Acta Universitatis Sapientiae

Electrical and Mechanical Engineering

Volume 9, 2017

Sapientia Hungarian University of Transylvania
Scientia Publishing House

Contents

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 9 (2017) 5-18

DOI: 10.1515/auseme-2017-0006

About the Profile Accuracy of the Involute Gear Hob

Márton MÁTÉ, Dénes HOLLANDA

Department of Mechanical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş,
e-mail: mmate@ms.sapientia.ro, hollanda@ms.sapientia.ro

Manuscript received September 10, 2017; revised October 19, 2017.

Abstract: Gear hobs are the most widely and frequently used gear cutting tools. During the time passed between the moment of invention (Schiele, 1876) and the present, gear hobs reached a considerable evolution regarding the geometry, the profile of the edge, the relieving technologies finalizing in the latest constructive and design solutions. This paper deals with the calculus of the edge profile in the case the basic worm of the hob has involute helicoid surfaces. In order to obtain a constant grinding allowance on the relief faces of the gear hob teeth it is necessary to compute the edge of the roughing relieving cutter. The equations are deduced considering that the provenience involute worm is a one teethed helical gear with shifted profile. The presented mathematical model proves that linearizing the relieving cutter profile is not an adequate solution if aspiring to higher precision.

Keywords: gear hob, involute worm, edge, roughing, optimizing, profile

1. Basic concepts regarding the precision of meshing with gear hobs

Gear hobs are the most widely used cutting tools in the gear industry. It was invented in 1856 by the German Christian Schiele. The application of the invention followed later because that time the existing manufacturing technology was not able to ensure the constancy of the cutting edge profile after the re-sharpening operation. This handicap was eliminated by the invention of the relieving lathe, accorded to Friederich Mueller, an American engineer from Hartford, Connecticut (US Patent 1299207 A, 1916.)

The widely use of this excellent gear cutting tool started with the invention of the German engineer Herman Pfauter, who built up in his factory situated in Chemnitz the first universal gear hobbing machine.

The mathematical models of the gear hobs were widely studied that after. Technical literature contains an immense quantity of studies, papers, dissertations regarding different aspects of the geometry, profile precision and constructive solutions. Nowadays a large diversification of constructive solutions are on hands. This solutions offer large machining possibilities with increased cutting performance including the novel and very popular dry cutting technology.

Regarding the mathematical model of the cutting kinematics, two major geometric approaches can be defined. The first model considers that gear hob's generating surfaces are permanently tangent to the surfaces of a generating rack, and the rack is moving along its pitch line due to the helix effect that appears while the hob is rotating about its own axis [1, 2, 5]. Introducing the axial feed, the edges of the hob will sweep the all rack tooth surface. This model is almost everywhere accepted. As a consequence the meshing with a gear hob and with a planning comb is considered often to be equivalent. Litvin has demonstrated that this supposition can be accepted only as an approximation, because the helix effect introduces certain profile modifications [15].

The second geometric hypothesis, more appropriate in opinion of the authors, considers that gear hob's generating surface coincides with a one to utmost 5 teethed helical gear's tooth surface. This helical gear constitute with the machined gear a hyperbolic gear pair. In case of cutting of an involute gear, the generating surface most coinciding with an involute worm, meshed with a standard generating rack [4, 12, 13, 14].

As long as modern computing and simulation methods and environments were developed, research regarding the profile of cutting edge and cutting wedge surfaces marked a new evolution. Different mathematical models were developed regarding the meshing between a gear and an arbitrary rack [3, 8]. Nowadays sophisticated numerical control based manufacturing technologies allow the achieving of the most complex surfaces [9, 10].

However, standards indicate only the basic cutter profile of the gear hob [6] admitting that this is equivalent with the basic rack profile. Gear hob profile is considered to be the task of the manufacturers.

Regarding the peculiar aspects mentioned in the synthesis above the goal of present paper is to answer in which cases the linearization of the roughing relieving cutter edge or the grinding wheel profile is possible. In the followings the hypothesis of the involute basic worm of the gear hob is admitted.

2. Theoretical and geometrical aspects of the basic worm's surface

Based on the theoretical achievements presented above, this paper starts by admitting the second hypothesis described above related to gear hobbing. As a consequence, generating surface of the gear hob is an involute helicoid that is meshing with a standard rack, whose standard dimensions are defined in the normal section [11]. The pitch cylinder of the hob's basic involute worm is tangent to the pitch plane of the rack, while the axis of the worm is perpendicular to the direction of linear motion of the rack in case of reciprocate meshing. Here the main helix of the involute thread closes with the pitch line the angle λ_0, equal with the pitch helix angle of the involute thread [13, 14]. As a consequence the rack teeth declination angle becomes $\beta_0 = \pi/2 - \lambda_0$. In case of one start thread the value of the declination angle must be smaller than 2°30' that leads to a much brooded rack profile whose profile angle approaches almost 84°. The consequence is that the difference between the radiuses of the basic circle and the pitch circle are improper large. This fact involves improper involute curve segments as real generatrix-point manifolds for the involute helicoid surfaces as it is shown in *Fig. 1*.

It can be observed that the properties of a worm thread profile limiting involute arch are completely different from the involute arches that limits a classical spur gear tooth profile [16]. The profile shifting modifies the tooth thickness on the pitch circle. In case of one starting thread worm and zero profile shifting this is equal to the half of the pitch circle circumference, as it is stated by the position of the involute curves 1 and 1'. If a positive profile shifting exists the tooth thickness is increasing. Therefore pitch circle points of the tooth profiles present not anymore diametric symmetry as it can be observed on curves 2 and 2'. Due to the particularly involute curve segments situated between the pitch and the addendum circle, in contrarious with the case of involute tooth, the topland width increases while increasing the profile shifting parameter. Due to this fact, increasing the profile shifting leads to a massive hob tooth that is disadvantageous as it generates large dedendum transition curves on the meshed gear tooth profile. From this reasons the basic worm of the gear hob is accepted as a helical involute one teethed Willis type gear.

The classical parametric equations of the involute curve [11, 15] are not advantageous for computing the useful subset of the involute helix surface due to the fact that the rolling parameter of the involute generating line – the φ angle – get there values starting from zero, and the useful subset mentioned before needs values grater then π radians more difficult to be perceptible. According to this, involute arch *1* is drawn by the extremity *A₁* of the segment *AA₁*, while the half-line opposite to *AA₁* rolls on the basic circle.

For an arbitrary position set by the value φ of the rolling parameter, the segment *AA₁* becomes *BB₁* and point where *B'* matches the following coordinates:

$$\begin{cases} x(\varphi) = R_b\left(\cos E(\varphi) + \left(\tan\alpha_{0t} + \varphi\right)\sin E(\varphi)\right) \\ y(\varphi) = jR_b\left(\sin E(\varphi) - \left(\tan\alpha_{0t} + \varphi\right)\cos E(\varphi)\right) \end{cases} \tag{1}$$

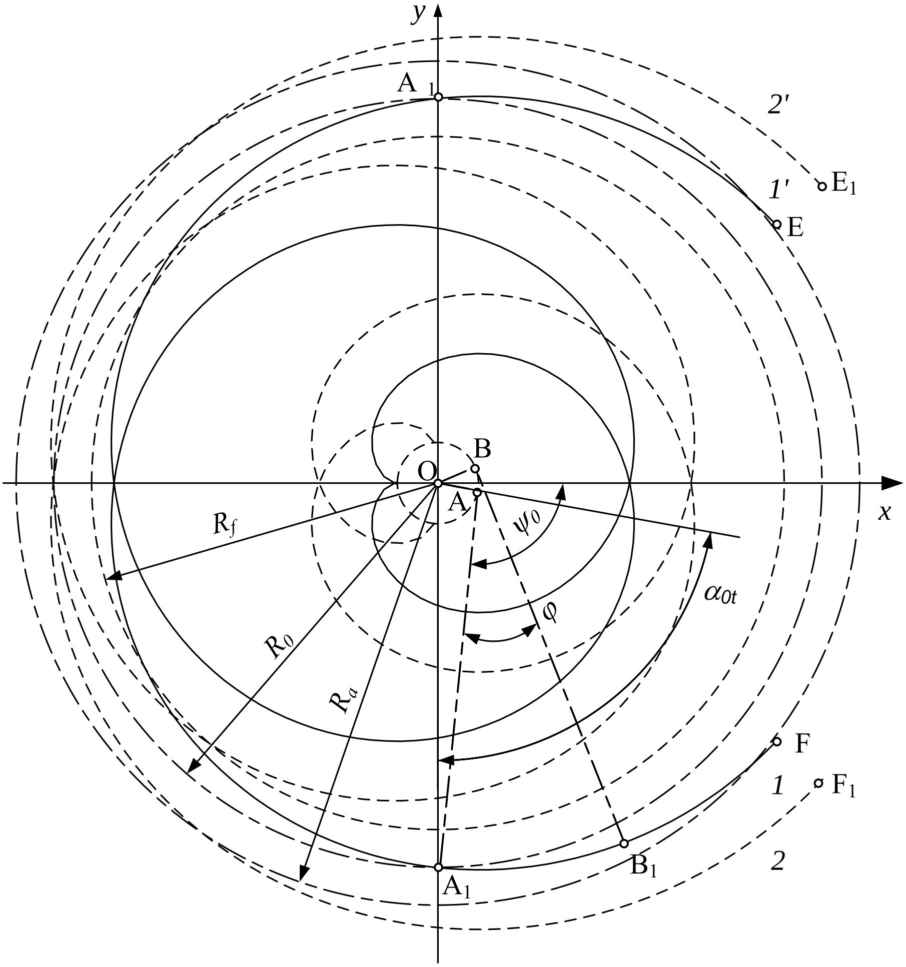

Figure 1: The complete generating curve of an involute worm (radial section)

Here the argument of sine and cosine function has the form

$$\begin{cases} E(\varphi) = \varphi - \psi_0(\Delta_x) + \alpha_{0t} \ , \\ \psi_0(\Delta_x) = \dfrac{\pi}{2} + 2\dfrac{\Delta_x}{m_t}\tan\alpha_{0t} \end{cases} \tag{2}$$

where Δ_x denotes the profile shifting.

Parameter *j* differentiates the involute arches who define the cross section of the worm. If $j=1$ equations (1) refer the arch *1*, for $j=-1$ they describe the opposite arch *1'*.

Using the second fundamental equation of the involute trigonometry [16] or the parametrical polar form of the involute,

$$\begin{cases} \rho(\varphi) = R_b\sqrt{1+\left(\tan\alpha_{0t}+\varphi\right)^2} \\ \theta(\varphi) = \left(\tan\alpha_{0t}+\varphi\right) - \arctan\left(\tan\alpha_{0t}+\varphi\right) - \psi_0\left(\Delta_x\right) \end{cases} \tag{3}$$

it can be proven that topland width increases with the profile shifting.

The equations of the involute worm surfaces are obtained through a roto-translation of the involute profile along the axis *z*. Excepting the transformation matrix [11] and the elementary calculus the equations result in the following form:

$$\begin{cases} x(\varphi,u) = R_b\left(\cos E_1(\varphi,u) + \left(\tan\alpha_{0t}+\varphi\right)\sin E_1(\varphi,u)\right) \\ y(\varphi,u) = jR_b\left(\sin E_1(\varphi,u) + \left(\tan\alpha_{0t}+\varphi\right)\cos E_1(\varphi,u)\right) \\ z(\varphi,u) = \dfrac{p_{ax}}{2\pi}u = \dfrac{m_n}{2\cos\lambda_0}u \end{cases} \tag{4}$$

$$E_1(\varphi,u) = E(\varphi) + ju,\ j\in\{-1;1\}$$

It is easy to observe that the structure of equations (4) is similar to equations (2) of the involute. However involute worm gear's equations can be re-written using the equations of the generating line that rolls on the basic helix [7], but in the following calculuses with this would lead to more complicated equations.

3. The gear hob derived from an involute worm

As defined in the literature gear hob is a worm that is endowed with cutting properties. Therefore cutting edges and cutting wedge surfaces must be created. The simple intersection of the worm surface with another helicoid is not enough while relief surfaces giving positive relief angles will not be produced. For achieve this, relieving operation must be applied. Nowadays relieving is realized on relieving lathes, both the roughing by cutting and the final grinding. The tooth resulted after the relieving operation present helicoid surfaces based on a conical helix directory. Practically, re-sharpening ensures the edge form constancy [7, 12]. Theoretically it was proven that the edge form of the spiroid gear hobs variates with the re-sharpening [11]. This effect appear at the cylindrical gear hob edges too, but it is neglectable since the modification is lower than 10^{-4} mm. The problem appears due to the fact that the re-sharpened edge cannot rebuild the original involute worm in the gearing process. Re-sharpening operations lead to diameter decreasing, helix angle increasing and as a consequence – to the modification of the curvature of helical surfaces of the equivalent worm. The equivalent worm can be defined as a worm whose helical surfaces include the edges of the re-sharpened hob. Errors of the involute profile obtained by hobbing, in case of admitting the perfect involute helical worm, are calculated in [15].

The main question that can be put regarding the theoretical and the geometrical peculiarities described before is how the best approximation of the involute worm can be obtained. The answer is given here by the model of the cutting edges. First cutting edges are obtained if intersecting the surfaces of the worm given by equations (4) with a helical rake face whose main helix line is perpendicular to the main helix line of the worm. Thus, the pitch p_C of the rake face results from the condition of perpendicularity applied on the outstretched pitch cylinder [7]:

$$\tan\lambda_0 = \frac{p_{ax}}{2\pi R_0} = \frac{2\pi R_0}{p_C} \tag{5}$$

The rake face of the hob is a linear helicoid whose generatrix is a straight line intersecting the axis and perpendicular to this. The directory of the surface is a cylindrical helix that fits the pitch cylinder and perpendicular to the pitch helix. Mathematically it can be described by the rototranslation of a line coincident with axis *x* using a helical direction opposite to the worm. After elementary calculus the equations of the rake face are the followings:

$$\begin{cases} x(t,v) = t\cos v \\ y(t,v) = t\sin v \\ z(t,v) = -\dfrac{p_C}{2\pi}v \end{cases} \tag{6}$$

Fig. 2 shows the helix surfaces of the involute worm intersected by the helical rake face.

The equations of the theoretical cutting edges result from equations (4) and (7). Here a dependence between the independent parameters φ, u of the helix surfaces Σ_l and Σ_r must be obtained. The easiest way is to transform the rake face's equations in an implicit form by eliminating parameters *u* and *t*:

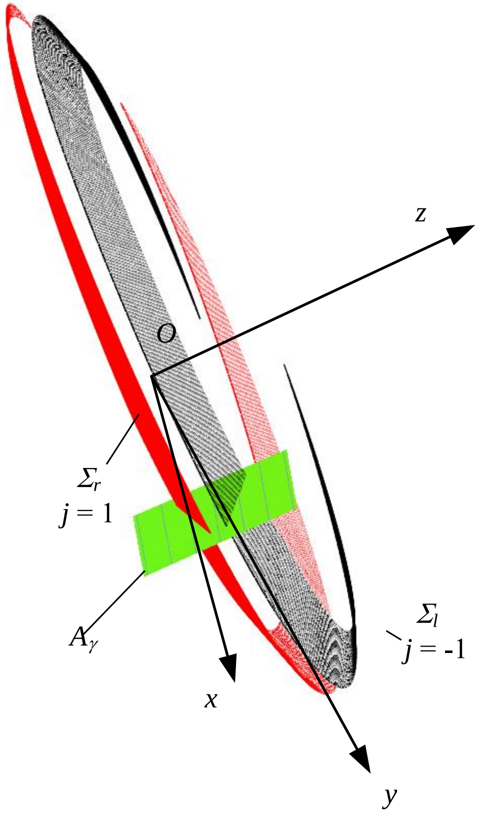

Figure 2: The involute worm thread limiting surfaces and the rake face

$$z + \frac{p_C}{2\pi}\arctan\frac{y}{x} = 0 \tag{7}$$

Now the coordinate functions (4) are implemented in equation (7) that becomes of form $\Phi(\varphi,u)=0$. In order to simplify the expression of the solution, some variable changings are necessary. Thus,

$$\begin{aligned} E_1(\varphi,u) &= \varphi - \psi_0 + \alpha_{0t} + ju = -\psi_0 + (\varphi + \tan\alpha_{0t}) - \tan\alpha_{0t} + \alpha_{0t} + ju = \\ &= -(\psi_0 + \operatorname{inv}\alpha_{0t}) + \varphi_1 + ju = -\zeta + \varphi_1 + ju \end{aligned} \tag{8}$$

and using this form the solution of $\Phi(\varphi,u)=0$ regarding to parameter u is the following:

$$u(\varphi_1; j) = j\frac{p_C\, p_{ax}}{p_C + p_{ax}}(\zeta + \arctan\varphi_1 - \varphi_1) \tag{9}$$

Finally, replacing the parameter u in (4) by the function (9) and using the transformations (8), the unified equations for both edges result as:

$$\begin{cases} x(\varphi_1; j) = R_b\left(\cos B(\varphi_1) + \varphi_1 \sin B(\varphi_1)\right) \\ y(\varphi_1; j) = jR_b\left(\sin B(\varphi_1) - \varphi_1 \cos B(\varphi_1)\right) \;, \\ z(\varphi_1; j) = j\dfrac{p_C p_{ax}}{p_C + p_{ax}}\left(\zeta + \arctan \varphi_1 - \varphi_1\right) \end{cases} \tag{10}$$

$$B(\varphi_1) = \frac{p_{ax}}{p_C + p_{ax}}\left(\varphi_1 - \zeta\right) + \frac{p_C}{p_C + p_{ax}} \arctan \varphi_1$$

The edges and the rake face are shown in *Fig. 3*.

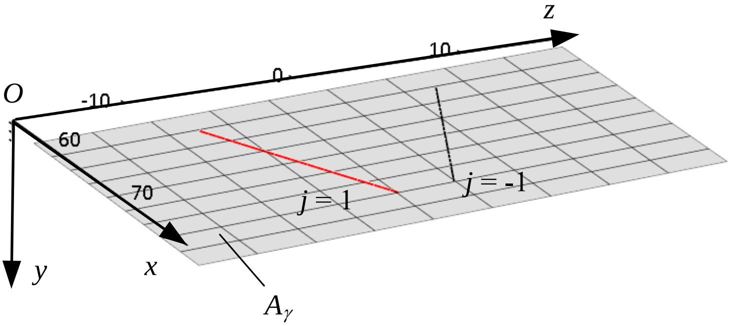

Figure 3: The theoretical cutting edges and the rake face

4. Manufacturing peculiarities

As well as known, relief faces of the gear-hob teeth are obtained by relieving turning and grinding, on the relieving lathe [1, 2, 5, 7, 17]. In order to reduce the manufacturing costs the investments in relieving operations are to be minimized. A fair solution can be obtained if the relief faces are linearized e.g. the generatrix of the surfaces must be a straight line. This leads to the aim of linearizing the cutting edges, while in this case the costs with the relieving cutter and the grinding wheel are considerably reduced. Even in case the precision of the edges are not acceptable still result a rough surface that presents a quasi-equal grinding allowance repartition that is advantageous. As a conclusion, the linearization of the cutting edges produce lowered manufacturing costs or a higher precision of the grinding realized with a wheel dressed about a curved profile.

Let’s suppose that the cutting edge results as intersection of two perpendicular planes

$$x+\alpha_i y+\beta_i y+\gamma_i=0,\ i\in\{1;2\},\ \alpha_i,\beta_i,\gamma_i\in\Re \tag{11}$$

Now let’s consider *N* equidistant points along the edge chosen as subject of linearization. The more the edgepoint is distanced from the line, the absolute value of the right side of equations (11) raises. As a consequence, the objective function can be defined as the sum of squares of left sides:

$$\mathrm{F}(\alpha_i,\beta_{i,},\gamma_i)=\sum_{i=1}^{2}\sum_{j=1}^{N}(x_j+\alpha_i y_j+\beta_i z_j+\gamma_i)^2\to\min \tag{12}$$

The objective function written in this form cannot be used, due to the symmetry related to index *i*. This causes the rank of the linear system of six equations built up of the partial derivatives of F is only 3 while the number of unknowns, $\alpha_i,\beta_i,\gamma_i\ ,i\in\{1;2\}$ is 6. The given situation can be overstepped by applying the Lagrangian multipliers method. Accepting the approximating line resulting as intersection of perpendicular planes, it is obvious that

$$1+\alpha_1\alpha_2+\beta_1\beta_2=0 \tag{13}$$

Using the constraint (13) whose subject is function (12) the new objective function results as

$$\mathrm{F}_1(\alpha_i,\beta_i,\gamma_i,\lambda)=\mathrm{F}(\alpha_i,\beta_i,\gamma_i)+\lambda(1+\alpha_1\alpha_2+\beta_1\beta_2)\to\min \tag{14}$$

The system built up of the partial derivatives of the objective function in 7 unknowns is the following:

$$\begin{cases}\dfrac{\partial\mathrm{F}_1}{\partial q}=0,q\in\{\alpha_i,\beta_i,\gamma_i,\lambda\},i\in\{1;2\}\\[2ex]\dfrac{\partial\mathrm{F}_1}{\partial\lambda}=0\end{cases} \tag{15}$$

This is a nonlinear system that is recommended to be handled numerically.

In order to minimize the number of iterations the initial position of the intersected planes is chosen in the vicinity of the edgepoint situated on the pitch cylinder. Let’s denote R_0 the radius of the pitch cylinder. Using the *x* and *y* coordinate functions (10), the value of the parameter φ_1 results as

$$\varphi_1^{(0)} = \sqrt{\frac{R_0^2}{R_b^2} - 1} \tag{16}$$

A good first approximation for the normal vectors of the planes can be deduced involving the geometrical elements presented in *Fig. 4*. Let's denote *A* the intersection point of the pitch cylinder generatrix $\kappa_0 - \kappa_0$ with that tangent line of the rake face's pitch helix $\tau_0 - \tau_0$ that intersects the axis *x*.

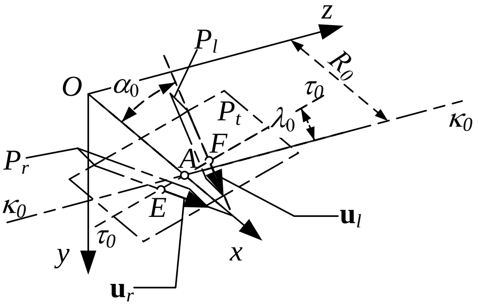

Figure 4: The geometrical elements involved in the computing of the first solution

The plane P_t is built on the axis *x* and the pitch helix tangent. Planes P_r respectively P_l are perpendicular to P_t intersecting this along the profile lines of the normal rake, declined by α_0 to the axis *x*. Thus $AE = AF = \dfrac{\pi m_n}{4} = a$, the coordinates of points *E* and *F* with respect of the values -1 and 1 of the switch parameter *j* are

$$\underline{\mathbf{r}} = \begin{pmatrix} R_0 & -ja\cos\lambda_0 & -ja\sin\lambda_0 \end{pmatrix}^T \tag{17}$$

In the same way the direction unit vectors of the profile lines can be written as

$$\underline{\mathbf{u}}_{r,l} = \begin{pmatrix} \cos\alpha_0 & -j\sin\lambda_0\sin\alpha_0 & j\cos\lambda_0\sin\alpha_0 \end{pmatrix}^T \tag{18}$$

Using the normal unit vector of the plane P_t of coordinates

$$\underline{\mathbf{n}}_t = \begin{pmatrix} 0 & -\cos\lambda_0 & -\sin\lambda_0 \end{pmatrix}^T \tag{19}$$

and expression (18), the normal vectors of the planes P_r respectively P_l can be computed as the cross product $\mathbf{n}_t \times \mathbf{u}_{r,l}$. Using the coordinates (17) of the points *E* respectively *F* the equations of the planes mentioned before can be written. Applying elementary transformations to these the first approximation of values $\alpha_i, \beta_i, \gamma_i,\ i=\overline{1,2}$ will be obtained.

5. The distribution of the distances from the theoretical edge to the approximant optimum.

Let's denote the solution of the system (16) built up for the left or the right edge with $\left(\alpha_1^0, \beta_1^0, \gamma_1^0, \alpha_2^0, \beta_2^0, \gamma_2^0\right)$. The value of λ is not affecting the position of the best approximant. Now solving the linear system (11) with coefficients replaced by the correspondent values given through the solution of the system (16), the coordinates of the approximant line's characteristic point *M* result as follows:

$$\begin{aligned} x^{(M)} &= \frac{\alpha_1^0\left(\beta_2^0 t+\gamma_2^0\right)-\alpha_2^0\left(\beta_1^0 t+\gamma_1^0\right)}{\alpha_2^0-\alpha_1^0} \\ y^{(M)} &= \frac{\left(\beta_1^0 t+\gamma_1^0\right)-\left(\beta_2^0 t+\gamma_2^0\right)}{\alpha_2^0-\alpha_1^0} \\ z^{(M)} &= t \end{aligned} \tag{20}$$

Here, for the simplifying of the formulae let's accept $t=0$.

The error is defined as the distance from the theoretical edgepoint to the approximant line. Recognizing here the classical analytical geometry problem of the distance from a given point to a line it can be written the distance as the module of the cross product computed with the unit vector **e** of the line and the vector binding the external point *A* of the theoretical edge with an arbitrary point of the line, in this case *M*. Thus, it can be written the distance as

$$d = \left|\mathbf{AM} \times \mathbf{e}\right| \tag{21}$$

where the unit vector's coordinates are

$$\underline{\mathbf{e}} = \frac{1}{\sqrt{\left(\alpha_1^0\beta_2^0-\beta_1^0\alpha_2^0\right)^2+\left(\beta_1^0-\beta_2^0\right)^2+\left(\alpha_1^0-\alpha_2^0\right)^2}}\begin{pmatrix} \alpha_1^0\beta_2^0-\beta_1^0\alpha_2^0 \\ \beta_1^0-\beta_2^0 \\ -\alpha_1^0+\alpha_2^0 \end{pmatrix} \tag{22}$$

6. Numerical results

The mathematical model described above was tested for a gear hob derived from a basic involute worm of one tooth, for a normal module value $m_n = 5\,\mathrm{mm}$ and a normal rack profile of $\alpha_0 = 20°$. Eight values of the pitch helix angle were considered in arithmetic progression, starting from $\lambda_0 = 2°$ till $\lambda_0 = 3°45'$ with an increment of $\Delta_{\lambda 0} = 15'$. The computed distributions of the errors are presented in *Fig. 5*.

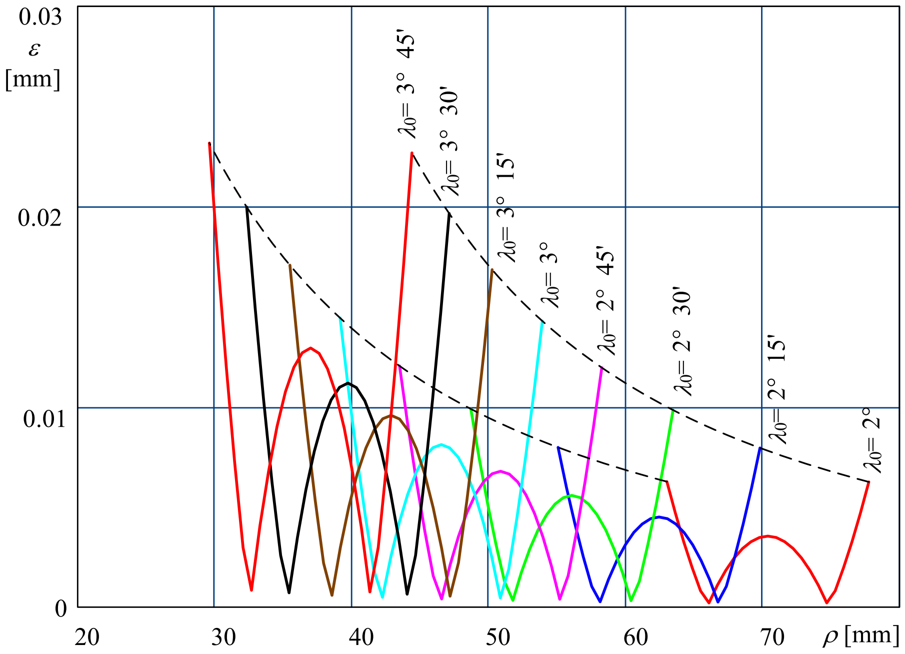

Figure 5: The geometrical elements involved in the computing of the first solution

Studying the shape of the error distribution curves the following remarks can be made:

- the best approximation lines doesn't intersect the theoretical edge;
- the error curves present a local maximum in the vicinity of the theoretical edgepoint situated on the pitch cylinder;
- the error's absolute maximum values are approximately equal and they are situated near the extremities of the theoretical edge;
- the maximum value of the error increases exponential with the pitch helix angle.

If computing the maximum error value for the intermediate value of the helix angle $\lambda_0 = 2°45'$ and for the module values of 1.25, 2.5, 5, 8 and 10 mm (according to DIN780) it will proved that the error is in linear dependence with the module.

The results above confirm the maximum error approximating formula must be of following form:

$$\varepsilon = C\lambda_0^x m^y \tag{23}$$

Using the data deduced on the parameter values described before, formula (23) becomes

$$\varepsilon = 1{,}324841\lambda_0^{2{,}077046}m \tag{23}$$

with a level of Pearson’s *r*-correlation of 0,999990362592.

3. Conclusion

As the mathematical model confirms, the pitch helix angle has a strong influence on the edge profile errors when the linearization is attempted.

The small values of the helix angles, e.g. $\lambda_0 < 2°$ the errors are acceptable, but the pitch diameter of the hob strongly increases that leads to exaggerated material consumption, especially in case of large modules.

By helix angles larger than $2°$, for finishing precision the linearization of the edges, the relieving using straight line profiled grinding wheels is not admitted. Here curved tooth profiles must be machined.

The linearized edges lead to irregular convolute worms, because they are not crossing lines.

The empirical formula can be used for computing the probable value of the maximal error while the edges are linearized. If the admissible value of error is given, formula (23) allows the computing of the maximum value of the pitch helix angle. Using this value, the frontal module and the pitch diameter of the theoretical involute worm can be computed. This led to the smallest possible diameters in case of linearized edges. If the addendum diameter exceeds the maximum admissible value, the decreasing is possible only if curved edges are applied.

References

[1] Liston, K., “Hob Basics. Part I.”, in *Gear technology magazine, vol. 10, nr.5*, 1993, pp. 46-52.

[2] Liston, K., “Hob Basics. Part II.”, in *Gear technology magazine, vol. 10, nr.6*, 1993, pp. 18-24.

[3] Sandeep M., Vijayakar, A. e.a., “Gear Tooth Profile Determination from Arbitrary Rack Geometry”, in *Gear technology magazine, vol. 5, nr. 6*, 1988, pp. 18-30.

[4] Innocenti, C., (2007). “Optimal Choice of a Shaft Angle for Involute Gear Hobbing”, in *Gear technology magazine, vol. 24, nr. 8*, 2007, pp. 42-50.

[5] Yefim, Kotlyar, (2000). “Precision Finish Hobbing”, in *Gear technology magazine, vol. 17, nr.4*, 2000. pp. 22-26. Retrieved on 16.04.2015, from http://www.lmt-tools.de/verzkat/page119.html #/124

[6] *** DIN 3972. Wälzfräser-Bezugsprofile

[7] Hollanda, D., “Bazele așchierii și generării suprafețelor II”. lecture notes, Petru Maior University of Tîrgu-Mureș, 1994.

[8] Dudás, I., Bányai, K., Varga, Gy., “Simulation of meshing of worm gearing”, in *American Society of Mechanical Engineers, Design Engineering Division (Publication) DE 88*, 1996, pp. 141-146.

[9] Dudás, I., Varga, Gy., Bányai, K., “Holonic manufacturing system for production of different sophisticated surfaces”, in *Proceedings of the IASTED International Conference on Modelling, Simulation, and Optimization*, 2004, pp. 72-75.

[10] Varga, Gy., Balajti, Zs., Dudás, I., “Advantages of the CCD camera measurements for profile and wear of cutting tools”, in *Journal of Physics: Conference Series 13* (*1*), 2005, pp. 159-162.

[11] Dudás, I., „The Theory and Practice of Worm Gear Drives”, Penton Press, 2005.

[12] Gyenge, Cs., “Determinarea profilului sculelor abrazive la detalonarea radială a frezelor-melc” in *Tehnologii Calitate Mașini Materiale Vol.7. Bucureşti: Editura Tehnică*, 1990, pp. 92-115.

[13] Gyenge, Cs., “Nagy pontosságú csigakerék-lefejtőmarók tervezése és gyártása” in *Gép 43:(11-12)*, pp. 385-394.

[14] Gyenge, Cs., “A Frenet-féle triéder alkalmazása a csavarfelületek gyártástervezésében”, in *Gépgyártástechnológia 32:(5-6)*, 1992, pp. 191-194.

[15] Litvin, F.L., Fuentes, A., “Gear Geometry and Applied Theory”, Cambridge University Press.

[16] Szeniczei, L., “Általános fogazás”, Műszaki Nehézipari Könyvkiadó, Budapest,1958.

[17] Radzewich, S. P., “Gear Cutting Tools”, CRC Press, NY, 2010.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 9 (2017) 19-32

DOI: 10.1515/auseme-2017-0007

Game Theory Based Radio Resource Management Algorithm for Packet Access Cellular Networks

Mihály VARGA, Zsolt Alfréd POLGÁR

Department of Communications,
Faculty of Electronics, Telecommunications and Information Technology,
Technical University of Cluj Napoca,
e-mail: Mihaly.Varga@com.utcluj.ro, Zsolt.Polgar@com.utcluj.ro

Manuscript received September 10, 2017; revised October 19, 2017.

Abstract: The goal of Radio Resource Management (RRM) mechanisms is to allocate the transmission resources to the users such that the transmission requests are satisfied while several constraints are fulfilled. These constraints refer to low complexity and power consumption and high spectral efficiency and can be met by multidimensional optimization. This paper proposes a Game Theory (GT) based suboptimal solution to this multidimensional optimization problem. The results obtained by computer simulations show that the proposed RRM algorithm brings significant improvement in what concerns the average delay and the throughput, compared to other RRM algorithms, at the expense of somewhat increased complexity.

Keywords: game theory, cellular network, bargaining theory, radio resource management, quality of service, service class

1. Introduction

Today’s wireless packet access communication networks have to deal with a challenging multi-user access issue: a large number of users located in the same geographical area use a large variety of services with various Quality of Service (QoS) requirements, such as voice, video, gaming, web browsing [1], and request high on-demand data rates in a finite bandwidth. Modern broadband wireless systems, such as 3GPP LTE, employ Orthogonal Frequency Division Multiple Access (OFDMA) as the basic multiple access scheme [2]. The OFDMA multiple access technique exploits both time and frequency diversity by allowing both time and frequency domain scheduling of the data packets [2] [3], [4]. Due to this, OFDMA presents the flexibility needed to accommodate many users with a broad range of services, bit rates, and QoS requirements. Several studies on time and frequency domain packet scheduling have been carried out in the last years [4], [5], [6]. Spatial multiuser OFDM based access techniques were considered also in [7].

The design of the RRM algorithms should consider that the traffic generated by the users is a mixture of Real-Time (RT) and Non-Real-Time (NRT) traffic, the parameters characterizing these types of traffic being presented in [8]. The purpose of these algorithms is to divide the network resources among the concurrent transmissions initiated by the users, subject to low complexity and power consumption, low call blocking probability, efficient spectrum usage, and high system capacity constraints. The mentioned issues are important both in cellular networks and in Wireless Local Area Networks [9]. The RRM entity also has to perform the selection of the Modulation and Coding Schemes (MCS). Adaptive MCS selection algorithms in fading affected and peak power limited radio channels are proposed in [10].

Assigning the transmission resources to the users of the network while fulfilling both network and service related performance criteria requires the definition of appropriate utility functions. The RRM entity will target the maximization of these functions and by this process, the optimal or close to optimal resource allocation to the users can be achieved. In [11] the authors propose a network utility maximization mechanism for optimizing multicast transmissions taking place in WLANs.

The design of RRM algorithms in OFDM cellular systems has attracted a lot of attention in the last years [4], [5]. The trade-off between spectral efficiency and fairness among users is one of the most challenging tasks and several papers propose RRM solutions for OFDMA networks based on “negotiation” strategies, thus transforming RRM into a game theory problem [12]. GT based RRM mechanisms have the potentials to achieve fairness between users while maximizing the overall system capacity [13], but its drawback is the increased complexity. The Nash Bargaining Solution (NBS) is considered in [14] together with coalition to find an optimal agreement among negotiating users.

This paper proposes a Bargaining Game (BG) theory based RRM algorithm for packet access cellular network, capable of ensuring the QoS requirements of RT and NRT type of services. Also, the appropriate utility functions are defined for each type of traffic considered. The paper is organized as follows: Section 2 presents the system model, Section 3 describes the modeling of the RRM process as a bargaining game and proposes the traffic dependent utility functions and Section 4 describes the proposed GT based RRM algorithm as well as the constrained optimization based RRM algorithm used as reference. The simulation scenarios, the numerical results obtained by the performed computer simulations and the analysis of these results are presented in Section 5, while Section 6 concludes the paper.

2. System model

The system model presented in *Fig. 1* consists of a cellular network with a variable number of users which access various RT and NRT services. The cell's access node (the eNB in 4G networks) runs the RRM algorithm responsible for the scheduling of the user's transmissions and the allocation of the transmission resources (divided into units called Resource Blocks (RB)) to the scheduled users. The scheduling process is executed during each Transmission Time Interval (TTI) and the link adaptation process, i.e. the selection of the modulation and coding scheme for each user, is a preliminary step of each scheduling round. The allocated RBs and the results of the link adaptation are signaled to the users on the specific control channels. Only due to evaluation reasons an OFDMA access technique is considered with transmission resources partitioned both in frequency and time domain and the RB represents the smallest time-frequency resource unit that the scheduler can assign.

It is considered that one user has only one running service at a given moment and that the type of the service is known by the scheduler. Note that a user device may run more than one service at a given moment and in this case, the system will consider each service, run by a given user, as a separate user having the same geographical position, the same speed, and motion pattern.

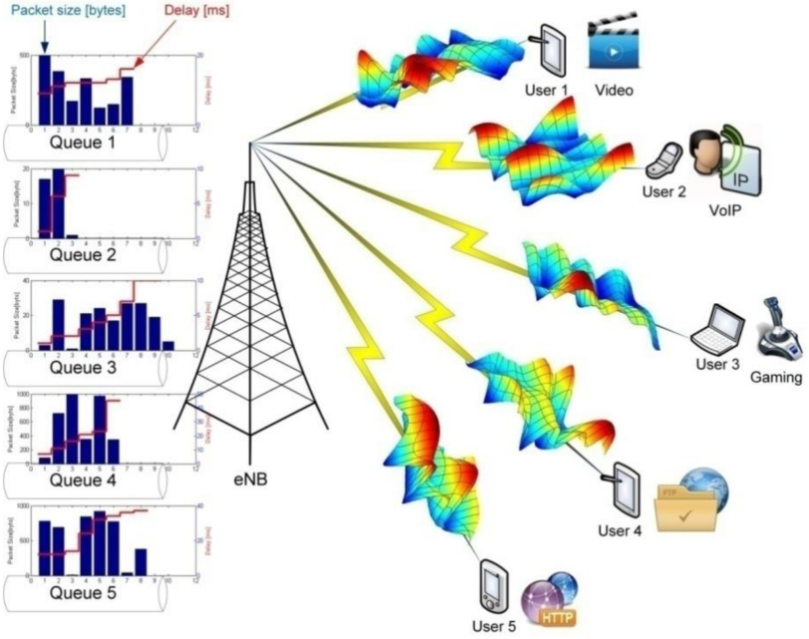

Figure 1: System model

In the cell's central node each user has an individual FIFO queue where the data packets are stored before transmission. The queue stores also information about the data packets (see *Fig. 1*) such as time stamp, packet length, type of service, information which constitutes the Queue State Information (QSI).

In order to perform the scheduling, the RB allocation, and the link adaptation the RRM process should exploit the information that characterizes the wireless links (Channel Quality Information (CQI)). The acquisition and representation of CQI are performed according to specifications given in [15].

3. RRM process modeling and definition of the utility functions

A. RRM process as a Bargaining Game

Let be $\mathbf{K}$ the set of indexes of the N active users located in a given cell, $|\mathbf{K}| = N$, and let be $k, k \in \mathbf{K}$, the index of an individual user. It is considered that each active user is represented by an *agent* which tries to fulfill the QoS requirements of the user's transmission while using the minimum number of RBs. By $\mathbf{RB} = \mathbf{RB_1} \ldots \cup \mathbf{RB_k} \ldots \cup \mathbf{RB_N}$ is denoted the set of available resource blocks and by $\mathbf{RB_k}$ the set of RBs assigned to user k. Let $\mathbf{A_K}$ denote the set of all possible agreement alternatives $\mathbf{a}$, each agreement being represented by the set of RBs allocated to each user, i.e. $\mathbf{a} = \{\mathbf{RB_1}, \ldots, \mathbf{RB_k}, \ldots, \mathbf{RB_N}\}$. Each agent has an upper bounded utility function $u_k(\mathbf{a}) : \mathbf{A_K} \rightarrow \mathbf{R}$ which describes the satisfaction of the user k if the negotiation result is agreement $\mathbf{a}$. The set of all utility functions that result from an agreement is denoted by $\mathbf{S_K} = \{u_1(\mathbf{a}), \ldots, u_i(\mathbf{a}), \ldots, u_N(\mathbf{a})\} \subset \mathbf{R}^N$, a non-empty convex and closed set [14]. If the agents fail to reach an agreement, then by $\mathbf{D}$ is denoted the outcome of this situation and by $\mathbf{d_0} = \{u_1(\mathbf{D}), \ldots, u_i(\mathbf{D}), \ldots, u_N(\mathbf{D})\} \subset \mathbf{R}^N$ the set of utilities achieved by the agents in this situation, referred as "*disagreement point*" [14]. The tuple $(\mathbf{S_K}, \mathbf{d_0})$ defines a bargaining problem. A mapping $f(\mathbf{S_K}, \mathbf{d_0}) \rightarrow \mathbf{A_K}$ is a Nash Bargaining Point (NBP) if some axioms presented in [14] are satisfied.

Let $\mathbf{A^0} = \{\mathbf{a} \in \mathbf{A_K} \mid \forall k, u_k(\mathbf{a}) \geq u_k(\mathbf{D})\}$ represent the set of agreements for which *all* agents achieve at least their minimum utilities (considered in this case to be the utilities from set $\mathbf{d_0}$. Let $\mathbf{J} = \{k \in \{1, \ldots, |\mathbf{K}|\} \exists \mathbf{a} \in \mathbf{A^0}, u_k(\mathbf{a}) \geq u_k(\mathbf{D})\}$ denote the set of users able to achieve a performance greater than or equal to their minimum performance. In this situation a unique NPB exists [14], [16]:

$$f(\mathbf{S_K},\mathbf{d_0}) = \arg\max \prod_{k\in J} \left(u_k(\mathbf{a}) \geq u_k(\mathbf{D})\right) \tag{1}$$

B. Utility functions for RT and NRT traffic

In the case of delay sensitive traffic, the time spent by a packet in the transmission chain is the main parameter which influences the QoS of the transmission. Let be $L(\mathbf{RB_k},\mathbf{CQI_k})$ the function which returns the number of payload bits nb_k which can be carried by the set of $\mathbf{RB_k}$ resource blocks assigned to user *k*. The $\mathbf{CQI_k}$ parameters of the RBs select the MCS schemes.

We denote by $\mathbf{T_k} = \left\{T_k^1, T_k^2, ..., T_k^{Np_k}\right\}$ the set of delays accumulated by the packets of user *k* in the transmission queue. Np_k represents the number of packets waiting in the queue and the set $\mathbf{B_k} = \left\{B_k^1, B_k^2, ..., B_k^{Np_k}\right\}$ represents the lengths of these packets. If we suppose that during several consecutive TTIs (with duration t_{TTI}) the instantaneous bit rate remain constant, the expected values of the delay, Te_k^j, accumulated in the network by a packet *j* of user *k* is:

$$Te_k^j = T_k^j + \left\lceil \sum_{i=1}^{j} B_k^i \Big/ nb_k \right\rceil \cdot t_{TTI} \tag{2}$$

Denoting by $\tau_k = \max_{j=1,..,Np_k}(Te_k^j)$ the maximum value of the expected delay we propose for delay sensitive (RT) traffic the following utility function:

$$u_k\left(\mathbf{RB_k},\mathbf{CQI_k},\mathbf{QSI_k}\right) = 10^{-\frac{\tau_k}{c_k}} \tag{3}$$

where $\mathbf{QSI_k} = (\mathbf{T_k},\mathbf{B_k})$ represents the Queue State Information of user *k*, and c_k characterizes the priority of the service accessed by user *k*.

In the case of delay tolerant traffic, the main parameter which influences the satisfaction of the user is the average call throughput. Let R_k^{call} denote the number of bits received by user *k* during the current call and t_i^{call} is the time elapsed from the beginning of the call. The instantaneous value of the average call throughput is given by (4) and the proposed utility function is given by (5):

$$R_k(\mathbf{RB_k},\mathbf{CQI_k}) = \frac{R_k^{call} + L(\mathbf{RB_k},\mathbf{CQI_k})}{t_k^{call}} \tag{4}$$

$$u_k\left(\mathbf{RB_k},\mathbf{CQI_k},\mathbf{QSI_k}\right)=\tanh\left(\frac{R_k\left(\mathbf{RB_k},\mathbf{CQI_k}\right)}{R_k^{av}}\right) \tag{5}$$

where R_k^{av} represents the average bit rate.

4. RRM algorithms based on BG and constrained optimization

A. Initial resource allocation

The proposed initial resource allocation algorithm represents the starting point of the bargaining process. This operation is implemented as a modified Round Robin algorithm which assigns to each user a number of RBs proportional to the ratio between the number of bits in the user's queue and the total number of bits waiting to be transferred to all users. The initial allocation algorithm assigns each available RB, but ignores the CQIs associated to the RBs.

$$\left|\mathbf{RB_k}\right| \approx |\,\mathbf{RB}\,| \cdot \sum_{j=1}^{Np_k} B_k^j \Bigg/ \sum_{i=1}^{|\mathbf{K}|} \sum_{j=1}^{Np_i} B_i^j \tag{6}$$

Algorithm 1 Initial resource allocation based on Round Robin algorithm

```
for $i$=1 to |K| do
    compute the initial number of RBs allocated to user $i$, $rb_i$, using (6)
    initialize $\mathbf{RB}_i = \varnothing$
end for
initialize $i$=1
for m=1 to |RB| do
    find an active user for which the number of allocated resource blocks is
    less than the computed number of initial resource blocks
    while $\left|\mathbf{RB}_i\right| \geq rb_i$ do
        $i = (i+1)_{\mathrm{mod}|\mathbf{K}|}$
    end while
    allocate resource block $m$ to user $i$
    $\mathbf{RB_i} = \mathbf{RB_i} \cup RB^m$
    $i = (i+1)_{\mathrm{mod}|\mathbf{K}|}$
end for
```

B. The bargaining game based resource management algorithm

In the proposed algorithm, after the initial assignment, each user $i \in \mathbf{K}$ will negotiate with each of the other users $j \in \mathbf{K}; i \neq j$, thus resulting $|\mathbf{K}| \cdot (|\mathbf{K}| - \mathbf{1})/2$ negotiations. For every pair (i, j), the two users merge the *"owned"* RBs and the agents negotiate to re-divide this set, $\mathbf{RB_{i,j}}$, of resources. For each RB in the set the ratio CQI_i / CQI_j is computed and the set is sorted decreasingly according to this ratio, as presented in *Fig. 2*. The RBs with low indexes in the set have good propagation conditions for the first user and worse conditions for the second user. Vice versa, the RBs with high indexes in the set have better propagation conditions for the second user and worse conditions for the first user. On the RBs at the middle of the sorted set both users experience almost the same CQIs, so it doesn't matter to which user will be allocated. This sorted set is denoted as $\mathbf{R\hat{B}_{i,j}}$, the k^{th} element of this set being $RB_{i,j}^{k}$.

$$\begin{aligned}
&\mathbf{a_{i,j}^{k}} = \left\{ \mathbf{RB_i^k} \cup \mathbf{RB_j^k} \right\}; k = 0, .., |\mathbf{R\hat{B}_{i,j}}| \\
&\mathbf{RB_i^k} = \left\{ RB_{i,j}^{1}, RB_{i,j}^{2}, ..., RB_{i,j}^{k} \right\}; \mathbf{RB_j^k} = \left\{ RB_{i,j}^{k+1}, RB_{i,j}^{k+2}, ..., RB_{i,j}^{|\mathbf{R\hat{B}_{i,j}}|} \right\}
\end{aligned} \tag{7}$$

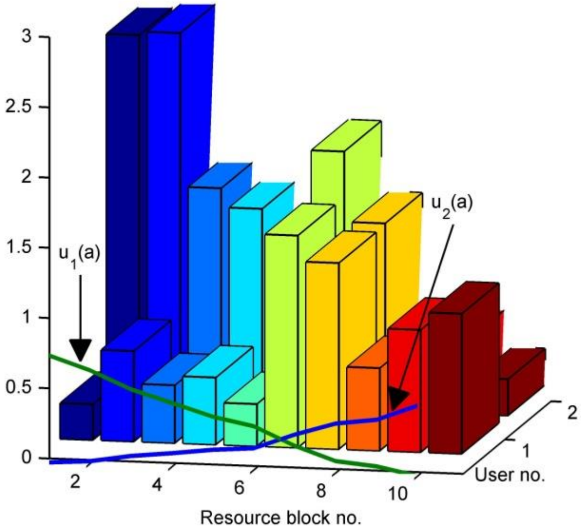

Figure 2: The bargaining process

C. The constrained resource management algorithm

This RRM algorithm adaptively assigns the RBs to the $|\mathbf{K}|$ active users and distributes the total power P_{tot} in order to maximize the ergodic weighted sum rate (8), satisfying the user's minimum rate and fairness requirements [17].

$$U_\gamma = E_\gamma\left\{\sum_{i=1}^{|\mathbf{K}|}\frac{1}{\left(R_i\right)^\alpha}\sum_{m=1}^{|\mathbf{RB}|}\rho_{\mathrm{i},m}\log_2\left(1+p_{\mathrm{i},m}\gamma_{\mathrm{i},m}\right)\right\} \tag{8}$$

where $\boldsymbol{\gamma} = \left[\boldsymbol{\gamma}_\mathbf{1}^\mathbf{T},\ldots,\boldsymbol{\gamma}_{|\mathbf{K}|}^\mathbf{T}\right]^\mathbf{T}$ with $\boldsymbol{\gamma}_\mathbf{i} = \left[\gamma_{i,1},\gamma_{i,2},\ldots,\gamma_{i,|\mathbf{RB}|}\right]$ and $\gamma_{i,j}$ is the effective SNR of user *i* at the *j*th resource block. $p_{\mathrm{i},m}$ denotes the power allocated to the user *i* on resource block *m*, $\rho_{\mathrm{i},m} \in \{0,1\}$ is an indicator which shows whether resource block RB^m is allocated to user *i* or not. Note that each RB can be assigned to at most one user at a given time, i.e. $\sum_{i=1}^{|\mathbf{K}|}\rho_{\mathrm{i},m} \in \{0,1\}$ for all *m*. The function $E_\gamma\{\cdot\}$ represents the statistical expectation with respect to γ, R_i is the user's average call throughput and α is an adjustable fairness parameter. Setting $\alpha = 1$ results in proportional fair allocation, while setting $\alpha = 0$ results in maximum throughput allocation of the available resources.

The constrained optimization problem can be stated as follows [17]:

$$f = \max_{\rho_{i,m},\, p_{i,m}}\left(U_\gamma\right) \tag{9}$$

Subject to:

$$E_\gamma\left\{\sum_{m=1}^{|\mathbf{RB}|}\rho_{\mathrm{i},m}\log_2\left(1+p_{\mathrm{i},m}\gamma_{\mathrm{i},m}\right)\right\} \geq R_i^{av} \;\&\; E_\gamma\left\{\sum_{i=1}^{|\mathbf{K}|}\sum_{m=1}^{|\mathbf{RB}|}\rho_{\mathrm{i},m}p_{\mathrm{i},m}\right\} \leq P_{tot} \tag{10}$$

Algorithm 2 Bargaining game based RRM algorithm

```
Run Algorithm 1 to perform the initial resource allocation to each user
for $i$=1 to $|\mathbf{K}|$ do
    for $j$=$i$+1 to $|\mathbf{K}|$ do
        merge user's $i$ and $j$ resource blocks: $\mathbf{RB}_{\mathbf{i,j}} = \mathbf{RB}_\mathbf{i} \cup \mathbf{RB}_\mathbf{j}$
        sort $\mathbf{RB}_{\mathrm{i,j}}$ decreasingly according to $CQI_\mathrm{i} / CQI_\mathrm{j}$ ratio to obtain $\hat{\mathbf{RB}}_{\mathrm{i,j}}$
        for $k$=0 to $\left|\mathbf{RB}_{\mathbf{i,j}}\right|$ do
            build a possible agreement $\mathbf{a}_{\mathbf{i,j}}^{\mathbf{k}}$ according to (7)
            $\mathbf{RB_i} = \varnothing \cup RB_{i,j}^{1} \cup \cdots \cup RB_{i,j}^{k}$; $\mathbf{RB_j} = \varnothing \cup RB_{i,j}^{k+1} \cup \cdots \cup RB_{i,j}^{\left|\mathbf{R\hat{B}_{i,j}}\right|}$
                compute the difference between the utility functions for $\mathbf{a_{i,j}^{k}}$
            $\psi_k = \left| u_i\left(\mathbf{RB_i}, \mathbf{CQI_i}, \mathbf{QSI_i}\right) - u_j\left(\mathbf{RB_j}, \mathbf{CQI_j}, \mathbf{QSI_j}\right) \right|$
        **end for**
        determine the NBS $k = \arg\min\left(\psi_k\right)$
        $\mathbf{RB_i} = \mathbf{RB_i^{NBS}} = \varnothing \cup RB_{i,j}^{1} \cup \cdots \cup RB_{i,j}^{k}$
        $\mathbf{RB_j} = \mathbf{RB_j^{NBS}} = \varnothing \cup RB_{i,j}^{k+1} \cup \cdots \cup RB_{i,j}^{\left|\mathbf{R\hat{B}_{i,j}}\right|}$
    **end for**
**end for**
```

By solving the problem described by (9), based on the Lagrange dual decomposition framework [17], block RB^m should be assigned to user k_m:

$$k_m = \arg\max_i \left(G_{i,m}\left(\tilde{p}_{i,m}\right) \right) \tag{11}$$

where $\tilde{p}_{i,m}$ is the optimal power allocation (12) and $G_{i,m}\left(\tilde{p}_{i,m}\right)$ is given by (13).

$$\tilde{p}_{i,m} = \max\left(0, \frac{\left(R_i\right)^{-\alpha} + \lambda_i}{\mu \ln 2} - \frac{1}{\gamma_{i,m}} \right) \tag{12}$$

$$G_{i,m}\left(\tilde{p}_{i,m}\right) = \frac{\left(R_i\right)^{-\alpha} + \lambda_i}{\mu \ln 2} \mathrm{e}^{\left(\frac{1}{\tilde{p}_{i,m}\overline{\gamma}_{i,m}}\right)} \cdot \int_1^{\infty} \frac{e^{-\frac{t}{\tilde{p}_{i,m}\overline{\gamma}_{i,m}}}}{t} dt - \tilde{p}_{i,m} \tag{13}$$

where λ_i and μ are the Lagrangian multipliers computed according to:

$$\lambda_i\left(\mu\right) = 2^{R_i^{av}} \frac{\mu \ln 2}{\overline{\gamma}_{i,m}} - \frac{1}{\left(R_i\right)^{\alpha}} \tag{14}$$

The optimum value of μ can be obtained through a one-dimensional search with a geometrical convergence of the convex function:

$$L_{\gamma}\left(\mu\right) = \frac{R_i^{av}}{\left(R_i\right)^{\alpha}} - \frac{2^{R_i^{av}}}{\overline{\gamma}_{i,m}} + \frac{\mu}{\overline{\gamma}_{i,m}} + \mu P_{tot} \tag{15}$$

Algorithm 3 Constrained optimization based RRM algorithm

```
Compute the optimal value of $\mu$ via one dimensional search
for $m$= 1 to |RB| do
    for $i$=1 to |K| do
    compute $\tilde{p}_{i,m}$ using (12) and compute $\lambda_i$ using (14)
    end for
    find user $k_m$ based on (11) to assign $RB^m$
end for
```

5. Numerical results

The evaluation scenario consists of an LTE cellular network (see *Fig. 1*) with 20MHz bandwidth allocated for downlink transmission [15]. An OFDM transmission scheme with 2048 subcarriers is used, out of which 1201 are modulated. The RB is represented by a frequency-time bin of 12 subcarriers and 7 OFDM symbols. The total number of RBs for downlink transmissions is $100/t_{TTI}$. The average speed of the users is 5km/h and each user follows a random walk movement pattern. The used channel model is the WINNER+ urban model [18] [19]. The parameters of the simulation scenarios are presented in Table 1 and the simulations were performed for 10^5 TTIs. As performance indicators, the Cumulative Density Function (CDF) of the packet delays (for RT services) and the CDF of the average instantaneous throughput (for NRT services) are used.

Table 1: The simulation scenarios

	Scenario 1	Scenario 2	Scenario 3
No. of users per cell	100	250	500
Traffic type / users	50% RT (VoIP&Video), 50% NRT (HTTP&FTP)		
R_i^{av} NRT traffic	1Mbps		
c_i RT traffic	10ms		

The CDFs of the delays suffered by the RT type traffic in the considered test scenarios and RRM algorithms are presented in *Fig. 3*. In all cases, the inserted delay has small and moderate values if the GT based RRM algorithm is used, even for a large number of active users in the cell. The constrained RRM algorithm has worse performance in all cases, compared to the GT algorithm.

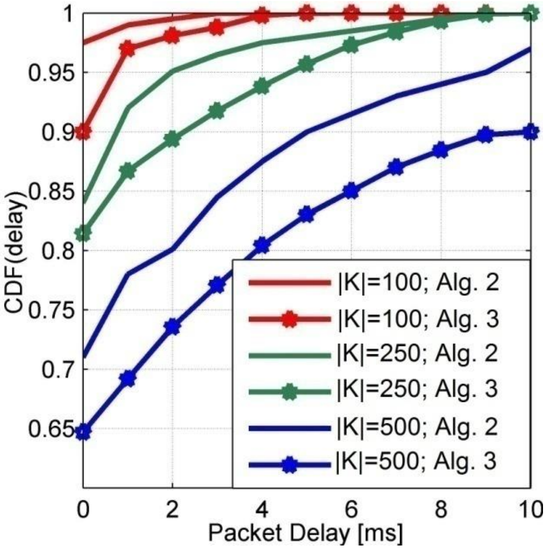

Figure 3: CDF of the packet delays for delay sensitive traffic

In *Fig. 4* it is presented the CDFs of the instantaneous throughput of the NRT type transmissions. The maximum achievable throughput depends on the number of users in the cell. The obtained results show that the GT based RRM algorithm ensures better performance, i.e. larger throughput, for the NRT transmissions.

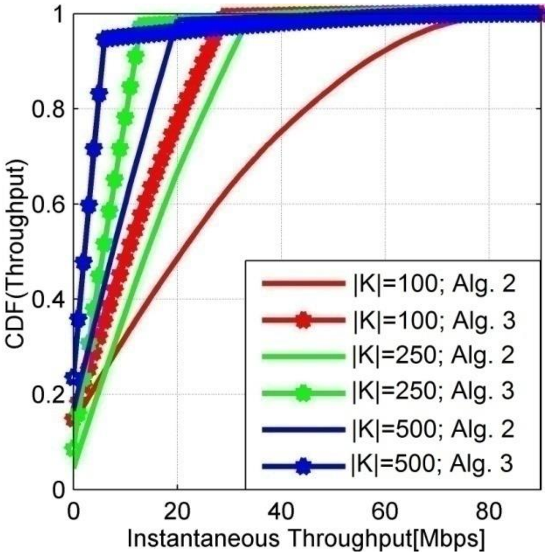

Figure 4: CDF of the instantaneous throughput for best effort traffic

A. Complexity analysis of the RRM algorithms

The RRM algorithm based on constrained optimization needs to run a one-dimensional search to compute the value of μ. After this step, for each user the value of multiplier λ_i and for each RB the power allocation $\tilde{p}_{i,m}$ has to be computed. This means that for every frame $|\mathbf{K}| \cdot |\mathbf{RB}|$ values of λ_i and $\tilde{p}_{i,m}$ should be computed using (14) and (12). Assuming that the search for μ requires I_μ operations the complexity of this algorithm is $O\left(|\mathbf{K}| \cdot |\mathbf{RB}| \cdot I_\mu\right)$.

The proposed GT based RRM algorithm involves $|\mathbf{K}| \cdot (|\mathbf{K}| - \mathbf{1}) / 2$ negotiations and the negotiating agents in each negotiation process share and sort $2|\mathbf{RB}|/|\mathbf{K}|$ RBs, operation which has a complexity of $O\left((2|\mathbf{RB}|/|\mathbf{K}|)^2\right)$. The values of the utility functions have to be computed for all agreements during the negotiation process and the computation of these functions has a linear variation with the number of RBs $O(2|\mathbf{RB}|/|\mathbf{K}|)$. Another search with complexity $O(2|\mathbf{RB}|/|\mathbf{K}|)$ is also necessary to find the NBS point, i.e. the best agreement between negotiating users. The overall complexity can be expressed as:

$$O\left(\left(2\frac{|\mathbf{RB}|^2}{|\mathbf{K}|} + |\mathbf{RB}|\right)\left(|\mathbf{K}| - 1\right)\right) \tag{16}$$

The variation of the required number of operations as a function of the number of active users and the number of resource blocks is presented in *Fig. 5*.

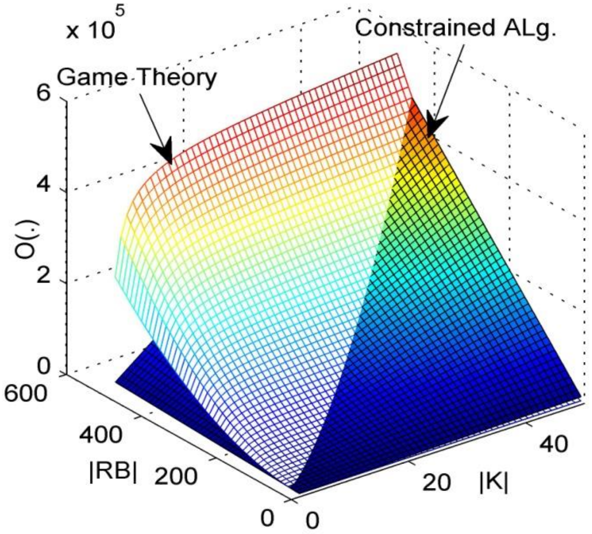

Figure 5: Complexity of the GT RRM and of the constrained RRM algorithms

Conclusion

The paper proposes a game theory based RRM algorithm which targets to find a close to optimal allocation of the transmission resources in cellular networks with OFDMA type multiuser access. The RRM problem in discussion is an NP-hard multidimensional optimization problem. The paper also proposes utility functions for the GT approach, separately for RT and NRT type traffic. The proposed RRM algorithm can ensure the QoS requirements of the user's services while providing high spectral efficiency of the wireless transmissions.

The results obtained using computer simulations show that the proposed RRM algorithm brings significant improvement in terms of average delay and throughput compared to other algorithms, like the constrained optimization based RRM algorithm, at the expense of somewhat larger complexity.

References

[1] Khan, A. H., et al., "4G as a Next Generation Wireless Network", in *Proc. 2009 International Conference on Future Computer and Communication*, *Kuala Lumpur*, *Malaysia*, April 2009, pp. 334-338.

[2] Luwrey, E., "Multiuser OFDM", in *Proc. 5th International Symposium on Signal Processing and its Applications, Brisbane, Australia*, August 1999, pp. 761-764.

[3] Morelli, M., et al., “Synchronization Techniques for Orthogonal Frequency Division Multiple Access (OFDMA): A Tutorial Review”, *Proceedings of the IEEE*, vol. 95, no. 7, pp. 1394-1427, 2007.

[4] Wang, X., and Giannakis, G. B., “Resource Allocation for Wireless Multiuser OFDM Networks”, *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4359-4372, 2011.

[5] Thanabalasingham, T., et al., “Joint Allocation of Subcarriers and Transmit Powers in a Multiuser OFDM Cellular Network”, *in Proc. 2006 IEEE International Conference on Communications*, *Istanbul*, *Turkey*, June 2006, pp. 269-274.

[6] Jang, J., and Lee, K. B., “Transmit Power Adaptation for Multiuser OFDM Systems”, *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171-178, 2003.

[7] Kim, J., and Cioffi, J. M., “Spatial Multiuser Access OFDM with Antenna Diversity and Power Control”, in *Proc. IEEE 52nd Vehicular Technology Conference*, *Boston*, *MA*, Sept. 2000, pp. 273-279.

[8] Esquerta-Soto, J. A., et al., “Performance Analysis of 3G+ Cellular Technologies with Mobile Clients”, *Journal of Applied Research and Technology*, vol. 10, no. 2, pp. 227-247, April 2012.

[9] Kim, S., and Choz, Y. J., “Adaptive Transmission Opportunity Scheme Based on Delay Bound and Network Load in IEEE 802.11e Wireless LANs”, *Journal of Applied Research and Technology*, vol. 11, no. 4, pp. 604-611, August 2013.

[10] Choi, W.-J., et al., “Adaptive Modulation with Limited Peak Power for Fading Channels”, in *Proc. IEEE 51st Vehicular Technology Conference*, *Tokyo*, *Japan*, May 2000, pp. 2568-2572.

[11] Chen, Y., et al., “The Convergence Scheme on Network Utility Maximization in Wireless Multicast Networks”, *Journal of Applied Research and Technology*, vol. 11, no. 4, pp. 533--539, August 2013.

[12] Fudenberg, D., and Tirole, J., “Game Theory”, MIT Press Cambridge, MA, 1991.

[13] Han, Z., et al., “Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining Solutions and Coalitions”, *IEEE Transactions on Communications*, vol. 53, no. 8, pp.1366-1376, 2005.

[14] Touati, C., et al., “Generalised Nash Bargaining Solution for Bandwidth Allocation”, *Computer Networks*, vol. 50, pp.3243-3263, 2006.

[15] 3GPP, “Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding”, 3GPP Tech. Rep. TS 36.212, Release 9, 2010. (online), Available from: http://www.etsi.org/deliver/etsi_ts/136200_136299/136212/09.02.00_60/.

[16] Yaiche, H., et al., “A Game Theoretic Framework for Bandwidth Allocation and Pricing in Broadband Networks”, *IEEE/ACM Transactions on Networking*, ,vol. 8, no.5, pp. 667-678, 2000.

[17] Brah, F., et al., “CDIT-Based Constrained Resource Allocation for Mobile WiMAX Systems”, *EURASIP Journal on Wireless Communications and Networking*, 2009.

[18] Hentia, L., et al.,”MATLAB implementation of the WINNER Phase II Channel Model ver1.1”, December 2007. (online), Available from: https://www.ist-winner.org/phase 2 model.html.

[19] Ikuno, J. C., et al., “System level simulation of LTE networks”, in *Proc IEEE 71st Vehicular Technology Conference*, *Taipei*, *Taiwan*, May 2010.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 9 (2017) 33-42

DOI: 10.1515/auseme-2017-0008

Modelling and Control of Bounded Hybrid Systems in Power Electronics

Áron FEHÉR, Dénes Nimród KUTASI

Department of Electrical Engineering,
Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş,
e-mail: fehera@ms.sapientia.ro, kutasi@ms.sapientia.ro

Manuscript received October 22, 2017; revised December 18, 2017.

Abstract: In this work, an explicit Model Predictive Control algorithm is devised and compared to classical control algorithms applied to a series resonant DC/DC converter circuit. In the first part, a model of the converter as a hybrid system is created and studied. In the second part, the predictive algorithm is applied and tested on the model. Finally, the designed control algorithm is compared to classical PI and sliding mode controllers.

Keywords: hybrid modelling, piecewise affine model, resonant converter, model predictive control

1. Introduction

These days nearly every time/response critical process is controlled with an embedded system. The high demands to achieve reliable performance in complex systems required the development of new control methods. To control a process, the properties of the said process must be known.

The term Hybrid Systems is relatively young in System theory. Systems which belong to this category cannot be categorised as fully continuous, nor fully discrete systems. From this definition we can conclude that hybrid systems are a mix of both, combining continuous and discrete events. Often these systems contain an analogue continuous-time process, some discrete stimuli, and a discrete controller.

A Hybrid system can be described as *Piecewise Affine* (PWA) system, which forms a particular class of nonlinear systems, where each state and output map are piecewise affine, or linear on each polyhedral partition of the state-input polytope [1]. Formula (1) shows the description of a PWA system, where $\underline{x}[k] \in$

$\mathbb{R}^n$, $\underline{y}[k] \in \mathbb{R}^p$, $\underline{u}[k] \in \mathbb{R}^m$ are the state, output and input vectors respectively, A_i, B_i, C_i, D_i, are the system matrices with the appropriate dimensions, f_i, g_i are constant vectors and P_i is the input-state polyhedron for i discrete states.

$$\begin{cases} \underline{x}[k+1] = A_i\,\underline{x}[k] + B_i\,\underline{u}[k] + f_i \\ \underline{y}[k] = C_i\,\underline{x}[k] + D_i\,\underline{u}[k] + g_i \end{cases}, \left(\underline{x}[k] \quad \underline{u}[k]\right)^T \in P_i, i = \overline{1,k} \qquad (1)$$

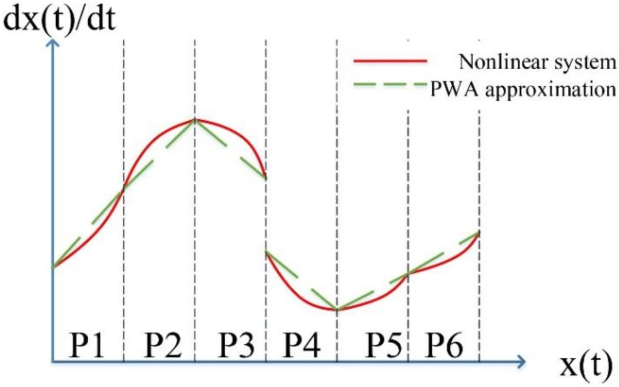

Figure 1: The figure shows the PWA approximation of a non-linear system containing discontinuities such as state changes or boundaries

Another description method is the *Mixed Logic Dynamic* (MLD) *System*, which is computer oriented and is widely used for controller synthesis. An MLD system can be written as shown in Equation (2), where $\underline{x}[k] \in \mathbb{R}^n x\{0,1\}^n$, $\underline{y}[k] \in \mathbb{R}^p x\{0,1\}^p$, $\underline{u}[k] \in \mathbb{R}^m x\{0,1\}^{mb}$ are the new concatenated state, input and output vectors, while $\underline{\delta} \in \{0,1\}^{rb}$, and $\underline{z} \in \mathbb{R}^{rr}$ are auxiliary variables.

$$\begin{cases} \underline{x}[k+1] = A_1\,\underline{x}[k] + B_1\,\underline{u}[k] + B_2\,\underline{\delta}[k] + B_3\,\underline{z}[k] \\ \underline{y}[k] = C_1\,\underline{x}[k] + D_1\,\underline{u}[k] + D_2\,\underline{\delta}[k] + D_3\,\underline{z}[k] \\ E_2\,\underline{\delta}[k] + E_3\,\underline{z}[k] \geq E_1\,\underline{x}[k] + E_1\,\underline{u}[k] + E_1 \end{cases} \qquad (2)$$

2. Series Resonant Converter model

In this section, the modelling of the *Series Resonant DC/DC Converter* (SRC) shown in *Fig*. 2 will be presented.

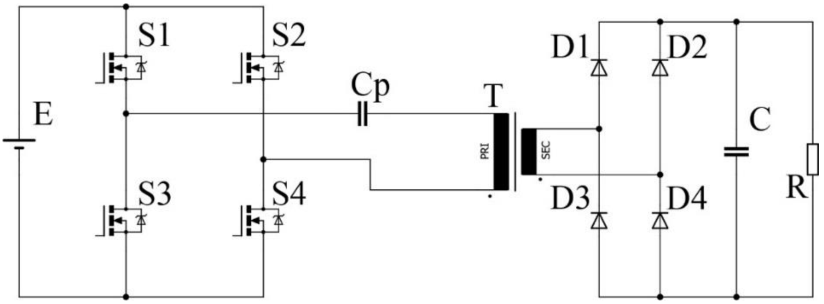

Figure 2: The figure shows an SRC circuit. The resonant tank and the output load acts as a voltage divider. The SRC can operate without load, but in that case, the output voltage can’t be regulated

An equivalent circuit must be devised for the transformer, to create the converter model. The transformer equivalent circuit is deduced neglecting the magnetizing and core loss currents.

Many description methods were successfully used to model resonant converters, for example, discrete time model [2], continuous time model based on averaging methods [3], or with the progressive analysis of circuit waveforms [4]. The models are nonlinear. Hence it is common to use the small signal linearized approximation of the model around the operating point. The problem with linearized model is the invariance to perturbation, input fluctuation and load changes cannot be revealed. We have chosen the hybrid modelling as a description technique of the resonant converter.

The SRC can be represented as shown in *Fig. 3* based on the equivalent circuit of the transformer. The H bridge switches are operated symmetrically (S1 and S4 are on, while S2 and S3 are off; S1 and S4 are off, while S2 and S3 are on).

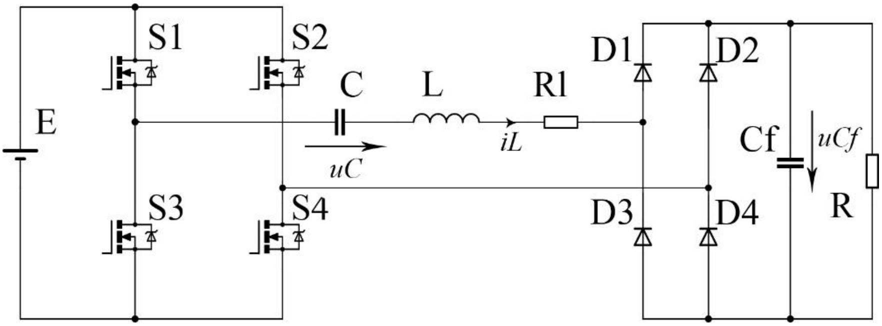

Figure 3: DC-DC SRC circuit with primary side approximate representation of the transformer

Let the state space vector be $\underline{x}(t) = \left(i_L \quad u_C \quad u_{Cf}\right)^T$, the PWA representation is shown in Equation (3), with variables shown in *Table 1.* as discussed in [6].

$$\begin{cases} \dot{\underline{x}}(t) = A_i \underline{x}(t) + f_i \\ i = R(u(t), \underline{x}(t)) \end{cases} \tag{3}$$

The described system must be discretized to design an MPC algorithm. The sampling period must be chosen according to Shannon's sampling theorem. The resonant frequency of the resonant tank can be calculated as $f_n = \dfrac{1}{2\pi\sqrt{LC}}$. In this study, the sampling period is chosen as $T_s = \dfrac{1}{10 f_s}$.

Let us consider the upper mentioned DC-DC SRC system with L=14.7μH, C=560nF, R_L=0Ω, C_f=1mF, E=48V parameters. The resonant frequency of the system is f_n=55.471kHz, so that the sampling period will be T_s=1.8μs. *Fig. 4* shows the open loop operation of the DC-DC converter created with Simscape Power Systems and Hysdel/MPT toolbox simulated with a 10kHz frequency and 50% duty cycle square wave input signal.

Table 1-PWA partition of the SRC based on the semiconductor states

i	A_i	f_i	*R*
1	$\begin{pmatrix} -\frac{R_L}{L} & -\frac{1}{L} & -\frac{1}{L} \\ \frac{1}{C} & 0 & 0 \\ \frac{1}{C_f} & 0 & -\frac{1}{RC_f} \end{pmatrix}$	$\begin{pmatrix} \frac{E}{L} & 0 & 0 \end{pmatrix}^T$	*If u(t)=1&x$_1$(t)>0*
2	$\begin{pmatrix} -\frac{R_L}{L} & -\frac{1}{L} & \frac{1}{L} \\ \frac{1}{C} & 0 & 0 \\ \frac{1}{C_f} & 0 & -\frac{1}{RC_f} \end{pmatrix}$	$\begin{pmatrix} \frac{E}{L} & 0 & 0 \end{pmatrix}^T$	*If u(t)=1&x$_1$(t)<0*

3	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\dfrac{1}{RC_f} \end{pmatrix}$	$\begin{pmatrix} \dfrac{E}{L} & 0 & 0 \end{pmatrix}^T$	*If u(t)=1&x₁(t)=0*
4	$\begin{pmatrix} -\dfrac{R_L}{L} & -\dfrac{1}{L} & -\dfrac{1}{L} \\ \dfrac{1}{C} & 0 & 0 \\ \dfrac{1}{C_f} & 0 & -\dfrac{1}{RC_f} \end{pmatrix}$	$\begin{pmatrix} -\dfrac{E}{L} & 0 & 0 \end{pmatrix}^T$	*If u(t)=0&x₁(t)<0*
5	$\begin{pmatrix} -\dfrac{R_L}{L} & -\dfrac{1}{L} & \dfrac{1}{L} \\ \dfrac{1}{C} & 0 & 0 \\ \dfrac{1}{C_f} & 0 & -\dfrac{1}{RC_f} \end{pmatrix}$	$\begin{pmatrix} -\dfrac{E}{L} & 0 & 0 \end{pmatrix}^T$	*If u(t)=0&x₁(t)>0*
6	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\dfrac{1}{RC_f} \end{pmatrix}$	$\begin{pmatrix} -\dfrac{E}{L} & 0 & 0 \end{pmatrix}^T$	*If u(t)=0&x₁(t)=0*

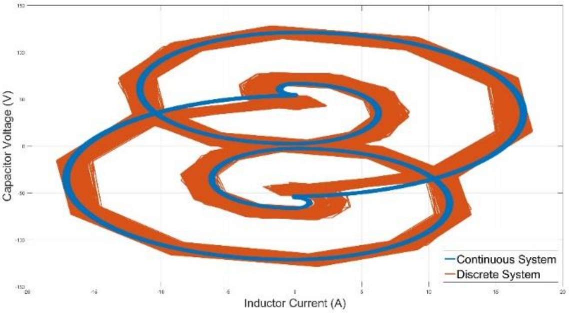

Figure 4: Comparison of the continuous SRC model created in Simscape Power Systems with the discretized version in Hysdel/MPT toolbox. The discrete model closely follows the state trajectory of the continuous model

3. Application of the explicit MPC

The Model Predictive Control (MPC) is a particular case of constrained optimal control which predicts the optimal control signal for a given system for a given horizon. An infinite horizon sub-optimal controller can be designed by repeatedly solving finite time optimal control problems in receding horizon fashion. The resulting controller is referred to as Receding Horizon Controller (RHC). An RHC where the finite time optimal control law is computed by solving the optimization problem online is called MPC [5]. Solving online the *Multi Parametric Quadratic Program* (MPQP) at each time sample is a compute-intensive operation. The explicit version of the MPC developed in order to decrease the computation requirements.

The state boundaries of the hybrid SRC were deduced with a step function as an input signal. The HYSDEL PWA model was imported to MPT (Multi Parametric Toolbox) and was transformed to MLD model, with the boundaries attached to the states, and input as shown in *Fig. 7*. The online MPC was generated by solving the Multi Integer MPQP. For reference tracking, we used linear cost function, because the quadratic one failed to translate to S-function. The explicit version of the controller was generated with the help of the MPT toolbox, and the generated control algorithm was optimized by concatenating polyhedral partitions with the same control signal.

A series of tests were made with the designed controller. In *Fig. 5* the closed loop system is tested with a constant load of 3Ω, while the reference voltage is changed from 10V to 12V (The converter parameters are the same as in the last chapter). We can see, that steady state error is 0 in every case, but the overshoot is present. In *Fig. 6* the reference voltage is kept at a constant 10V while the load is changed. We can see, that the transient overshoot is present, but the controller tracks the reference with zero steady state error. Also, by increasing the prediction horizon, the overshoot decreases and the response time of the controller decreases.

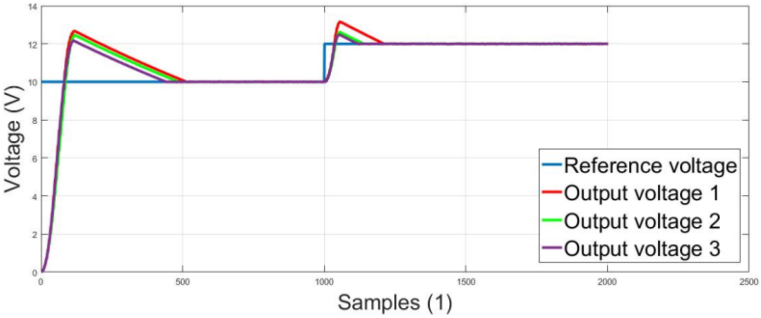

Figure 5: MPC applied to the SRC. In this scenario the reference voltage is perturbed, while the load is constant. The output voltages correspond to a controller with N = 3, 5, and 10 prediction and control horizons respectively

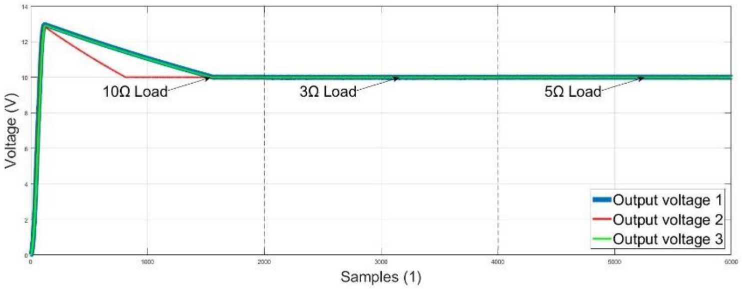

Figure 6: MPC applied to the SRC. In this scenario the reference voltage is constant, while the load is changed. The output voltages correspond to a controller with N = 3, 5, and 10 prediction and control horizons respectively.

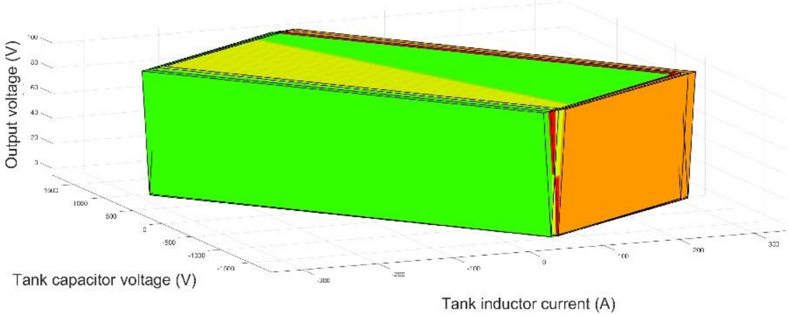

Figure 7: The figures show the reachable set of the SRC in the state space. With an input voltage of 48V, and a transformer ratio of 17:1

4. Comparison of different control methods

For comparison purposes, we designed two controllers for the converter. The first controller was a PI controller tuned to the linearized model of the SRC devised with the help of the Fourier series expansion of the nonlinear terms, keeping only the fundamental ones [7]. *Fig. 8* shows the behaviour of the controlled system, from where we can conclude, the PI controller can be used for a given operating point.

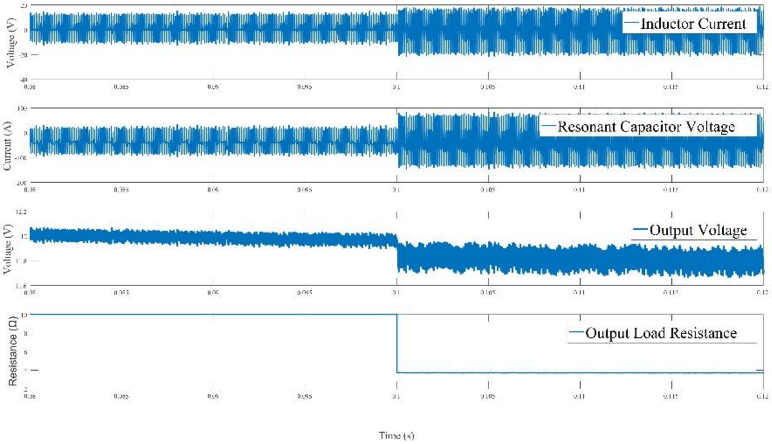

Figure 8: SRC controlled with PI controller. The operating point of the circuit was 48V input voltage, 12V output voltage at 10Ω output load

The second applied controller was a Sliding Mode Controller (SMC) with the first two states as the sliding plane. In *Fig.* 9 the load perturbation, while in *Fig. 10* the reference perturbation is shown. The response time of the closed loop system can be measured with the time constant of the controlled system. In case of the SMC this time is $T_{SMC} = 3{,}2426$ms.

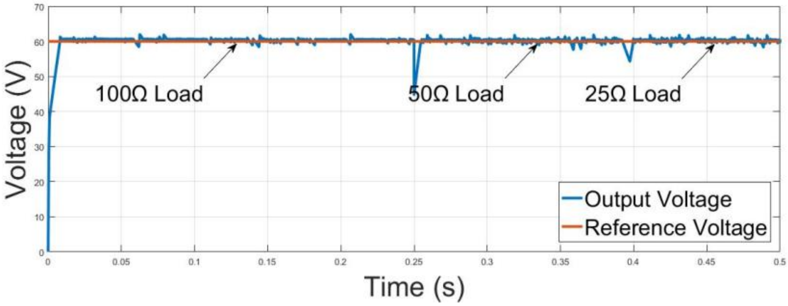

Figure 9: The load stability testing of the SRC

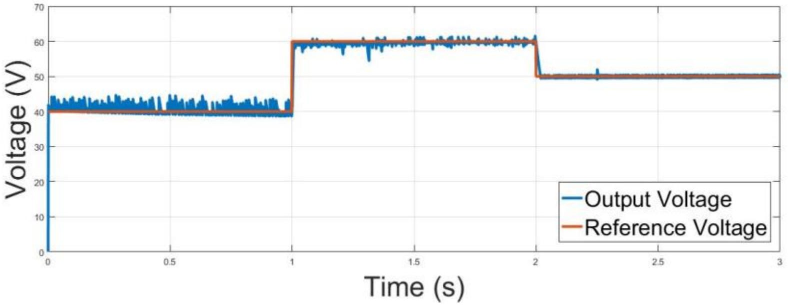

Figure 10: The reference tracking dynamics of the SRC

5. Results and conclusions

In this research we've studied the novel modelling technique named hybrid modelling, which incorporates continuous and discrete event design, to create a more precise representation of the described system. With the hybrid modelling technique, we successfully created a precise representation of the SRC circuit, which was tested and compared against the continuous model (Simscape/Power systems).

With the help of the hybrid model, we created controller structures of basic PI, SMC, and MPC, which were successfully tested both on the hybrid systems and the continuous models.

As final note we can state that the model predictive controller is ideal to track the reference voltage, and to create immunity to parameter perturbations, but not very adequate for fast transient response, and control without overshoot. If we want to achieve the same response time in case of the MPC as in case of the SMC the control horizon time should match the control time of the SMC, this time can be expressed in horizon samples ($f_s * T_{SMC} = 1800$) which would require an infeasible generation time and memory.

References

[1] Lazar, M., Heemels, W. P. M. H., Weiland S., and Bemporad, A., "Stabilizing Model Predictive Control of Hybrid Systems," in *IEEE Transactions on Automatic Control*, vol. 51, no. 11, Nov. 2006, pp. 1813-1818.

[2] Witulski, A. F., Hernandez, A. F., and Erickson, R. W., "Small signal equivalent circuit modeling of resonant converters," in *IEEE Transactions on Power Electronics*, vol. 6, no. 1, Jan 1991, pp. 11-27.

[3] Sun, J., and Grotstollen, H., "Averaged modeling and analysis of resonant converters," *Power Electronics Specialists Conference, 1993. PESC '93 Record, 24th Annual IEEE*, Seattle, WA, 1993, pp. 707-713.

[4] Chung, H. S. H., Ionovici, A., and Zhang, J., "Describing functions of power electronics circuits using progressive analysis of circuit waveforms," in *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 7, Jul 2000, pp. 1026-1037.

[5] Borrelli, F., Bemporad, A., and Morari, M., "Predictive control for linear and hybrid systems", *Cambridge University Press*, 2017.

[6] Afshang, H., Tahami, F., and Molla-Ahmadian, H., "A novel hybrid modeling of DC-DC series resonant converters," *IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society*, Vienna, 2013, pp. 280-286.

[7] Salem, M., Jusoh, A., Idris, N. R. N., and Alhamrouni, I., "Modeling and simulation of generalized state space averaging for series resonant converter", in *2014 Australasian Universities Power Engineering Conference (AUPEC)*,, Sept 2014, pp 1-5.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 9 (2017) 43-55

DOI: 10.1515/auseme-2017-0009

Multilevel Distributed Embedded System for Control of the DC Magnetron Sputtering Process

Albert-Zsombor FEKETE, András KELEMEN,
László JAKAB-FARKAS

Department of Electrical Engineering,
Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş,
e-mail: zsombor.fekete@tetronic.ro, kandras@ms.sapientia.ro, jflaci@ms.sapientia.ro

Manuscript received September 15, 2017; revised October 19, 2017.

Abstract: The paper presents in detail a unique multilevel control architecture designed for the comprehensive management of the DC magnetron sputtering process and of all subsystems of the sputtering equipment. The ultimate goal is to increase the repeatability, stability and the controllability of the complex process. The presented topics include embedded and distributed electronics, data acquisition and supervisory control, networking, data management, redundant local and remote data-archiving. There are presented platform independent algorithms managing the data exchange between computational devices, and conclusions are drawn regarding the efficiency of the various algorithms used.

Keywords: sputtering process, multilevel distributed system, embedded electronics, networking, data management

1. Introduction

The DC magnetron sputtering has been used for several decades for producing a wide variety of thin film coatings, serving purely decorative or more functional purposes, providing versatile enhanced mechanical properties, such as increased wear resistance.

The literature in the field is rich and there exist many valuable contributions to the development of different versions of the magnetron sputtering equipment and of the thin film deposition process [8-10].

The process itself consists of the deposition on a surface, named substrate, of different compounds formed in a chamber with controlled atmosphere. This atmosphere consists of an inert gas (usually Ar) used for bombing a surface named target, particles sputtered from the target, and different reactive gases expected to participate in the composition of the thin film deposited on the substrate. Special power supplies are required to form the plasma containing the bombing ions and particles at different energy levels. High vacuum has to be created in the sputtering chamber and the admission and evacuation of different gases has to be controlled in order to ensure the proper composition of the atmosphere. These control tasks are carried out by customized embedded systems, such as the dynamic pressure, the substrate temperature and the various mass flow controllers. Special measurement devices like vacuum gauges, flow meters, mass spectrometer, thin film growth rate monitor, different temperature sensors, sputtering voltage and current measuring units are used to make possible output feedback and state estimation.

The experimental equipment contains two interconnected vacuum chambers. One of them is the location of the sputtering process, while a significantly smaller secondary chamber provides the ultra-low pressure operating conditions for the mass spectrometer, which is used to determine the partial pressures of the gases and the composition of the gas mixture formed. Both chambers are equipped with rotary and turbomolecular vacuum pumps, water cooling systems, safety elements, as well as several control and monitoring units.

Ensuring the repeatability of the process assumes the presence of a controlled environment, achieved by monitoring and controlling as many process parameters as possible with the help of proper data acquisition and control electronics developed for these purposes.

2. The multilevel distributed control system

Even from the first step of the automation process, it became quite obvious that a multilevel control system [7] was needed for the adequate and comprehensive management of the sputtering equipment, due to the complexity of the process. The aim was to create a modular, easily expandable, well-structured, low cost system, which offers redundant data-archiving and remote access to the various subsystems.

When creating a multilevel control system, beyond the basic requirements for data transfer rates, computational capacity, power consumption, compatibility between different units, there are several other criteria that need to be taken into consideration such as implementation and maintenance costs, delivery time, platform flexibility and development time.

The initial approach was to adapt an industrial network such as ProfiBus or ProfiNet. The first test bench included a Unigate Deutschmann ProfiBus PBDPX - V3704 type slave unit acting as a protocol converter and a Vipa CPU-115 type of programmable logic controller (PLC) acting as master unit and as a central data processing unit. In the second setup, a Unigate Deutschmann

ProfiNet PN - V3804 module and a Siemens S7-300 type of PLC was used. The presented Unigate modules feature galvanic isolation, high reliability, small size, increased electromagnetic noise rejection, reduced development time, simple hardware demands, with only a few external components needed.

Thorough tests were carried out in order to determine the highest data transfer rate of the various units in one specific circumstance: constant length user packets of 30B and full message processing in order to perform protocol conversion. Even though the Unigate embedded units feature a high speed fieldbus interface, the global transfer rate was limited by the slow onboard data processing to 1.5kB/s in case of the PBDPX - V3704 module and to 4kB/s in case of the PN - V3804 module. Another significant disadvantage is the relatively high installation cost (PLCs: 800-900€/unit, Unigate modules: 130€/unit, accessories: 15€/unit). The purchasing time is also high, varying from two to three weeks depending on the supplier`s stock. Based on these results, the proposed architecture has proven to be unuseful in our application field.

In order to fulfill the imposed requirements, a four-level, low cost microcontroller based architecture was proposed which can be seen in *Fig. 1*.

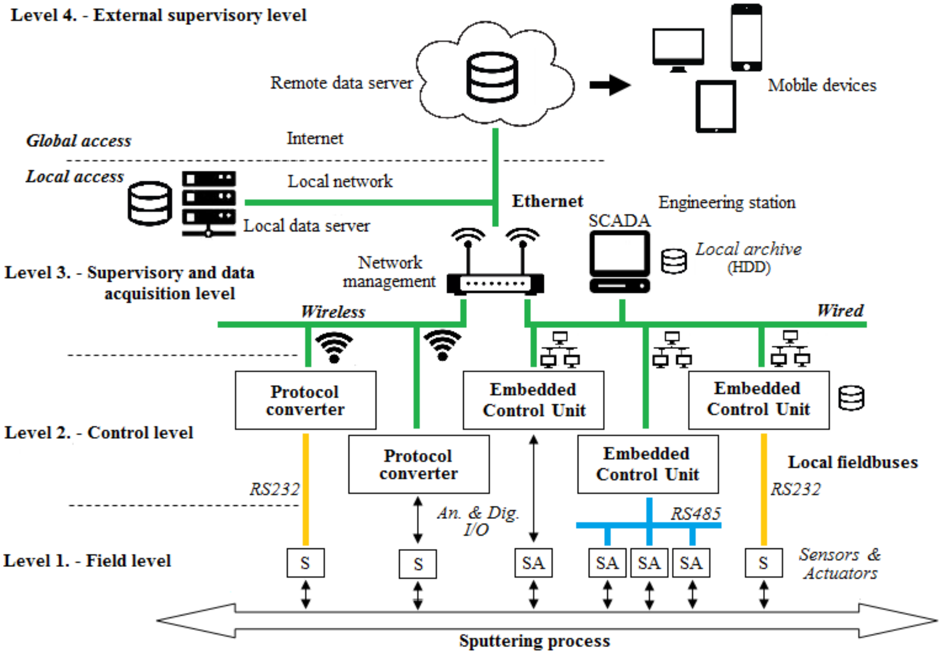

Figure 1: The multilevel distributed system developed for magnetron sputtering

On a small scale, the basic architecture and operation principle of the proposed multilevel control system was previously developed for the mass spectrometer unit [5,6]. Based on the results obtained [5,6] and after the revision of the entire communication stack of the mass spectrometer unit by increasing the modularity of the various algorithms, the small scale version was adapted to the entire DC magnetron sputtering process.

In the presented architecture there can be distinguished four levels: a field level, a control level, a supervisory and data acquisition level and an external supervisory level [1]. All the components can be classified into one of the levels presented. The field level consists of the different vacuum equipments, sensors and actuators which ensure a link between the sputtering process itself and the different embedded control systems situated at the inner level. At this level, generally there are implemented processes like data acquisition, local data processing and fast control (e.g. dynamic pressure controller, substrate temperature controller). At level three, all data and information available can be accessed through different communication channels, enabling local monitoring, data-archiving and the control of slow process parameters through online TCP connection with a dedicated Matlab application capable of executing several complex control tasks. The external supervisory level provides the same features as level three, providing the possibility of access control.

As presented in *Fig. 1*, the supervisory and data acquisition level incorporates a local Ethernet network with star topology [1]. This means that the network is managed by a router [2] enabling wired and wireless connectivity, as well as providing the possibility of online extension, without system shutdown. Also, if one or more nodes (embedded devices) are disconnected from the router, the network remains operational. This gives high reliability to the system in hand [1].

In contrast to the star topology found at level three, at lower levels the majority of the available measuring equipments demanded low speed (9600 and 19200 Baud) local buses and point to point topologies [1]. Typical examples are the different pressure gauges (e.g. Pfeiffer Vacuum MPT100, CMR365 and PKR251) mounted on the vacuum chambers and the digital multimeters (e.g. Appa 103N for measuring the polarization voltage of the substrate and Appa 305 for measuring the temperature of the substrate with a thermocouple). These units or components provide only serial RS232 or RS485 communication interfaces, thus low level, low speed topologies are needed. Furthermore, the baud rates are well defined by the manufacturers and limited to 9600 bps or 19200 bps.

Another typical equipment where the baud rates and communication topologies are restricted by the manufacturers is the Quadrupole Mass Spectrometer (QMS). In order to operate correctly and safely the measuring unit of the QMS, a multi-microcontroller based system must carefully supervise the vacuum pumps, the safety valve, the pressure levels, the various cooling conditions (water flow and temperature, equipment temperature), the power consumptions and the different equipment states. Each safety function is solved with the help of individual embedded units and different sensors connected to a bus type RS485 network [5, 6], featuring a master-slave type of serial communication. In order to increase the safety factor of the installation, cross monitoring is introduced between the units presented above [5]. This means that in case of the malfunction of one or more subunits, the bus network stays functional and the remaining electronics can trigger the shutdown procedure due to the cross-monitoring.

The presented safety mechanism stays functional even if only one unit or subassembly remains operational. The conclusion is that the bus topology in this case provides the necessary conditions to implement the much needed cross-monitoring [1]. As a side benefit, the local network created is easily expandable and the packet collisions are eliminated due to the master-slave type of communication.

Usually a point to point connection [1] is used between a smart field device (e.g. MPT100 full range pressure gauge, APPA 304N digital multimeters, and temperature sensors) and a local data processing and control unit.

The quality of a network strongly depends on the speed, reliability and the security of the data-exchange [1,2]. The speeds of the different networks presented reach from 9600kb/s (field level) to 100Mb/s (supervisory and data acquisition level). The reliability is increased by eliminating ground loops [3] and introducing proper cable shielding [1-3], line isolation and overvoltage protections, as well as different EMI filters in order to assure noise immunity. All the networks use S/FTP type cables, which feature double shielding for noise cancellation. The surge protection mainly consists of transient voltage suppressor (TVS) diodes, which feature low capacitance. This means that high speed data lines are not influenced by the presence of these diodes and every node (every device or unit) can incorporate them [3]. The security of the data-exchange is achieved by restricting global access to different communication channels.

In comparison with the first approach of forming a control network, the presented multilevel architecture presents a lower installation cost (Ethernet network: existent, wired embedded microcontroller units: 30-50€/unit, wireless embedded microcontroller units: 12-20€/unit, accessories: 1-5€/unit). The prices for the embedded units, in contrast to the ProfiBus and ProfiNet modules, include housing, internal wiring, power supplies and local Human-Machine Interfaces.

3. Embedded systems

One of the requirements regarding the development of the embedded systems is to create independent modules with both wired and wireless Ethernet connectivity and with isolated ground planes and power supplies in order to eliminate interdependencies between sensitive measuring circuits such as low voltage (e.g. 800mV) high resolution (e.g. 24bit) analog to digital converters used for Pirani type pressure gauges.

For the various tasks, three types of microcontroller based embedded electronics have been designed and built. Regardless of the chosen architecture, the modules feature a central digital processing unit and all the task specific peripheral circuits (e.g. analog and digital signal conditioner circuits, power electronics) are connected in form of add-on or extension cards. This method provides modularity to the systems in discussion. Every embedded software mainframe uses the same core which means that the development time for each subsystem is optimized and reduced significantly. The main difference between the circuits is the computational capacity of the microcontroller used. For less complex tasks (e.g. protocol converters, sputtering process state monitors, water cooling and substrate heating controllers) there have been used two slightly different ESP8266 based circuits featuring wireless connectivity, onboard RS232, digital I/O pins and a high resolution LTC2410 sigma-delta type analog to digital converter with dedicated reference source. The more complex tasks (e.g. Quadrupole mass spectrometer controller) require a multi-core distributed embedded system [5], based on PIC32MX795F512L microcontroller for wired communication and dsPIC33EP512MC806for high speed signal processing and PID or Fuzzy control algorithms.

Regardless of the architecture used, the software mainframes are constructed using the cooperative multitasking principle [4], limiting the execution time of each section to a maximum of 30ms in order to avoid partial or total lockups inside the software`s vital parts, resulting in an unwanted restart of the microcontroller. Because there exists no higher level arbitration mechanism to monitor the execution times, all the tasks that have to be performed by the microcontroller are distributed in the development phase and are executed based on a predefined order or sequence [4]. The execution timeout was determined empirically. To further conserve valuable microcontroller resources, the data transfer between the internal memory and the communication peripheral is executed by the DMA controller resulting in a 30µs time saving for a 30B packet in every sampling period. The values presented are valid for a 70MIPS, 32bit platform using a sampling frequency of 10kHz. This method is quite useful when a 4kB packet is sent containing the measurement results of the mass spectrometer.

Every connection type featured (TCP, UDP) is implemented in the form of different network services, each having his dedicated socket and port assigned to it. The TCP connection [2] is used to exchange data with the centralized supervisory control application, the UDP connection [2] is used to stream data to a remote data-server for archiving. The different IoT protocols (e.g. MQTT) are utilized to establish unidirectional connection with two separate Cloud type of data-servers. Note that every connection forwarded outside of the local Ethernet network is restricted to a unidirectional connection, only to serve remote monitoring and supervisory functions such as the embedded Web servers.

In order to meet the redundant archiving requirement, data and event archiving is implemented on different levels: control level, local and external supervisory and data acquisition levels. This means that every type of data or event is stored at least on two different computational systems, reducing the risk of irreversible data loss. Note that by moving upwards on the multilevel architecture the storing capacity is increasing. Therefore, on the control level the embedded units only store fault states and important events with or without timestamp. On the other hand, the storage speed is the highest at the supervisory levels, peaking at 1kB/s. The data stored on remote data-servers can be accessed from within the local network or through the Internet with mobile devices as shown in *Fig. 1*.

4. Communication management

By introducing a network connection, it became necessary to develop a unified, well-structured and configurable high level communication protocol which can be used on all embedded systems and computers regardless of their architecture and communication module. The proposed structure is shown in *Fig. 2*.

The developed stack features three layers: user layer, packet management layer and physical layer [6]. When implementing the stack on different platforms, only the physical layer needs to be adapted. The inner layer is responsible for assembling and dissembling the packets, transferring data from and to the memory via DMA, checking for faults and for scheduling the outgoing messages selected by the user.

The top layer is probably the simplest, because the user only needs to set the priority bit of the desired message (‘0’ for excluding the message, ‘1’ for one time and ‘2’ for cyclic transmission). This automated packet forming and sending mechanism assumes the presence of predefined message templates. This communication management protocol enabled us to send measurements periodically and systems events (faults, warnings, states) only upon change.

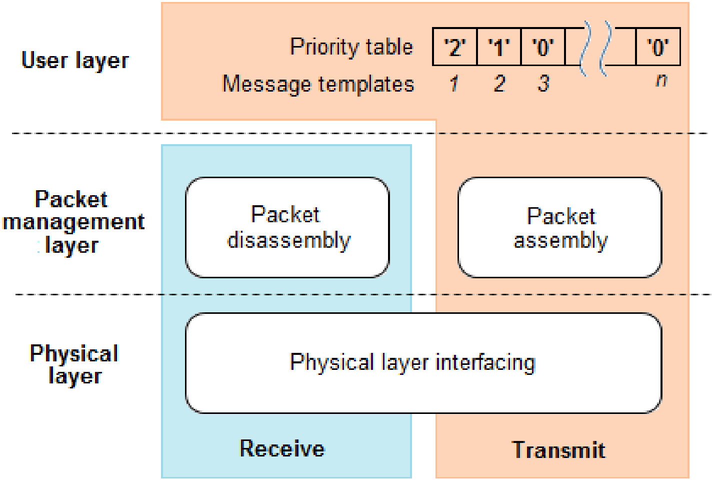

Figure 2: Platform independent communication algorithm structure

5. Supervisory control and data acquisition application

The main idea behind the multilevel distributed control system is to have access to all the data and system parameters at a superior level. This claim assumes the presence of a supervisory control and data acquisition application (SCADA). The mainframe of the application in discussion was developed in CVI Labwindows from National Instruments.

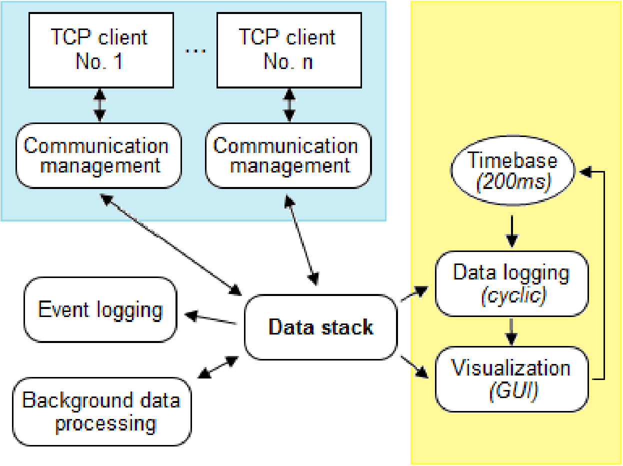

Figure 3: Simplified structure of the mainframe

One of the main criteria regarding the development of the software was to create an efficient application, by distribution of the tasks and by executing them with well-structured algorithms. The simplified structure of the mainframe is shown in *Fig. 3*. The software developed and executed on a process computer has three main functions or features which correspond to the basic components of the application: establishing TCP connection with all the embedded units, data processing and visualization, and data archiving.

Despite of the fact, that all the embedded electronics use the same message structure and communication protocol presented earlier, an iterative TCP/IP model was used. This way, one socket and one port is permanently allocated to every device and connection used. The reason for choosing this type of model is to separate every connection in order to increase program transparency, to implement independent packet processing algorithms and to facilitate the extension procedure of the mainframe with new connections and devices.

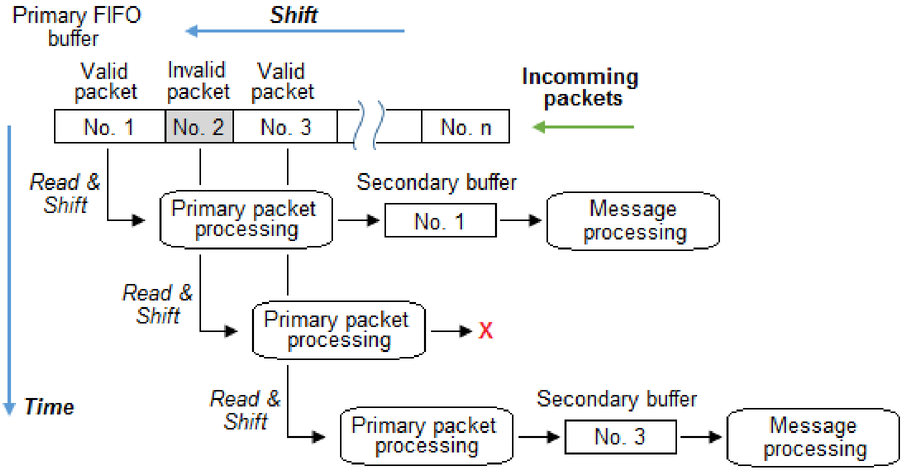

Figure 4: Supervisory control and data acquisition application – TCP buffer structure

It is important to emphasize that every TCP connection uses two separate buffers for storing incoming data messages: primary FIFO buffer, containing the incoming raw packets and a secondary buffer containing valid, demarcated messages which can be individually processed by the corresponding algorithm. The main FIFO type buffer can contain multiple consecutive messages, hence the need for the primary packet processing and the presence of the secondary software buffer or array as presented in *Fig. 4*. The message processing algorithm uses the top and inner layers of the generalized communication management algorithm (*Fig. 2*).

The incoming and derived data are divided into two groups: data that are refreshed periodically and data that are refreshed only upon change. This is available for the corresponding graphical elements of the graphical user interface (GUI) as well. Periodically received data are subcategorized into two classes: fast changing data (e.g. pressures, sputtering voltage and power) and slowly changing data (e.g. cooling and temperature parameters, thin film growth rate), which are determined in function of the variation speed of process parameters. The borderline between the two classes is set to 2s, and the adequate refresh rate is derived from the 200ms time-base. The background data processing is also simplified due to the fact that the majority of the received data is already pre-processed. This method of categorizing the data, the software structure and refreshing particular graphics only if there is a value change contributes to the efficiency of the application by optimizing the utilization of the computer's hardware resources. The average processor utilization on a single core 2.4GHz computer is less than 12%. By turning off the presented method and by keeping the same 200ms refresh and packet rate, the average utilization increases to 75%.

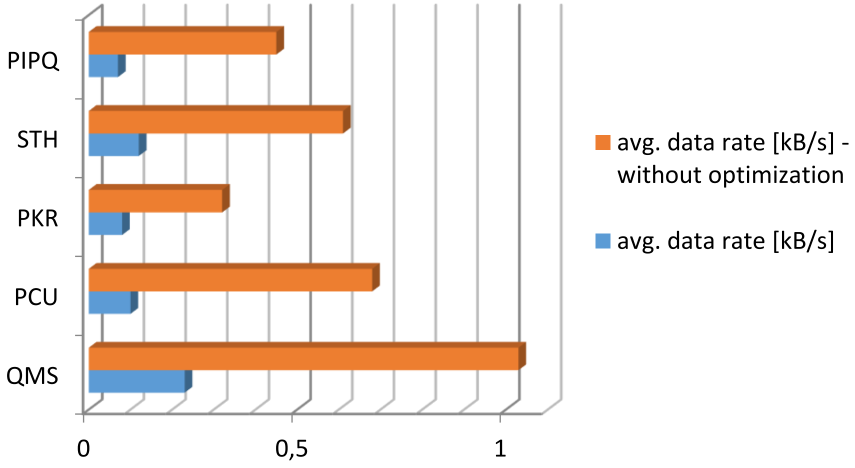

Figure 5: Average data rate with and without traffic optimization

The supervisory application features a built-in TCP traffic monitoring function, with the help of which the user can easily keep track of different network related data traffic. Due to the fact, that the outgoing TCP data rate does not have the same order of magnitude as the incoming data rate, in the following only the latter will be discussed. The positive effect of the optimization presented above is shown in *Fig. 5* through 5 different embedded units (PIPQ – Pirani pressure gauge, STH – Substrate thermometer and controller, PKR – Full range pressure gauge, PCU – dynamic pressure controller, QMS – Quadrupole mass spectrometer). The obtained traffic data presented in *Fig. 5* shows that without the different optimization algorithms, data management and data message types introduced above, the incoming average data rate would increase by approximately 5 times, resulting in unnecessary network and microcontroller CPU load.

The GUI is split into two well delimited zones as shown in *Fig. 6*. Zone 1 is basically a permanent frame which contains graphic elements displaying high priority information like measured and derived data, alarms, events, gauge states, system faults and important system messages. Events which may alter the course of the ongoing experiment and need the immediate attention of the user, such as communication errors, hardware malfunctions, cooling and power failures, process or system parameters out of the allowed operating range.

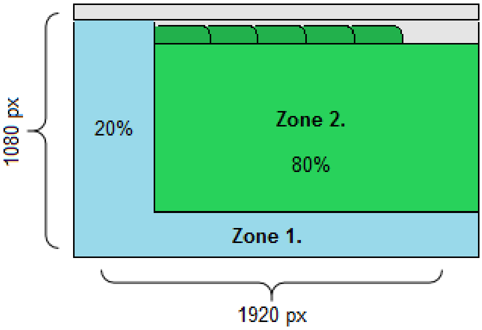

Figure 6: Subdivision of the graphical user interface created for the supervisory control and data acquisition application

All the received and derived data are further categorized in function of the application area (e.g. water-cooling, substrate temperature management) and subsystem (e.g. dynamic pressure controller, Quadrupole mass analyzer, networking, power grid and power distribution, high voltage DC power sources), and they are displayed in separate tabs, which make up Zone number 2. Each tab contains a graphic illustration or process diagram of the sub-process in hand; making it easier to identify the location of the different sensors and actuators, as well as the different systems and components that take part is the construction of the complex sputtering equipment. There are a total of 12 tabs, featuring a main tab containing a user-configurable chart or trend for the various system parameters, with the help of which the user can monitor the variation in time of the different measurements, helping in identifying patterns and dependencies between the process parameters.

The last, and possibly one of the most important features is the local data-archiving. The application creates automatically three types of data tables on the hard drive of the process computer: event log, scan log and process log. The first type contains time stamped system events, states, faults, errors, system shutdowns in chronological order as a result of background data processing. There are well over 100 different event types registered. The log itself has proven to be helpful in case of equipment troubleshooting. The second type is generated only when there is a valid scan result regarding the composition of the gas mixture inside the sputtering chamber. The third type of log contains measurements and derived values. Every 200ms a new entry is created with over 35 selected system and process parameters/states. The application is archiving locally at a rate of 1kB/s. The event and process logs are created 24 hours a day and contribute to the redundant data-archiving presented earlier. These structured logs are indispensable in the offline data processing, based on which mathematical models are created and fine-tuned with the goal of better understanding the complex phenomenon during a sputtering process.

6. Conclusion

The developed multilevel control system has proven to be a valuable asset in the comprehensive management of the complex reactive magnetron sputtering process and of all its subsystems, providing a stable environment for both online and offline data processing, as well as for the control and the monitoring of various interdependent process parameters and states.

Taking advantage of the fact, that the distributed system can be easily updated by connecting online new systems to it, the developed embedded control units, as well as the multilevel architecture ensures a modular and expandable platform. The ideas regarding the presented software, network topology and the implemented electronics can be used for the control of other complex processes.

The distributed system incorporates a total of 11 embedded control and data acquisition units developed over the last 6 years, monitoring 30 process parameters and measurements, over 80 electronic system states and parameters, approximately 200 systems event messages and controlling a total of 18 processes and system parameters combined. The constantly improving thin film structure and composition properties, as well as the increased process reproducibility justify the necessity and the usefulness of the system in discussion.

It is important to emphasize that the system developed can serve educational purposes as well, granting a fully functional environment, where multilevel control systems, embedded hardware, microcontroller programming, data acquisition, supervisory control and remote management can be tested under real circumstances.

Acknowledgements

The authors thank the members of the Thin Solid Films Research Group from the Sapientia Hungarian University of Transylvania, Domokos Biró, Sándor Papp and István Szöllösi for their indispensable contributions to the development of the research work presented.

References

[1] Westermo Handbook 5.0, "Industrial Data Communication – Theoretical and General Applications", Westermo, Sweden, 2004.

[2] Westermo Handbook 5.0, "Industrial Data Communication – Industrial Ethernet", Westermo, Sweden, 2004.

[3] Kugelstadt, T., "Protecting RS-485 Interfaces Against Lethal Electrical Transients", in *Application report, SLLA292A-May 2009-Revised March 2011*.

[4] *** TCP/IP application note: Microchip TCP/IP Stack Help (version: 5.42.08 - 2013).

[5] Fekete, A. Zs., Jakab-Farkas, L., "Development of an Embedded System for Accessing Mass Spectrometry Measurements through Ethernet Network", *in Proceedings of the XXIth International Scientific Conference of Young Engineers*, Cluj-Napoca, Romania, March 17-18, 2016, pp. 161-164.

[6] Fekete, A. Zs., "Automation of the reactive magnetron sputtering process", *in Proceedings of the XVIIth International Conference of Technical Sciences*, Cluj-Napoca, Romania, November 26, 2016, pp. 79-84.

[7] Kopetz, H., "Real-Time Systems: Design Principles for Distributed Embedded Applications", Second Edition, Springer Science + Business Media, 2011.

[8] Jonsson, L. B., Nyberg, T., and Berg, S., "Target compound layer formation during reactive sputtering", *J. Vac. Sci. Technol. A 17(4)*, Jul/Aug 1999, pp. 1827-1831.

[9] Görgy, K., „Cercetări privind dezvoltarea unor electrotehnologii pentru depunerea straturilor metalice subţiri", *Teză de doctorat, Universitatea Tehnică din Cluj-Napoca*, 2010.

[10] Berg, S., Katardijev, I. V., „Preferential sputtering effects in thin film processing", *J. Vac. Sci. Technol. A 17(4).*, Jul/Aug 1999, pp. 1916-1925.

DE GRUYTER OPEN

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 9 (2017) 57-77

DOI: 10.1515/auseme-2017-0010

Solving of the Modified Filter Algebraic Riccati Equation for H-infinity fault detection filtering

Zsolt HORVÁTH[1], András EDELMAYER[2,3]

[1] School of Postgraduate Studies of Multidisciplinary Technical Sciences, Faculty of Technical Sciences, Széchenyi István University, Győr, e-mail: zsolt2.horvath@audi.hu
[2] Department of Informatics Engineering, Faculty of Technical Sciences, Széchenyi István University, Győr, e-mail: edelmayer@sze.hu
[3] Systems and Control Laboratory, Institute for Computer Science and Control, Hungarian Academy of Sciences, Budapest, e-mail: edelmayer@sztaki.mta.hu

Manuscript received September 29, 2017; revised December 18, 2017.

Abstract: The objective of this paper is solving of the Modified Filter Algebraic Riccati Equation (MFARE) for calculating of the filter gain. The results are used for model-based fault detection filtering of faults in the air path of diesel engines. The H-infinity optimization approach requires the solution of a linear-quadratic optimization problem that leads to the solution of MFARE. In our paper two basic concepts for solving MFARE are examined, namely the analytically implemented gamma-iteration and casting the problem as a convex optimization problem based on Linear Matrix Inequalities (LMIs).

The algorithms are implemented in MATLAB. Each algorithm has to ensure the condition for a global convergence and also has to deliver an optimal solution. Not at least, the computational cost has to be as small as possible.

Keywords: modified Filter Algebraic Riccati Equation, linear-quadratic optimization problem, H-infinity optimization, gamma-iteration, LMI

1. Introduction

With the increasing complexity of combustion engines in current automotive vehicles, the early detection of failures for engine diagnostics plays an increasingly important role. Possible faults are due to actuator, sensor and component failures, which can lead to engine malfunctions or even damages in the worst case. The subject of our investigation is a robust model-based fault detection filtering of faults in the air path of diesel engines. The filter robustness is ensured by the application of a design trade-off that is made between the worst-case disturbance and the L_2 norm of the filter error. This method requires the solution of a linear-quadratic optimization problem that leads to the solution of the Modified Filter Algebraic Riccati Equation (MFARE), see e.g. in [1], [2], [3] and [4].

Combustion engines can typically be characterized by highly nonlinear processes that may have very fast dynamics. This property poses additional requirements for the fault detection filter implementation. On the one hand, the filter should be capable of running recursively, in real-time, in few millisecond cycles, by taking the constrained computational capability of on-board microcontrollers into account. On the other hand, the computational complexity of the model might need processing power usually not available for the specific application. For this reason, finding an efficient algorithm to an optimal solution of the MFARE, which is definitely the core of the fault detection filter, is of great importance.

Several investigations have been carried out in the past two decades for using LMI to issues of robust control see e.g. [5], [6], [7]. So, it has been already proven, that LMI-s are effective and powerful tools for handling complex, but standard problems, such as fast computing of global optimum with some pre-specified accuracy. This has to be done by solving of the H-infinity optimization problem. While the analytically computed gamma-iteration represents the first step to solving MFARE, we have been first off, all interested in the efficiency and robustness of the solution based on LMI, which should, in our assumption, produce a better performance.

This paper is organized as follows: after the introduction, in Section II we shortly revisit the problem of H-infinity optimization and describe briefly the derivation of MFARE. In Section III MFARE is converted to an optimization problem based on LMI. In Section IV an algorithm called gamma-iteration is implemented to solve MFARE analytically. Then it is formulated as a linear objective minimization problem using LMI. Finally, each algorithm is evaluated to measure convergence, computation cost and at last but not at least, practicability.

2. Deriving the Modified Filter Algebraic Riccati Equation for robust H-infinity detection filtering

2.1 The optimal H-infinity detection filtering problem

The goal of H-infinity filtering is minimizing the magnitude of the effects of perturbations on the filter output and maximizing the magnitude of the transfer function from failure modes to the filter error, through the appropriate choice of filter gain. This estimation problem can be represented as a mixed H_2 / H_∞ filtering problem (Edelmayer, 2012) [8].

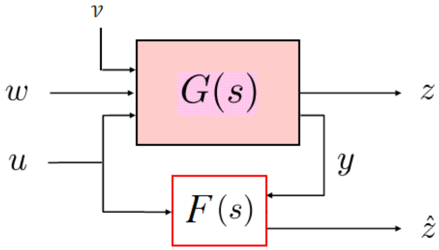

Figure 1: A standard setup for a robust H_∞ filtering synthesis problem
(G: Generalized Plant, F: Filter)

According to the study in [7], the linear time-invariant system (LTI-system) subjected to disturbance and unknown faults can be represented in state space form as follows:

$$\dot{x}(t)=Ax(t)+Bu(t)+B_\omega\omega(t)+\sum_{i=1}^{k}L_i\nu_i(t),$$
$$y(t)=Cx(t). \tag{1}$$

In (1) x $\mathbb{R}^n$, $y\,\epsilon\,\mathbb{R}^p$, $u\,\epsilon\,\mathbb{R}^m$, and $\omega\,\epsilon\,\mathbb{R}^p$ denotes the process disturbance in $L_2\,[0,T]$. *A, B, C* and B_ω are appropriate constant matrices. It is assumed, that *(A, C)* is an observable pair. $B_\kappa=[B_{w,}L_\Delta]$ is the worst-case input direction and $\kappa(t)\in L_2\,[0,T]$ is the input function for all $t\in\mathbb{R}_+$ representing the worst–case effects of modelling uncertainties and external disturbances. It is to be noted, that the equation does not include parametric uncertainty [8]. The cumulative effect of a number of k faults appearing in known directions L_i of the state space is modelled by an additive linear term, $\sum L_i\nu_i(t)$. $L_i\,\epsilon\,\mathbb{R}^{nxs}$ and $\nu_i(t)$ are the fault signatures and failure modes respectively. $\nu_i(t)$ are arbitrary unknown time functions for $t\ \geq t_{ji}$, $0\leq t\leq T$, where t_{ji} is the time instant when the i-th fault appears and $\nu_i=0$, if $t\ <t_{ji}$. If $\nu_i(t)=0$, for every i, then the plant is assumed to be fault free. Assume, however, that only one fault appears in the system at a time [8].

For the purpose of explanation of the concept of the H-infinity filter, consider the system representation given in *Fig.1.*, where $z\,\epsilon\,\mathbb{R}^p$ denotes the output signal. Based on the LTI-system model (1), the state estimate can be obtained as

$$\dot{\hat{x}}(t)=A\hat{x}(t)+K(C(x(t)-\hat{x}(t)))+Bu(t), \tag{2}$$
$$\hat{y}(t)=C\hat{x}(t),$$
$$\hat{z}(t)=C_z\,\hat{x}(t).$$

In (2), $\hat{x} \in \mathbb{R}^n$ represents the observer state, $\hat{y} \in \mathbb{R}^p$represents the output estimate, and $\hat{z} \in \mathbb{R}^p$represents the weighted output estimate, K is the observer gain matrix and C_z is the constant estimation weight (see in [8]).

The filter error system can be derived as

$$\begin{aligned} \dot{\tilde{\mathrm{x}}}(t) &= (A-KC)\,\tilde{\mathrm{x}}(t) + B_w w(t) + \sum_{i=1}^{k} L_i \nu_i(t), \\ \varepsilon(t) &= \mathrm{C}_z \tilde{\mathrm{x}}(t). \end{aligned} \tag{3}$$

In (3), $\tilde{x}$(t) and (t) are the state error and weighted output error, respectively, defined as

$$\begin{aligned} \tilde{\mathrm{x}}(t) &= x(t) - \hat{\mathrm{x}}(t), \\ \varepsilon(t) &= z(t) - \hat{z}(t). \end{aligned} \tag{4}$$

In the presence of faults, the estimation error does not converge asymptotically to zero, but converges asymptotically to a subspace which is different from zero [8].

In the following we have to choose the filter gain, by minimizing the magnitude of the effects of perturbations on the output of the filter, which has to maximize the magnitude of the transfer function from failure modes to the filter error.

2.2 Solution to a H-infinity filtering

Based on the representation in *Fig.1,* the performance measure considered as a quadratic cost function of the minimax method is defined as

$$J(w,\nu,\hat{z}) = \frac{1}{2}\left[\left\|z-\hat{z}\right\|_2^2 - \gamma^2\left(\left\|w\right\|_2^2 + \left\|\nu\right\|_2^2\right)\right], \tag{5}$$

where $\gamma > 0$ is a positive rational constant.

According to the H-infinity filtering problem the quadratic cost function to be minimized is defined as

$$\sup_{w,\nu,a_i} J(w,\nu,\hat{z}). \tag{6}$$

The performance can be formulated as a min-max problem. That is, minimizing the H-infinity norm of the transfer function, denoted by $H_{\varepsilon\kappa}$, of the worst-case disturbance to the filter output. The worst-case performance is given by

$$J(\mathrm{K},\kappa) = \sup \frac{\left\|z-\hat{z}\ \right\|_2}{\left\|\kappa\ \right\|_2} = \left\|H_{\varepsilon\kappa}(s)\right\|_\infty . \tag{7}$$

The filter gain K can be obtained by solving a linear-quadratic optimization problem, using the procedure presented below (see also in [8]).

With substitution of the decision variable $Q \in R^{nxn}$ which is a positive definite matrix, the observer equation can be described as

$$\dot{\hat{\mathrm{x}}}(t) = (A - QC^T C)\,\hat{\mathrm{x}}(\mathrm{t}) + Bu(t) + QC^T y(t),$$
$$\hat{\mathrm{z}}(\mathrm{t}) = \mathrm{C}_z\,\hat{x}(t). \tag{8}$$

The goal of the linear-quadratic optimization is to obtain the smallest L_2 -gain of the disturbance input of the system that is guaranteed to be less than a specified positive constant γ_{min}, and in the same time to increase filtering sensitivity as much as possible (Edelmayer, 2012). The algorithm, which is used to find an optimal solution for Q, iteratively reduces γ until Q has no longer a positive definite solution. Note that the γ_{min} obtained this way is within a given arbitrarily small tolerance $\varepsilon > 0$.

The procedure is based on the solution of the Modified Filter Algebraic Riccati Equation (MFARE). From the bounded-real lemma, we have $\|H_{\varepsilon\kappa}\|_\infty < \gamma$ if and only if there exists $Q \geq 0$ such that

$$AQ + QA^T - Q(C^T C - \frac{1}{\gamma^2} C_z^T C_z)Q + B_K B_K^T = 0. \tag{9}$$

After solving equation (9) and getting a solution for Q, the filter gain matrix can be obtained as

$$K = Q\,\mathrm{C}^T. \tag{10}$$

With the use of γ_{min} the detection threshold of the filter can be given as

$$\tau(C_z) = \gamma_{\min} \|\kappa\|_2\ . \tag{11}$$

It is important to note, that the failure modes, which have the magnitude smaller than that of the detection threshold, cannot be detected by the filter.

3. Solving MFARE by LMI

Originally the problem was introduced in about 1890 by the Russian mathematician Aleksandr Mikhailovich Lyapunov. Linear Matrix Inequalities (LMIs) have become nowadays effective and powerful tools for solving complex optimization problems. The applicability of LMI is really wide, starting e.g. from classical Lyapunov stability analysis of linear time variant and invariant systems, going through traditional Linear Quadratic Gaussian (LQG) control, up to the synthesis of modern robust H-infinity state feedback. The reason for it is that many problems can be cast as convex optimization problems. What is more, most of them can be converted to a standard LMI problem such as computing of global optimum with some pre-specified accuracy, even if it is to be done in our case by solving of H-infinity optimization problem. The main benefit of the LMI formulation is that it defines a convex constraint with respect to the variable vector. For that reason, it has a convex feasible set which can be found guaranteed by convex optimization.

A detailed survey about the theory of LMI can also be found in the mathematical literature, see e.g. in [9], [10] and also in textbooks for control engineering e.g. in [11], [12], [13] and [14].

3.1 Standard problems involving LMIs

A linear matrix inequality is a matrix inequality of the form

$$F(x) \overset{\Delta}{=} F_0 + \sum_{i=1}^{m} x_i F_i > 0, \tag{12}$$

where $x \in \boldsymbol{R}^m$ is the vector of decision variables, and
$F_i = F_i^T \in \boldsymbol{R}^{n\,x\,n},\ i = 0, \cdots, m$ are symmetric matrices.

Let $A(x)$, $B(x)$ and $C(x)$ be symmetric matrices that depend affinely on $x \epsilon \mathbb{R}^m$. Then, in addition to the canonical from in (12) standard LMI problems can be formulated in three different ways (see e.g. in [13]):

1. Feasibility problem with the task of finding a solution for decision variable x so that the constraint
$$A(x) < 0 \tag{13}$$
is sufficient.

2. Linear objective minimization i.e. searching for x which minimizes the linear function subject to an LMI.

   That is, minimize $c^T x$ subject to $A(x) < 0$. (14)

3. Generalized eigenvalue minimization problem i.e. minimizing the maximum generalized eigenvalue of a pair of matrices, that depend affinely on a variable, subject to an LMI constraint.
   The task is to minimize λ subject to an LMI constraint:

$$\begin{gathered} A(x) < \lambda B(x) \\ B(x) > 0 \\ C(x) < 0\,. \end{gathered} \tag{15}$$

Unfortunately, most of the control synthesis problems are not formulated as an LMI, but the nonlinear (convex) inequalities can be converted to an LMI form using the Schur complements' lemma (Boyd et. al. in 1994) [13].

According to this lemma the expressions (16) and (17) are equivalent.

$$\begin{bmatrix} Q(x) & S(x)^T \\ S(x) & R(x) \end{bmatrix} < 0, \tag{16}$$

$$R(x) < 0, \quad Q(x) - S(x)^T R(x)^{-1} S(x) < 0. \tag{17}$$

$Q(x) = Q(x)^T,\ R(x) < 0$, and $S(x)$ depend affinely on x.

In this manner the set of nonlinear inequalities in (17) can be represented as the LMI in (16).

Back to our problem of quadratic optimization we have to solve the MFARE as

$$AQ + QA^T - Q(C^T C - \frac{1}{\gamma^2} C_z^T C_z)Q + B_K B_K^T = 0. \tag{18}$$

To transform (18) into an LMI, at first, we rewrite it in form of inequalities. For this let $R = Q^{-1}$, so we get

$$A^T R + RA - C^T C + \frac{1}{\gamma^2} C_z^T C_z + RB_K B_K^T R < 0\ ,\ R > 0. \tag{19}$$

Applying the Schur complement lemma (17) for (19) yields to

$$\underbrace{A^T R + RA - C^T C}_{Q(x)} - \underbrace{\begin{bmatrix} C_z^T & RB_K \end{bmatrix}}_{S^T(x)} \underbrace{\begin{bmatrix} -\gamma^2 I & 0 \\ 0 & -I \end{bmatrix}^{-1}}_{R^{-1}(x)} \underbrace{\begin{bmatrix} C_z \\ B_K^T R \end{bmatrix}}_{S(x)} < 0. \tag{20}$$

Finally, by using the Schur complement lemma in (16) we obtain the LMI for the MFARE as

$$\begin{bmatrix} RA + A^T R - C^T C & C_z^T & RB_\kappa \\ C_z & -\gamma^2 I & 0 \\ B_\kappa^T R & 0 & -I \end{bmatrix} < 0, \tag{21}$$

which has a solution $R = R^T \in \mathbb{R}^{n\, x\, m}$ and $\gamma > 0$.

Consequently, we can solve the MFARE by minimizing γ with respect to $R \succ 0$ subject to (21).

The corresponding Hamiltonian matrix

$$H_\gamma = \begin{bmatrix} A^T & -(C^T C - \frac{1}{\gamma^2} C_z^T C_z) \\ -B_\kappa B_\kappa^T & -A \end{bmatrix}, \tag{22}$$

has no eigenvalue on the imaginary axis.

In most cases it is possible to solve the Algebraic Riccati Equation also through similarity transformation of the Hamiltonian matrix see e.g. in [15]. Although this method is not for solving MFARE as an optimization problem, so it won't lead to an expected result, it may be useful to check a result obtained via optimization.

The method is described as follows, see in [15]. First the $(2n, n)$ matrix V is built which contains the eigenvectors corresponding to the eigenvalues with negative real parts (stable invariant subspace) of the Hamiltonian matrix:

$$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}. \tag{23}$$

We can get the solution for matrix Q_H as

$$Q_H = V_2 V_1^{-1}. \tag{24}$$

4. Calculation of the filter gain based on the LTI -model of the air path of the diesel engines

For the investigation of fault detection filtering problem, we are interested in the efficiency and robustness of the optimal solution. Thus, two different methods for solving MFARE are compared. First, an algorithm called gamma-iteration is implemented to solve MFARE analytically, then it is formulated as a linear objective minimization problem, solved via LMI.

4.1 LTI-model for the air path of diesel engines

As mentioned in the introduction of the robust fault detection filter design methodologies that we apply in our investigation, it is required to use the LTI-model. Here we refer to a simplified nonlinear model of the air path which was first suggested by Jankovic and Kolmanovsky in 1998 [16] and later by Jung [17] for the purpose of robust control of the diesel engines. In our earlier investigation [18] we have already linearized this model around a specified operating point (Herceg, 2006) [19]. For the sake of simplification, we have considered the fuelling of diesel oil as a constant input, and not as a disturbance, furthermore the disturbance was modelled as the fluctuating change of the engine speed.

As a result, we derive the following LTI-model in the chosen operating point [18]

$$A=\begin{bmatrix}-5.2643 & 4.7316 & 28.5021\\ 50.7697 & -156.9827 & 0\\ 0 & 0.4287 & -9.0909\end{bmatrix},$$

$$B=\begin{bmatrix}1.6111\cdot10^{9} & 0 & 0\\ -1.5720\cdot10^{10} & 8.3514\cdot10^{4} & 1.46083\cdot10^{8}\\ 0 & -141.6484 & 0\end{bmatrix}, \tag{25}$$

$$B_{\omega}=\begin{bmatrix}-47.7946\\ 466.3408\\ 0\end{bmatrix},\quad C=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 3.924\cdot10^{-5}\end{bmatrix},$$

where *A, B, C* and B_ω are appropriate constant matrices, B_ω is the matrix for the disturbance acting on the system.

4.2 Solution of the MFARE by a gamma-iteration algorithm

This section discusses a conventional numerical method called gamma-iteration to get an optimal solution of MFARE. It has to be noted, that this method is often referred to, see e.g. in [1], [2] and [20], [21], but we have not found any algorithm about it. This has been the motivation for its description.

For the start of the explanation, the estimation weight of the filter is chosen arbitrarily, according to the methodology described in [8]:

$$C_Z = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 25 \end{bmatrix}. \tag{26}$$

The MFARE is written again as

$$AQ + QA^T - Q(C^T C - \frac{1}{\gamma^2} C_z^T C_z)Q + B_K B_K^T = 0. \tag{27}$$

Arranged for the use of the MATLAB function ***care*** [22], the equation becomes:

$$AQ + QA^T - Q\begin{bmatrix} C_z^T & C^T \end{bmatrix} \underbrace{\begin{bmatrix} -\gamma^2 I & 0 \\ 0 & I \end{bmatrix}^{-1}}_{R_{care}} \begin{bmatrix} C_z \\ C \end{bmatrix} Q + B_\kappa B_\kappa^T = 0. \tag{28}$$

It is important to note, that the function ***care*** is typically used for solving the H-infinity Riccati Equation for control problems. However, according to the principle of duality between controllers and observers the ***care*** function can be parameterized to be used for a filter in the form:

[Q L Gr report] = care (A', CC, B$_\kappa$\* B$_\kappa$', R$_{care}$, 'report'),

where $CC = [C_Z^T \quad C^T]$.

The function ***care*** returns the optimal value for the decision variable, denoted by *Q*.

Of course the R_{care} - matrix contains γ, but this has a constant value for a specified level of the disturbance attenuation. It results that the function ***care*** cannot be directly used for a quadratic minimization problem, that is, the value of γ is to be iteratively reduced and the decision variable minimized. In this manner, in order to get the γ_{min} value, and so the corresponding optimal solution for Q, we implemented an algorithm called gamma-iteration in which an interval halving method is used iteratively. The algorithm reduces the value of γ until Q has no longer positive definite solution. The γ_{min}, which is reached, is within the limits given by an arbitrarily small tolerance $\varepsilon > 0$.

The gamma-iteration algorithm can be formulated as follows.

The inputs for the method are the *A*, *B$_d$*, *C*, *C$_z$* matrices, which define the LTI-system, *eps* as the relative accuracy of the solution, *maxgamma* as the right limit of the interval (the left limit is zero).

a, *b* and *i* are secondary variables, they stand for assignation of interval and counting cycle respectively.

The outputs are: matrix *Q* as a positive definite decision variable, the *gamma* as step size (midpoint), the *mingamma* variable, which contains the value of gamma at the end of an iteration, the *minigamma* contains the *gamma* value when the iteration is finished.

Each iteration performs the following steps:

1. Calculate *gamma*, the midpoint of the interval, which is assigned by *a* and *b*, that is *gamma = a+(b-a)/2*;
2. Call the MATLAB function *care* which returns the matrix *Q* and the "*report*";
3. Calculate the eigenvalues of *Q* , called *Lambda*;
4. If the convergence criteria of the iteration are not satisfied, namely: *Q* is NOT positive definite, i.e. *prod(Lambda)<=0 or t*he associated Hamiltonian matrix (22) that contains γ has eigenvalues on or very near the imaginary axis, then the upper and lower bounds of interval are changed;
   Otherwise the value of *gamma* is saved, that is *mingamma = gamma* and the iteration is continued;
5. Examine whether the new interval defined by *b-a* reached the relative accuracy of the solution, called *epsilon*. If not, the iteration is repeated, if yes, the iteration is finished and the filter gain is calculated based on the previous value of *gamma* (*mingamma*).

The algorithm is implemented in MATLAB and the script is given below (the example is based on the LTI-system defined by (25)).

```
% matrices of the proposed LTI-system

A=[ -5.2643, 4.7316, 28.5021; 50.7697, -156.9827, 0; 0 , 0.4287, -9.0909 ];

B=[1.6111e+009, 0, 0; -1.5720e+010, 8.3514e+004, 146083000; 0, -141.6784, 0];

C=[1, 0 ,0 ; 0, 1, 0 ; 0, 0, 3.924e-005]; Cz=[5, 0, 0; 0 ,5, 0; 0, 0, 25];

Bd=[-47.7946 0 0  ; 466.3408 0 0  ; 0 0 0 ];
```

```
eps =1e-2;                              % the relative accuracy of the solution
CC =[Cz', C'];                          % building the output matrix
m1 = size(Cz',2);                       % building submatrices for the Rcare
m2 = size(C',2);                            diagonally matrix
maxgamma =1100;                         % the upper limit of the interval
gamma = maxgamma;                       % the step size (midpoint)
b= maxgamma;                            % the initial upper limit of the interval
a=0;                                    % the initial lower limit of the interval
i=0;                                    % initialization of the step counter

while (b-a) >eps                        % examine whether the new interval
                                        reached the relative accuracy
            gamma = a+(b-a)/2;          % interval-halving
            i = i+1;                    % step counting

           % calculation of the Rcare diagonal matrix containing the gamma value
           Rcare = [-(gamma )^2*eye(m1) zeros(m1, m2) ; zeros(m2, m1) eye(m2)];

           % solving of the MFARE using the function care
            [Q L Gr report] = care(A', CC, Bd*Bd', Rcare, 'report')
           Lambda = eig(Q);             % calculation of the eigenvalues of Q

           % reports:
           % if it is < 0, then the associated Hamiltonian matrix has its
           eigenvalues on or very near the imaginary axis, which results in failure
           % if prod(Lambda)<=0, then Q is not positive definite
           if (report==-1 || report==-2 || prod(Lambda)<=0)
                      a = gamma;        % the lower bound is changed to gamma
           else
                      b = gamma;        % the upper bound is changed to gamma
                      mingamma = gamma; % saving gamma value
           end                          % the iteration is continued
end                                     % the iteration is finished
gammamin = mingamma                     % the obtained γmin
K=Q*C'                                  % the obtained filter gain
```

Repeating the γ-iteration 21 times, the optimal value of $\gamma_{min} = 4.9698$ is obtained. Using (10), the corresponding filter gain results as:

$$K = \begin{bmatrix} 257.2236 & -39.2216 & -0.0000 \\ -39.2216 & 699.2298 & 0.0000 \\ -0.7934 & 1.6744 & 0.0000 \end{bmatrix}. \tag{29}$$

It has to be noted, that in steps 8,11,13 and 20 we did not get solution because ***care*** returned with a report = -1. This means that the associated Hamiltonian matrix (22) had its eigenvalues on or very close to the imaginary axis which results in failure, see in [22]. According to the interval halving algorithm, in these steps the upper and lower bounds of the interval are changed in order to keep the solution away from the imaginary axis.

In order to prove the filter performance for disturbance attenuation, the transfer function of the disturbance to a filter residual for the obtained filter gain K is

$$\mathrm{H}_{\varepsilon\omega}(\mathrm{s}) = C_Z(\mathrm{sI} - \mathrm{A} + \mathrm{KC})^{-1}\,\mathrm{B}_{\omega}\,. \tag{30}$$

The evolution of the disturbance attenuation during the iteration steps can be observed on the value of $\|H_{\varepsilon\omega}(s)\|_\infty$, calculated in MATLAB and plotted in *Fig. 2*.

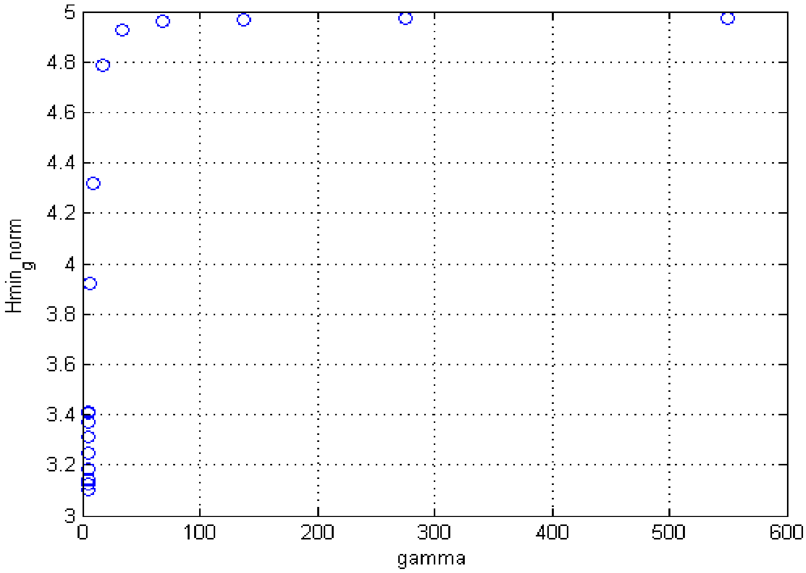

Figure 2: The variation of the $\|H_{\varepsilon\omega}(s)\|_\infty$ value as a function of gamma values during the iteration

The optimal value obtained at the end of the iteration is for $\|\mathrm{H}_{\varepsilon\omega}(s)\|_\infty = 3.3737$.

4.3 The impact of increasing the value of γ_{min}

As known, γ is a measure for the filter sensitivity [8]. In the following it is examined the impact of increasing the value of $\gamma_{\min}$.

In case if γ_{min} reached its upper limit (here $\gamma_{min} = \infty$) the term depending on γ dropped out and (9) was reduced to the form

$$AQ + QA^T - QC^T CQ + B_K B_K^T = 0. \qquad (31)$$

Table 1: The impact of increasing the value of γ_{min}

gamma	matrix K	Eigenvalues of Q	$\left\|H_{\varepsilon K}(s)\right\|_\infty$
γ_{min}	257.2236 -39.2216 -0.0000 -39.2216 699.2298 0.0000 -0.7934 1.6744 0.0000	0.0875 253.7718 702.6875	3.4047
10 γ_{min}	13.5339 -39.8412 0.0000 -39.8412 330.4657 0.0000 -0.2148 0.2480 0.0000	0.0017 8.6061 335.3976	4.9492
100 γ_{min}	13.4326 -39.6856 0.0000 -39.6856 329.3445 0.0000 -0.2129 0.2461 0.0000	0.0017 8.5275 334.2637	4.9717
determinist. Kalman Filter	13.4328 -39.6857 0.0000 -39.6856 329.3445 0.0000 -0.2129 0.2461 0.0000	0.0017 8.5275 334.2637	4.9719

The magnitude of the transfer functions of the disturbance to a filter residual for increased γ_{min} values is shown in *Fig. 3*.

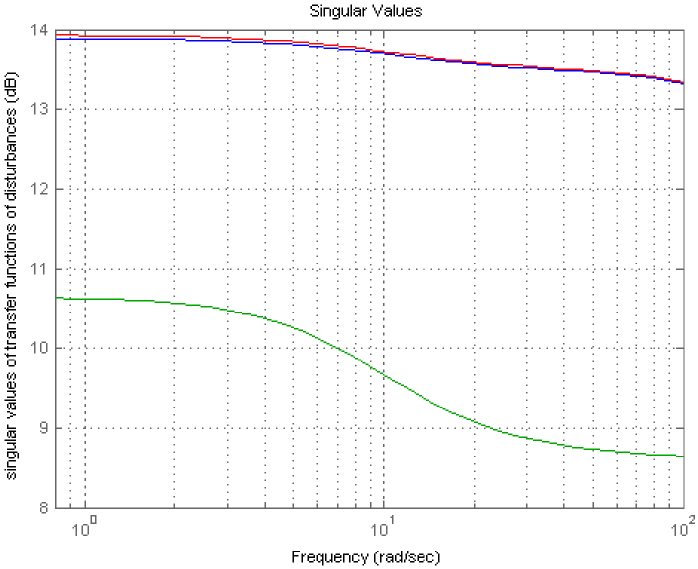

Figure 3: The magnitude (maximal singular values) of transfer functions:
$H_{\varepsilon\varpi}$ (γ_{min}): green line , $H_{\varepsilon\varpi}$ ($10\ \gamma_{min}$): blue line , $H_{\varepsilon\varpi}$ ($100\ \gamma_{min}$): red line

As it can be seen in Table 1, the smallest value for $\|H_{\varepsilon\varpi}\ (s)\|_\infty$ is 3.4047 (10.6415 dB) so the best filter sensitivity against worst case disturbance can be achieved in case of γ_{min} as it is shown in *Fig.3.*

In case of $100\,\gamma_{min}$ and the deterministic Kalman-filter there exists no significant difference between the magnitudes of the transfer functions, as it can be seen in *Fig.3* (blue and red lines). For $100\,\gamma_{min}$ we get a $\|H_{\varepsilon\varpi}\ (s)\|_\infty$ = 4.9717 (13.93 dB), which results in lower disturbance attenuation.

It can be concluded, that the more the value of γ_{min} is increased, the less filter sensitivity can be achieved. In this sense, getting an optimal γ_{min} value is of great importance.

It is conceivable, that the H-infinity filter becomes a deterministic Kalman-filter by reaching its upper limit at $\gamma_{min} = \infty$. This can be also proven easily based on (31).

Of course, the H-infinity filter ensures that the energy gain from the disturbances to the estimation error is always less than a pre-specified level γ^2. Thus it is less conservative than the deterministic Kalman-filter. This is its main advantage from designer's point of view.

4.3 Verification of the solution obtained for the MFARE via the Hamiltonian-matrix

It is possible to verify the solution for the decision variable also via calculating the eigenvectors of the Hamiltonian-matrix of MFARE as it was explained in Subsection 3.1.

The resulting Hamiltonian-matrix for MFARE in case of $\gamma_{\min} = 4.9698$ is

$$H_\gamma = 10^5 \begin{bmatrix} -0.0001 & 0.0005 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.0016 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0003 & 0.0000 & -0.0001 & 0.0000 & 0.0000 & 0.0003 \\ -0.0228 & 0.2229 & 0.0000 & 0.0001 & -0.0000 & -0.0003 \\ 0.2229 & -2.1747 & 0.0000 & -0.0005 & 0.0016 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0000 & 0.0001 \end{bmatrix}.$$

At first we calculate the eigenvalues and the corresponding eigenvectors of the Hamiltonian-matrix via a similarity transformation. The resulting matrix, containing the eigenvalues is

$$diag\,\lambda_i(H_\gamma) = diag\left[150.0393,\ \text{-}150.0393,\ 6.8273,\ 0.4622,\ \text{-}0.4622,\ \text{-}6.8273\right].$$

Secondly, we have to build a $(2n, n)$ matrix V, which contains the eigenvectors of the Hamiltonian matrix corresponding to the eigenvalues with negative real parts (23).

The submatrices of V, which contain the eigenvectors, are:

$$V_1 = \begin{bmatrix} 0.0005 & 0.0039 & 0.0029 \\ -\ 0.0014 & 0.0001 & -0.0003 \\ 0.0004 & 0.0062 & -0.1194 \end{bmatrix}, V_2 = \begin{bmatrix} 0.1749 & 0.9986 & 0.8475 \\ -0.9846 & -0.0516 & -0.5171 \\ -0.0027 & -0.0023 & -0.0139 \end{bmatrix}.$$

Let Q_H denote a solution calculated using the Hamiltonian-matrix, which has a solution

$$Q_H = V_2 V_1^{-1} = \begin{bmatrix} 258.0838 & -32.9691 & -0.7468 \\ -33.7848 & 691.7261 & 1.7722 \\ -0.7809 & 1.6763 & 0.0932 \end{bmatrix}.$$

From the gamma-iteration in Subsection 4.2 we got an optimal solution as

$$Q = \begin{bmatrix} 257.2236 & -39.2216 & -0.7934 \\ -39.2216 & 699.2298 & 1.6744 \\ -0.7934 & 1.6744 & 0.0934 \end{bmatrix}.$$

It can be stated , that matrices Q_H and Q are slightly different. This leads to the conclusion of plausibility of an optimal solution Q obtained by the gamma-iteration.

4.4 Solution for the MFARE by LMI

In Section 3 we introduced the method for finding the optimal solution for MFARE implemented analytically as an interval halving algorithm. However, the task of minimization results in the task of computing a system of matrix equations which is not always convex [8].

Thus, let us now consider the problem of finding the optimal solution for the filter gain by solving of MFARE formulated as a LMI.

To handle it, several commercial software tools can be chosen. In this study the LMI Control Toolbox of MATLAB has been used, which provides a set of convenient functions to solve problems involving LMIs [23].

Generally, the solution of LMIs is carried out in two stages in MATLAB. At first, the decision variables of the LMI are defined, then it is defined the system of LMIs based on these decision variables. These are mostly represented in matrix form. In the second stage, the optimization problem is solved numerically using the chosen solvers as it is explained in Section 2.

In our case study the LMI in (21) is formulated as a linear objective minimization problem. That is, the task is to minimize a linear function of x subject to an LMI constraint:

$$\min_{x} \left\{ c^T x : F(x) \succ 0 \right\}. \tag{32}$$

The LMI for the MFARE derived in Section 2 is described in (21).

In the following it is presented the MATLAB script for the linear objective minimization problem of the MFARE

```
% matrices of the proposed LTI-system
A=[ -5.2643, 4.7316, 28.5021; 50.7697, -156.9827, 0; 0 , 0.4287, -9.0909 ];
B=[1.6111e+009, 0, 0; -1.5720e+010, 8.3514e+004, 146083000; 0, -141.6784, 0];
C=[1, 0 ,0 ; 0, 1, 0 ; 0, 0, 3.924e-005]; Cz=[5, 0, 0; 0 ,5, 0; 0, 0, 25];
Bd=[-47.7946 0 0  ; 466.3408 0 0  ; 0 0 0 ];
I=eye(3);
    % specifying the matrix variables of the LMI
    setlmis([]);
    R = lmivar(1, [size(A, 1) 1]);
    % constructing the system of the LMI
    gamma2 = lmivar(1, [1, 1]);
```

```
lmiterm([1, 1, 1, R], 1, A, 's');        % R’A+AR
lmiterm([1, 1, 1, 0], -C'*C);            % -C’C
lmiterm([1, 2, 1, 0], Cz);               % Cz
lmiterm([1, 2, 2, gamma2], -1, I);       % -gamma^2I
lmiterm([1, 2, 3, 0], 0);                % 0
lmiterm([1, 3, 1,R], Bd', 1);            % Bd’R
lmiterm([1, 3, 2, 0], 0);                % 0
lmiterm([1, 3, 3, 0], -1);               % -I
lmiterm([-2, 1, 1,R], 1, 1);
lmiterm([-3, 1, 1, gamma2], 1, 1);
% obtaining the system of the LMI
lmimingfilt5 = getlmis;
c = mat2dec(lmimingfilt5, zeros(size(A, 1), size(A, 1)), 1);
% the relative accuracy of the solution
 options = [1e-3 , 0, 0, 0, 0];
% solving LMI
[alpha, popt] = mincx(lmimingfilt5, c,  options);
% the optimal value for the decision variable "R"'
Ropt = dec2mat(lmimingfilt5, popt, R);

% the optimal value of the gamma
gopt = dec2mat(lmimingfilt5, popt, gamma2);
% the obtained $\gamma_{min}$
gammaopt=sqrt(gopt)
% the optimal solution of the LMI
Qopt = inv(Ropt);
% the calculated filter gain
K = Qopt*C'
```

4.5 Comparison of the performance of the LMI with the performance of the gamma-iteration

The efficiency and robustness of the optimal solution are interesting aspects of the fault detection filtering problem. Thus, two different methods for solving MFARE are compared, namely the LMI formulated as a linear objective minimization problem and the numerically implemented gamma-iteration.

The results of the MATLAB simulations are shown in Table 2.

Table 2: Comparison of the different solutions for the MFARE

Performance	LMI as an linear objective minimization problem	gamma-iteration
γ_{min}	4.9704	4.9698

K	278.80 -52.70 0.0000
	-52.70 1308.4 0.0000
	-0.500 0.600 0.0000
		257.2236 -39.2216 0.0000
		-39.2216 699.2298 0.0000
		-0.7934 1.6744 0.0000

Eigenvalues of Q	0.3 276.1 1311	0.0017 8.5275 334.2637
$\|\|H_{\varepsilon\omega}(s)\|\|_{\infty}$	4.4345	3.4047
number of iterations	9	21
computation cost (sec)	0.1	1

From the simulation and results of the comparison of the two different methods it can be concluded that each one gives an optimal solution. To be more precise, the minimization algorithm has been applied until the satisfaction of the positive definiteness. As it can be seen in Table 2, the smallest γ_{min} value could be reached using the simple gamma-iteration, but the result obtained this way is just slightly different from the result obtained using LMI. However, the higher filter gain obtained in case of LMI suggests that the filter may be faster but less effective against disturbance. On other hand the burden of successive numerical computation of the quadratic matrix equality resulted in a significant computation cost. It has disadvantages despite its simplicity. From the results it is visible that modern computation methods as LMI are more capable to handle such complex mathematical problems as the solution of the MFARE. From the results mentioned above, it is conceivable that LMI-s are effective and powerful tools for handling complex but standard problems such as rapidly computing of a global optimum with some specified accuracy.

The technique of gamma-iteration, despite its slowness, is easy to be handled. Concretely it gives more flexibility to examine the solution for MFARE. For example, it is easy to analyze the impact of the *gamma* value on the number of iteration steps or the impact of changing of the disturbance on the optimal solution.

One can easily perform experiments and get answers e.g. to the following questions: How does the iteration converge? How do the eigenvalues of the decision variable change? How close are they to the imaginary axis? How are they distributed? How does the filter gain change by reduction of the value of gamma? All these issues can be easily examined, step by step during the iterations.

4. Conclusion

In our paper we performed a benchmark based on collected concepts for solutions of MFARE by conventional gamma-iteration and LMI. From the simulation results of LMI, it can be concluded that it is well capable for computing the global optimum of the quadratic cost function rapidly with some specified accuracy even if this is to be done in the case of MFARE. Both methods, i.e. the gamma-iteration and the LMI formulation as a linear objective minimization problem, are capable for solution of MFARE. Moreover, they deliver only slightly different results. However, the LMI leads to an optimal solution faster, in about 100ms.

The analytically implemented gamma-iteration, despite its slowness, gives much more flexibility to examine the minimization process. For example, it is easy to examine the impact of the iteration steps or the impact of changing of the disturbance on the optimal solution. For this reason we propose the use of both approaches, that is, using the gamma-iteration in the preliminary stage in order to perform an analysis and using LMI in the stage of the synthesis to perform the implementation. Our further work will include an extension of our LMI approach to a switched linear system.

References

[1] Edelmayer, A., Bokor, J., Keviczky, L., "An H_∞ Filtering Approach to Robust Detection of Failures in Dynamical Systems", *in Proc. 33th Annual Decision and Control, Conf.*, pp. 3037-3039, Buena Vista, USA, 1994.

[2] Edelmayer, A., Bokor, J., Keviczky, L., "An H_∞ Filter Design for Linear Systems: Comparison of two Approaches" *IFAC 13th Triennial World Congress, San Francisco, USA*, 1996.

[3] Yaesh, I. , Shaked, U., " Game Theory Approach to Optimal Linear State Estimation and Its Relation to the Minimum H_∞ -norm Estimation", *IEEE Trans. Aut. Control, AC-37(6)*, pp. 828-831, 1992.

[4] Chen, J., Patton, R. J., "Robust Model-Based Fault Diagnosis for Dynamic Systems", First Edition, Springer Science & Business Media, New York, 1999.

[5] Matusu, R., "Linear Matrix Inequalities and Semidefinite Programming: Applications in Control", Internal Journal of Mathematical Models and Methods in Applied Sciences, Vol. 8, 2014.

[6] Gahinet, P, Apkarian, P., "A linear matrix inequality approach to H_∞ control", International Journal of Robust and Nonlinear Control, Vol. 4, pp. 421–448, 1994.

[7] Iwasaki, T., Skelton, R. E., "All controllers for the general H_∞ control problem: LMI existence conditions and state space formulas", *Automatica*, Vol. 30, pp. 1307–1317, 1994.

[8] Edelmayer A., "Fault detection in dynamic systems: From state estimation to direct input reconstruction", Universitas-Győr Nonprofit Kft., Győr, 2012.

[9] Chong, E., K., P., Zak, S., H., "An Introduction to Optimization", 4th Edition, Wiley, New Jersey, 2013.

[10] Ostertag, E., "Mono- and Multivariable Control and Estimation: Linear, Quadratic and LMI Methods" , Matematical Engineering, Vol.2, Springer, Berlin, Heidelberg, 2011.

[11] Ankelhed, D., "On the design of low order H_∞ controllers", Ph.D. These, Linköping University, Linköping, 2011.

[12] Duan, G., R., Yu, H., H., "LMIs in Control Systems: Analysis, Design and Applications", CRC Press, Boca Raton, 2013.

[13] Boyd, S., Ghaoui, L. E., Feron, E., and Balakrishnan, V., "Linear Matrix Inequalities in System and Control Theory", SIAM, Philadelphia, 1994.

[14] Bokor, J., Gáspár, P., Szabó, Z., " Robust Control Theory", Typotex, Budapest, 2013.

[15] Lunze, J., "Regelungstechnik 2 – Mehrgrößensysteme, Digitale Regelung", Springer, 7. Auflage, 2013.

[16] Jankovic, M., Kolmanovsky, I., "Robust Nonlinear Controller for Turbocharged Diesel Engines", *Procedings of the American Control Conference, Philadelphia*, 1998.

[17] Jung, M., "Mean-value modelling and robust control of the airpath of a turbocharged diesel engine", Ph.D. These, University of Cambridge , 2003.

[18] Horvath, Zs., Edelmayer, A., "LTI-modelling of the Air Path of Turbocharged Diesel Engine for Fault Detection and Isolation", *Mechanical Engineering Letters*, Vol. 14, pp. 172-188, Gödöllő, 2016.

[19] Herceg, M., "Nonlinear Model Predictive Control of a Diesel Engine with Exhaust Gas Recirculation and Variable Geometry Turbocharger", Diploma Thesis, Slovak University of Technology in Bratislava, Bratislava, 2006.

[20] Yung, C. F., "Reduced-order H_∞ controller design: An algebraic Riccati euqation approach", Automatica, Vol. 36, pp. 923–926, 2000.

[21] Lanzon, A., Feng, Y., Anderson, B.,D.,O. and Rotkowitz, M., " Computing the Positive Stabilizing Solution to Algebraic Riccati Equations With an Indefinite Quadratic Term via a Recursive Method", *IEEE Trans. Aut. Control*, Vol. 53, pp. 2280–2291, NO. 10. November, 2008.

[22] https://de.mathworks.com/help/control/ref/care.html?requestedDomain=www.mathworks.com

[23] Gahinet, P., Nemirovski, A., Laub, A.J., and Chilali, M., " LMI Control Toolbox for Use with Matlab", The MathWorks, Natick, Messachusetts, 1995.

Acta Universitatis Sapientiae

The scientific journal of Sapientia University publishes original papers and surveys in several areas of sciences written in English.
Information about each series can be found at
`http://www.acta.sapientia.ro`.

Editor-in-Chief
László DÁVID

Main Editorial Board

Laura NISTOR
Ágnes PETHŐ
Zoltán KÁSA
András KELEMEN
Emőd VERESS

Acta Universitatis Sapientiae Electrical and Mechanical Engineering

Executive Editor
András KELEMEN (Sapientia University, Romania)
kandras@ms.sapientia.ro

Editorial Board
Vencel CSIBI (Technical University of Cluj-Napoca, Romania)
Dénes FODOR (University of Pannonia, Hungary)
Dionisie HOLLANDA (Sapientia University, Romania)
Maria IMECS (Technical University of Cluj-Napoca, Romania)
Levente KOVÁCS (Óbuda University, Hungary)
Zsolt LACZIK (University of Oxford, United Kingdom)
András MOLNÁR (Óbuda University, Hungary)
Géza NÉMETH (Budapest University of Technology and Economics, Hungary)
Csaba SIMON (Budapest University of Technology and Economics, Hungary)
Gheorghe SEBESTYÉN (Technical University of Cluj-Napoca, Romania)
Iuliu SZÉKELY (Sapientia University, Romania)
Imre TIMÁR (University of Pannonia, Hungary)
Mircea Florin VAIDA (Technical University of Cluj-Napoca, Romania)
József VÁSÁRHELYI (University of Miskolc, Hungary)

Sapientia University

DE GRUYTER OPEN

Scientia Publishing House

ISSN 2065-5916
http://www.acta.sapientia.ro

Information for authors

Acta Universitatis Sapientiae, Electrical and Mechanical Engineering publishes only original papers and surveys in various fields of Electrical and Mechanical Engineering. All papers are peer-reviewed.

Papers published in current and previous volumes can be found in Portable Document Format (PDF) form at the address: `http://www.acta.sapientia.ro`.

The submitted papers must not be considered to be published by other journals. The corresponding author is responsible to obtain the permission for publication of co-authors and of the authorities of institutes, if needed. The Editorial Board is disclaiming any responsibility.

The paper must be submitted both in MSWord document and PDF format. The submitted PDF document is used as reference. The camera-ready journal is prepared in PDF format by the editors. In order to reduce subsequent changes of aspect to minimum, an accurate formatting is required. The paper should be prepared on A4 paper (210 × 297 mm) and it must contain an abstract of 200-250 words.

The language of the journal is English. The paper must be prepared in single-column format, not exceeding 12 pages including figures, tables and references.

The template file from http://www.acta.sapientia.ro/acta-emeng/emeng-main.htm may be used for details.

Submission must be made only by e-mail (`acta-emeng@acta.sapientia.ro`).

One issue is offered to each author free of charge. No reprints are available.

Contact address and subscription:
Acta Universitatis Sapientiae, Electrical and Mechanical Engineering
RO 400112 Cluj-Napoca
Str. Matei Corvin nr. 4.
E-mail: acta-emeng@acta.sapientia.ro

Printed by Digital Color Company
www.digitalcolorcompany.ro