise [32], 3333odour, or weather induced sweating reduce the speed of exploration with the devices. As environmental factors cannot be controlled, the developers can take cautious steps to minimize the accessibility burden.

7 Discussion

The user evaluation of the obtained data strongly confirms the hypothesis that usability can be elucidated through the action performed on each process of mental model.

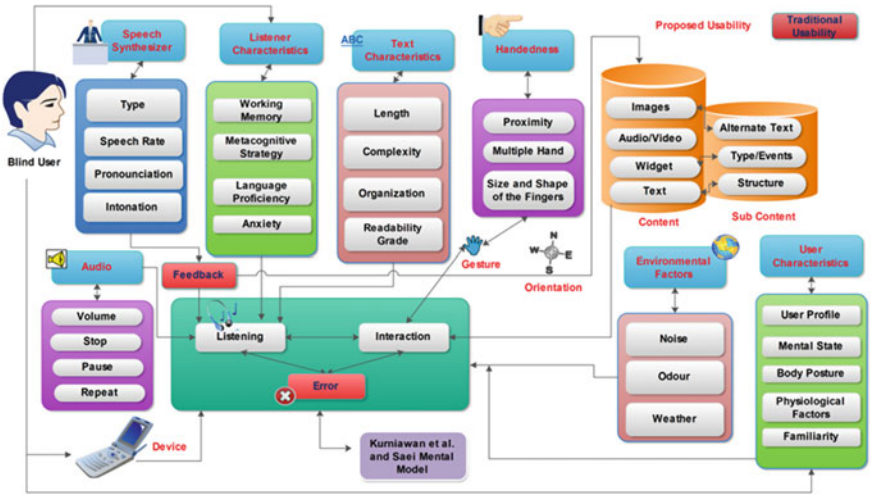


Fig. 1 Proposed mental model with elucidate usability and its metrics

The proposed mental model is the extension of existing Kurniawan and Sutcliffe's [17] mental model for the blind users with respect to touch screen technologies. The proposed conceptual mental model includes listening and interaction components which support each stage in the mental model. These components are necessary in order to assist the developers to minimize the gap between the developers and the blind person.

The mental models are always unstable [33]. In addition, the added components such as listening and interaction are stable. The techniques used for these components are unstable. The interaction with the system differs with the advancement of technology. For instance, the keyboard is used in desktop computing to navigate through the items. On the other hand, swipe gesture is used in touch screen technologies. Below we discuss some findings derived from this study.

- **The usability elucidated with the mental model is less than 100 % of the most complex applications**. These differences are due to the fact that the complexity of usability calls for the proper identification of usable features in the affected parts of the system. The usability elucidated in this paper is universal in nature and can be applied for all the 'apps' (applications) if needed. Specific usability problems can be identified with stackholders and users.
- **Text characteristics are good. But users are not able to comprehend what was spoken**. This impediment is due to the speech synthesizer. The speech synthesizer used by the user should be evaluated based on the usability problem such as: what type of synthesizer is used, speech rate, understandable pronunciation and suitable intonation (see Fig. 1). Another impediment is listener characteristics. If the user has poor working memory or less metacognitive strategy or poor proficiency in the spoken language or is anxious due to some

reason (see Fig. 1), the listener (user) has to be excluded from testing or to accommodate some features in the application to alleviate this problem.

- **Some of the elucidate usability features did identify the testing condition of usability**. For instance, developers need to think about the problems faced by blind users during the different types of weather and climatic conditions. For example, if developers feel that noise due to thunder will have an adverse impact on the interaction, than an audio control such as volume increase and/or decrease can be included in the software requirement.
- **Some of the usability features overlap with each other**. For instance, the developer felt that during a negative mood phase (mental state—user characteristics), the blind user may perform any of these undesirable two actions: firstly, the user can press the target forcefully, giving more pressure during the touch. It can lead to the non-execution of the touch event or the execution of a long click event; secondly, the user can hit the incorrect target, which may lead to the execution of an undesired event.

To solve these, the developer can adjust the requirement to get the user to confirm before the event is executed.

- **Usability is the subset of accessibility. We cannot elucidate any major usability problems out of the inaccessible part of the system**. Currently, the blind user is not interacting with the content such as the widget except by button control. They navigate through the pages with the swipe gesture in a static layout. Thus, the minor usability problem arises in a static layout where direct interaction with the content is not experienced.

Although our findings are specific to our prototype developed, it needs more checking based on the requirements of the target application. These findings give reliability in eliciting usability features to the knowledge warehouse that is beneficial in the process of asking the right questions by the novice developers to the stakeholders (blind users) and to confining accurate usability requirements for software development without an HCI expert on the development team.

8 Conclusion

It is highly important to define usability requirements at the earliest stage of software development process such as in the requirement stage. This is a difficult task as usability features are more difficult to specify as it requires a lot of discussion among stakeholders, especially the blind users or to approach HCI expertise to perform this. However, non-availability of HCI expertise or high cost to be paid to the HCI experts will cause the developer to find an alternative solution to elicit usability requirements. Our work takes a step in this direction, suggesting that usability features should be dealt with at the requirements stage.

This paper has focused on eliciting usability based on the proposed mental model. This analysis will reduce the burden of novice developers to understand the mental model of blind users. The list of usability features suggested by the paper is not intended to be exhaustive; there are a number of confounding variables that need to be considered: metrics to measure listening load and complexity of interaction with the touch screen system. However, our future studies will address these factors and should help interpret the proposed metrics within these constraints.

References

1. P.N. Johnson-Laird, Mental models and thought, in *The Cambridge Handbook of Thinking and Reasoning (Cambridge Handbooks in Psychology)*, ed. by K.J. Holyoak, R.G. Morrison (Cambridge University Press, Cambridge, 2005), pp. 185–208. ISBN: 13: 978-0521531016
2. M.J. Davidson, L. Dove, J. Weltz, Mental models and usability. Cogn. Psychol. **404**, (1999). Available: http://www.lauradove.info/reports/mental%20models.htm
3. S. Weinschenk, The secret to designing an intuitive UX: match the mental model to the conceptual model, Article No. 513, 8 Oct 2011. Available: http://uxmag.com/articles/the-secret-to-designing-anintuitive-user-experience
4. ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 11: Guidance on usability, 15 March, 1998. Available: http://www.it.uu.se/edu/course/homepage/acsd/vt09/ISO9241part11.pdf
5. N. Juristo, A.M. Moreno, M.-I. Sanchez-Sugura, Analysing the impact on usability on software design. J. Syst. Soft. **80**(9), 1506–1516 (2007). doi:10.1016/j.jss.2007.01.006. ISSN: 0164-1212. Available: http://www.sciencedirect.com/science/article/pii/S0164121207000088
6. L. Bass, B.E. John, J. Kates, Achieving Usability Through Software Architecture, Tech. R. CMU/SEI-2001-TR-005, ESC-TR-2001-005, Software Engineering Institute, Carnegie Mellon University, pp. 29-30, http://resources.sei.cmu.edu/asset_files/TechnicalReport/2001_005_001_13859.pdf
7. L. Bass, B.E. John, Linking usability to software architecture patterns through general scenarios. J. Syst. Soft. **66**(3), 87–197 (2003). doi:10.1016/S0164-1212(02)00076-6
8. D.A. Norman, *The Design of Everyday Things, Revised and Expanded Edition* (Basic Books, New York, 2013). ISBN-13: 978-0465050659
9. A. Kurtz, Mental Models—A Theory Critique, (2003). Available: http://mcs.open.ac.uk/yr258/ment_mod/
10. S. Makri, A. Blandford, J. Gow, J. Rimmer, C. Warwick, G. Buchanan, A library or just another information resource? A case study of users' mental models of traditional and digital libraries. J. Am. Soc. Info. Sci. Tech. **58**(3), 433–445 (2007). doi:10.1002/asi.20510
11. K. Potesnak, Mental models: helping users understand software. Software IEEE **6**(5), 85–86 (1989). doi:10.1109/52.35592
12. M.C. Puerta-Melguizo, C. Chisalita, G.C. Van der Veer, Assessing users mental models in designing complex systems. Syst. Man Cybern. IEEE Int. Conf. **7**, 6 (2002). doi:10.1109/ICSMC.2002.1175734
13. E. Murphy, R. Kuber, G. McAllister, P. Strain, W. Yu, An empirical investigation into the difficulties experienced by visually impaired Internet users. Univers. Access Inf. Soc. **7**(1–2), 79–91 (2008). doi: 10.1007/s10209-007-0098-4. (Springer)
14. J.P. Bigham, A.C. Cavender, J.T. Brudvik, J.O. Wobbrock, R.E. Lander, WebinSitu: a comparative analysis of blind and sighted browsing behavior, in *ASSETS'07 Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*,

pp. 51–58, 15–17 Oct 2007. doi:10.1145/1296843.1296854. ISBN: 978-1-59593-573-1. Available: http://webinsight.cs.washington.edu/papers/webinsitu.pdf
15. K. Shinohara, J.D. Tenenberg, A blind person's interaction with technology. Comm. ACM **52**(8), 58–66 (2009)
16. H. Takagi, S. Saito, K. Fukuda, C. Asakawa, Analysis of navigability of Web Applications for improving blind usability. ACM Trans. Comput. Hum. Interact. **14**(3), 13 (2007). doi:10.1145/1279700.1279703
17. S.H. Kurniawan, A. Sutcliffe, Mental models of blind users in the windows environment, in *Computers Helping People with Special Needs*. Lecture Notes in Computer Science, vol. 2398 (Springer, Berlin, 2002), pp. 568–574. doi: 10.1007/3-540-45491-8_109
18. S.N.S.M. Saei, S. Sulaiman, H. Hasbullah, Mental model of blind users to assist designers in system development. 2010 Int. Symp. Info. Tech. (ITSim), **1**, 1–5 (2010). doi:10.1109/ITSIM.2010.5561350
19. Y.S. Lee, S.W. Hong, T.L. Smith-Jackson, M.A. Nussbaum, K. Tomioka, Systematic evaluation methodology for cell phone user interfaces. Interact. Comput. **18**(2), 304–325 (2006). doi: http://dx.doi.org/10.1016/j.intcom.2005.04.002. ISSN: 0953-5438. Available: http://www.sciencedirect.com/science/article/pii/S0953543805000366
20. M. Fakrudeen, M. Ali, S. Yousef, A.H. Hussein, Analysing the mental model of blind users in mobile touch screen devices for usability, in *Proceedings of The World Congress on Engineering 2013*. Lecture Notes in Engineering and Computer Science, vol. II, WCE 2013, London, pp. 837–842, 3–5 July 2013. Available: http://www.iaeng.org/publication/WCE2013/WCE2013_pp837-842.pdf
21. Y. Zhang. The influence of mental models on undergraduate students' searching behavior on the Web. Inf. Process. Manage.**44**(3), 1330–1345 (2008).
22. M. Sasse, Eliciting and Describing Users' Models of Computer Systems, Ph.D. University of Birmingham, 1997
23. P. Adank, B.G. Evans, J. Stuart-Smith, S.K. Scott, Comprehension of familiar and unfamiliar native accents under adverse listening conditions. J. Exp. Psychol. Hum. Percept. Perform. **35**(2), 520–529 (2009). doi:10.1037/a0013552. ISSN: 0096-1523. Available: https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-man-scw:29214&datastreamId=POST-PEER-REVIEW-PUBLISHERS-DOCUMENT.PDF
24. K. Papadopoulos, E. Katemidou, A. Koutsoklenis, E. Mouratidou, Differences among sighted individuals and individuals with visual impairments in word intelligibility presented via synthetic and natural speech. Augmentative Altern. Comm. **26**(4), 278–288 (2010). doi:10.3109/07434618.2010.522200
25. D. Crystal, A Dictionary of Linguistics and Phonetics, 6th edn. (Wiley-Blackwell, Malden, 2008). ISBN: 978-1-4051-5296-9
26. M. Harrington, M. Sawyer, L2 working memory capacity and L2 reading skill. Stud. Second Lang. Acquisition **14**(1), 25–38 (1992). doi:10.1017/S0272263100010457. (Cambridge University Press)
27. L. Vandergrift, C. Goh, C. Marescha, M. Tafaghodtar, The metacognitive awareness listening questionnaire: development and validation. Lang. Learn. **56**(3), 431–462 (2006)
28. M. Sasaki, Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. Lang. Test. **17**(1), 85–114 (2000). doi:10.1177/026553220001700104
29. G.R. Kiany, S. Jalali, Relationship between modality, types of passages, and performance of advanced EFL learners on listening comprehension tests. Iran. J. Appl. Linguist. **9**(2), 79–99. Available: http://www.sid.ir/en/VEWSSID/J_pdf/87620060202.pdf
30. E.H. Jung, The role of discourse signaling cues in second language listening comprehension. Mod. Lang. J. **87**(4), 562–577 (2003). doi:10.1111/1540-4781.00208
31. S.A. Crossely, D.B. Allen, D.S. McNamara, Text readability and intuitive simplification: a comparison of readability formulas. Read. Foreign Lang. **23**(1), 84–101 (2011). Available: http://files.eric.ed.gov/fulltext/EJ926371.pdf. ISSN: 1539-0578

32. N. Golestani, S. Rosen, S.K. Scott, Native-language benefit for understanding speech-in-noise: The contribution of semantics. Biling. (Camb. Engl.) **12**(3), 385–392 (2009). doi: 10.1017/S1366728909990150. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2999832/
33. D.A. Norman, Some observations on mental models, in *Mental Models*, ed. by D. Gentner, A.L. Stevens (Lawrence Erlbaum Associates, New Jersey, 1983), pp. 7–14. ISBN-13: 978-0898592429

Numerical Solution and Stability of Block Method for Solving Functional Differential Equations

Fuziyah Ishak, Mohamed B. Suleiman and Zanariah A. Majid

Abstract In this article, we describe the development of a two-point block method for solving functional differential equations. The block method, implemented in variable stepsize technique produces two approximations simultaneously using the same back values. The grid-point formulae for the variable steps are derived, calculated and stored at the start of the program for greater efficiency. The delay solutions for the unknown function and its derivative at earlier times are interpolated using the previous computed values. Stability regions for the block method are illustrated. Numerical results are given to demonstrate the accuracy and efficiency of the block method.

Keywords Block method · Delay differential equation · Functional differential equation · Neutral delay differential equation · Polynomial interpolation · Stability region

F. Ishak (✉)
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia
e-mail: fuziyah@tmsk.uitm.edu.my

M. B. Suleiman
Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia,
43400 Serdang, Selangor, Malaysia
e-mail: mohamed@math.upm.edu.my

Z. A. Majid
Mathematics Department, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
e-mail: zanariah@math.upm.edu.my

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_43,

1 Introduction

Differential equation of the form

$$y'(x) = f(x, y(x), y(\alpha(x)), y'(\alpha(x))) \tag{1}$$

appears in many real life applications and has been investigated by many authors in recent years. The classical case is when $\alpha(x) = x - \tau$, τ a constant. When the right hand side of (1) does not depend on the derivative of the unknown function y, the equation is known as delay differential equation. Otherwise, it is known as neutral delay differential equation. In this article, we consider numerical solution for functional differential equation of the form:

$$\begin{cases} y'(x) = f(x, y(x), y(qx), y'(qx)), & 0 < x \leq T, \\ y(0) = y_0, \end{cases} \tag{2}$$

where $0 < q < 1$, and

$$\begin{cases} y'(x) = f(x, y(x), y(x-\tau), y'(x-\tau)), & 0 < x \leq T, \\ y(x) = \phi(x), & x \leq 0. \end{cases} \tag{3}$$

Equation (2), known as the pantograph equation arises in many physical applications such as number theory, electrodynamics, astrophysics, etc. Detailed explanations can be found in [1–3]. Numerical solutions for (2) and (3) have been studied extensively, see for example [4–11] and the references cited therein. These methods produce one approximation in a single integration step. Block methods, however produce more than one approximation in a step. Block methods have been used to solve wide range of ordinary differential equations as well as delay differential equations (see [12–16] and the references cited therein).

The functional differential equations are solved using a two-point block method in variable step. In a single integration step, two new approximates for the unknown function are obtained using the same stepsize. New approximates for the next block are obtained by keeping the stepsize constant, doubled or halved depending upon the local approximation error. In any variable stepsize method, the coefficients of the method need to be recalculated whenever the stepsize changes. In order to avoid the tedious calculation, the coefficients based on the stepsize ratio are calculated beforehand and stored at the start of the program.

The organization of this article is as follows. In Sect. 2, we briefly describe the development of the variable step block method. Stability region for the block method is discussed in Sect. 3. Numerical results for some functional differential equations are presented in Sect. 4 and finally Sect. 5 is the conclusion.

2 Method Development

Referring to (1), we seek a set of discrete solutions for the unknown function y in the interval [0, T]. The interval is divided into a sequence of mesh points $\{x_i\}_{i=0}^{t}$ of different lengths, such that $0 = x_0 < x_1 < \cdots < x_t = T$. Let the approximated solution for $y(x_n)$ be denoted as y_n. Suppose that the solutions have been obtained up to x_n. At the current step, two new solutions y_{n+1} and y_{n+2} at x_{n+1} and x_{n+2} respectively are simultaneously approximated using the same back values by taking the same stepsize. The points x_{n+1} and x_{n+2} are contained in the current block. The length of the current block is $2h$. We refer to this particular block method as two-point one-block method. The block method is shown in Fig. 1.

In Fig. 1, the stepsize of the previous step is viewed in the multiple of the current stepsize. Thus, $x_{n+1} - x_n = h$, $x_{n+2} - x_{n+1} = h$ and $x_{n\text{-}1} - x_{n-2} = x_n - x_{n-1} = rh$. The value of r is either 1, 2, or $\frac{1}{2}$, depending upon the decision to change the stepsize. In this algorithm, we employ the strategy of having the stepsize to be constant, halved or doubled.

The formulae for the block method can be written as the pair,

$$\begin{aligned} y_{n+1} &= y_n + h\sum_{i=0}^{4} \beta_i(r) f(x_{n-2+i}, y_{n-2+i}, \bar{y}_{n-2+i}, \hat{y}_{n-2+i}), \\ y_{n+2} &= y_n + h\sum_{i=0}^{4} \beta_i^*(r) f(x_{n-2+i}, y_{n-2+i}, \bar{y}_{n-2+i}, \hat{y}_{n-2+i}), \end{aligned} \tag{4}$$

where $\bar{y}_n$ and $\hat{y}_n$ are the approximations to $y(\alpha(x_n))$ and $y'(\alpha(x_n))$ respectively. For simplicity, from now on we refer to $f(x_n, y_n, \bar{y}_n, \hat{y}_n)$ as f_n. The coefficient functions $\beta_i(r)$ and $\beta_i^*(r)$ will give the coefficients of the method when r is either 1, 2, or $\frac{1}{2}$.

The first formula in (4) is obtained by integrating (1) from x_n to x_{n+1} while replacing the function f with the polynomial P where $P(x)$ is given by

$$P(x) = \sum_{j=0}^{4} L_{4,j}(x) f_{n+2-j},$$

and

$$L_{4j}(x) = \prod_{\substack{i=0 \\ i\neq j}}^{4} \frac{(x - x_{n+2-i})}{(x_{n+2-j} - x_{n+2-i})}, \quad \text{for} \quad j = 0, 1, \ldots, 4.$$

Similarly, the second formula in (4) is obtained by integrating (1) from x_n to x_{n+2} while replacing the function f with the polynomial P. The value of $\bar{y}_n$ is obtained by the interpolation function such as,

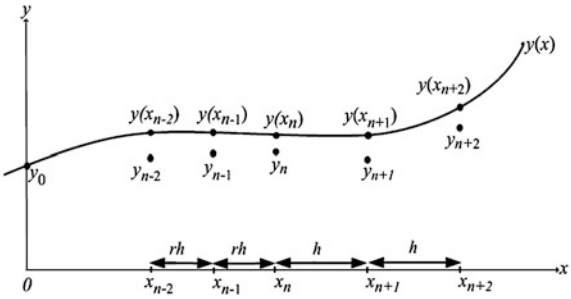


Fig. 1 Two-point one-block method

$$\bar{y}_n = y[x_j] + (\alpha(x_n) - x_j)y[x_j, x_{j-1}] + \cdots + (\alpha(x_n) - x_j)\cdots(\alpha(x_n) - x_{j-3})y\,[x_j, \ldots, x_{j-4}],$$

where

$$y[x_j, x_{j-1}, \ldots, x_{j-4}] = \frac{y[x_j, \ldots, x_{j-3}] - y[x_{j-1}, \ldots, x_{j-4}]}{x_j - x_{j-4}},$$

provided that $x_{j-1} \leq \alpha(x_n) \leq x_j$, $n \geq j$, $j \geq 1$. We approximate the value of $\hat{y}_n$ by interpolating the values of f, that is,

$$\hat{y}_n = f[x_j] + (\alpha(x_n) - x_j)f[x_j, x_{j-1}] + \cdots + (\alpha(x_n) - x_j)\cdots(\alpha(x_n) - x_{j-3})f[x_j, \ldots, x_{j-4}],$$

where

$$f[x_j, x_{j-1}, \ldots, x_{j-4}] = \frac{f[x_j, \ldots, x_{j-3}] - f[x_{j-1}, \ldots, x_{j-4}]}{x_j - x_{j-4}}.$$

The formulae in (4) are implicit, thus a set of predictors are derived similarly using the same number of back values. The corrector formulae in (4) are iterated until convergence.

For greater efficiency while achieving the required accuracy, the algorithm is implemented in variable stepsize scheme. The stepsize is changed based on the local error that is controlled at the second point. A step is considered successful if the local error is less than a specified tolerance. If the current step is successful, we consider either doubling or keeping the same stepsize. If the same stepsize had been used for at least two blocks, we double the next stepsize. Otherwise, the next stepsize is kept the same. If the current step fails, the next stepsize is reduced by half. For repeated failures, a restart with the most optimal stepsize with one back value is required. For variable step algorithms, the coefficients of the methods need to be recalculated whenever a stepsize changes. The recalculation cost of these coefficients is avoided by calculating the coefficients beforehand and storing them

at the start of the program. With our stepsize changing strategy, we store the coefficients $\beta_i(r)$ and $\beta_i^*(r)$ for r is 1, 2 and $\frac{1}{2}$.

3 Region of Absolute Stability

In the development of a numerical method, it is of practical importance to study the behavior of the global error. The numerical solution y_n is expected to behave as the exact solution $y(x_n)$ does as x_n approaches infinity. In this section, we present the result of stability analysis of the two-point one-block method when they are applied to the delay and neutral delay differential equations with real coefficients.

For the sake of simplicity and without the lost of generality, we consider the equation

$$\begin{array}{ll} y'(x) = ay(x) + by(x-\tau) + cy'(x-\tau), & x \geq 0, \\ y(x) = \phi(x), & -\tau \leq x < 0, \end{array} \tag{5}$$

where a, b, $c \in R$, τ is the delay term such as $\tau = mh$, h is a constant stepsize such that $x_n = x_0 + nh$ and $m \in Z^+$. If $i \in Z^+$, we define vectors $\mathbf{Y}_{N+i} = \begin{bmatrix} y_{n-3+2i} \\ y_{n-2+2i} \end{bmatrix}$ and $\mathbf{F}_{N+i} = \begin{bmatrix} f_{n-3+2i} \\ f_{n-2+2i} \end{bmatrix}$. Then, the block method (4) can be written in matrix form such as,

$$A_1\mathbf{Y}_{N+1} + A_2\mathbf{Y}_{N+2} = h\sum_{i=0}^{2} B_i(r)\mathbf{F}_{N+i}, \tag{6}$$

where $A_1 = \begin{bmatrix} 0 & -1 \\ 0 & -1 \end{bmatrix}$, $A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $B_i(r)$ is a matrix that contains the coefficients $\beta_i(r)$ and $\beta_i^*(r)$. Applying method (6) to (5), we get

$$\begin{aligned} A_1\mathbf{Y}_{N+1} + A_2\mathbf{Y}_{N+2} =& H_1\sum_{i=0}^{2} B_i(r)\mathbf{Y}_{N+i} + H_2\sum_{i=0}^{2} B_i(r)\mathbf{Y}_{N+i-m} \\ &+ cA_2\mathbf{Y}_{N+2-m} + cA_1\mathbf{Y}_{N+1-m}, \end{aligned}$$

where $H_1 = ha$ and $H_2 = hb$. Rearranging, we have

$$\sum_{i=0}^{2} (A_i - H_1B_i(r))\mathbf{Y}_{N+i} = \sum_{i=0}^{2} (H_2B_i(r) + cA_i)\mathbf{Y}_{N+i-m}, \tag{7}$$

where A_0 is the null matrix. Characteristic polynomial for (7) is given by $C_m(H_1, H_2, c; \zeta)$ where C_m is the determinant of

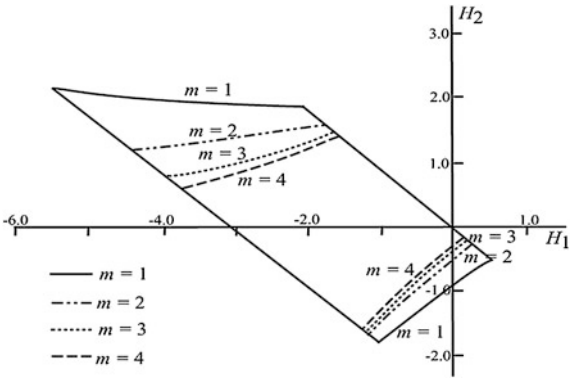


Fig. 2 Stability regions for the block method with $c = 0$, $r = 1$

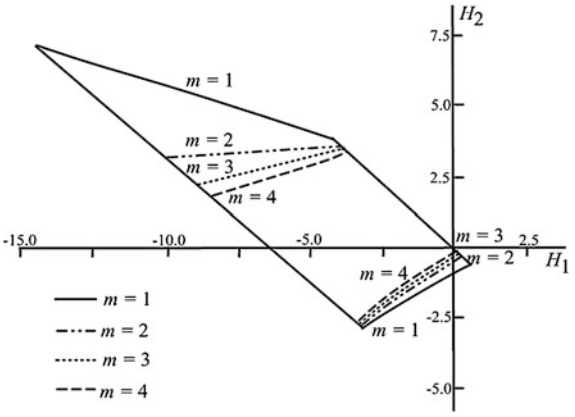


Fig. 3 Stability regions for the block method with $c = 0$, $r = 2$

$$\sum_{i=0}^{2} (A_i - H_1 B_i(r))\zeta^{m+i} - \sum_{i=0}^{2} (H_2 B_i(r) + cA_i)\zeta^i = 0. \tag{8}$$

The numerical solution (7) is asymptotically stable if and only if for all m, all zeros of the characteristic polynomial (8) lie within the open unit disk in the plane. The stability region is defined as follows:

Definition 1 For a fixed stepsize h, a, $b \in R$, and for any, but fixed c, the region S in the $H_1 - H_2$ plane is called the stability region of the method if for any $(H_1, H_2) \in S$, the numerical solution of (5) vanishes as x_n approaches infinity.

In Figs. 2, 3, 4 the stability regions for $c = 0$ are depicted with $r = 1$, $r = 2$, and $r = \frac{1}{2}$, respectively, see also [15]. In Fig. 5, the stability regions for $m = 1$ and $c = 0.5$ are illustrated. We use the boundary locus technique as described in [17, 18]. The regions are sketched for $r = 1$, $r = 2$, and $r = \frac{1}{2}$.

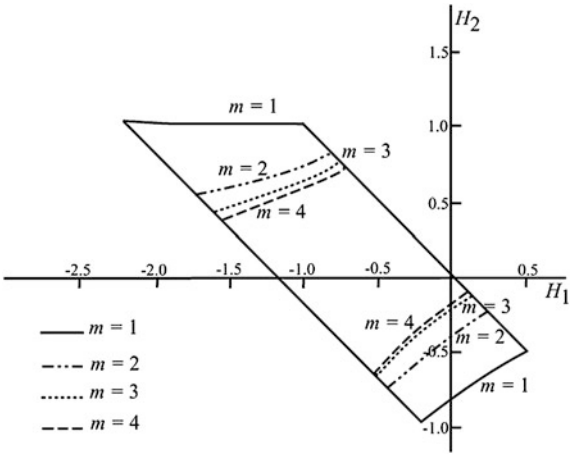


Fig. 4 Stability regions for the block method with $c = 0$, $r = 0.5$

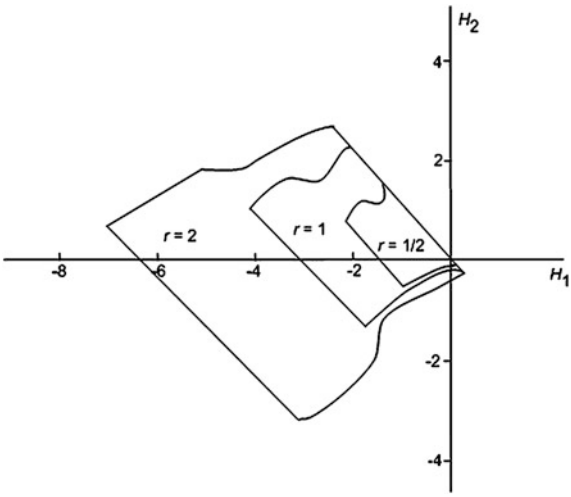


Fig. 5 Stability regions for the block method with $c = 0.5$

The coefficient matrices are given as follows:
For $r = 1$:

$$B_0 = \begin{bmatrix} 0 & \frac{11}{720} \\ 0 & \frac{-1}{90} \end{bmatrix}, \quad B_1 = \begin{bmatrix} \frac{-74}{720} & \frac{456}{720} \\ \frac{4}{90} & \frac{24}{90} \end{bmatrix}, \quad \text{and} \quad B_2 = \begin{bmatrix} \frac{346}{720} & \frac{-10}{720} \\ \frac{124}{90} & \frac{29}{90} \end{bmatrix}.$$

For $r = 2$:

$$B_0 = \begin{bmatrix} 0 & \frac{137}{14400} \\ 0 & \frac{-1}{900} \end{bmatrix}, \quad B_1 = \begin{bmatrix} \frac{-335}{14400} & \frac{7455}{14400} \\ \frac{5}{900} & \frac{285}{900} \end{bmatrix}, \quad \text{and} \quad B_2 = \begin{bmatrix} \frac{7808}{14400} & \frac{-565}{14400} \\ \frac{1216}{900} & \frac{295}{900} \end{bmatrix}.$$

For $r = \frac{1}{2}$:

$$B_0 = \begin{bmatrix} 0 & \frac{145}{1800} \\ 0 & \frac{-20}{225} \end{bmatrix}, \quad B_1 = \begin{bmatrix} \frac{-704}{1800} & \frac{1635}{1800} \\ \frac{64}{225} & \frac{15}{225} \end{bmatrix}, \quad \text{and} \quad B_2 = \begin{bmatrix} \frac{755}{1800} & \frac{-31}{1800} \\ \frac{320}{225} & \frac{71}{225} \end{bmatrix}.$$

Referring to Figs. 2, 3, 4, 5, the stability regions are closed region bounded by the corresponding boundary curves. It is observed that the stability region shrinks as the stepsize increases.

4 Numerical Results

In this section, we present some numerical examples in order to illustrate the accuracy and efficiency of the block method. The examples taken and cited from [8, 19] are as follows:

Example 1

$$y'(x) = \frac{1}{2}y(x) + \frac{1}{2}e^{x/2}y\left(\frac{x}{2}\right), \quad 0 \le x \le 1,$$
$$y(0) = 1.$$

The exact solution is $y(x) = e^x$.

Example 2

$$y'(x) = -\frac{5}{4}e^{-x/4}y\left(\frac{4}{5}x\right), \quad 0 \le x \le 1,$$
$$y(0) = 1.$$

The exact solution is $y(x) = e^{-1.25x}$.

Example 3

$$y'(x) = -y(x) + \frac{q}{2}y(qx) - \frac{q}{2}e^{-qx}, \quad 0 \le x \le 1,$$
$$y(0) = 1.$$

The exact solution is $y(x) = e^{-x}$.

Example 4

$$y'(x) = ay(x) + by(qx) + \cos x - a\sin x - b\sin(qx), \quad 0 \le x \le 1,$$
$$y(0) = 0.$$

The exact solution is $y(x) = \sin x$.

Table 1 Numerical results for Example 1

TOL	STEP	FS	AVERR	MAXE
10^{-2}	20	0	7.02683E−01	9.68012E−05
10^{-4}	27	0	4.15905E−07	9.05425E−07
10^{-6}	35	0	3.09498E−07	4.21310E−07
10^{-8}	48	0	9.55941E−09	1.28084E−08
10^{-10}	75	0	7.68855E−11	9.88663E−11

Table 2 Numerical results for Example 2

TOL	STEP	FS	AVERR	MAXE
10^{-2}	21	0	9.76063E−08	1.17093E−07
10^{-4}	27	0	4.95202E−09	1.27731E−07
10^{-6}	35	0	1.06955E−09	1.21167E−08
10^{-8}	50	0	3.12606E−11	1.43148E−10
10^{-10}	79	0	5.60297E−13	1.51457E−12

Example 5

$$y'(x) = -y(x) + \frac{1}{2}y\left(\frac{x}{2}\right) + \frac{1}{2}y'\left(\frac{x}{2}\right), \quad 0 \le x \le 1,$$
$$y(0) = 1.$$

The exact solution is $y(x) = e^{-x}$.

Example 6

$$y'(x) = -y(x) + 0.1y(0.8x) + 0.5y'(0.8x) + (0.32x - 0.5)e^{-0.8x} + e^{-x}, \quad 0 \le x \le 10,$$
$$y(0) = 0.$$

The exact solution is $y(x) = xe^{-x}$.

Example 7

$$y'(x) = y(x) + y(x-1) - \tfrac{1}{4}y'(x-1), \quad 0 \le x \le 1,$$
$$y(x) = -x, \quad x \le 0.$$

The exact solution is $y(x) = -\frac{1}{4} + x + \frac{1}{4}e^{x}$.

Example 8

$$y'(x) = y(x) + y(x-1) - 2y'(x-1), \quad 0 \le x \le 1,$$
$$y(x) = -x, \quad x \le 0.$$

The exact solution is $y(x) = -2 + x + 2e^{x}$.

Numerical results for Example 1–8 are given in Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. The following abbreviations are used in the tables, TOL—the chosen tolerance,

Table 3 Numerical results for Example 3, $q = 0.2$

TOL	STEP	FS	AVERR	MAXE
10^{-2}	20	0	1.54360E−05	1.91824E−05
10^{-4}	27	0	1.15315E−07	5.14934E−07
10^{-6}	35	0	6.83967E−08	8.52745E−08
10^{-8}	47	0	2.03160E−09	2.57931E−09
10^{-10}	74	0	1.50180E−11	2.00260E−11

Table 4 Numerical results for Example 3, $q = 0.8$

TOL	STEP	FS	AVERR	MAXE
10^{-2}	20	0	1.57507E−07	2.30455E−06
10^{-4}	27	0	2.90818E−08	4.90683E−07
10^{-6}	35	0	4.03006E−10	3.90288E−09
10^{-8}	48	0	2.46267E−11	1.51318E−10
10^{-10}	74	0	3.09422E−13	1.74260E−12

Table 5 Numerical results for Example 4, $a = -1$, $b = 0.5$, $q = 0.1$

TOL	STEP	FS	AVERR	MAXE
10^{-2}	20	0	1.61996E−06	1.15114E−05
10^{-4}	27	0	1.08039E−07	1.15418E−06
10^{-6}	35	0	8.23406E−09	1.15448E−07
10^{-8}	48	0	5.74175E−10	1.15451E−08
10^{-10}	79	0	4.12730E−11	1.15451E−09

Table 6 Numerical results for Example 4, $a = -1$, $b = 0.5$, $q = 0.5$

TOL	STEP	FS	AVERR	MAXE
10^{-2}	20	0	6.25587E−06	4.77895E−05
10^{-4}	27	0	4.53165E−07	4.78257E−06
10^{-6}	35	0	3.48942E−08	4.78294E−07
10^{-8}	50	0	2.91205E−09	4.78298E−08
10^{-10}	79	0	1.83823E−10	4.78299E−09

Table 7 Numerical results for Example 5

TOL	STEP	FS	AVERR	MAXE
10^{-2}	20	0	7.31358E−04	1.83509E−03
10^{-4}	27	0	1.88577E−05	2.92294E−05
10^{-6}	71	0	9.02824E−06	2.12659E−05
10^{-8}	166	2	1.14939E−06	2.13569E−06
10^{-10}	236	4	4.56047E−08	5.25142E−08

Table 8 Numerical results for Example 6

TOL	STEP	FS	AVERR	MAXE
10^{-2}	70	0	1.22749E−02	4.54860E−02
10^{-4}	97	0	3.92866E−04	1.14032E−03
10^{-6}	118	0	1.61615E−06	4.86641E−06
10^{-8}	173	0	2.72657E−07	4.87304E−07
10^{-10}	300	3	1.72160E−08	3.97650E−08

Table 9 Numerical results for Example 7

TOL	STEP	FS	AVERR	MAXE
10^{-2}	21	0	9.06905E−08	8.55589E−07
10^{-4}	27	0	1.68103E−08	3.01837E−07
10^{-6}	34	0	7.73475E−10	1.19985E−08
10^{-8}	43	0	4.71906E−11	5.91029E−10
10^{-10}	60	0	1.41839E−12	9.05609E−12

Table 10 Numerical results for Example 8

TOL	STEP	FS	AVERR	MAXE
10^{-2}	22	0	3.41065E−07	3.95940E−06
10^{-4}	29	0	1.77364E−08	1.88569E−07
10^{-6}	36	0	9.95212E−10	1.52315E−08
10^{-8}	48	0	4.20781E−11	3.12677E−10
10^{-10}	72	0	9.49656E−13	3.90998E−12

STEP—the total number of steps taken, FS—the number of failed steps, AVERR—the average error, and MAXE—the maximum error. The notation 7.02683E−01 means 7.02683×10^{-1}.

From Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, it is observed that for the given tolerances, the two-point block method achieves the desired accuracy. When the tolerance becomes smaller, the total number of steps increases. In order to achieve the desired accuracy, smaller stepsizes are taken, thus resulting in the increase number of total steps taken.

5 Conclusion and Future Work

In this paper, we have discussed the development of a two-point block method for solving functional differential equations of delay and neutral delay-types. The block method produces two approximate solutions in a single integration step by using the same back values. The algorithm is implemented in variable stepsize technique where the coefficients for the various stepsizes are stored at the

beginning of the program for greater efficiency. Stability regions for a general linear test equation are obtained for a fixed, but variable stepsizes. The numerical results indicate that the two-point block method achieves the desired accuracy as efficiently as possible.

In the future, the focus for the research should include the implementation of the block method on parallel machines. The efficiency of the block method can be fully utilized if the computation for each point can be divided among parallel tasks.

References

1. R.D. Driver, *Ordinary and Delay Differential Equations* (Springer, New York, 1977)
2. J.R. Ockendon, A.B. Taylor, The dynamics of a current collection system for an electric locomotive. Proc. R. Soc. Lond., Ser. A **322**, 447–468 (1971)
3. A. Iserles, On the generalized pantograph functional differential equation. Eur. J. Appl. Math. **4**, 1–38 (1992)
4. F. Ishak, M.B. Suleiman, Z.A. Majid, Block method for solving pantograph-type functional differential equations, in Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2013, WCE 2013, 3–5 July 2013, London, UK, pp. 948–952
5. D.J. Evans, K.R. Raslan, The Adomian decomposition method for solving delay differential equation. Int. J. Comput. Math. **82**(1), 49–54 (2005)
6. W.S. Wang, S.F. Li, On the one-leg θ-methods for solving nonlinear neutral functional differential equations. Appl. Math. Comput. **193**, 285–301 (2007)
7. I. Ali, H. Brunner, T. Tang, A spectral method pantograph-type delay differential equations and its convergence analysis. J. Comput. Math. **27**(2–3), 254–265 (2009)
8. S. Sedaghat, Y. Ordokhani, M. Dehghan, Numerical solution of the delay differential equations of pantograph type via Chebyshev polynomials. Commun. Nonlinear Sci. Numer. Simul. **137**, 4815–4830 (2012)
9. F. Samat, F. Ismail, M. Suleiman, Phase fitted and amplification fitted hybrid methods for solving second order ordinary differential equations. IAENG Int. J. Appl. Math. **43**(3), 95–105 (2013)
10. C. Yang, J. Hou, Numerical method for solving volterra integral equations with a convolution kernel. IAENG Int. J. Appl. Math. **43**(4), 185–189 (2013)
11. K.M. Hsiao, W.Y. Lin, F. Fuji, Free vibration analysis of rotating Euler beam by finite element method. Eng. Lett. **20**(3), 253–258 (2012)
12. T.A. Anake, D.O. Awoyemi, A.O. Adesanya, One-step implicit hybrid block method for the direct solution of general second order ordinary differential equations. IAENG Int. J. Appl. Math. **42**(4), 224–228 (2012)
13. Z.A. Majid, Parallel block methods for solving ordinary differential equations. Ph.D. thesis, Universiti Putra Malaysia (2004)
14. F. Ishak, M. Suleiman, Z. Omar, Two-point predictor-corrector block method for solving delay differential equations. Matematika **24**(2), 131–140 (2008)
15. F. Ishak, Z.A. Majid, M. Suleiman, Two-point block method in variable stepsize technique for solving delay differential equations. J. Mater. Sci. Eng. **4**(12), 86–90 (2010)
16. F. Ishak, Z.A. Majid, M.B. Suleiman, Development of implicit block method for solving delay differential equations, in Proceedings of the 14th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering, Malta, 2012, pp. 67–71
17. C.T.H. Baker, C.A.H. Paul, Computing stability regions-Runge-Kutta methods for delay differential equations. IMA J. Numer. Anal. **14**, 347–362 (1994)

18. A.N. Al-Mutib, Stability properties of numerical methods for solving delay differential equations. J. Comp. Appl. Math. **10**, 71–79 (1984)
19. C.A.H. Paul, A test set of functional differential equations, in Numerical Analysis Report No. 243, February 1994

Semi Supervised Under-Sampling: A Solution to the Class Imbalance Problem for Classification and Feature Selection

M. Mostafizur Rahman and Darryl N. Davis

Abstract Most medical datasets are not balanced in their class labels. Furthermore, in some cases it has been noticed that the given class labels do not accurately represent characteristics of the data record. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. This is because they aim to optimize the overall accuracy without considering the relative distribution of each class. The class imbalance problem can also affect the feature selection process. In this paper we propose a cluster based under-sampling technique that solves the class imbalance problem for our cardiovascular data. Data prepared using this technique shows significant better performance than existing methods. A feature selection framework for unbalanced data is also proposed in this paper. The research found that ReliefF can be used to select fewer attributes, with no degradation of subsequent classifier performance, for the data balanced by the proposed under-sampling method.

Keywords Class imbalance · Clustering · Over sampling · ReliefF · SMOTE · Under sampling

1 Introduction

A well balanced training dataset is very important for creating a good training set for the application of classifiers. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. They aim to optimize the overall accuracy without considering the relative

M. M. Rahman (✉) · D. N. Davis
Department of Computer Science, University of Hull, Kingston upon Hull, UK
e-mail: mmrbappy@gmail.com; M.M.Rahman@2009.hull.ac.uk

D. N. Davis
e-mail: D.N.Davis@hull.ac.uk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_44,

distribution of each class [1]. Typically real world data are usually imbalanced and it is one of the main causes for the decrease of generalization in machine learning algorithms [2]. Conventional learning algorithms do not take into account class imbalance; giving the same attention to a data record irrespective of whether it is from the majority class or the minority class. When the imbalance is massive, it is hard to build a good classifier using conventional learning algorithms [3]. Actually, the cost in mis-predicting minority classes is higher than that of the majority class for many class imbalance datasets; this is particularly so in medical datasets where high risk patients tend to be the minority class. Furthermore, in many cases the class labels do not accurately reflect the nature of a patient. Some patients die for some reason other than the target cause and some patients are alive by chance. Therefore, there is a need of a good sampling technique for such datasets where the target classes are not balanced and the given labels are not always appropriate.

Feature selection is the process of selecting a subset of relevant features for use in model construction. Feature selection is also useful as part of the data analysis process. However, the class imbalance problem can affect the feature selection process. Our research found that very little work is done in this area, to find and address the issues of class imbalance in feature selection. In the work of Al-Shahib et al. [4], the author used random under-sampling to balance their data for SVM classification. The author also used feature selection on balanced data. They found that SVM performed well on the balanced data, but the paper does not provide any analysis of the effect of the class imbalance problem in feature subset selection.

Sampling strategies have been used to overcome the class imbalance problem by either eliminating some data from the majority class (under-sampling) or adding some artificially generated or duplicated data to the minority class (over-sampling) [5]. Over-sampling techniques [6] increase the number of minority class members in the training set. The advantage of over-sampling is that no information from the original training set is lost since all members from the minority and majority classes are kept. However, the disadvantage is that the size of the training set is significantly increased [6]. Random over-sampling is the simplest approach to over-sampling, where members from the minority class are chosen at random; these randomly chosen members are then duplicated and added to the new training set [7]. Chawla et al. [6] proposed an over-sampling approach called SMOTE in which the minority class is over-sampled by creating "synthetic" examples rather than over-sampling with duplicated real data entries.

In summary, over-sampling may cause longer training time and over-fitting [8]. The alternative to over-sampling is under-sampling. If we do not consider the time taken to resample, under-sampling betters over-sampling in terms of time and memory complexity [1]. Drummond and Holte [9] showed that random under-sampling yields better minority prediction than random over-sampling. Under-sampling is a technique to reduce the number of samples in the majority class, where the size of the majority class sample is reduced from the original datasets to balance the class distribution. One simple method of under-sampling (random under-sampling) is to select a random subset of majority class samples and then combine

them with minority class sample as a training set [8]. Many researchers have proposed more advanced ways of under-sampling the majority class data [8, 10, 11].

In the rest of this paper we present a semi supervised cluster based under-sampling technique to balance cardiovascular data for classification and feature selection.

2 Cluster Based Under-Sampling

Down-sizing the majority class results in a loss of information that may result in overly general rules. In order to overcome this drawback of the under-sampling approach Yen and Lee [8] proposed cluster-based under-sampling. Their approach is to first cluster all the training samples into K clusters then choose appropriate training samples from the derived clusters. The main idea is that there are different clusters in a dataset, and each cluster seems to have distinct characteristics, their approach selects a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the derived cluster.

They first cluster the full data to K clusters. A suitable number (M) of majority class samples from each cluster are then selected by considering the ratio of the number of majority class samples to the number of minority class samples in the cluster. The number M is determined by Eq. 1. The M number of majority class samples are randomly chosen from each cluster. In the ith cluster ($1 \leq i \geq \mathrm{K}$) the $Size^{i}_{MA}$ will be

$$Size^{i}_{MA} = (\mathrm{m\ x}\ Size_{MI})\ \mathrm{x}\ \frac{Size^{i}_{MA}/Size^{i}_{MI}}{\sum_{i-1}^{K} Size^{i}_{MA}/Size^{i}_{MI}} \tag{1}$$

This approach may be suitable for datasets where class labels are confidently defined and truly reflect the properties of the labeled class record. But in some cases, especially for medical datasets, there is no guarantee that the class labels are truly reflecting the actual characteristics of that record [11].

Our approach to under-sampling is different to the approach of Yen and Lee [8]. The aim is not to derive a majority and minority class ratio of 1:1; but just to reduce the gap between the numbers of majority class samples to the numbers of minority class samples. As shown in Fig. 1, we first separated the data into two sets; one subset has all the majority class samples and the other subset has the entire minority class sample. Next we cluster the majority class samples to make $K(K > 1)$ subsets, where each cluster is considered to be one subset of the majority class. All the subsets of the majority class are separately combined with the minority class samples to make K different training data sets. The value for K is dependent on the data domain, in our implementation the final K value used was 3. All the combined datasets are classified with decision tree [12] and Fuzzy Unordered Rule Induction Algorithm [13]. We kept the datasets that gave the highest accuracy with the majority of the classifiers for further data mining processes.

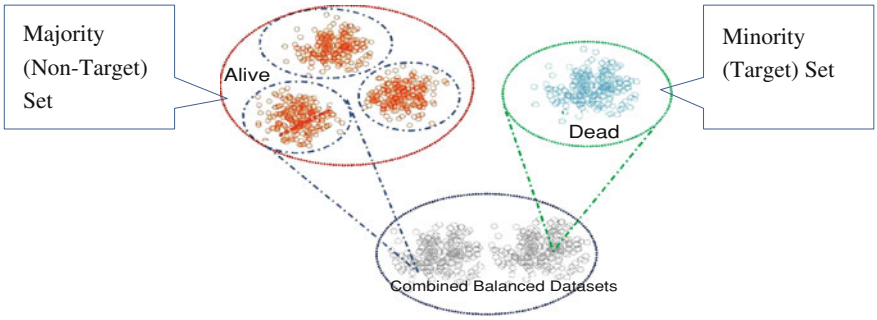


Fig. 1 Example custer based under-sampling

Table 1 The descriptions of the datasets

Data	Ratio	Description
D1:	2:1	Data consist of all the minority class samples ("dead") and one cluster of majority class records out of three clusters made by K-Mean
D2:	2.4:1	Data consist of combination of two clusters of the minority class samples and one cluster of majority class samples. Clusters are made with simple k-mean for both of the classes (K = 3)
D3:	3:1	Data consist of combination of all the minority class samples with randomly (random cut 1) selected samples from majority class sample
D4:	3:1	Data consist of combination of all the minority class samples with randomly (random cut 2) selected samples from majority class sample
D5:	6:1	Original data with full samples
D6:	1.8:1	Majority samples of the data set D2 are clustered into 3 cluster and each clusters are combined with the minority samples
K3M1Yen	1:1	Majority and minority ratio 1:1 (M = 1) using Yen and Lee [8]
K3M2Yen	2:1	Majority and minority ratio 2:1 (M = 2) using Yen and Lee [8]

For experiments we prepared several datasets presented in Table 1, using k-means clustering and classified using decision tree. The experimental outcomes are discussed in the result section.

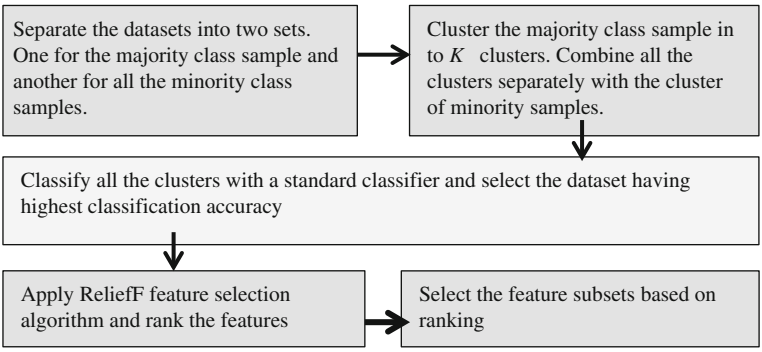


Fig. 2 Feature selection framework for unbalanced dataset

3 Feature Selection Framework for Unbalanced Data

A framework of feature selection for unbalanced clinical data sets is proposed in this research. The framework is based on k-means clustering and instance based feature selection algorithm. First the cluster based under-sampling is used to balance the datasets; later the ReliefF algorithm is used for feature ranking. The steps of the feature selection process are given in Fig. 2.

4 Experiments and Algorithms

We have used two cardiovascular datasets from Hull and Dundee clinical sites. k-means [14] clustering is used for under-sampling the minority class and ReliefF is used for feature selection. For choosing the best subset, we have used j48 decision tree [14] and Fuzzy Unordered Rule Induction Algorithm (FURIA) [15] as classifiers.

4.1 Overview of FURIA

Fuzzy Unordered Rule Induction Algorithm (FURIA) is a fuzzy rule-based classification method. Fuzzy rules are obtained through replacing intervals by fuzzy intervals with trapezoidal membership functions [15]:

$$\mu_{r^F}(x) = \prod_{i=1\ldots k} i_i^F(x_i) \tag{2}$$

For fuzzification of a single antecedent, only relevant training data D_T^i is considered and data are partitioned into two subsets and rule purity is used to measure the quality of the fuzzification [15]:

$$D_T^i = \{x = (x_1 \ldots x_k) \in D_T^i | I_j^F(x_j) > 0 \quad for\ all\ j \neq i\} \subseteq D_T \tag{3}$$

The fuzzy rules $r_1^{(j)} \ldots r_k^{(j)}$ are learnt for the class λ_{j}, the support of this class is defined by Maimon and Rokach [15]:

$$s_j(x) \stackrel{\mathrm{df}}{=} \sum_{i=1 \ldots k} \mu_{r_i^{(j)}}(x) \cdot CF\left(r_i^{(j)}\right) \tag{4}$$

where, the certainty factor of the rule is defined as

$$CF\left(r_i^{(j)}\right) = \frac{2\frac{|D_T^{(j)}|}{|D_T|} + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}}(x)}{2 + \sum_{x \in D_T} \mu_{r_i^{(j)}}(x)} \tag{5}$$

4.2 Decision Tree

The decision tree classifier is one of the most widely used supervised learning methods. A decision tree is expressed as a recursive partition of the instance space. It consists of a directed tree with a "root" node with no incoming edges and all the other nodes have exactly one incoming edge [14]. Decision trees models are commonly used in data mining to examine the data, and to induce the tree and its rules that will be used to make predictions.

Ross Quinlan introduced a decision tree algorithm (known as Iterative Dichotomiser (ID 3) in 1979. C4.5, as a successor of ID3, is the most widely-used decision tree algorithm [16]. The major advantage to the use of decision trees is the class-focused visualization of data. This visualization is useful in that it allows users to readily understand the overall structure of data in terms of which attribute mostly affects the class (the root node is always the most significant attribute to the class). Typically the goal is to find the optimal decision tree by minimizing the generalization error. The algorithms introduced by Quinlan [17] have proved to be an effective and popular method for finding a decision tree to express information contained implicitly in a data set. WEKA [18] makes use of an implementation of C4.5 algorithm called J48 which has been used for all of our experiments.

4.3 Relief: An Instance Base Approach to Feature Selection

Kira and Rendell [19] introduced an algorithm called Relief that uses instance based learning to assign a relevance weight to each feature. Relief is a simple yet efficient procedure to estimate the quality of attributes. The key idea of the Relief is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other. Given a randomly selected instance R_i from class L, Relief searches for k of its nearest neighbours from the same class called nearest hits H, and also k nearest neighbours from each of the different classes, called nearest misses M. It then updates the quality estimation W_i for the ith attribute based on their values for R_i, H, and M. If instance R_i and those in H have different values on the ith attribute, then the quality estimation W_i is decreased. On the other hand, if instance R_i and those in M have different values on the ith attribute, then W_i is increased.

$$W[A] := W[A] - diff(A, R_i, H)/m + diff(A, R_i, M)/m \tag{6}$$

Where A is the current attribute; $W[A]$ is the weight of the currently considered attribute; R_i is the ith sample; H is the "hit"; M is the "miss"; *diff*() is the probability function; and m is number of the neighbours.

The Relief algorithm is limited to classification problems with two classes. The ReliefF algorithm [20] is an extension of Relief algorithm that can deal with multi-class problems. ReliefF is a simple yet efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes. In practice, ReliefF is usually applied in data pre-processing as a feature subset selection method.

There are many other extensions of the Relief and ReliefF proposed by many researchers. Details about the algorithms and their application can be found in work of Robnik et al. [20].

Algorithm 1. ReliefF

Input:	For each training instance a vector of attribute values and the class value.
Output:	the vector W of estimations of the qualities of attributes
Step 1:	Set all weights $W[A]$: = 0.0;
Step 2:	**for** i := 1 to m **do begin**
	randomly select an instance R_i; find k nearest hits H_j;
Step 3:	for each class C 6 = class(Ri) do
	from class C find k nearest misses $M_j(C)$;
Step 4:	**for** A := 1 **to** a **do**
	$W[A] := W[A] - \sum_{j=1}^{K} diff(A, Ri, Hi)/(m-k) + \sum_{c \neq class(Ri)} \left[\frac{P(C)}{1 - P(class(Ri)} \sum_{j=1}^{K} diff(A, Ri, Mj(C)) \right] /(m-k);$
Step 5:	end;

4.4 Cardiovascular Data

We have used two datasets from Hull and Dundee clinical sites. The Hull site data includes 98 attributes and 498 cases of cardiovascular patients and the Dundee site data includes 57 attributes, and 341 cases from cardiovascular patients. After combining the data from the two sites, 23 matched attributes are left.

Missing values: After combining the data and removing redundant attributes we found that out of 23 attributes 18 attributes have a missing value frequency from 1 to 30 % and out of 832 records 613 records have 4–56 % missing values in their attributes.

From these two data sets, we prepared a combined dataset having 23 attributes with 823 records. Out of 823 records 605 records have missing values and 218 records do not have any missing values. Among all the records 120 patients are dead (High risk) and 703 patients are alive (Low risk). For this experiment according to clinical risk prediction model (CM1) [21], patients with status "Alive" are consider to be "Low Risk" and patients with status "Dead" are consider to be "High Risk".

4.5 Classifier Evaluation

The performance of the classification is evaluated by accuracy (ACC); sensitivity (Sen); specificity (Spec) rates, and the positive predicted value (PPV) and negative predicted value (NPV), based on values residing in a confusion matrix.

Assume that the cardiovascular classifier output set includes two typically risk prediction classes as: "*High risk*", and "*Low risk*". Each pattern x_i ($i = 1, 2 \ldots n$) is allocated into one element from the set (P, N) (positive or negative) of the risk prediction classes. Hence, each input pattern might be mapped into one of four possible outcomes: true positive—true high risk (TP)—when the outcome is correctly predicted as High risk; true negative—true low risk (TN)—when the outcome is correctly predicted as Low risk; false negative—false Low risk (FN)—when the outcome is incorrectly predicted as Low risk, when it is High risk (positive); or false positive—false high risk (FP)—when the outcome is incorrectly predicted as High risk, when it is Low risk (negative). The set of (P, N) and the predicted risk set can be built as a confusion matrix (Fig. 3).

The accuracy of a classifier is calculated by:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

		Predicted classes	
		High risk	Low risk
Expected or Actual classes	High risk	TP	FN
	Low risk	FP	TN

Fig. 3 Confusion matrix

The sensitivity is the rate of number correctly predicted "High risk" over the total number of correctly predicted "High risk" and incorrectly predicted "Low risk". It is given by:

$$Sen = \frac{TP}{TP + FN} \tag{8}$$

The specificity rate is the rate of correctly predicted "Low risk" over the total number of expected/actual "Low risk". It is given by:

$$Spec = \frac{TN}{TN + FP} \tag{9}$$

Higher accuracy does not always reflect a good classification outcome. For clinical data analysis it is important to evaluate the classifier based on how well the classifier predicts the "High Risk" patients. In many cases it has been found that the classification outcome is showing good accuracy as it can predict well the low risk patients (majority class) but failed to predict high risk patients (the minority class).

5 Results

We tried different methods in preparing a closely balanced datasets through clustering as outlined above. The method never runs with the aim of having class ratio 1:1. Our aim was to reduce the ratio gap between the majority and minority classes. The results are presented in Tables 2 and 3.

We made six datasets with different combinations of the clusters from majority and minority class samples and named as D1…D6, as described in Table 1. For exploring different alternatives we also tried to reduce further the ratio gap of majority class samples to minority class samples. In order to understand the quality of the training sample we also cluster the minority samples into three clusters and group them by different combinations with the clusters of majority class samples. An example of such a dataset is D2. We took the dataset D2 that has the best classification sensitivity among all the other datasets, we further cluster the majority class samples of D2 and select one cluster out of three clusters and combine with the minority class sample of the D2 and made another sample datasets called "D6".

Table 2 Classification outcome of FURIA

Data sets	ACC (%)	SEN (%)	SPEC (%)	PPV (%)	NPV (%)
D1	85.89	64.17	98.12	95.06	82.94
D2	92.11	79.78	97.21	92.21	92.07
D3	74.68	11.67	96.29	51.85	76.07
D4	70.82	15.83	89.52	33.93	75.78
D5	66.71	30.00	72.97	15.93	85.93
D6	96.39	91.01	99.38	98.78	95.21
K3M1Yen	61.48	67.50	55.65	59.56	63.89
K3M2Yen	60.39	22.50	79.66	36.00	66.90

Table 3 Classification outcome of decision tree

Data sets	ACC (%)	SEN (%)	SPEC (%)	PPV (%)	NPV (%)
D1	84.08	67.50	93.43	85.26	83.61
D2	92.05	83.15	95.77	89.16	93.15
D3	67.66	35.83	78.57	36.44	78.13
D4	66.60	33.33	77.90	33.90	77.46
D5	79.59	20.00	89.76	25.00	86.80
D6	97.59	93.26	100	100	96.39
K3M1Yen	51.64	52.50	50.81	50.81	52.50
K3M2Yen	59.55	39.17	69.92	39.83	69.33

We also made two more datasets using the under-sampling by clustering method proposed by Yen and Lee [8] The first dataset (K3M1Yen) was produced by separating the full data to 3 clusters and collected the majority class samples using Eq. 1 with the majority and minority ratio 1:1 (M = 1). The second dataset (K3M2Yen) was produced by separating the full data to 3 clusters and collected the majority class samples using Eq. 1 with the majority and minority ratio 2:1 (M = 2). The datasets are classified using J48 and FURIA and results are presented in Tables 2 and 3.

From the Tables 2 and 3 we can see that the original unbalanced dataset D5 has accuracy of 66.71 % with FURIA classification and 79.59 % with decision tree classification. But for both of the classifiers the sensitivity value is very poor (30 and 20 %). The accuracy is high because the classifier was able to classify the majority class (*Alive*) sample well (72.97 and 89.76 %) but failed to classify the minority. Dataset D1 where data are balanced by clustering the majority class samples and combining all the minority samples shows better classification outcome than the original unbalance data. With the FURIA and decision tree classification of the D1 dataset, we found the sensitivity value 64.2 % with the decision tree and 67.5 % with the FURIA. The classification outcome of the D1 is 2–3 times higher than the original datasets. The datasets prepared by the method proposed by Yen and Lee [8] could show some increase in the sensitivity value but the accuracy was dropped and overall performance was not good. Under-sampling

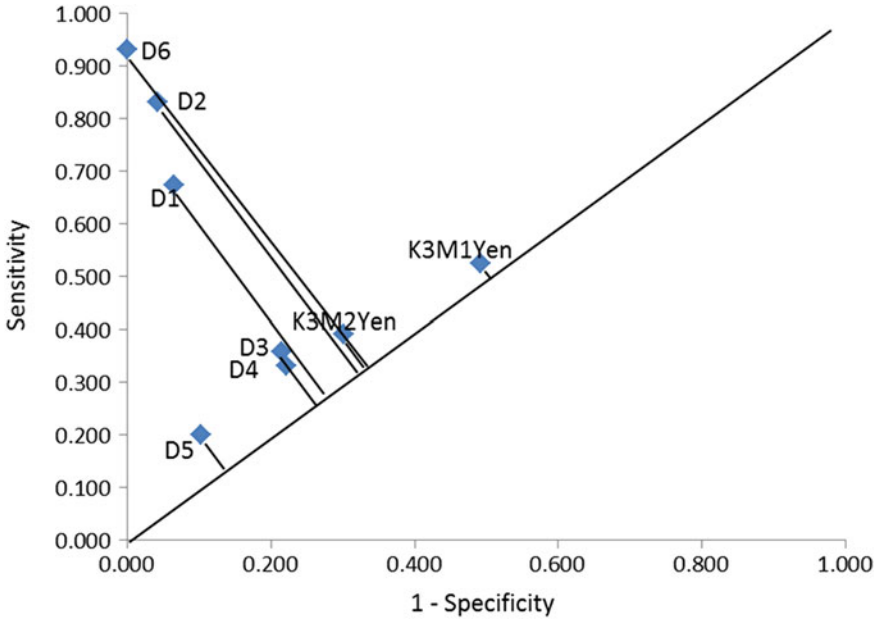


Fig. 4 ROC of decision tree classification

by random cut D3 and D4 also disappointed with its accuracy and sensitivity values.

It is observed from the experiments that the majority and minority ratio is not only the issue in building a good prediction model. There is also a need of good training simple, which should display the true properties of the class label assigned to them. As we discussed earlier, that for some data records, the class labels of clinical dataset do not accurately reflect the true properties of the class. The majority and minority ratio of D1, D2 and D6 are very close but the classification outcomes are not similar. Although the majority minority ratio is almost same, there is a big difference in the classification accuracy, sensitivity and specificity of D1 and D6, as can be noticed in Tables 2 and 3. The dataset "K3M1Yen" prepared by the method proposed by Yen and Lee [8] has 1:1 ratio but still has poorer classification outcome than other datasets.

If we analyse the ROC [22] space for all datasets classified with decision tree plotted in Fig. 4 and FURIA plotted in Fig. 5, we will find that overall accuracy of all the datasets are above the random line and the datasets D1, D2 and D6 which are prepared by our proposed method display the highest accuracy of all the datasets.

An experiment was made based on the proposed feature selection framework. The aim of the experiment was to observe the effect of the class imbalance problem on feature selection with ReliefF. ReliefF feature selection was applied to rank the unbalance data (D5) and data the balanced by the proposed method (D6), see Fig. 6.

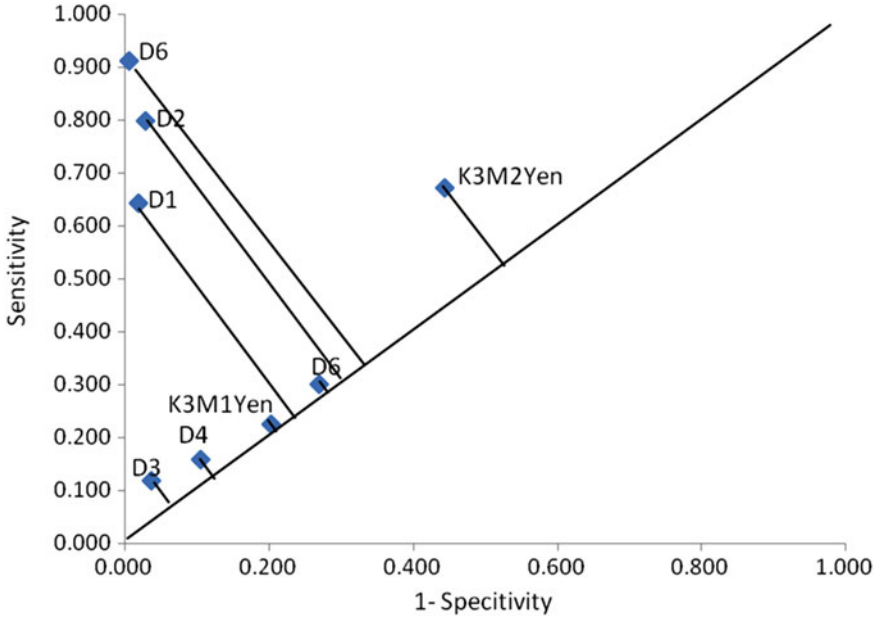


Fig. 5 ROC of FURIA classification

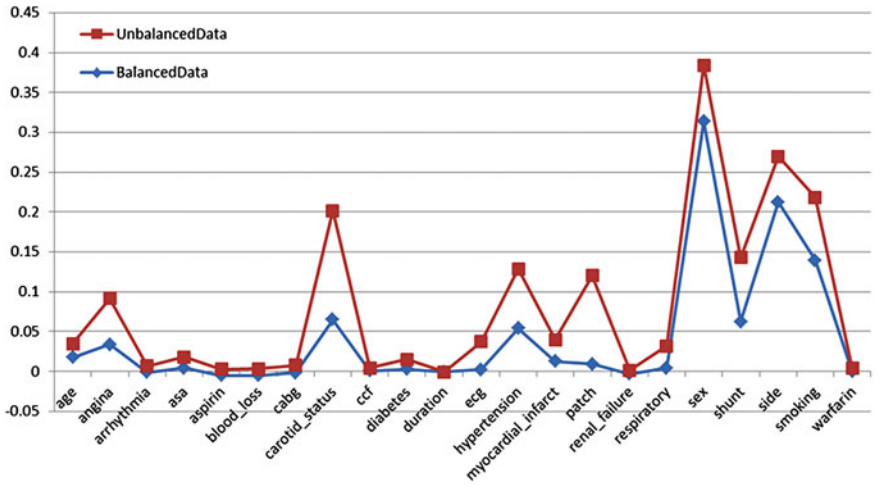


Fig. 6 Attribute ranking by ReliefF for the balanced and unbalanced data

The datasets were later classified by the decision tree classifier. From the experiment it is observed that, for the balanced data out of 23 attributes only 9 attribute were needed to keep the highest sensitivity of the data for decision tree

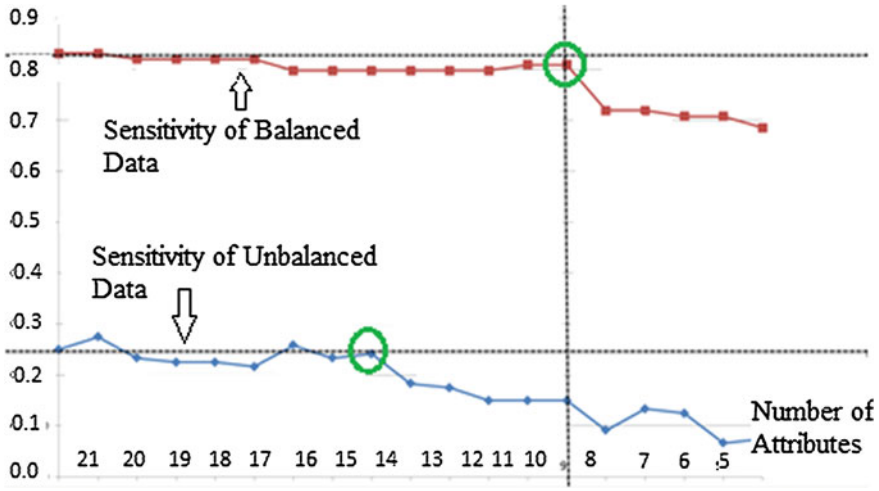


Fig. 7 Sensitivity value of unbalanced and balanced data, for different attribute subsets prepared by ReliefF attribute ranking

classification which is 40 % of the total attributes. On the other hand, for the unbalanced data out of 23 attributes, a minimum of 14 attributes were needed to keep the highest sensitivity of the data which is 63 % of the total attributes. The finding (see Fig. 7) shows that ReliefF can perform better with the data balanced by the proposed under-sampling method and can select a fewer number of attributes with no degradation in classifier performance.

6 Conclusion

Most medical datasets are not balanced in their class labels. In some cases it has been noticed that class labels do not represent a true property of the record. If we consider the cardiovascular risk based on dead or alive status of previous patients records, some of the patients may have died with some other cause and some are alive by chance. The proposed method is found to be useful for preparing unbalanced datasets where the given class labels are not always appropriate and fail to truly reflect the underlying characteristics of the patient record.

Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. Sampling strategies have been used to overcome the class imbalance problem by either over-sampling or under-sampling. Many researchers proposed different methods of under-sampling the majority class sample to balance the data. We proposed a cluster based under-sampling method that not only can balance the data but also can chose good quality training set data for building classification models.

The research found that class imbalance not only affects the classification process but also has effect on feature selection process. This is because most of the feature selection methods use the class label of the dataset to select the attribute subset.

In summary, we suggest the techniques used here are of benefit for problematic data and can help to alleviate the class imbalance problems typically found in clinical datasets and data from other domains.

Acknowledgments The authors gratefully acknowledge SEED Software in the Department of Computer Science of The University of Hull, UK, for funding this research project.

References

1. Y. Liu, X.H. Yu, J.X. Huang, A.J. An, Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf. Process. Manage. **47**, 617–631 (2011)
2. M.-S. Kim, An effective under-sampling method for class. Imbalance data problem, in Presented at the 8th International Symposium on Advance intelligent System (ISIS 2007), 2007
3. Z. Yan-Ping, Z. Li-Na, W. Yong-Cheng, Cluster-based majority under-sampling approaches for class imbalance learning, in 2010 2nd IEEE International Conference on Information and Financial Engineering (ICIFE), 2010, pp. 400–404
4. Al-Shahib, R. Breitling, D. Gilbert, Feature selection and the class imbalance problem in predicting protein function from sequence. Appl Bioinformatics **4**, 195–203 (2005)
5. R. Laza, R. Pavon, M. Reboiro-Jato, F. Fdez-Riverola, Evaluating the effect of unbalanced data in biomedical document classification. J. Integr. Bioinformatics **8**, 177 (2011)
6. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Y. Zhai, N. Ma, D. Ruan, B. An, An effective over-sampling method for imbalanced data sets classification. Chin. J. Electron. **20**, 489–494 (2011)
8. S.-J. Yen, Y.-S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. **36**, 5718–5727 (2009)
9. C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in Presented at the Workshop on Learning from Imbalanced Data Sets II, 2003
10. Y.-M. Chyi, Classification analysis techniques for skewed class distribution problems. Master, Department of Information Management, National Sun Yat-Sen University (2003)
11. M.M. Rahman, D.N. Davis, Cluster based under-sampling for unbalanced cardiovascular data, in Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013, London, 2013, pp. 1480–1485
12. R.C. Barros, M.P. Basgalupp, A.C.P.L.F. de Carvalho, A.A. Freitas, A survey of evolutionary algorithms for decision-tree induction. IEEE. Trans. Syst. Man Cybern. Part C: Appl. Rev. **42**, 291–312 (2012)
13. F. Lotte, A. Lecuyer, B. Arnaldi, FuRIA: an inverse solution based feature extraction algorithm using fuzzy set theory for brain-computer interfaces. IEEE Trans. Signal Process. **57**, 3253–3263 (2009)
14. O. Maimon, L. Rokach, *Data mining and knowledge discovery handbook* (Springer, Berlin, 2010)

15. F. Lotte, A. Lecuyer, B. Arnaldi, FuRIA: a novel feature extraction algorithm for brain-computer interfaces using inverse models and fuzzy regions of interest, in Presented at the 3rd International IEEE/EMBS Conference on Neural Engineering, CNE '07, 2007
16. I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.F. Chang et al., Data mining in healthcare and biomedicine: a survey of the literature. J. Med. Syst. **36**, 2431–2448 (2012)
17. R. Quinlan, *C4.5: programs for machine learning* (Morgan Kaufmann, San Mateo, 1993)
18. R.R. Bouckaert, E. Frank, M.A. Hall, G. Holmes, B. Pfahringer, P. Reutemann et al., WEKA-experiences with a java open-source project. J. Mach. Learn. Res. **11**, 2533–2541 (2010)
19. K. Kira, L.A. Rendell, A practical approach to feature selection, in Presented at the Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, United Kingdom, 1992
20. M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. **53**, 23–69 (2003)
21. D.N. Davis, T.T.T. Nguyen, Generating and verifying risk prediction models using data mining (A case study from cardiovascular medicine), in Presented at the European Society for Cardiovascular Surgery 57th Annual Congress of ESCVS, Barcelona Spain, 2008
22. T. C. W. Landgrebe, R. P. W. Duin, Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **30**(5), 810–822, (2008)

Claims-Based Authentication for an Enterprise that Uses Web Services

William R. Simpson and Coimbatore Chandersekaran

Abstract Authentication is the process of determining whether someone or something is, in fact, who or what they are declared to be. The authentication process uses credentials (claims) containing authentication information within one of many possible authentication protocols to establish the identities of the parties that wish to collaborate. Claims are representations that are provided by a trusted entity and can be verified and validated. Of the many authentication protocols, including self-attestation, username/password and presentation of credentials, only the latter can be treated as claims. This is a key aspect of our enterprise solution, in that all active entities (persons, machines, and services) are credentialed and the authentication is bi-lateral, that is, each entity makes a claim to the other entity in every communication session initiated. This paper describes authentication that uses the TLS protocols primarily since these are the dominant protocols above the transport layer on the Internet. Initial user authentication may be upgraded to multi-factor as discussed in the text. Other higher layer protocols, such as WS-Security, WS-Federation and WS-Trust, that use a Public Key Infrastructure credential for authentication, integrate via middleware. This authentication is claims based and is a part of an enterprise level security solution that has been piloted and is undergoing operational standup.

Keywords Authentication · Bi-lateral authentication · Claims-based identity · Enterprise processes · Multi-factor authentication · Public key infrastructure · Transport layer security · Web services

W. R. Simpson (✉) · C. Chandersekaran
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

C. Chandersekaran
e-mail: cchander@ida.org

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_45,

1 Introduction

This paper is based in part on a paper published by WCE [1]. Authentication is a system function that establishes a level of confidence in the truth of a claim (e.g., a user's identity or the source and integrity of data). The authentication process includes the presentation of one or more credential(s), validation of the credential(s)l, proof of the claimed binding, determination of authentication assurance level (includes multiple factors), and the completion of the authentication decision by the establishment of a communications channel with the entity claiming the identity. In this paper we discuss an aspect (authentication) of an enterprise process that has been developed as a part of an integrated security approach for the enterprise. This Enterprise Level Security (ELS) has been piloted and is currently undergoing stand-up throughout the enterprise.

2 Active Entities in the Enterprise Context

Entities within the enterprise environment may be active or passive. Passive entities include information packages, static files, and reference data structures. They are the target of activities. They do not initiate activities and cannot provide the role of requestor or provider. Active entities are those entities that change or modify passive entities, request or provide services, or participate in communication flows. Active entities are users, hardware, and services. All active entities in the enterprise have DoD certificates, and their private keys are stored in tamper-proof, threat-mitigating storage. Communication between active entities in the enterprise requires full bi-lateral, Public Key Infrastructure (PKI), end-to-end authentication [2–4].

Active entities must be named in accordance with the enterprise naming instruction [5]. Authentication in the enterprise environment is implemented as a verifiable claims-based attestation process. Figure 1 displays two active entities performing authentication and Active Entity B retrieving content from a passive entity.

3 Credentials in the Enterprise

A credential is a claim (in this case of identity) that can be verified as accurate and current. Credentials must be provided for all active entities that are established in the enterprise in order to perform authentication. Prior registration as an active entity with a confirmable entity name is required. The forms of credentials in use include certificates, Kerberos tickets, and hardware tokens. The details of generating, escrowing/retrieval, distributing, validating, and revoking certificates are

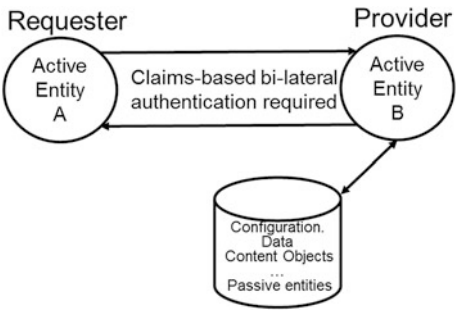


Fig. 1 Communication between active entities

discussed in specifications for Certificate Authority Certificates. Users are issued hardware tokens (Smart Cards) that have Certificate Authority Certificates stored on them with the private keys stored in hardware on the card. Machines and services are issued software certificates that contain the public key with the private key generated and remaining in hardware storage modules.

4 Authentication in the Enterprise

Authentication is responsible for establishing the identity of an entity. Authentication is achieved by receiving, validating, and verifying the identity credential(s). For certificates, validation is achieved by encrypting a message with the private key of the requester and transmitting it to the provider. The provider can then validate that it was sent by the requester by decrypting it with the requester's public key. This assures that the requester is the holder of the private key. Verification is achieved by verifying the trusted agent that issued the certificate, this authentication is two-way (the requestor authenticates the provider and the provider authenticates the requestor). In certain cases additional claims may be examined (multi-factor identification) including bio-metric measures.

4.1 Certificate Credentials

The required credential for enterprise personnel is an enterprise-issued X.509 (currently version 2.1), RSA-based certificate. X.509 certificates are used to bind an entity name to a public key in the PKI and to hold additional attributes (such as organizational unit data, and other data encoded in the distinguished name (DN)). They are used by authentication and authorization services, digital signing, and other cryptographic functions. Enterprise certificate credentials for users must be obtained through designated trusted Certificate Authorities (CAs). The CA provides the enterprise PKI credentials for users, devices, and services. Certificate credentials contain non-secret (publicly available) information. A hardware token

that contains the certificate is preferred to software-only certificates. For enterprise users, the method of credential storage is an enterprise-issued card with a highly secure tamper-proof hardware store, which is FIPS 140-2 level 2 validated for cryptographic tokens [3].

Software certificates (used in addition to hardware tokens) are in the PKCS#12 [2] formats and must be installed in certificate storage associated with the entity that owns the certificate or its host device (which must also be credentialed). A user may have a software certificate issued by a designated CA that is installed in certificate storage in the user's host device. For devices and services that are established in the enterprise, a software certificate is acquired from a designated CA and is installed in certificate storage on the device itself and on the host device. [For hardware elements outside the enterprise, PKCS#12 files may be maintained as backup offline—but, in general, should not be stored on the hardware device attached to the network.] The certificate credential for an entity must contain the enterprise-unique and persistent identifier in the certificate subject DN field; for users this is the extended common name; for devices and services this is the Universally Unique Identifier (UUID) in accordance with the enterprise naming standard.

4.2 Registration

The registration function is a service that creates and maintains the information about the identities of entities in the enterprise. There are three main issues to consider as discussed below:

4.2.1 Kerberos Tickets

Kerberos is a network authentication protocol originally developed by the Massachusetts Institute of Technology, and now documented in several Internet Engineering Task Force (IETF) Internet Drafts and RFCs [6]. Kerberos tickets are used with enterprise Active Directory (AD) forests.

4.2.2 Authentication and Attribute Assertion Tokens

Once authentication is established, the attributes of the identities are used to produce authorization claims. The primary method for expressing authorization claims in the enterprise uses derived credentials based on attribute assertion tokens at the message layer. These tokens contain security assertions and are obtained from a Security Token Service (STS). These tokens are based on the Security Assertion Markup Language (SAML) (current version) standard [7]. The use of SAML tokens in this context is discussed in [8]. Although the standard allows for

authentication elements in the SAML token, they are not used in this formulation. SAML is used only for authorization, and the only link to authentication is the binding to the requester by a holder-of-key (HOK) check (see [9] for the definition of this check and how it is performed).

4.2.3 Interoperability of Credentials

Public key cryptography depends on the ability to validate certificates against a trusted source. The use of PKI is discussed in [10]. External information sharing includes authentication based upon a federation agreement that specifies approved primary and derived credentials. The credentials will be configured for such federations.

4.3 Authentication

The enterprise supports two general methods for authentication: Kerberos-based and Direct PKI. Authentication relies on certificates.

4.3.1 Devices and Services Authentication PKI

Devices and Services are configured to authenticate themselves to the identity provider of the enterprise using bi-lateral Transport Layer Security (TLS) [11]. The authentication relies on enterprise-issued PKI certificates.

4.3.2 User Initial Authentication to the Domain

The user authenticates using the PKI-enabled logon program, which asks the user for a passcode that is, in turn, used as an index to a Kerberos key. This is a hybrid approach where the hardware token is read and user ownership is sought by presenting an input screen for the passcode associated with the hardware token. PKINIT is invoked, completing the authentication by PKI (Kerberos supports both password based user authentication and PKI based principal authentication with the PKINIT extension) using the certificate stored on the card. The Kerberos-based authentication uses the PKINIT and Kerberos protocols. For enterprise operations, users authenticate to the Identity Manager with the enterprise hardware token.

The hardware token credential is only used by human users, and either soft certificates or certificates stored in hardware storage modules are used for other entities. The user authenticates to the domain controller using a smartcard logon program such as the CAC or another approved active card and authenticates using the hardware token and a user-supplied Passcode. Multi-factor authentication is

not currently implemented, but may be used at this point. Biometric measures will add an increase in the strength of authentication. The PKI Initiation program is invoked completing the authentication by PKI. External users (users communicating from outside the enterprise) are then provided a virtual private network (VPN) tunnel and treated as if they were within the domain. Kerberos supports both password-based user authentication and PKI-based principal authentication (with the PKINIT extension), however, the enterprise uses only PKI-based principal authentication. Successful completion of the logon procedure signifies successful authentication of the user to the domain controller (a timeout will occur at pre-configured period more details are provided in [2]).

4.3.3 User Authentication to Services Using PKI

It is assumed at this point that the user has successfully authenticated to the Identity Manager using PKI. If the user wishes to access any other web service through the web browser, he does so using HTTPS. All entity drivers will be configured to use TLS mutual authentication. This additionally provides Transport Layer Confidentiality compression, and integrity (through message authentication) for subsequent message layer traffic over https. This validates the user's certificate and passes the certificate to the web service being accessed.

4.3.4 Service-to-Service Authentications

Requesters make requests for capabilities from Web services. In all cases, any capability request is preceded by TLS mutual authentication. Services may request other web services for capabilities (service providers). Services may include web services, utility services, and others.

5 Infrastructure Security Component Interactions

Figure 2 shows the basic authentication flows required prior to all interactions. This flow is the basic TLS setup.

When a requester wishes to use another service, there are four active entities that come into play. Details are provided in Figs. 3 and 4. The active entities are listed below.

1.(a) For a user:

The user (Requester) web browser—This is a standard web browser that can use the HTTP and HTTPS drivers (including the TLS driver) on the platform.

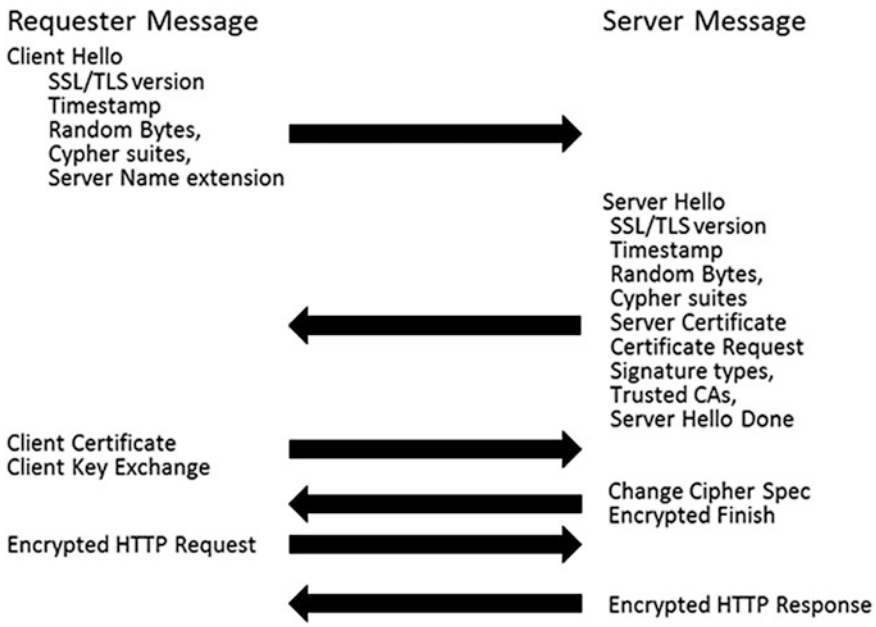


Fig. 2 Authentication flows

(b) For a service:

The requester host platform.

2. The Security Token Service (STS) in the requester's domain.
3. The Enterprise Attribute Store.
4. The requested service (application server) in the resource application environment.

5.1 Interactions Triggered by a User Request for Service

The user first makes a request to STS. Included in that request is an identifier (the Uniform Resource Identifier (URI) [12]) or a token referring to this identifier of the target service. The STS will generate the SAML credentials and return them to the browser with instructions to redirect to the service and post the SAML in this request to the application server (see Fig. 3). If HTTPS messages are used, then bi-lateral authentication takes place based on configuration of the servers and the web browsers.

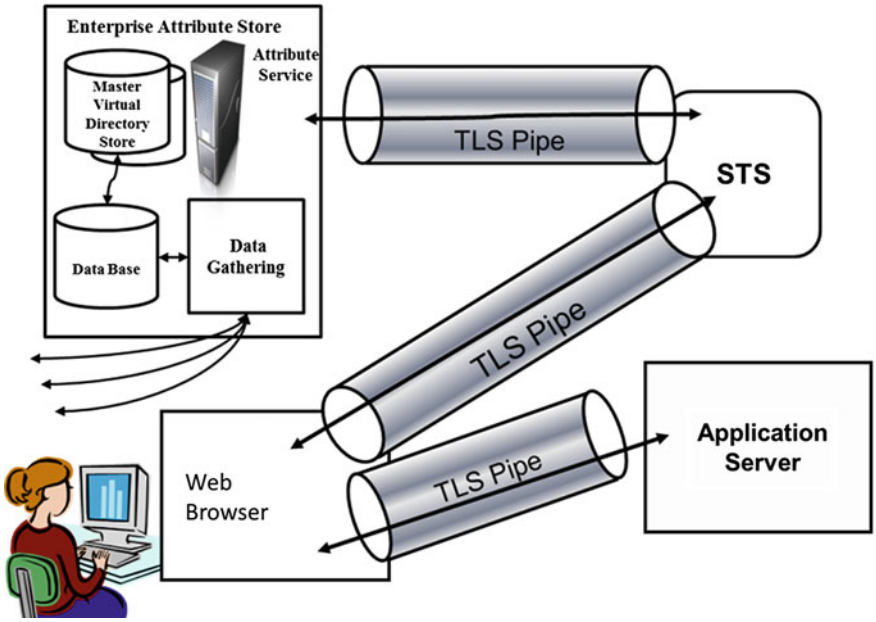


Fig. 3 Web browser request for service message flows

5.2 Interactions Triggered by a Service Request

This is similar to the flow in Fig. 3 except that instead of a browser requesting a service, another service formulates and makes the request. Note that authentication has nothing to do with Authorization which is performed via a SAML token. Authentication tokens are issued to ALL active entities and are called Identity Certificates.

The web application or service (on application server 1) will send a service request to the web service (on application server 2) as shown in Fig. 4. All communications shown in this figure are preceded by a bi-lateral authentication triggered by an HTTPS message.

6 Compliance Testing

Authentication testing verifies that the bi-lateral PKI-based authentication is working properly in the enterprise. This includes testing TLS on every connection in the security flows. Packet captures are done on nodes in the flow and then TLS traffic is checked for certificate exchanges and encryption. Checks for OCSP (the Online Certificate Status Protocol (OCSP) is an Internet protocol used for obtaining the revocation status of an X.509 digital certificate). Calls and returns

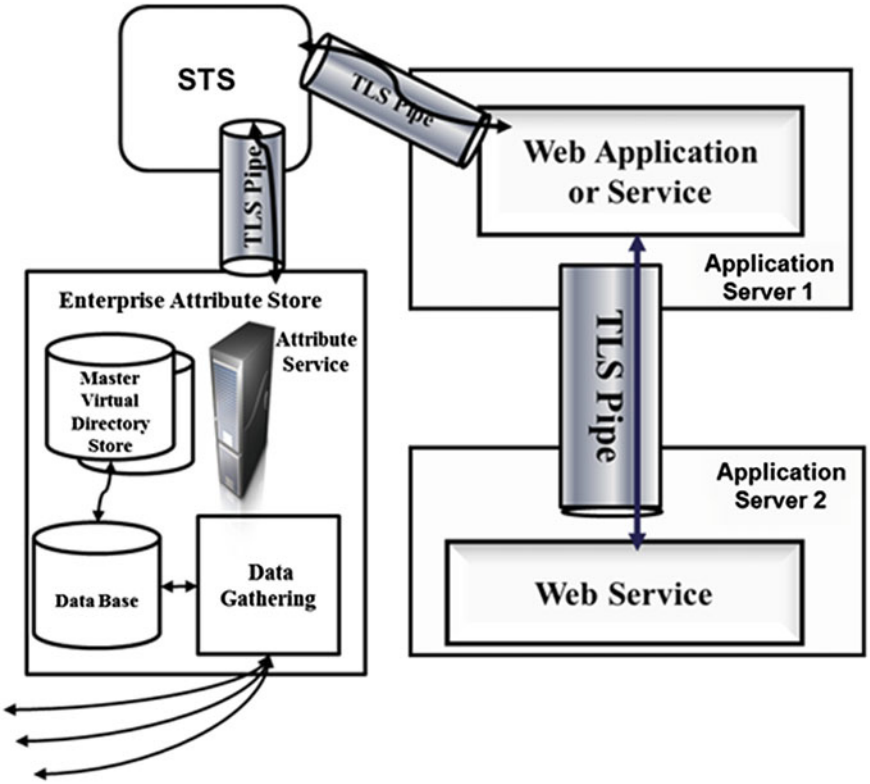


Fig. 4 Web service request for service message flows

verify that certificate status is being checked correctly. The packet captures are executed for a request to the STS. Authentication testing covers revoked and expired certificates as well as certificates that have been modified or tampered. Captures will show OCSP traffic for the revoked certificate.

7 Federated Authentication

Federated communications must meet all of the enterprise requirements, including:

- Naming PKI certificates,
- Certificates issued by a recognized certificate issuer,
- Valid, not-revoked, dates,
- TLS mutual authentication,
- Multifactor authentication as required, and
- SAML tokens from designated authorized STSs that meet all of the above requirements.

The federation partner must present a PKI certificate that meets enterprise requirements as described below and is issued by a trusted certificate authority. Trust is signified by including the certificate authority in the trust list. When required, enforcement of multi-factor authentication will be undertaken at this point. In cases where a trusted certificate authority cannot be found, the federation partner must be issued an enterprise PKI certificate and be included in the enterprise attribute stores. Details of this store are given in [13].

7.1 Naming and Identity

Identity is established by the enterprise or the requesting agency as agreed to in the federation agreement. In the enterprise, this is primarily through the enterprise naming contained in the enterprise-issued X.509. These names should be standardized throughout the enterprise and satisfy the property of uniqueness over space and time. For people, this name is the enterprise standardized name, but for other certificate authorities, their naming schemes are accepted based on federation agreements. The identity used by all federated exchanges is the Distinguished Name as it appears on the primary credential provided by the certificate authority. If there is a collision, mapping of federation names will be required.

Credentials are an integral part of the federation model. Each identity requiring access is credentialed by a trusted credentialing authority. Further, the STS used for generating SAML [7, 14–21] tokens, is also credentialed (as are all active entities in the enterprise). The primary exchange medium for setting up authentication of identities and setting up cryptographic flows is the PKI embodied in an X.509 certificate. The certificate authority must use known and registered (or in specific cases defined) certificate revocation and currency-checking software.

7.2 Translation of Claims or Identities

Identities are translated as indicated in the federation agreement. For simple federation, where requests are across the enterprise domains, there is no mapping, as the identities are in the appropriate form already. In any event, the mappings will be rare if distinguished names are used, and will only be needed when anonymity is a requirement or a collision occurs between the names provide by designated certificate authorities.

7.3 Data Requirements

Configuration files are developed and maintained as specified in enterprise requirements.

All configuration files and stored data are appropriately protected using cryptographic services. Even though these files are distributed for proximity to the relevant service, they are centrally maintained by an appropriate service agent mechanism.

7.4 Other Issues

All code that is generated is subject to code assurance review/tools with risks identified and resolved.

WS-Reliable Messaging [22], WS-Secure Conversation [23] shall be used for communication between active entities. The selection of either WS-Reliable Messaging or WS-Secure Conversation shall be based on session efficiency.

The registering of recognized STS and claim mapping must be promulgated in an enterprise policy memorandum after ratification of the federation agreement. The federation agreement may be an attachment to such a policy memorandum. This memorandum must be distributed to the appropriate organization for implementation by the Enterprise Attribute Store (EAS) and STS Administrators for incorporation in the trusted STS store. This maintains the lines of authority. A more complete discussion of federation, including a sample federation agreement is provided in [24].

8 Maturing Guidance

Related changes to OASIS, W3C, and IETF standards will necessarily be cause to reconsider and possibly modify these processes when appropriate. Because these standards tend to be backward compatible, or allow appropriate sunset periods, it can be assumed that phased-in implementation of changes will take place. Authentication for HTTPS protocol using Representation State Transfer (REST), Asynchronous JavaScript and XML (AJAX) /JavaScript Object Notation (JSON) has not yet evolved to a standard and is to be addressed in future revisions. As such, these and other technologies (e.g., wireless authentication, biometrics) are being pursued.

9 Summary

We have presented an authentication process for identity management and bi-lateral authentication between requesters and providers in an enterprise environment. The enterprise environment providers include web application and web services. The authentication is the beginning of a claims-based process that will

include SAML claims for authorization and may include multiple factors for identity verification. The authentication process is part of an enterprise solution and architecture for high assurance that is web-service based and driven by commercial standards. Portions of this architecture are described in Refs. [8, 9, 13, 24–35].

References

1. W.R. Simpson, C. Chandersekaran, in WCE 2013: *Claims-Based Authentication for a Web-Based Enterprise*. Proceedings World Congress on Engineering, London, July 2013. Lecture Notes in Engineering and Computer Science (3–5 July 2013), pp. 524–529
2. Public Key Cryptography Standard, PKCS #1 v2.1: RSA Cryptography Standard, RSA Laboratories, 14 June 2002
3. FIPS PUB 140, Security Requirements for Cryptographic Modules. National Institute of Standards, Gaithersburg, Maryland, 25 May 2001
4. Internet Engineering Task Force (IETF) Standards. RFC 2459: "Internet X.509 Public Key Infrastructure Certificate and CRL Profile", January 1999
5. Standard for Naming Active Entities on DoD IT Networks, Version 3.5 (or current), 23 September 2010
6. Internet Engineering Task Force (IETF) Standards. RFC 4120: The Kerberos Network Authentication Service V5), updated by RFC 4537 and 5021
7. S. Cantor et al. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005
8. C. Chandersekaran, W.R. Simpson, in *IMETI2010*: *A SAML Framework for Delegation, Attribution and Least Privilege*. The 3rd International Multi-Conference on Engineering and Technological Innovation, Orlando, FL, July 2010, vol. 2, pp. 303–308
9. W.R. Simpson, C. Chandersekaran, in *IMETI2010*: *Use Case Based Access Control*. The 3rd International Multi-Conference on Engineering and Technological Innovation, Orlando, FL, July 2010,vol. 2, pp. 297–302
10. FPKI-Prof Federal PKI X.509 Certificate and CRL Extensions Profile, Version 6, 12 October 2005
11. Internet Engineering Task Force (IETF) Standards. RFC 5246: "The Transport Layer Security (TLS) Protocol Version 1.2", August 2008
12. Internet Engineering Task Force (IETF) Standards. STD 66 (RFC3986) Uniform Resource Identifier (URI): Generic Syntax, T. Berners-Lee, R. Fielding, L. Masinter, January 2005
13. C. Chandersekaran, W.R. Simpson, in *WCE 2012*: *Claims-Based Enterprise-Wide Access Control*. Proceedings World Congress on Engineering 2012, 30 June–July 2012, London. Lecture Notes in Engineering and Computer Science, pp. 524–529
14. N. Ragouzis et al., Security Assertion Markup Language (SAML) V2.0 Technical Overview. OASIS Committee Draft, March 2008
15. P. Madsen et al., SAML V2.0 Executive Overview. OASIS Committee Draft, Apr 2005
16. P. Mishra et al. Conformance Requirements for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005
17. S. Cantor et al. Bindings for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005
18. S. Cantor et al. Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005
19. S. Cantor et al. Metadata for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005
20. F. Hirsch et al. Security and Privacy Considerations for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005

21. J. Hodges et al. Glossary for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, March 2005
22. WS-ReliableMessaging Specification 1.2. http://docs.oasis-open.org/ws-rx/wsrm/200702 OASIS, 2 February 2009
23. WS-SecureConversation Specification 1.4. http://docs.oasis-open.org/ws-sx/ws-secureconversation/200512 OASIS, February 2009
24. W.R. Simpson, C. Chandersekaran, in *IMETI2009*: *Information Sharing and Federation*. The 2nd International Multi-Conference on Engineering and Technological Innovation, Orlando, FL, July 2009, vol. 1, pp. 300–305
25. W.R. Simpson, C. Chandersekaran, A. Trice, in *IMET 2008*: *Cross-Domain Solutions in an Era of Information Sharing*. The 1st International Multi-Conference on Engineering and Technological Innovation, Orlando, FL, June 2008, vol. 1, pp. 313–318
26. C. Chandersekaran, W.R. Simpson, in *W3C*: *The Case for Bi-lateral End-to-End Strong Authentication.* World Wide Web Consortium Workshop on Security Models for Device APIs, London, December 2008, pp. 4
27. C. Chandersekaran, W.R. Simpson, in *CCSIT-2011*: *A Model for Delegation Based on Authentication and Authorization*. The First International Conference on Computer Science and Information Technology. Lecture Notes in Computer Science (Springer, Heildleberg, 2011), 20 pp
28. W.R. Simpson, C. Chandersekaran, in *CCT2011*: *An Agent Based Monitoring System for Web Services*. The 16th International Command and Control Research and Technology Symposium, Orlando, FL, April 2011, vol. 2, pp. 84–89
29. W.R. Simpson, C. Chandersekaran, An agent-based web-services monitoring system. Int. J. Comput. Technol. Appl. **2**(9), 675–685 (2011)
30. W.R. Simpson, C. Chandersekaran, R. Wagner, in *WCECS 2011*: *High Assurance Challenges for Cloud Computing*. Proceedings of World Congress on Engineering and Computer Science 2011, San Francisco, October 2011. Lecture Notes in Engineering and Computer Science, pp. 61–66
31. W.R. Simpson, C. Chandersekaran, in *WCE 2012*: *Assured Content Delivery in the Enterprise.* Proceedings World Congress on Engineering 2012, 30 June–July 2012, London. Lecture Notes in Engineering and Computer Science, pp. 555–560
32. C. Chandersekaran, W.R. Simpson, in *Co-Existence of High Assurance and Cloud-Based Computing*. Book Chapter, IAENG Transactions on Engineering Technologies—Special Edition of the World Congress on Engineering and Computer Science 2011. Lecture Notes in Electrical Engineering 170. DOI: 10.1007/978-94-007-4786-9_16, ISBN: 978-94-007-4785-2, May 2012, Chap. 16 (Springer Science+Business Media, Dordrecht 2012) 14 pp
33. W.R. Simpson, C. Chandersekaran, in *WCECS2012*: *Enterprise High Assurance Scale-up*. Proceedings World Congress on Engineering and Computer Science 2012, 24–26 October 2012, San Francisco, USA. Lecture Notes in Engineering and Computer Science, pp. 54–59
34. C. Chandersekaran, W.R. Simpson, A uniform claims-based access control for the enterprise. Int. J. Sci. Comput. **6**(2), 1–23 (2012). ISSN: 0973-578X
35. C. Chandersekaran, W.R. Simpson, in *WCECS2013*: *Cryptography for a High-Assurance Web-Based Enterprise*. Proceedings World Congress on Engineering and Computer Science 2013, San Francisco, USA. Lecture Notes in Engineering and Computer Science, pp. 23–28

Multilevel Verification and User Recognition Strategy for E-mail Clients

Artan Luma, Bujar Raufi and Burim Ismaili

Abstract User authentication and identification have always represented a challenge in web-based e-mail systems. The text-based authentication and user identification are not sufficient to address the security issues facing web-based e-mail systems. This sort of security is completely retrograde and obsolete for current security threats that easily undermine authentication, identification and non-repudiation. In this paper, a security increase in e-mail client is proposed by introducing multiple-level authentication and identification in e-mail clients. The proposed multilevel authentication and identification consist of four levels, where level-1 is the text-based authentication, level-2 involves an image based authentication and finally level-3 and level-4 use a specific algorithm that exploits the powerful properties of two mathematical operators called *Pentors* and *UltraPentors* applied against the image in level-2.

Keywords Algorithms · Authorisation · Cryptography · Cryptosystems · E-mail security · User authentication

1 Introduction

User authentication and identification in e-mail clients have always represented a challenge in the Web. Email based user authentication and identification represent emerging techniques that appear as an alternative to the standard Public-Key-Infrastructure (PKI) and furthermore these approaches allow securing users from faulty

A. Luma (✉) · B. Raufi · B. Ismaili
South East European University, Ilindenska no. 335, 1200 Tetovo, Macedonia
e-mail: a.luma@seeu.edu.mk

B. Raufi
e-mail: b.raufi@seeu.edu.mk

B. Ismaili
e-mail: bi12858@seeu.edu.mk

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_46,

impersonations and identity thefts [1]. However, the authentication and identification process in the web has not changed over the last twenty years and is mainly based on password identification and cookies [2]. The report from Google on email account security indicates that in 2011 and 2012 there is an increase in Google account blocking as a result of account hijacking and identity thefts [3]. The most widely used authentication strategy represents the text-based password scheme where users enter their login names and passwords. Despite their popularity, textual passwords suffer from several drawbacks. Although simple and straightforward textual passwords are easy to remember and maintain, they are mostly vulnerable to attacks. While complex and arbitrary passwords render the system substantially more secure, resisting the brute force search and dictionary attacks, they are difficult to guard and memorize [4]. Another aspect that advances the textual authentication is the graphical authentication which (compared to words) is easier to remember. Accordingly, it is difficult to formulate and orchestrate attacks for graphical authentication considering that the password space of graphical authentication extends more than that of textual passwords and makes them harder to crack and brute force attack resistant. Still, graphical authentication suffers from the so called shoulder-surfing which represents a hazard of intruder scrutinizing passwords by recording user sessions or directly supervising the users [5]. Some other related work regarding multilevel authentication is elaborated in [6] where the authors propose 3-level authentication based on textual, image based and one-time password fashion. However this kind of approach does not involve image encryption and their safe storage for avoiding direct compromise of the data used for authentication and identification. Other work involve the definition of a strict authentication system by introducing a multi-level authentication technique that generates a password in multilevel instances for accessing and using cloud services inside of which, an e-mail cloud service can reside as well [7]. The main goal of this paper is to focus mainly on securing user authentication and identification by the use of specifically designed encryption algorithms which is applied on the image while identifying the e-mail client.

The rest of this paper is organized as follows: Sect. 2 introduces some related work regarding cryptographic algorithms used today, Sect. 3 elaborates the introduction on two mathematical operators called *Pentor* and *UltraPentor* and examines closely their properties, Sect. 4 introduces the multilevel authentication and user identification methodology by using the above mentioned mathematical operators, Sect. 5 gives a real life scenario of the above mentioned methodology with a use case, Sect. 6 elaborates some aspects of cryptosystem's strength and finally, Sect. 7 concludes this paper.

2 Related Work

Many encryption algorithms utilized today use proprietary methods to generate keys and therefore are useful for various applications. Here, we introduce details for some of these encryption algorithms. Strong side of these algorithms lies in the length of the key that is generated and used.

RSA algorithm [8, 9], for example, is based on the generation of two simple large numbers p and q, multiplied in the form $n = p \cdot q$. The algorithm also selects an encryption exponent e, as $gcd(e, (p-1) \cdot (q-1)) = 1$ and the pair (n, e) is sent to the recipient. The recipient on the other side will now generate cryptic message in the form $c \equiv m^e(modn)$. This encrypted message, then can be decrypted after the component d is found, which is easy considering that the sender has p and q from where it finds d as following $d \equiv 1(mod(p-1) \cdot (q-1))$. Decryption process is performed as $m \equiv c^d(modn)$. The problem is that the algorithm is based on the generation of large prime numbers which is time consuming and computationally intensive.

Another approach that belongs to Online authentication is the TEA (Tiny Encryption Algorithm) [10–12]. This algorithm generates random numbers that will be sent to users that request authentication. From this random number together with user's secret key, a ciphertext message is generated. After the server receives the encrypted message, it decrypts it using the random number sent earlier. The disadvantage of this approach is that the secret key is previously used for securing the communication line established between the user and the server rather than directly for authentication. Another aspect of this approach is that it is not clear in which way the secret key is sent or at least generated by the user.

In addition to the above mentioned, let us present two operators given as mathematical models called *Pentor* and *Ultra Pentor* [13]. These operators can easily be generated from any number and can be used for encryption purposes. The power of the proposed cryptosystem lies in the irreversibility trait that these two operators have during the authentication process. Once operators are generated, it is extremely difficult to find numbers from which these operators are derived. This irreversible feature is used to create online authentication scheme which uses exactly three steps of encryption algorithm. The power of the cryptosystem justifies the proposed approach for its potential application in online authentication systems. In the following section, their definitions and properties will be analyzed.

3 Cryptography with Pentor and Ultra Pentor Operators

In [14, 15] a mathematical definition for *Pentor* and *Ultra Pentor* is introduced. A *Pentor* of a number is given as an integer number with base n and for every natural and integer number n there exists one *Pentor* for the given base B. In order to represent this operator mathematically, we start from the modular equation for *Pentor* of an integer number n with base B that fulfills the condition $gcd(n, B) = 1$. From the aforementioned conditions the following was acquired [14]:

$$B^m P(n) \equiv 1(mod\, n) \tag{1}$$

where B represents the base of the integer number n, $P(n)$ is the *Pentor* of the integer number, whilst n and m represent the order of the *Pentor* for the given integer number. The modular expression (1) was transformed to the equality expression of the form:

$$B^m P(n) = 1 + nk \tag{2}$$

$$P(n) = \frac{1 + nk}{B^m} \tag{3}$$

where k is an integer number that fulfills the condition for the fraction to remain an integer number. For example if we want to find the *Pentor* of the first order than $m = 1$, the Pentor of the second order than $m = 2$ and so on [14].

Likewise, the *UltraPentor* of a number n with base B in which for every natural and integer number n there exists an *UltraPentor* for the given base B [14]. In order to represent this operator mathematically, we start from modular equation for *UltraPentor* of integer number n with base B that fulfills the condition $gcd(n, B) = 1$. Considering the above mentioned conditions, the modular equation for *UltraPentor* will look like:

$$B^m \equiv 1 (mod\, n) \tag{4}$$

where m is an integer number. The modular expression [1], was transformed to the equality expression by applying logarithmic operations on both sides and finding the *UltraPentor* as follows:

$$B^m = 1 + nl| \cdot \log_B \tag{5}$$

$$\log_B B^m = \log_B (1 + nl) \tag{6}$$

$$m \log_B B = \log_B (1 + nl) \tag{7}$$

where $\log_B = 1$ and there is:

$$m = \log_B (1 + nl) \tag{8}$$

If $m = UP(n)$ then *UltraPentor* of an integer number n with base B can be written as:

$$UP(n) = \log_B (1 + nl) \tag{9}$$

where l is an integer number that fulfills the condition for $(1 + nl)$ to be written as B^a, where a is also an integer number [15]. The power of the above mentioned operators lie in their properties of irreversibility of retrieving the *ID* from the *Pentor* or *UltraPentor* itself which in our designed cryptosystem is kept secret on the user's side.

4 Multi-level Authentication for E-mail Applications

Based on the properties of the above mentioned two mathematical operators, a web application for an email client can be designed. This web application will be able to send/receive two types of e-mails:

- Send/receive regular (non-authenticated and non-identified e-mail).
- Send/receive authenticated and identified e-mail where the source of the sender is verified.

The aim of this web application is to authenticate and identify users while sending e-mails. The milestone of this application lies in the power of the two mathematical operators (Pentor and Ultra Pentor) as well as in the so called "Pentoric Attack" procedure.

The "Pentoric attack" is based on the following procedure. If we take, for example, the value for $ID = 13$ and based on the above mentioned formulas for Pentor 1 and Ultra Pentor 4 we can retrieve values for $Pentor(ID) = 4$ and $UltraPentor(ID) = 6$. Furthermore, if we take a particular value for a username such as $Username = art$ and by converting into an ASCII code we receive $Username = 97114116$ out of which we receive a vector by multiplying with the value of ID as follows:

$$Vector = 97114116 \cdot 13$$
$$Vector = 1262483508$$

Considering that Vector is consisted of 10-digit sequence which is greater than the value of the *UltraPentor*, we divide the sequence into 6-digit chunks starting from the right as illustrated below:

$$Vector = 1262|483508$$

By summing these two values, a new vector is acquired as follows:

$$Vector = 1262 + 483508 = 484770$$

This represent a 6-digit sequence which is smaller or equal to the value of Ultra Pentor and the "Pentoric Attack" procedure can be applied in following way:

```
     4 8 4 7 7 0 <- 4
+            0
-------------
     4 8 4 7 7 <- 4
+        2 8
-----------
     4 8 7 5 <- 4
+      2 0
---------
     5 0 7 <- 4
+    2 8
-------
     7 8 <- 4
+ 3 2
-----
   3 9 <- 4
+ 3 6
-----
   3 9
```

From the example, it is seen clearly that the final value from the "Pentoric Attack" is $N = 39$ which should be fully divisible by the value of ID, i.e. 13|39. From here it can be certainly concluded that the vector originates from user with *Username* = *art* and *ID* = 13.

The logic of authentication and identification process for this application is based on the following procedure. The user that wishes to use the services of the application initially sends the credentials such as *Name*, *Surname*, *Username* and an *Image* of its choice. During the registration process, in the database values for the user such as *Name*, *Surname*, *Image* and *Vector* are stored. In the *Image* attribute, the location of the image is stored where initially the image is converted into an RGB matrix $M(R, G, B)$ and encrypted with the following formula:

$$M_c(R, G, B) \equiv M(R, G, B) \cdot ID \cdot UltraPentor(ID)$$

The *Vector* attribute on the other hand is received by converting the username characters to their respective ASCII counterparts and by multiplying with the user's ID.

$$R = Username_{ascii} \cdot ID$$

This value of R should be divided into sequence chunks the size of *UltraPentor* starting from the right side as follows:

$$R = R_n|R_{n-1}|\ldots|R_2|R_1$$

After each separation R_i, the divided chunks are added together $R = R_n + R_{n-1} + \cdots + R_2 + R_1$ and if the size of *Vector* is greater than the value of *UltraPentor*, the procedure is repeated until the length is less or equal to that of *UltraPentor*. After registration, a secret *ID* is sent to the client and now the user is ready to use the web application by performing a simple text-based authentication with *username* and *password* input which is validated against user credentials on the database. If the user exists and is registered, access is granted, on the contrary the user is rejected or is asked to register. After successful login, a user interface for sending/receiving e-mail is introduced. The characteristic of this e-mail client web application is the two types of emails that it supports. The first one is the regular non-authenticated e-mail and the second one is the authenticated e-mail sending. Of particular interest is the authenticated version of e-mails which is based on the *Pentor* and *UltraPentor* mathematical operators.

When the user composes an authenticated e-mail, it fulfills the *Messageto* box together with *Subject* field where he enters the subject of the e-mail. The user is also allowed to choose the *Emailtype* field for non-authenticated or authenticated e-mails. If an authenticated e-mail is chosen the user should provide its *Image* that it used while registering and the *ID* that was sent during registration phase. After completing the Message field the user tries to send the e-mail and during this phase a special authentication algorithm is used based on the above mentioned operators in the following order.

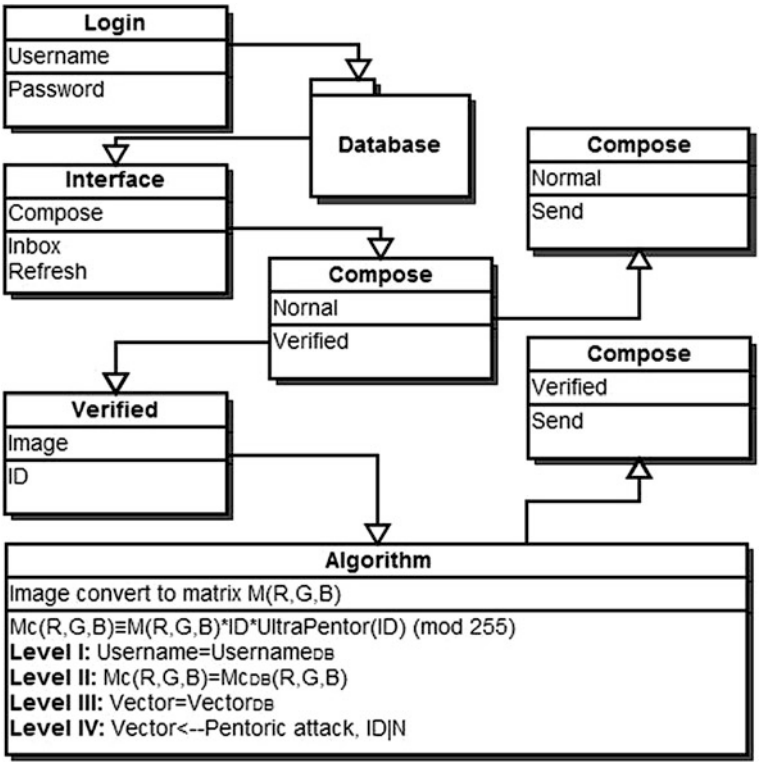


Fig. 1 Multi-level authentication with the use of Pentor and ultra Pentor operators

Initially, the *Image* provided is converted into an RGB matrix $M(R, G, B)$ and furthermore this matrix is encrypted in the following way:

$$M_c(R,G,B) \equiv M(R,G,B) \cdot ID \cdot$$
$$\cdot\, UltraPentor(ID) mod(255)$$

Further on, we introduce multiple level of authentication enumerated as below:

In level one, the *Username* provided by the client while initial sign-on is checked against user credentials in the database. Therefore,

$$Username = Username_{db}$$

The second level checks the encrypted matrix $M_c(R, G, B)$ generated on the fly while the user is authenticated against the encrypted matrix stored in the database.

$$M_c(R,G,B) = M_c db(R,G,B)$$

Fig. 2 Image provided by the user

The third level checks whether the *Vector* generated from the *Username* and *ID* of the client is equal with the *Vector* value stored in the database.

$$Vector = Vector_{db}$$

Finally, the fourth level of security is enforced by the condition $ID|N$, where the value of N is acquired by performing a "Pentoric Attack" against the *Vector* as follows:

$$N = Vector \leftarrow Pentor(ID)$$

The whole process with the levels of authentication is illustrated as in Fig. 1. After all these levels, the user is allowed to send an authenticated e-mail and the person receiving this e-mail will also receive a text attached at the end of the message stating that the message is authenticated with the above mentioned algorithm.

5 Multi-level E-mail Authentication: A Case Study

The above mentioned procedures and levels can be illustrated with a real example. Suppose that we would like to register a user with the following credentials:

$$Name = Artan$$
$$Surname = Luma$$
$$Username = a.luma@seeu.edu.mk$$

Image as given below in Fig. 2 based on this data, *Name*, *Surname*, *Username* and *Vector* will be stored in a database, where in the *Image* attribute the path of the encrypted *Image* will be stored. The process of encrypting the image as shown in the previous section is done by initially converting the image into a RGB matrix where each pixel is represented through RGB colors as given below with an image composed of 150 $\times$ 150 pixels.

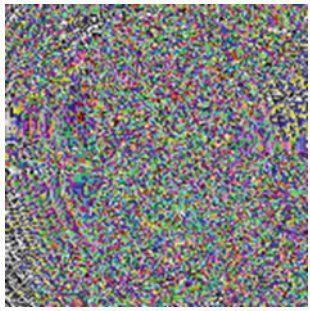


Fig. 3 Encrypted image of the user

$$M(R,G,B) = \begin{pmatrix} M_{1,1}(85,83,84) & \dots & M_{1,150}(88,86,87) \\ M_{2,1}(88,86,87) & \dots & M_{2,150}(41,40,35) \\ \vdots & \dots & \vdots \\ M_{150,1}(56,56,56) & \dots & M_{150,150}(70,52,68) \end{pmatrix}$$

This matrix is encrypted by having the ID of the client which is generated by the administrator and sent later to the user after the Vector generation. The encrypted matrix is acquired as follows:

$$M_c(R,G,B) \equiv M(R,G,B) \cdot ID \cdot \\ \cdot UltraPentor(ID) mod(255)$$

$$M_c(R,G,B) = \begin{pmatrix} M_{1,1}(0,99,177) & \dots & M_{1,150}(63,240,105) \\ M_{2,1}(234,78,156) & \dots & M_{2,150}(138,60,180) \\ \vdots & \dots & \vdots \\ M_{150,1}(111,111,111) & \dots & M_{150,150}(105,231,204) \end{pmatrix}$$

The encrypted matrix represents the encrypted Image depicted as in Fig. 3. This encrypted Image of the user is stored in a folder on the server's side of the web application. The calculation of the Vector is done by multiplying the ASCII version of the Username with the ID as shown in Sect. 3.

$$Vector = Username_{ascii} \cdot ID$$
$$Vector = 974610811710997641151011 0\dots \\ \dots 11174610110011746109107 \cdot 13$$
$$Vector = 126699405522429693349631431452 69\dots \\ \dots 931430152699418391$$

Considering that the result is a 50-digit sequence, it should be "chopped" into 6-digit sequences starting from the right side as follows:

$$Vector = 12|669940|552242|969334|963143|145269|\dots \\ |931430|152699|418391$$

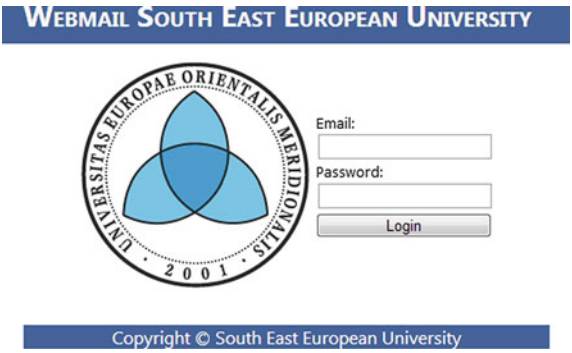


Fig. 4 Initial user login form

Fig. 5 User's mailbox after authentication

Summing these sequences altogether results in the value for *Vector* given as below:

$$\begin{aligned} Vector &= 12 + 669940 + 552242 + 969334 + 963143 + 145269 + \ldots \\ &\quad + 931430 + 152699 + 418391 \\ Vector &= 4802460 \end{aligned}$$

Considering that *Vector* is a 7-digit sequence which is greater than the value of Ultra Pentor, the process is repeated once again yielding the following result:

$$\begin{aligned} Vector &= 4|802460 \\ Vector &= 4 + 802460 \\ Vector &= 802464 \end{aligned}$$

This value of the Vector which is unique for each user is stored in a database and the value of the ID is sent to the user which is kept secret. The overall process of client activity on the web application consists of the following steps. At the beginning the user authenticates with a simple text-based authentication method by providing Username and Password (the Login step in Fig. 1). The Login form is depicted as in Fig. 4. After successful login, a window with user's mailbox appears where he can check for mails and compose new ones as illustrated in Fig. 5. If the user wants to send an email, by clicking in the Compose button a new form for email composing and sending appears. In this form the user has to fill and choose

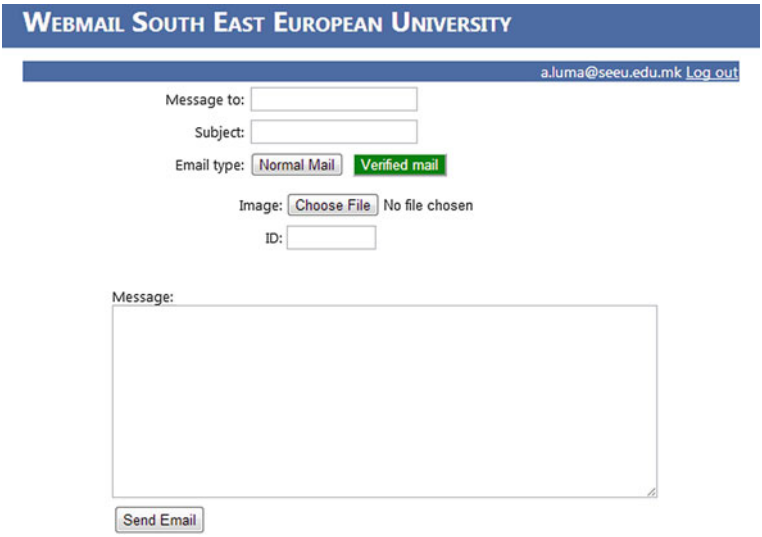


Fig. 6 Composition of verified e-mails

the following options: *Messageto*, *Subject*, *Emailtype* (normal, verified e-mails), *Image* (where user chooses its Image), *ID* and *Message*. This is illustrated in Fig. 6. When the Send Email button is clicked, the 4-level authentication process is initiated. At the beginning its image is converted into a matrix as given below:

$$M(R,G,B) = \begin{pmatrix} M_{1,1}(85,83,84) & \dots & M_{1,150}(88,86,87) \\ M_{2,1}(88,86,87) & \dots & M_{2,150}(41,40,35) \\ \vdots & \dots & \vdots \\ M_{150,1}(56,56,56) & \dots & M_{150,150}(70,52,68) \end{pmatrix}$$

The encryption process of this matrix is done in the following way:

$$M_c(R,G,B) \equiv M(R,G,B) \cdot ID \cdot \\ \cdot UltraPentor(ID) mod(255)$$

$$M_c(R,G,B) = \begin{pmatrix} M_{1,1}(0,99,177) & \dots & M_{1,150}(63,240,105) \\ M_{2,1}(234,78,156) & \dots & M_{2,150}(138,60,180) \\ \vdots & \dots & \vdots \\ M_{150,1}(111,111,111) & \dots & M_{150,150}(105,231,204) \end{pmatrix}$$

After the encryption, this matrix is stored as an encrypted image. In the first level, the *Username* given by the user and the one stored on the server's side is checked and if this *Username* is identical to *a.luma@seeu.edu.mk*, level two is initiated where on-the-fly encrypted matrix is checked against the encrypted matrix

stored in the database. If these two values coincide level three is introduced. In level three, the generated *Vector* from the user is checked against the *Vector* value stored in the database which after being identical with the one in the database, the level four kicks off in which "Pentoric Attack" against $Vector = 802464$ is performed. The attack results in value $N = 39$ which is divisible with Ultra Pentor, i.e. 13|39 and the a-mail is sent. In any other case, if one of the conditions would not be fulfilled the verified e-mail procedure would fail and the user rejected.

6 Strength of the Cyrptosystem

The power of the proposed cryptosystem lies in the fact that in order to break it's user credentials, all four authentication elements are required:

$$(Username, Password, Images, ID)$$

Assume that Person *B* will have access to the online system for sending electronic mail with verification to Person *A*. If in any way the intruder *C* acquires *Username* and *Password*, he still can not testify it's authentication, which as mentioned above, is created by the *Username*, *Password*, *Images* and *ID*. Even if it could create the *Vector* from *Username* and *Password*, this will fail later during the "Pentoric Attack" which ill not satisfy the requirement $ID|N$. It is worth noting that the algorithm is irreversible, which means that once values are transformed, it can not be traced back. For example, if intruder *C* founds the *Pentor* and *UltraPentor*, there is no easily computable way to find *ID*, as it is not known the sequence of *Pentor* and *UltraPentor*, and the values of k and l. Furthermore, it is concluded that different numbers can have the same values for *Pentor* and *UltraPentor*, but the distribution of these numbers is random, which makes it impossible to find the ID, even for system administrators, because the only person who knows ID is the user. The system generates *ID*, but does not preserve it in the internal database.

Finally, the power is substantially elevated in the third step. If the intruder *C* acquires *Pentor* and *UltraPentor*, he should perform a "Pentoric Attack" on the ciphertext generated by him, which, as explained above, from the beginning is faulty.

7 Conclusion and Future Work

In this paper we have introduced a method of user authentication and identification in the process of sending verified e-mails. The methodology uses a multi-level approach of identifying the user while sending verified e-mails and it consists of the following steps:

1. The first step represents a classical text-based authentication.
2. The second step involves an image based authentication where the user's image is encrypted by multiplying it with a secret key provided to the user.
3. Finally step three is concentrated around level-3 and level-4 elaborated earlier that use a specific algorithm that exploits the powerful properties of two mathematical operators called *Pentor* and *UltraPentor* applied against the image in step two.

Further research is needed on *Pentor* and *UltraPentor* properties and their application in the development of various cryptosystems [15]. One direction of this application would be the possibility of digitally signing the emails by using *Pentor* and *UltraPentor* and this is currently the focus of our further research.

References

1. S.L. Garfinkel, E-mail based authentication and identification: an alternative to PKI. IEEE Comput. Soc. **1**(6), 20–26 (2003)
2. M. Dietz, A. Czeskis, D.S. Wallach, D. Balfanz, Origin-bound certificates: a fresh approach to strong client authentication for the web, in *Proceedings of the 21st Usenix Security Symposium*, 2012
3. M. Hern, An update on our war against account hijackers. The Google Blog (2013). Available via GOOGLE Online Security. http://googleonlinesecurity.blogspot.com/2013/02/an-update-on-our-war-against-account.html of subordinate document. Cited 15 June 2013
4. S. Balaji, Authentication techniques for engendering session passwords with colors and text. Adv. Inf. Technol. Manage. **1**(2), 71–78 (2012)
5. H. Zhao, X. Li, S3PAS: a scalable shoulder-surfing resistant textual-graphical password authentication scheme, in *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW 07)* (2007), pp. 467–472
6. S. Anand, P. Jain, Nitin, R. Rastogi, Security analysis and implementation of 3-level security system using image based authentication, in *Computer Modelling and Simulation (UKSim)* (2012), pp. 547–552
7. H.A. Dinesha, V.K. Agrawal, Multi-level authentication technique for accessing cloud services, in *International Conference on Computing, Communication and Applications (ICCCA)* (2012), pp. 1–4
8. R. Rivest, A. Shamir, L. Adleman, A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
9. D. Boneh, Twenty years of attacks on the RSA cryptosystem. Not. Am. Math. Soc. **46**(2), 203–213 (1999)
10. D.J. Wheeler, R.M. Needham, TEA, a tiny encryption algorithm. Lecture Notes in Computer Science (LNCS) (Leuven, Belgium: Fast Software Encryption: Second International Workshop), vol. 1008 (1994), pp. 363–366
11. J. Kelsey, B. Schneier, D. Wagner, Related-key cryptanalysis of 3-WAY, Biham-DES, CAST, DES-X NewDES, RC2, and TEA. Lecture Notes in Computer Science (LNCS) vol. 1334 (1997), pp. 233–246
12. A. Bogdanov, M. Wang, Zero-correlation linear cryptanalysis with reduced data complexity. Lecture Notes in Computer Science (LNCS) (Fast Software Encryption 2012), vol. 7549 (2012), pp. 29–48

13. A. Luma, B. Ismaili, B. Raufi, Multilevel user authentication and identification scheme for e-mail clients. in *Proceedings of the world congress on engineering, WCE 2013*, 3–5 July 2013. Lecture notes in engineering and computer science, London, UK (2013), pp. 1221–1225
14. A. Luma, B. Raufi, New data encryption algorithm and its implementation for online user authentication, in *International Conference on Security and Management*, (CSREA Press, USA, 2009), pp. 81–85
15. A. Luma, B. Raufi, Xh Zenuni, Asymmetric encryption decryption with Pentor and ultra Pentor operators. Online J. Sci. Technol. (TOJSAT) **2**(2), 9–12 (2012)

Securing Information Sharing Through User Security Behavioral Profiling

Suchinthi A. Fernando and Takashi Yukawa

Abstract This paper presents a method of minimizing the human-related information security problem of improper sharing of information by insiders with outsiders or unauthorized insiders. As opposed to most currently available information security solutions, this system does not rely solely on technological security measures, but adapts a mixture of social and technological solutions. The system presented through this research detects users' observance of security best practices and behavioral patterns using both automatic and personal monitoring methods. It then creates user security behavioral profiles and thus identifies users who might potentially pose threats to the organization's information security and determines and schedules the level and type of security education and training to be given to identified users.

Keywords Human behavior · Information security · Insider threat · Profiling · Social · Technological

1 Background and Introduction

Despite the overall acknowledgement during the past decade that the human factor should be taken into consideration in information security management (ISM), most security solutions available today still rely on purely technical measures to

S. A. Fernando (✉)
Department of Information Science and Control Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan
e-mail: s095191@stn.nagaokaut.ac.jp

T. Yukawa
Department of Management and Information Systems Science, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan
e-mail: yukawa@vos.nagaokaut.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_47,

enforce information security. Most identified information security breaches occur because of human errors [1], resulting from the lack of proper knowledge and training, ignorance, and failure to follow procedures. People's beliefs and expectations may lead to mistakes and misjudgments of risks [2]. Thus, being the weakest link in the chain of security, people may unintentionally reveal confidential information to others. Schneier [3] explains how the perception of security diverges from its reality and how people feel secure as long as there is no visible threat. This human weakness is exploited in most present-day attacks, such as social engineering, spear phishing, and collusion from an insider, which require a human element to succeed [4].

The percentage of insiders wittingly or unwittingly involved in an attack originating from the inside is said to be at least 60–80 % [5, 6]. An insider threat is defined as trusted users with legitimate access abusing system privileges [7] or as intentionally disruptive, unethical, or illegal behavior by individuals possessing internal access to an organization's information assets [8]. Insider attacks are indistinguishable or difficult to distinguish from normal actions as insiders have authorization to access and use the system and these actions are less likely to differ from the norm [7]. Though difficult to detect until after the damage has been done, since most insider attacks are planned, there is a window of opportunity during which people can intervene before the attack occurs and prevent the attack or limit its damage [8].

Effective information security uses physical, technical, and operational controls, where operational controls concern the conduct of employees with regard to information security [9]. Even though information systems security auditing ensures that an organization's security policies, procedures, and regulations are effective, employees' adherence to these audited policies is automatically assumed [9]. Although somewhat sufficient to keep the outside attacks at bay, technical measures alone are clearly insufficient to ward off insider attacks, since people may easily bypass technological restrictions such as access control by revealing their authentication information to others. Sabett [10] states that security systems should be designed by accepting that the malicious attackers are already inside the system. A holistic approach blending people, process, and technology by focusing on behaviors and activities appearing to be risky using a combination of risk management, functional analysis of insider behaviors, and risk mitigation is recommended [8, 11].

Human behavior, which is performed according to the personality of the individual, can be categorized [9]. Observable behaviors include cyber activities, which only provide limited insight into intent and character, but are easier to collect, process, and correlate automatically, as well as personal conduct, which is observed through background checks and personal observations [8]. Employees may be divided based on their level of awareness of information security objectives, or according to job category, function, their knowledge about information processing, and technology used [12]. Accidents will not normally happen if security measures stay above a certain threshold and the risk is kept below the accident zone [13]. Perceived risk gradually declines when accidents do not occur

as a consequence of improved security, leading to a decline in the compliance with security measures until system becomes vulnerable again. Thus, risk perception renewals through properly scheduled interventions such as security awareness programs are needed to sustain an appropriate level of risk perception [13]. A proactive and sustainable security program requires: preventive (credentialing and restricting access through authorization), detective (auditing, monitoring, and referrals to validate allegation), corrective (additional monitoring or auditing, updating credentials, access restriction, or access removal), and feedback (dynamic, reactive, and planned feedback and creating and implementing solutions) components [14].

This research addresses the problem of improper sharing of information within an organization, and presents a solution by blending social and technological solutions to detect the levels of observance of security best practices by its employees by monitoring their cyber and non-cyber activities, detecting patterns among these behaviors, and using this information together with background information and job details to create security behavioral profiles of users, in order to identify users whose actions could potentially lead to ISM problems and therefore require special education and training in ISM.

2 Proposed System

The system proposed through this research to secure information sharing within an organization is explained briefly in this section. The detailed explanation of this system is available in [15].

Curtailing or limiting the personal browsing ability of employees is detrimental to their productivity [16]. Yet, depending on the projects they are working on and the criticality of the business information they have to access, it is sometimes mandatory to restrict access to the Internet in order to protect the security of the business information. This system addresses this problem by providing two separate modes: the "strict" mode, which is the default mode, and the "relaxed" mode. Only pre-specified, work-related programs and services are allowed during the "strict" mode, and all activities are monitored and logged, while personal browsing, e-mails, or instant messaging, etc. are disallowed, and all information exchanges (e-mail contents, attachments, file-sharing, etc.) are recorded.

During the "relaxed" mode, personal browsing, personal e-mails, instant messaging, etc. are allowed, and are not monitored to protect the user's privacy, while access to work-related information is disallowed. Figure 1 depicts the top-level architectural design of the system. This system constantly monitors for extraordinary behavior: excessive or untimely access to information, services, or systems, access from remote terminals, attempts to access data of a higher classification level than the user's security clearance level, or data for which the user has no Need-to-Know. Employees' observance of best practices is monitored

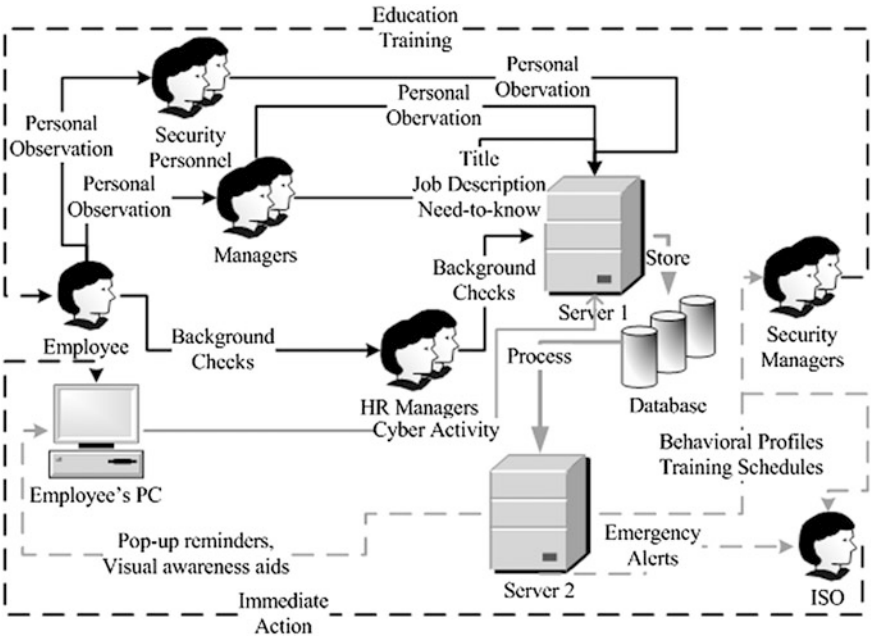


Fig. 1 Top-level architectural diagram

regularly in the areas of password security behavior, data backup behavior, data sanitization behavior, network security behavior, and physical security behavior.

Cyber activities such as password renewal frequency, reuse of former passwords, password strength, data backup frequency, etc., are automatically monitored by the system. Non-cyber activities such as whether confidential documents are left unattended, whether doors are locked, whether credentials are validated before revealing information to others, etc. are personally observed by managers or the security personnel of the organization. Information from background checks before employment and periodically during employment such as contact details, financial status and stability, number of dependents, educational level, criminal record, etc. are inputted to the system by human resource managers. Employees' job descriptions are updated by their managers according to the projects they are currently working on. Together, this information will be used for profiling and for detecting each employee's behavioral type.

The resulting security behavioral profiles will include the security consciousness of the employee, the extent of understanding and the value given to ISM rules and procedures, the extent of adherence to policies, how easily an employee can be enticed or tricked into revealing information, employee's ambitiousness and drive to move ahead in their career, sociability, capability to work in a team, and respect gained by peers, the employee's potential to intentionally or unintentionally reveal or improperly share confidential information, and whether the employee has any

motive or incentive (financial, career-wise, social, psychological or personal) to access unauthorized information or reveal information to others.

Based on these behavioral profiles, the system will identify potentially problematic employees and determine the level of security awareness, guidance, or training to be given: planned and scheduled awareness and training programs for identified potentially problematic users, and randomly scheduled awareness and training programs for all users, periodically, as risk perception renewals to maintain the desired level of security awareness. Real-time alerts will be sent to the information security officer (ISO) if extensively problematic behavior is detected, allowing the ISO to take immediate action. Security managers and the ISO can view behavioral profiles in summarized, detailed or graphical form, and view security training schedules. The ISO can additionally request separate views of personally inputted (non-cyber-activity-related) data and automatically monitored (cyber-activity-related) data and use his personal judgment to avoid any bias of managers or security personnel towards employees.

3 Profiling

For a better understanding of the security profiling techniques adapted in this system, it is important to explore those currently used in other prevailing areas in the field of security, such as in criminal investigations. Thus, an understanding about criminal profiling will provide insight into profiling techniques which may be adaptable to information security.

Criminal profiling is an investigative approach based on the premise that the crime scene provides details about offense and offender [17]. It is used in homicide, sexual assault, arson, etc. Criminal profiling is defined as the careful evaluation of physical evidence for systematically reconstructing the crime scene and developing a strategy to capture the offender, by weeding out suspects, developing investigative strategy, linking crimes and suspects, and assessing risk [18]. Assuming that every criminal works to a certain set of values, criminal profiling is used to classify behavioral patterns and predict the next move [19]. The developed offender description contains: psychological variables (personality traits, psychopathologies, and behavior patterns), and demographic variables (age, race, gender, emotional age, level of education, arrest and offense history, etc.) [20]. Criminal profiling uses geographic or psychological typologies to isolate offender characteristics [17]. The system presented through this research uses a psychologically-based technique, which compiles psychological background using observable behaviors of offender's traits. Behavior is interpreted from the presence or absence of forensic elements, offender's behavioral choices, modus operandi, signature behaviors, knowledge of crime scene's dynamics, etc. [17]. Inductive criminal profiling entails broad generalization and statistical reasoning, whereas deductive behavioral evidence analysis is a dynamic process which helps to

capture successful criminal whose methods either become more refined or deteriorate over time [21].

Since programs to detect various aspects of data sanitization behavior and network security behavior are currently available, this research focuses mainly on the monitoring of password security behavior, data backup behavior, and physical security behavior. The following aspects of this system were implemented and tested through this research:

- *Strict mode*
- *Password Security Behavior*: password strength (an existing common algorithm was reused), password modifying frequency, password reuse
- *Data Access and Backup Behavior*: data backup frequency, attempts to read data over clearance, attempts to read data without need to know
- *Personal observations*: forgetting keycards or Personal Identification Numbers (PINs), leaving items unattended, sociability, ambitiousness, writing passwords down, lending keycards or PINs, security consciousness, understanding and valuing ISM rules
- *Information obtained through background checks*: marital status, number of dependents, academic record, financial status record, criminal record
- *Creation of user security behavioral profiles*: displayed in summarized, detailed, and graphical versions, and as separate views
- *Scheduling security awareness, education and training.*

Each of these areas will be examined in detail in the rest of this section.

3.1 Password Security Behavior

Password Modifying Frequency: The system counts the number of password modifications since joining the organization and during the past 1 year. If the numbers are the same, then the employee might have joined less than a year ago, thus the system checks the number of password modifications in the past 10 months, 8 months and so on until the past month. If the employee joined less than a month ago, it is too new to determine their password modification frequency. If not, the system checks if the password is modified infrequently, few times a year, monthly, every two weeks, weekly, or excessively. If password modification has not been frequent, but has suddenly picked up pace, then it is determined to be a recent activity. Listed in Table 1 are the algorithms used for determining password modifying frequency. The functions for 10, 8, 6, 4, and 2 months are omitted for brevity.

Password Reuse: The system counts the number of total passwords used by the employee in the past 1 year, the number of passwords reused once or twice, the number of passwords reused three-to-five times, the number of passwords reused six-to-nine times, and the number of passwords reused ten times or more in the past year to determine the employee's inclination to reuse passwords.

Table 1 Algorithm for determining password modifying frequency

```
Algorithm
    Start
        modFreq (modifying frequency), yearly count = count pw changes within last 1 year
        total pw = count all password changes since joining organization
            if (total pw <= yearly count)                                   //joined less than 1 year ago
                ten month count = count pw changes within last 10 months
                if (total pw <= ten month count)                            //joined less than 10 months ago
                    eight month count = count pw changes within last 8 months
                    ……
                        if (total pw <= monthly count)                      //joined less than 1 month ago
                            modFreq = "Too new to determine"
                        else
                            if(monthly count > 1)
                                modFreq = do 1m○nth
                            else if(monthly count == 1)
                                modFreq = "Monthly"
                            else                                            //monthly count < 1
                                modFreq = "Infrequent"
                    ……
                else
                    if( ten month count > 3)
                        modFreq = do 10m○nths
                    else if(ten month count == 3)
                        modFreq = "Few times yearly"
                    else                                                    //ten month count < 3
                        modFreq = "Infrequent"
            else
                if( yearly count > 4)
                    modFreq = do 1year
                else if(yearly count == 4
                    modFreq = "Few times yearly"
                else                                                        //yearly count < 4
                    modFreq = "Infrequent"
        Return modFreq
    Stop
    1 year: Start
        modFreq
        yearly count = count pw changes within last 1 year
            if (yearly count <10)
                modFreq = "Few times yearly"
            else                                                            //more than 10 times in the past 1 year
                ten month count = count pw changes within last 10 months
                if (ten month count < yearly count)
                    modFreq = do 10m○nths
                else
                    modFreq = "Recent activity"
        Return modFreq
    Stop
```

continued

continued

```
Algorithm
    1month: Start
        modFreq
        Read monthly count
            If (monthly count < 2)
                modFreq = "Monthly"
            else                                                    //more than once a month
                two week count = count pw changes within last 14 days
                if (two week count < monthly count)
                    modFreq = do 2weeks
                else
                    modFreq = "Recent activity"
        Return modFreq
    Stop
    Two weeks: Start
        modFreq
        Read two week count
            If (two week count <2)
                    modFreq = "Every 2 weeks"
            else                                                    //more than once every 2 weeks
                weekly count = count pw changes within last 7 days
                if (weekly count < two week count)
                    if (weekly count < 2)
                        modFreq = "Weekly"
                    else                                            //more than once a week
                        modFreq = "Excessively
                else                                    //weekly count is the same as two week count
                    modFreq = "Recent activity"
        Return modFreq
    Stop
```

3.2 Backup Frequency

The system counts the number of total backups performed by the employee since joining the organization and the number of backups performed in the past month. If these two numbers are the same, then the employee might have joined the organization less than a month ago and it is thus too new to determine their backup frequency. If the employee joined earlier, however, then the system determines if the backup frequency is infrequent, weekly, daily, or excessive. If the employee used to perform backups at a slower pace, but has recently started backing up more frequently, then the system determines the backup frequency to be a recent activity.

3.3 Request for Behavioral Profiles

The system compiles a user security behavioral profile containing the relevant behavioral characteristics for each observable behavioral pattern concerning personally observed non-cyber activities, automatically monitored cyber activities, and background information, by checking the current profile for its characteristics

and adding the new characteristics if they are not already listed. The system allows these rules to be configured by the ISO to be aligned with the organization's business objectives. Table 2 lists default values for personally observed non-cyber activities. "N" depicts not having that characteristic, while "Y" depicts having that characteristic. Thus, according to the default values, the security behavioral profile for an employee who leaves items unattended, for example, will contain the characteristics of not being security conscious, easily revealing information, not valuing or understanding ISM rules, and having a potential for improper sharing of information.

3.4 Scheduling Security Education and Training

The system reads the database to check for any existing security training schedules. If so, the system checks if any of those schedules are yet to come. If there are no available schedules, or if all the available schedules are past, then the system recomputes the new schedules. The random schedule for periodic risk perception renewal is set in 4 weeks from the coming Tuesday for all employees. This will likely consist of a pop-up presentation about security best practices followed by a questioning session to check the employee's understanding of security awareness. For employees who have a potential for improper information sharing, a hands-on security workshop conducted by external security professionals is scheduled in 2 weeks from the coming Wednesday. If an employee has the potential for unauthorized access to information, the system schedules a security seminar by security managers and legal officials in 1 week from the coming Wednesday. For employees who are deemed to have a motive for engaging in improper information sharing or unauthorized access, the system schedules closer inspection and background checks in 2 weeks from the coming Thursday.

4 Results and Interpretations of Hypothetical Cases

To test this system, the authors created ten hypothetical test case scenarios as shown in Table 3. Table 4 provides an example for computing password modifying frequency for employee Claire McCormick (Emp0007), who joined the organization on 2013/4/2. Password modification frequency is computed on the last date of modifying the password, which is 2013/9/23 for employee Claire McCormick. Table 5 shows the resulting password security behavior for Claire McCormick, which shows that she modifies her password "Monthly" and that out of a total of 7 passwords used, she had reused no passwords 10 times or more, 0 passwords were reused six-to-nine times, 1 password was reused three-to-five times and 4 passwords were used once or twice. Table 6 provides an example for computing backup frequency for employee Gavin Fields (Emp0009), who joined

Table 2 Behavioral characteristics for observable non-cyber behavioral patterns

Activity	Security conscious	Reveals information	Values/ Understands ISM rules	Sociable	Ambitious	Technical knowledge	Easy hack target	Suspicious behavior	Social incentive	Career-wise incentive	Personal motive	Financial motive	Psychological motive	Improper sharing potential	Unauthorized access potential
Forgets keys	N	–	–	–	–	–	–	–	–	–	–	–	–	Y	–
Does not forget keys	Y	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Leaves items unattended	N	Y	N	–	–	–	–	–	–	–	–	–	–	Y	–
Does not leave items	Y	N	–	–	–	–	–	–	–	–	–	–	–	–	–
Sociable	–	–	–	Y	–	–	–	–	–	–	–	–	–	–	–
Not sociable	–	–	–	N	–	–	–	–	Y	–	–	–	–	–	–
Ambitious	–	–	–	–	Y	–	–	–	–	Y	–	–	–	–	Y
Not ambitious	–	–	–	–	N	–	–	–	–	–	–	–	–	–	–
Writes down passwords	N	Y	–	–	–	–	–	–	–	–	–	–	–	Y	–
Does not write passwords	Y	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Lends keys/ PINs	–	Y	–	–	–	–	–	–	Y	–	–	–	–	Y	–
Does not lend keys/PINs	–	N	–	–	–	–	–	–	–	–	–	–	–	–	–
Security conscious	Y	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Not security conscious	N	–	–	–	–	–	–	–	–	–	–	–	–	Y	–
Understands/ values ISM rules	–	–	Y	–	–	–	–	–	–	–	–	–	–	–	–
Does not understand/ value rules	–	–	N	–	–	–	–	–	–	–	–	–	–	Y	–

Table 3 Hypothetical employee data

ID	Name	Designation	Marital status	Dependents	Academic record	Financial status	Criminal record
Emp0001	Martha Hall	Accountant	Unmarried	0	BA: Accounting	Steady	None
Emp0002	Monica White	Software Engineer	Married	1	BS: Computer Science	Steady	None
Emp0003	Shaun Mills	Computer Operator	Divorced	1	Computer Tech Certification	Low	Juvenile breaking and entering
Emp0004	John Flynn	Software Engineer	Widowed	2	MS: Computer Engineering	Steady	Teenaged federal DB hacking
Emp0005	Jacob Call	Computer Operator	Married	3	Computer Tech Certification	Low	None
Emp0006	Faith Stellar	Software Engineer	Divorced	1	MS: Computer Engineering	Steady	None
Emp0007	Clair McCormick	Accountant	Unmarried	0	BA: Accounting	Steady	None
Emp0008	Samantha Colt	Computer Operator	Unmarried	1	Computer Tech Certification	Low	Juvenile shoplifting
Emp0009	Gavin Fields	Accountant	Divorced	3	BA: Accounting	Steady	None
Emp0010	Sarah Mason	Software Engineer	Widowed	2	MS: Computer Engineering	Steady	None

Table 4 Password changes for employee Claire McCormick (Emp0007)

Password change ID	Date	Password	Strength
2013-04-02_emp0007_03:40:18	2013-04-02	4cMc7LrI	Medium
2013-04-26_emp0007_03:48:28	2013-04-26	LcM7cC01	Medium
2013-05-31_emp0007_21:24:16	2013-05-31	RaI007lC	Medium
2013-06-28_emp0007_21:16:28	2013-06-28	LcM7cC01	Medium
2013-07-19_emp0007_21:24:43	2013-07-19	cL7MM92c	Medium
2013-08-22_emp0007_19:25:26	2013-08-22	LcM7cC01	Medium
2013-09-23_emp0007_01:41:07	2013-09-23	cCmC7k05	Medium

Table 5 Password security behavior of Claire McCormick (Emp0007)

Employee ID	Password strength	Reuse	Password modifying frequency
Emp0007	Medium	7_0_0_1_4	Monthly

Table 6 Data backup by employee Gavin fields (Emp0009)

Data backup ID	Employee ID	Date
2013-03-22_emp0009_20:18:23	Emp0009	2013-03-22
2013-04-26_emp0009_21:55:46	Emp0009	2013-04-26
2013-05-24_emp0009_21:58:47	Emp0009	2013-05-24
2013-06-28_emp0009_19:55:30	Emp0009	2013-06-28
2013-07-22_emp0009_21:55:53	Emp0009	2013-07-22
2013-08-23_emp0009_21:55:54	Emp0009	2013-08-23
2013-09-06_emp0009_21:55:56	Emp0009	2013-09-06
2013-09-06_emp0009_21:55:57	Emp0009	2013-09-06
2013-09-09_emp0009_17:36:48	Emp0009	2013-09-09
2013-09-13_emp0009_21:55:59	Emp0009	2013-09-13
2013-09-22_emp0009_20:56:01	Emp0009	2013-09-22
2013-09-22_emp0009_20:56:02	Emp0009	2013-09-22
2013-09-22_emp0009_21:56:03	Emp0009	2013-09-22
2013-09-22_emp0009_22:02:56	Emp0009	2013-09-22
2013-09-30_emp0009_13:06:20	Emp0009	2013-09-30
2013-09-30_emp0009_13:06:22	Emp0009	2013-09-30
2013-09-30_emp0009_14:05:25	Emp0009	2013-09-30
2013-09-30_emp0009_14:05:30	Emp0009	2013-09-30
2013-09-30_emp0009_16:08:34	Emp0009	2013-09-30
2013-09-30_emp0009_16:53:09	Emp0009	2013-09-30

the organization on 2009/10/1. The data backing up of the last 6 months by Gavin Fields is displayed. Backup frequency is computed on the last date of performing data backup, which is 2013/9/30 for employee Gavin Fields. By examining the data in Table 6, it can be seen that Gavin Fields used to perform backups monthly until August 2013, but has since been performing backup more frequently, including six times on 2013/09/30. This is categorized as a “Recent activity” as shown in Table 7.

Table 7 Backup behavior of Gavin fields (Emp0009)

Employee ID	Backup frequency
Emp0009	Recent activity

Table 8 Cyber activity

ID	Password strength	Password reuse	Password modifying frequency	Backup frequency	Access over clearance	Access without need-to-know
Emp0003	Weak	20_0_1_2_2	Excessive	Excessive	2	1
Emp0006	Strong	8_0_0_0_8	Monthly	Daily	0	0
Emp0007	Medium	7_0_0_1_4	Monthly	Weekly	0	0

Table 9 Personal views on non-cyber activity

ID	Manager's view	Security personnel's view
Emp0003	Writes down passwords, leaves items unattended	Forgets keycards
Emp0006	Security conscious, understands and values ISM rules, ambitious	–
Emp0007	Lends keycards and PINs, does not value ISM rules	–

Table 8 displays the automatically monitored and computed cyber activity for Shaun Mills (Emp0003), Faith Stellar (Emp0006), and Claire McCormick (Emp0007). Table 9 shows personal views about non-cyber activities of these employees observed by managers and security personnel.

Table 10 depicts the resulting profiles of the above three employees obtained through the security behavioral profiling system on 2013/9/30. These results show that employee Shaun Mills (Emp0003), has security behavioral flaws that could lead to information security problems along with motives or incentives, and thus needs the hands-on training workshop, security educational seminar and closer inspection, along with the random security awareness, whereas, employee Faith Stellar (Emp0006) does not engage in any wrongful security behavior, but her knowledge about computers and background information show that she still requires the security seminar showing the legal aspects of security violations, along with closer inspection and the random security awareness. Employee Claire McCormick (Emp0007), however, is an example of a case where the personal views of her manager might be biased. Her cyber activities and background information show that she does not engage in any wrongful security behavior, but the personal views state otherwise. In this instance, the ISO can request separate views of her security profile, and upon seeing that the personal observations by her manager contradict the rest of her security traits determined by the system, can use his or her own personal judgment to avoid any bias this employee's manager might

Table 10 Computed security behavioral profiles, security status and training schedules

ID	Profile	Security status	Random schedule	Workshop schedule	Seminar schedule	Inspection schedule
Emp0003	Not security conscious. Information revealed easily. Does not understand or value ISM rules. May have personal motives. May have social incentives. May have financial motives. May have psychological motives and potential. Suspicious behavior. Easy hack target. Ambitious	Has unauthorized access potential. Has improper sharing potential. Has motives/incentives	2013_10_29	2013_10_16	2013_10_9	2013_10_17
Emp0006	Ambitious. May have career-wise incentives. Security conscious. Understands and values ISM rules. May have personal motives. Has technical knowledge about computers.	Has unauthorized access potential. Has motives/ incentives	2013_10_29	None	2013_10_9	2013_10_17
Emp0007	Information revealed easily. May have social incentives. Does not understand or value ISM rules	Has improper sharing potential. Has motives/incentives	2013_10_29	2013_10_16	None	2013_10_17

have towards her, and thereby decide whether she requires the hands-on training workshop, or whether closer inspection and the random security awareness program are sufficient.

5 Conclusions and Future Work

In conclusion, it can be stated that by examining the automatically monitored cyber activities of the employees, their personally observed non-cyber activities, and their background information, the system compiles security behavioral profiles showing which of the employees could potentially engage in which wrongful activities that could present a threat to the organization's information security. Accordingly, the system also determines and schedules the level and type of security education and training to be given to each individual employee. By allowing observable information about employees' behavior to be inputted personally by managers and security personnel, and through automatic monitoring of cyber-activities of employees, this system attempts to handle the human-related problem of improper information sharing using both technological and social information gathering methods. It also provides a mixture of technological and social solutions by means of automatic access control, logging, and risk perception renewals by the system, along with hands on security awareness and training workshops conducted by security professionals, and the allowing of the use of personal judgment by the ISO. The system thereby helps to overcome the weaknesses of a purely technological solution to this human-related problem of information security.

Through the satisfactory results obtained by testing the system presented above with the hypothetical test cases, it can be stated that this system can be used for effective prediction of security infractions by employees within an organization to a certain extent.

As future work, currently existing common algorithms could be reused with modifications and integrated to the implementation of this system to cover the rest of the areas of monitoring of security behavior proposed through this research. Deploying and putting the system to use on real people in order to obtain real test results would help to further evaluate the system's functionality.

References

1. M. Bean, Human error at the centre of IT security breaches (2008), Available: http://www.newhorizons.com/elevate/network%20defense%20contributed%20article.pdf
2. E. Pronin, Perception and misperception of bias in human judgment. J. Trends Cogn. Sci. **11**, 37–43 (2006)
3. B. Schneier, The psychology of security (2011), Available: http://www.schneier.com/essay-155.html

4. B.R. Williams, Do it differently. J. Inf. Syst. Secur. Assoc. **9**(5), 6 (2011)
5. D.M. Lynch, Securing against insider attacks, Information Security and Risk Management (2012), pp. 39–47. Available: http://www.csb.uncw.edu/people/ivancevichd/classes/MSA%20516/Supplemental%20Readings/Supplemental%20Reading%20for%20Wed,%2011-5/Insider%20Attacks.pdf
6. R.A. Grimes, How to thwart employee cybercrime, insider threat deep drive—combating the enemy within, infoworld—special report (2012), pp. 2–7. Available: http://resources.idgenterprise.com/original/AST-0001528_insiderthrea_2_v1.pdf
7. A. Liu, C. Martin, T. Hetherington, S. Matzner, A comparison of system call feature representations for insider threat detection, in *Proceedings of the 2005 IEEE Workshop on Information Assurance*, United States Military Academy, West Point, NY, 2005
8. R.F. Mills, M.R. Grimaila, G.L. Peterson, J.W. Butts, A scenario-based approach to mitigating the insider threat. J. Inf. Syst. Secur. Assoc. **9**(5), 12–19 (2011)
9. C. Vroom, R. Von Solms, Information security: auditing the behavior of the employee, in *Security and Privacy in the Age of Uncertainty*, *IFIP TC11 18th International Conference on Information Security (SEC2003)*, Athens, Greece ed. by D Gritzalis, S. De Capitani di Vimercati, P. Samarati, S. Katsikas (Kluwer Academic Publishers, Norwell, MA, 2003), pp. 401–404
10. R.V. Sabett, Have you seen the latest and greatest 'security game changer'? J. Inf. Syst. Secur. Assoc. **9**(5), 5 (2011)
11. T. Asai, *Information Security and Business Activities* (Kameda Book Service, Niigata, Japan, 2007)
12. T.R. Peltier, *Information Security Policies, Procedures and Standards: Guidelines for Effective Information Security Management* (Auerback Publications, Boca Raton, FL, 2002)
13. J.J. Gonzalez, A. Sawicka, A framework for human factors in information security, in *Proceedings of the 2002 World Scientific and Engineering Academic Society International Conference on Information Security*, Rio de Janeiro, 2002
14. K. Foley, Maintaining a proactive and sustainable security program while hosting and processing personally identifiable information. J. Inf. Syst. Secur. Assoc. **9**(5), 25–32 (2011)
15. S.A. Fernando, T. Yukawa, Internal control of secure information and communication practices through detection of user behavioral patterns, in *Proceedings of The World Congress on Engineering 2013*, Lecture Notes in Engineering and Computer Science, WCE 2013, London, U.K., pp. 1248–1253, 3–5 July 2013
16. D. Lacey, *Managing the Human Factor in Information Security: How to win over staff and influence business* (Wiley, West Sussex, England, 2009)
17. T.M. Young, S. Varano, *Profiling pros and cons: an evaluation of contemporary criminal profiling methodologies, Final report—Honors Program* (Northeastern University, Boston, MA, 2006)
18. M. Thompson, An introduction to behavioral evidence analysis (2012), Available: http://colbycriminaljustice.wikidot.com/criminal-profiling
19. J. Claridge, Criminal profiling and its use in crime solving (2012), Available: http://www.exploreforensics.co.uk/criminal-profiling-and-its-use-in-crime-solving.html
20. L. Winerman, Criminal profiling: the reality behind the myth. Am. Psychol. Assoc. **35**(7), 66–69 (2004)
21. B. Turvey, Criminal profiling: an introduction to behavioral evidence analysis. Am. J. Psychiatry **157**, 1532–1534 (2000)

Filtering of Mobile Short Messaging Service Communication Using Latent Dirichlet Allocation with Social Network Analysis

Abiodun Modupe, Oludayo O. Olugbara and Sunday O. Ojo

Abstract In this study, we introduce Latent Dirichlet Allocation (LDA) with Social Network Analysis (SNA) to extract and evaluate latent features arising from mobile Short Messaging Services (SMSs) communication. This would help to automatically filter unsolicited SMS messages in order to proactively prevent their delivery. In addition, content-based filters may have their performance seriously jeopardized, because SMS messages are fairly short and their meanings are generally rife with idioms, onomatopoeias, homophones, phonemes and acronyms. As a result, the problem of text-mining was explored to understand the linguistic or statistical properties of mobile SMS messages in order to improve the performance of filtering applications. Experiments were successfully performed by collecting time-stamped short messages via mobile phones across a number of different categories on the Internet, using an English language-based platform, which is available on streaming APIs. The derived filtering system can in the future contribute in optimal decision-making, for instance, in a scenario where an imposter attempts to illegally gain confidential information from a subscriber or an operator by sending SMS messages.

Keywords Dirichlet · Filtering · Message · Mining · Mobile · Network · Topic

A. Modupe (✉)
College of Science, Engineering and Technology, School of Computing, Johannesburg, Florida 1709, South Africa
e-mail: abiodunmodupe@gmail.com

O. O. Olugbara
Department of Information Technology, Durban University of Technology, Durban 4001, South Africa
e-mail: oludayoo@dut.ac.za

S. O. Ojo
Faculty of Information and Communication Technology, Tshwane University of Technology, Pretoria 0001, South Africa
e-mail: ojoso@tut.ac.za

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8_48,

1 Introduction

The worldwide growth in the numbers of mobile phone users has led to a dramatic increase of unsolicited SMS messages. A recent report, clearly indicates that the volume of unsolicited SMSs on the public mobile phone networks is dramatically increasing year by year. In practice, combatting unsolicited SMS aberration is difficult, which is attributable to several factors, including the relatively low cost of SMS messaging that allows mobile phone operators to ignore the issue of spam-filtering software that could help to proactively detect malicious SMSs—with the aim of maintaining an uncontaminated online infrastructure.

The overarching objectives with the current research study were firstly to introduce a practical application of a Latent Dirichlet Allocation (LDA) method for filtering unsolicited messages communicated through mobile phones (SMSs), and secondly to use the Social Network Analysis (SNA) to integrate extracted latent features with the intention of understanding key users' interest interactions or relationships. Nowadays, it is predictable that cyber-criminals are in search of unprotected means to search for peers with the same interests. In this realm, they can share and comment on their feelings and interests with their peers—that support their ideas—by texting or chatting and thus using the value-added features of mobile devices, the so-called or Short Messages Services (SMSs). Lately, it seems that virtually the entire world is texting. Particularly, Africa has seen an exponential growth in the SMS services that support the rapid exchange of information, with the accompanying unprecedented leverage in interpersonal connections [1].

Prominent examples of the escalating networking include the popular mobile applications ('apps'), such as WhatsApp, tweeting and blogs. A certain Internet site boasts that it conveys over 90 million tweets per day via their SMS service, typically in English [2]. SMSs are usually used for commenting on, or debating a topic. Other emerging text-based participative manifestations provide new opportunities of web-browsing, such as advertising, personalized information services, propaganda, and a growing tendency of (undesirable) interaction between ideological groups. The growth in SMS services can be compared to the availability of low-priced bulk pre-paid packages and the fact that delivery is guaranteed [3]. The penetration of the mobile phone market is at a staggering 3.2 billion worldwide [4] and contributes to a dramatic increase in mobile spam messages. Mobile phone spam is a form of spam directed at the text messaging or other communications services of mobile phones. It is described as mobile spamming, SMS spam, text spam or m-spam. According to the International Telecommunication Union (ITU), the SMS market has become a massive commercial industry, valued at 11.3–24.7 % of the Gross National Income (GNI) per capita in developing countries in early 2013 [5].

One of the key defining moments of mobile phones' value-added services (VAS, i.e. non-core services provided by the telecommunications industry) that remains, is how to make sense of the high volume of short message streams sent electronically

in an unstructured social stream—along with the copious unsolicited and unwanted messages. There are very insufficient proactive spam-filtering techniques available to block unsolicited SMS effectively. The current methods used by Global System for Mobile Communications (GSM) operators to detect SMS spam—called Anti-Spoofing and faking—are successful in identifying phoney SMS messages of which the source has been manipulated to avoid charges [6].

Unlike traditional e-mail, SMS spamming is evolving to the mobile phone networks, due to concept drift and a lack of contextual information. Sparse feature representation poses tough challenges to effectively design proactive solutions for optimizing decision-making. This pertains to the aim of successfully filtering incoming and outgoing SMS streams—with the purpose of improving security and practically increase confidence in communication—using modern technology solutions.

Stemming from the above, the two central prevalent research questions are: (1) What is the user's interest in mobile messaging that can be exploited to characterize malicious text messages and (2) how latent features can be used to produce new interactions, leading to better messaging personalization. These two questions necessitate the efficient feature detection of SMSs and the development of a topic-based model to recognize security threats caused by some suspect mobile subscribers. In the study at hand, probable answers to these questions were pursued by grouping unsolicited text messages, based on their topic interest. The results were then combined with social network analyses to determine the interaction between the nodes and edges—based on topic similarities.

The body of this chapter is structured as follows. In Sect. 2, related research is discussed and beneficial conclusions drawn that can assist the study at hand. In Sect. 3, the authors discuss a proposed methodology to explore the topic-based extraction problems. Section 4 deals with the experimental design as well as a discussion of the results of the study. To conclude, the main deductions are presented and recommendations are made for future research.

2 Related Work

Topic-based Social Network Analysis (SNA) is an approach used for tracking themes of a mobile phone subscriber's interest, with the aim of identifying unsolicited messages—and new relationships—which can be related to malicious information and threats. For example, an extremist user on a social network uses bootstrapping methods to recognize users with similar ideological thoughts. Social network services could recommend some existing users to new users with shared interests.

In order to analyze relationships in a social network, a well-known probabilistic model, the Latent Dirichlet Allocation (LDA), was used to cluster the nodes and edges of (possible) into multiple topics [7]. Each user in a social network was considered as a node and each follower as a directed edge between the nodes.

The topics of the conversations were determined by grouping the sets of nodes. The contents of colloquial messages—such as on Twitter or Instant Messaging (IM)—were scrutinized with the intention of identifying online conversations, allowing the development of authors and generate topics suitable for online conversations [8]. In [9], the goal was to use machine learning techniques—Support Vector Machine (SVM) with Entropy Weighted Genetic Algorithm (EWGA), a hybridized genetic algorithm that incorporates the information gain heuristic to model linguistic characteristics of Arabic for feature selection in Dark Web forums to improve the classification of messages that contains malicious sensitive information on extremist' opinions and sentiments.

For every identified conversation, the Latent Dirichlet Allocation (LDA) was applied to analyze and extract documents—so as to detect latent topics from the websites of radicals—that exploit the ubiquity of the Internet to form virtual communities at fairly low costs. With the intention of gaining insight into the structure and properties of organized crime activities on the Internet [10], LDA-based methods were employed—rather than traditional Information Retrieval (IR)—in an attempt to develop effective combating strategies against such malicious behavior. Moreover, other applications were correspondingly applied to detect domestic web forums [11] and the creation of a social network—by mapping and identifying their structure and cluster affinities [12]. Finally, web forums were examined in order to determine whether an identified community had been involved in illegitimate activities, by using automated and semi-automated measures for gathering and analyzing the information [12].

Latent Semantic Analysis (LSA) has been applied previously in Dark-Web applications, such as [13] where Latent Semantic Indexing (LSI) was implemented to connect nodes to certain topics in social network construction. Furthermore, Principle Component Analysis (PCA) for the parameterization of mobile (SMSs) and entropy term-weighting schemes were used [14], while Artificial Neural Networks (ANNs) are used to classify mobile SMS into some predefined categories. The goal was to analyze the concept of a new classification model to classify SMS with the application of text classification. Also, authorship analysis [15] was applied to groups' authorship tendencies and tackled the anonymity problem associated with virtual communities.

Along this vein, it is important to note that numerous investigations have been done on the use of evolutionary filtering of SMS spam. In such studies, it was reported that the average time to classify a message—using supervised learning algorithms—was a mere fraction of a second. On the other hand, most evolutionary classifiers required three to four seconds for classification. The fastest classifier proved to be the Supervised Classifier System (SCS) at 1–2 s. As far as the performance of classification algorithms on SMS spam data was concerned, a slight variation in the set, including orthogonal word bigrams, improved the classification time even more [16].

The contents of mobile SMSs present many impediments: the messages are brief; the language used are rich in morphology; spellings are phonetic reductions; punctuation is poor, resulting in limited data for training—which is unlike

conventional messages typically used in e-mails [17]. Spam filtering is significantly more effective for e-mail messaging than for SMSs. E-mail comprises of contextual information, headings and sub-headings [18, 19], while SMSs contain far less appropriate information—without headings and paragraphs—and there is consequently less context to make analyses simpler.

3 Proposed Methodology

The main research question of the current study was how to enhance the exploration of cellular phone users' interests and their interactions, grounded on a topic-based social network analysis. The first approach was aimed at obtaining a reduced, or filtered presentation of the mobile SMS datasets that functions within a social community. The datasets had to be created in such way that the core information collected was substantial enough to discover key users' interests, and secondly, to apply social network analysis methods with the intention of attaining social interactions—based on corresponding topics of interest. As a result, the authors obtained all of the users' interest patterns and were thus able to reveal the latent features of activists on key topics of the ideology of the transmitters or receivers.

It is noteworthy, from a research point of view, to observe the linguistic or statistical properties when users are texting malicious messages, generated by certain mobile phone users. These properties assisted the researchers in improving the existing technology and models, namely to discover hidden features and relationships between the users generating mobile SMS messages. The messages were partitioned into segments, and the disclosed topics in each segment were disseminated—in order to reduce the latent features—and thereby establishing what the users were texting about, as well as the potential creation of new relationships—as a result of the text conversation.

By applying a hybrid approach—SNA and LDA—to determine a cell phone user's unique pattern of interest, the researchers were able to identify certain malicious or threatening topics from the SMSs. Hence specific analyses could be performed on the contents of each message in order to automatically filter contents over time.

3.1 Basic Notation

In Sect. 3, the observed concepts that had been implemented are presented. Suppose the size of simple text document is V, which defines the vocabulary of the words present in a text message. In the study at hand, a single word was denoted as $w = (w_1, w_2, \cdots\cdots, w_N)$; and was the basic unit of discrete data; where N represented the length of the word in a SMS message, indexed by $\{1, \ldots\ldots, |V|\}$ as a dimensional vector, where a single component equalled one (1) and the other

equalled zero (0); that is to say, if the cluster existed in the document. Therefore, in a collection of text documents, D given a class $c = (w_1, \ldots\ldots w_{|d|})$; it was found that the weight representation of a given word was more significant than another one in a collection denoted as follows:

$$tf - idf = tf(w, d) \times idf(w, D) \tag{1}$$

The term frequency $tf(w, d)$ simplified the occurrence of a word w in a simple text document d and the inverse document frequency $idf(w, D)$ measured whether the word occurred frequently—or rarely—across the entire document, by the using the logarithm of the quotient as follows:

$$tf - idf = \left((w, d) \times \frac{|D|}{|\{d \in D : w \in d\}|}\right), \quad \text{where } tf(w, d) \neq 0 \tag{2}$$

3.2 Topic Model

The model described in the current study can be considered as a probabilistic model, inspired by Probabilistic Latent Semantic Indexing (PLSA) [20] that relates to documents and words by means of variables which represent the main topics inferred from the document. In this context, a document can be considered as a mixture of topics, represented by probability distributions which generate the words in a single message. The inferring process of the latent variables, or topics, is the key component of this model, and the main objective is to infer—from the text data—the distribution properties or linguistic structure of the underlying topics in a given corpus of text documents, such as SMSs.

The proposed model for the analyses of the data in the present study was the Latent Dirichlet Allocation (LDA) [21, 22]. The LDA is an extension of the Bayesian statistical model where latent topics, i.e. unobserved 'hidden' variables in mobile SMS documents, are inferred from the estimated probability distributions of the dataset. The documents were henceforth tokenized as a stream of words, phrases and symbols in a training dataset, applying Eq. (2) as a probability distribution over the set of words, represented by the vocabulary $(w \in V)$. This formed a number of (K) probabilities, a dissimilar collection, named topics (T). Each distribution segment of words—typically assigned as topics—were sampled from the multinomial Dirichlet processes. It therefore implies that each document could be interpreted with different number of topics. It utilizes θ and ϕ distributions from the previous iteration as a prior probability for subsequent iterations. This step assisted in classifying whether the message could be filtered successfully as a legitimate mobile SMS, or a malicious and threatening messages.

For Latent Dirichlet Allocation (LDA) analyses, a given mobile user u and a number k of topics, in the SMS of a user u can be represented as a multinomial

distribution ϕ_k over topics—drawn from a Dirichlet prior—with a smoothing parameter α.

The distribution of topics was represented by a smoothing parameter β_k drawn from a Dirichlet prior β. The rationale was to determine the joint probability distribution to extract words from a SMS document d of a mobile user u from a set of topics $T \approx z_{d,u} \in \{1, \ldots\ldots, K\}$ drawn from θ_d, composed by a set of words, represented by $w_{d,u} \in \{w_1, \ldots\ldots, w_{|V|}\}$, which was drawn from the distribution $\beta_{z_{d,u}}$,

$$p(w, z, \theta, \phi|\alpha, \beta) = p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z, \phi) \tag{3}$$

The hidden features could be deduced by integrating Eq. (3) in order to offer a more simplified model, when a corpus $z \in T$ was given over the random smoothing parameters α and β. Therefore, the objective functions of the model parameters and inferred distribution of the latent features were calculated as:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{k=1}^{K} \sum_{z_{d,u} \in T} p(z_k|\theta) p\left(w^k|z_k, \beta\right) \right) d\theta \tag{4}$$

The ultimate aim was to optimize the distribution when the number of SMS messages increased and the generated token was higher than in a typical document—hence the Gibbs sampling '*Markov Chain Monte Carlo*' (MCMC) [23] model was implemented. This model assisted in building a compact representation from a high dimensionality of learning topics in a low-dimensionality document, in order to find the latent distribution of each SMS message across topics. Therefore, each SMS of a user was modelled as a topic-vector, where each dimension represented the probability value to produce the topic and the probability generative process of the LDA [16]. The process is summarized in Table 1.

3.3 Social Network Configuration

In order to construct a social network relating to the present study, unique topics of interest to a mobile cell phone user u, as described in Sect. 3.2, must be taken into consideration. Generally, the interests of a mobile user u could be observed by noting his or her participation in a SMS chatting forum. This was particularly evident when the participation occurred when some mobile user u initiated a text conversation in the forum. A mobile user's (u) interest patterns in such a forum could be inferred from the conversational interaction (i.e. the content of communication extracted)—using topic-based methods (Sect. 3.2), and expressed as the core member's participation.

The social network configurations were constructed as follows: *Nodes* were the topic labels, and *arcs* represented interaction, i.e. the topic similarity between the two. The clustering of the topics of interest (i.e. the extracted latent features, as

Table 1 LDA Algorithm

Step 1: Data Preprocessing-Term-Document Matrix of frequencies
- Extract SMS texting as a Corpus Object document from TwitterAPI in R environment covering four categories Politics, Technology, Sport and Entertainment
- Tokenize SMS text documents and removed white-space or other delimiting character (hyphens), convert all upper case to lower case
- Removed Stop-words, Punctuation and Numbers, then Apply Stemming algorithms [18]
- Compute Term-Document Matrix (TDM) to words, which occurred in at least five documents for Vocabulary size (V), Total length of the document (N)

Step 2: Estimate coarse topic

Set D as the total number of Mobile SMS documents

Set V as the vocabulary size in the document (D)

Determine and estimate latent Topics (K)

Step 3: Determine the Topic distribution per Mobile SMS document

For each topic $k = 1, \ldots\ldots, K$ where $\alpha_k > 0$;

Then, Drawn a distribution over words $\phi_k \approx Dir(\alpha)$

Step 4: Topic to words assignments and term distribution per topic of the whole Mobile SMS document (D)

For each document d
- Draw the global topic proportions $\theta_d \approx Dir(\alpha)$
- For each document word i;
  - Draw topic proportions $z_{d,i} \approx Mult(\theta_d), z_{d,n} \in \{1, \ldots\ldots, K\}$
  - Draw a word $w_{d,i} \approx Mult\left(\varphi_{z_{d,i}}\right), w_{d,i} \in \{1, \ldots\ldots, V\}$

described in Sect. 3.2) and the measuring of malicious topic '*cohesiveness*' to the larger network, were two of the main concerns covered in the present study.

In this way, two networks were constructed, namely (1) the oriented-node and (2) the reply-edge network in order to characterise social interaction within a mobile SMS community via Twitter API,[1] following the replying schema of members:

1. An oriented-nodes network can be described as a malicious topic interest, i.e. the threatening conversation originated at the core member within a mobile SMS community, and every response—i.e. replies—created at the individual nodes.
2. A reply-edges network can be defined as every reply to a threats' response—or tweeted and re-tweeted simultaneously.

The interactions within a mobile SMS community were characterized by edges (i.e. arcs), as the replies—based on the topic interests of the members and nodes u—i.e. the topic interest that initiated the posts conversation. Therefore, the

[1] https://dev.twitter.com/docs/streaming-apis

weights of the arcs were calculated as the probability distribution associated from one given node response to another.

In order to determine the arcs' weight—based on the nodes (i.e. topic interests), linked to the underlying ideology of the mobile user u, and within the SMS forum, as well as filtering malicious messages—topic-based algorithms were applied (as proposed in Sect. 3.2).

Topic-based methods were used to detect and filter all the mobile user u replies, or communication content, that were considered as threats, according to the topics of interest. Subsequently, Latent Dirichlet Allocation (LDA), was implemented to extract a list of latent features, with the intent of determining a mobile user's core malicious interest patterns, and later on, used to build a social network graph.

3.4 Topic-Based Network Measurement

In this section, the authors used the approach described in Sect. 3.2 to organize latent features in mobile messaging—with the intent of identifying unsolicited (malicious) text messages—that could consequently contribute in curbing the problematic multi-faceted sociological behavior, as well as other social issues. The inferred patterns of social engagement reflected how information was exchanged among mobile users, leading to better message personalization, composed of sets of hidden latent features. The rationale was to measure *modularity*, which reflected the distribution while calibrating a mobile the users' posted messages.

If the modularity was above a certain threshold θ, an interaction could be measured between nodes as a set of positive numbers ranging from -1 to 1. Subsequently, these principles were used to avoid irrelevant overlapping of a node (i.e. latent features) that could belong to more than one mobile user forum (i.e. a social circle).

By way of illustration: Two nodes of a mobile SMS community, where d_i and d_j represented the degree of the vertex (topics), and posted by a user j that interacted, or replied by the user i and the expected number of m edges between the two in a social network graph $d_i d_j / 2m$. The modularity measures [24], i.e. how far the interaction is that had occurred between them, was calculated by means of Eq. (5),

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i - c_j) \tag{5}$$

where A_{ij} is the score of topic k between the posted text messages between user i and j, and $\delta(c_i - c_j) = 1$ is a Kronecker Delta value [25, 26], which determine the modularity in the network configuration described in Sect. 3.3.

The algorithm described in Table 2 was applied to build the social network graph. Firstly, the weights matrix is built according to the LDA algorithm in Table 1 to initialize the semantic weight according to topics k. Then, the network

Table 2 Network Construction Algorithm

Input: Input P as total mobile user messages
Output: Network graph $g = (V, E)$

1. Construct latent features matrix according to LDA Algorithm in Table 1
2. Initial graph element $V = \{\}$, and $E = \{\}$
3. **for** each $i \in P$**do**
4. $V \leftarrow V \cup i$
5. **for** each $i \in P$ **do**
6. **for** each $j \in \{i.response\ or\ replies\}$, $i \neq j$ (exclusive mutual) **do**
7. **if** $d_w(P_i, P_j) \geq \theta$ then
8. $a_{i,j} = a_{i,j} + 1$
9. $E = A \cup a_{i,j}$
10. **end if**
11. **end for**
12. **end for**
13. **end loop**

is built by considering all mobile SMSs posted based on the structure of oriented-nodes and reply-edges network presented in Sect. 3.3. That is, for each mobile SMS posted created as i, the arc weight $a_{i,j}$ is increased according to the interested mobile user that response to the posted message j, which could be greater or equal to the threshold θ.

4 Experimental Setup and Results

Section 4 covers a description of the analyses of the extracted topics, using LDA (discussed in Sect. 3.2)—and network-topology construction, by using both nodes and edges from the mobile SMSs.

4.1 Data Collection

For the current study, an extensive SMS collection of the authors (i.e. collecting time-stamped short messages from across a number of different categories on the Internet [27]) was utilized. Additionally, the authors collected a '*ground-truth*', domain-specific Twitter stream via mobile phones on the Internet—using a Twitter API approach [28] by '*crawling*' a program that extracts user tweet via Mobile phone and other information to create entries for mobile SMS documents—of prominent accounts pertaining to four domains, i.e. technology, entertainment, politics and sports. Based on the API method, a domain-specific stream of 'tweet' messages was generated from mobile phones, totalling 38,506 tweets,

i.e. approximately 1,500 tweets per hour, between September and October 2013. The collections of SMS documents were texted in English. Only tagged messages, according to the message categories, were extracted.

4.2 Topic-Based Extraction

After the SMSs were '*crawled*' from a social network of twitter messages, the LDA was firstly calculated by the extraction of topics (Sect. 3.2). The models were then evaluated in order to automatically determine proportions of topics that had been identified in the *corpus* dataset—as well as the associated word instances—as a realistic representation of a user's key interest. Pre-processing was the next step, such as extracting stop-words, punctuation and stemming algorithms [29] from the 33,525 mobile SMSs and over 48,597 words in the vocabulary of the SMSs V—using a C++ Gibbs sampling—based on implementation of LDA[2] (described in Sect. 3.2).

Each topic and distribution in mobile SMSs were learned meaningfully in a non-parametric (i.e. unsupervised) manner. The parameters used in the current investigation were the number of topics ($K = 100$) and the number of iterations—using Gibbs sampling—to fit the inference over a single instance of mobile SMSs at 1,000th iteration.

The hyper-parameters used in the LDA for the Dirichlet prior is respectively $\alpha = 50/K$ and $\beta = 0.1$, which are common settings denoted in the literature. For the extracted topics, illustrations were made of the top five topical words, conditionally generated from the topics for the mobile SMS datasets.

Table 3 reveals that the overall concept proposed is '*Internet betting*' as its main latent interest features were related to gambling activities. The proposed topic names were '*bid*', '*casino*', '*happi*', *and* '*action*'.

Table 4 is presented as an '*Online friendly deal*', as its main topics were related to '*cash*', '*bank*', '*won*', '*balance*', '*secur*' and '*money*'.

Finally, as shown in Table 5, the overall concept proposed is '*Chargeback fraudster*', where topics such as '*skilful*', '*hog*', '*gyp*', '*valid*', and '*charg*' were included.

Subsequently, the topics were grouped into concepts, using Gephi [30] for social network analysis.

4.3 Topic-Based Social Network Visualization

In Fig. 1, the social network-filtered graph is presented, using the topic-based method described in Sect. 3.2. Figure 1 can be interpreted as the complete

[2] http://gibbslda.sourceforge.net/

Table 3 Five most relevant words with their respective probability values for five topics associated "internet betting" concept

Topic 27	Topic 83	Topic 68	Topic 23	Topic 35
bid(0.03260)	casino(0.23831)	happi(0.11123)	action(0.09326)	send(0.02196)
match(0.02177)	quizzer(0.09764)	watch(0.04912)	house(0.05664)	home(0.14294)
draw(0.01094)	win(0.02789)	act(0.02426)	roll(0.05331)	keno(0.05411)
winner(0.00991)	becalm(0.01634)	match(0.01243)	loan(0.04332)	cal-bankr(0.04442)
tender(0.00012)	custom(0.01594)	shift(0.01184)	sold(0.01335)	close(0.022585)

Table 4 Five most relevant words with their respective probability values for five topics associated with the "online friendly deal" concept

Topic 1	Topic 2	Topic 3	Topic 21	Topic 79
money(0.20503)	won(0.21830)	salari(0.48208)	week(0.13721)	bank(0.02415)
prize(0.07431)	collect(0.14061)	chennai(0.10468)	smile(0.13260)	direct(0.116678)
contact(0.06208)	ppm(0.011101)	consum(0.05077)	wont(0.04958)	balance(0.01156)
receiv(0.05926)	app(0.03703)	card(0.02540)	cash(0.038059)	poor(0.011740)
dollr-urgentnt(0.005173)	secur(0.03331)	citi(0.02223)	forward(0.03229)	chat(0.01118)

Table 5 Five most relevant words with their respective probability values for five topics associated with the "chargeback fraudster" concept

Topic 40	Topic 71	Topic 39	Topic 69	Topic 73
free(0.51810)	ill(0.13539)	dinner(0.11876)	gyp(0.07186)	messag(0.20289)
interview(0.03703)	sick(0.08203)	rent(0.046363)	valid(0.05227)	meet(0.09343)
motiv(0.00796)	shock(0.04340)	break(0.03188)	paym(0.03921)	skilful(0.06136)
crack(0.00531)	credit(0.03743)	transfer(0.02319)	felt(0.03050)	pay(0.04253)
hog(0.002669)	telamon(0.036638)	legal(0.017404)	review(0.01308)	dream(0.03696)
				charg(0.01534)

network—constructed by using the complete word-to-topic structure—where the edges were defined as the '*mobile user interest*' and the nodes represented as the topics that interconnected the user's interest with that of another user. For convenience, the notations of the social network were defined using graphs $G = (V, E)$, where $V = 157$ and $E = 155$ respectively, and the connections were represented by the probability distribution (known as the adjacency $matrix_{(i,j)} \neq 0$) between the nodes and the edges. The graph (Fig. 1) illustrates a large density reduction, which suggests that visualization techniques can provide better visual compactness—based on the extracted latent features of mobile SMS documents—to calculate the average shortest path (ASP = 2.43) among the top five words in the topics of the $matrix_{(i,j)}$.

Furthermore, in Fig. 1, the centres of interaction (i.e. new participation) are clearly observed, based on the ideology of modern online communication processes. The graph furthermore contributes in identifying the influence and power of extracting latent features in mobile SMS documents.

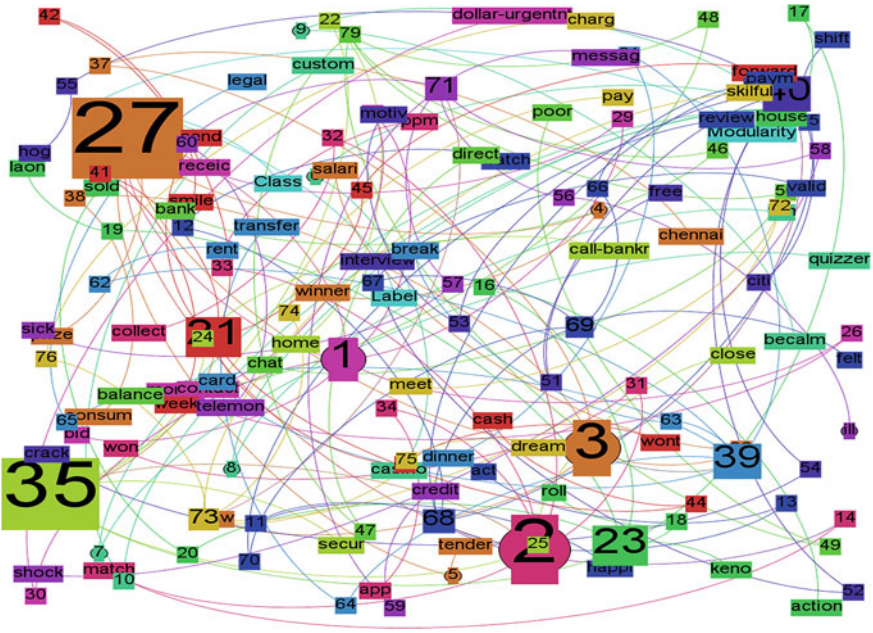


Fig. 1 A topic-based social network analysis demonstrating the density between topics

Fig. 2 Modularity distribution maximized

While, in Fig. 2, we optimize the structure of mobile SMS documents by measuring the modularity [31], the strength of the social network on the extracted latent topics with $Q = 0.922$ as modularity value selected randomly.

5 Conclusion and Future Work

In the current chapter, the authors proposed to combine the traditional Social Network Analysis (SNA) with a probabilistic generative topic-based model, the Latent Dirichlet Allocation (LDA). The proposed method, deriving from the current study, was firstly applied to filter malicious SMS communication and secondly to show that topic-based algorithms can effectively detect distinctive latent features, in order to support automatic content-filtering simultaneously. Consequently, by using different network analysis measures, latent features were evaluated using a benchmark in social networking—with the aim of building new interactions of SMS documents. The resulting outcomes of the study enabled the researchers to gain insight into the curiosity of a mobile SMS user and, moreover, to measure some social aspects which have not been considered by applying SNA alone.

Consequently, the writers were able to filter malicious SMS dialogues and themes on independent mobile phones, as well as to identify linguistic communities that were characterized by circumspect malicious social activities. It was accordingly possible to suggest topological characteristics for mobile SMS documents, by measuring the modularity.

By applying Latent Dirichlet Allocation (LDA) and using Average Shortest Path (ASP)—in addition to modularity, we were capable to obtain 15 topics (most relevant) closes to 20 words on each topic, but chose to select the top five most significant words, based on the probability strength on each topic. After close inspection, the rest of the topics were discarded—as the information it contributed was worthless—its visual representation did not contribute in identifying any new interest, or relationship, in the mobile SMS *corpus*.

However, it should be kept in mind that social network analysis (SNA) on its own, proved to be inadequate to identify mobile users' key interests in conversations, specifically about some online-distributed malicious topics. For that reason, combining advanced topic-based text-mining methods ***and*** social network analysis, the authors were able to disclose the linguistic structures of online communities' key users' interest patterns. This allowed for the application of improved analyses on online social networks.

References

1. K.Y. Kamath, J. Caverlee, Expert-driven topical classification of short message streams. Paper presented at the privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom), 2011
2. T. Chen, M.-Y. Kan, Creating a live, public short message service corpus: the NUS SMS corpus. Lang. Res. Eval. **47**, 1–37 (2013)
3. S.J. Delany, M. Buckley, D. Greene, Review: SMS spam filtering: methods and data. Expert Syst. Appl. **39**(10), 9899–9908 (2012). doi:10.1016/j.eswa.2012.02.053

4. Page, M., Molina, M., & Gordon, J., The Mobile Economy (2013), http://www.atkearney.com/documents/10192/760890/The_Mobile_Economy_2013.pdf. Accessed 15 Nov 2013
5. International Telecommunication Union, *The World in 2011: ICT Facts and Figures* (ITU, 2011)
6. I. Fette, N. Sadeh, A. Tomasic, Learning to detect phishing emails. Paper presented at the proceedings of the 16th international conference on world wide web, 2007
7. Y. Cha, J. Cho, Social-network analysis using topic models. Paper presented at the proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, 2012
8. G. Inches, F. Crestani, Online conversation mining for author characterization and topic identification. Paper presented at the proceedings of the 4th workshop on workshop for Ph.D. students in information and knowledge management, 2011
9. A. Aizawa, An information-theoretic perspective of tf–idf measures. Inf. Process. Manage. **39**(1), 45–65 (2003)
10. L. Yang, F. Liu, J.M. Kizza, R.K. Ege, Discovering topics from dark websites. Paper presented at the IEEE symposium on computational intelligence in cyber security, 2009 (CICS'09)
11. Y. Zhou, E. Reid, J. Qin, H. Chen, G. Lai, US domestic extremist groups on the web: link and content analysis. IEEE Intell. Syst. **20**(5), 44–51 (2005)
12. E. Reid, J. Qin, Y. Zhou, G. Lai, M. Sageman, G. Weimann, H. Chen, Collecting and analyzing the presence of terrorists on the web: a case study of Jihad websites. Intelligence and security informatics (Springer, 2005), pp. 402–411
13. R.B. Bradford, Application of latent semantic indexing in generating graphs of terrorist networks. Intelligence and security informatics (Springer, 2006), pp. 674–675
14. D. Patel, M. Bhatnagar, Mobile SMS classification. Int. J. Soft Comput. Eng. (IJSCE) (2011). ISSN:2231-2307
15. A. Abbasi, H. Chen, Applying authorship analysis to extremist-group web forum messages. IEEE Intell. Syst. **20**(5), 67–75 (2005)
16. D.M. Blei, J. Lafferty, *Topic Models* (illustrated ed. vol. 10). (Taylor & Francis, London, England, 2009)
17. C. Kobus, F. Yvon, G. Damnati, Normalizing SMS: are two metaphors better than one? Paper presented at the proceedings of the 22nd international conference on computational linguistics, vol. 1, 2008
18. A. Modupe, O.O. Olugbara, S.O. Ojo, in *Comparing Supervised Learning Classifiers to Detect Advanced Fee Fraud Activities on Internet*. Advances in Computer Science and Information Technology. Computer Science and Information Technology (Springer, 2012), pp. 87–100
19. C.-C. Lai, An empirical study of three machine learning methods for spam filtering. Knowl.-Based Syst. **20**(3), 249–254 (2007)
20. T. Hofmann, Probabilistic latent semantic indexing. Paper presented at the proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, 1999
21. D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
22. D.M. Blei, J.D. McAuliffe, Supervised topic models. arXiv preprint arXiv:1003.0783, 2010
23. B. Walsh, Markov Chain Monte Carlo and Gibbs Sampling, Lecture Notes for EEB 581, University of Arizona (2004), http://nitro.biosci.arizona.edu/courses/EEB581-2004/handouts/Gibbs.pdf. Accessed Oct 13 2013
24. M.E. Newman, Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
25. L. Šubelj, M. Bajec, Robust network community detection using balanced propagation. Eur. Phys. J. B **81**(3), 353–362 (2011)
26. L. Tang, H. Liu, Graph mining applications to social network analysis, in *Managing and Mining Graph Data* (Springer, 2010), pp. 487–513

27. A. Modupe, O.O. Olugbara, S.O. Ojo, Investigating topic models for mobile short messaging service communication filtering. Lecture notes in engineering and computer science: Proceedings of The World Congress on Engineering, WCE 2013, 3 July–5 July, 2013, London, U.K., pp. 1197–1199
28. A. Bifet, E. Frank, in *Discovery Science*. Sentiment knowledge discovery in twitter streaming data. (Springer, Berlin, 2010), pp. 1–15
29. P. Willett, The Porter stemming algorithm: then and now. Program: Electron. Libr. Inf. Syst. **40**(3), 219–223 (2006)
30. M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks. In ICWSM, May 2009
31. V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), 10008 (2008)

Author Index

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8,

Subject Index

Note: Page numbers followed by "f" and "t" indicate figures and tables respectively

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-8832-8,

N

O

P

Q

R

S