

Experiment-based Performance Improvement of State Feedback Control Systems for Single Input Processes

Mircea-Bogdan Rădac¹, Radu-Emil Precup¹, Emil M. Petriu²,
Stefan Preitl¹

¹Department of Automation and Applied Informatics, “Politehnica” University of Timisoara, Bd. V. Parvan 2, RO-300223 Timisoara, Romania
E-mail: mircea.radac@aut.upt.ro, radu.precup@aut.upt.ro, stefan.preitl@aut.upt.ro

²School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward, Ottawa, Ontario, Canada, K1N 6N5
E-mail: petriu@eecs.uottawa.ca

Abstract: This paper gives an extension to the Iterative Feedback Tuning (IFT) approach that ensures the performance improvement of state feedback control systems for single input processes. IFT employs sensitivity functions and the experiments conducted on the real-world control system in order to provide an efficient way to deal with the nonlinear or ill-defined processes when the model-dependent Linear-Quadratic Regulator (LQR) is not successful. An experimental setup is suggested to implement the real-time iterative calculation of the gradients in the minimization of the LQR's objective function. The experimental results validate the performance of the proposed IFT algorithm in a mechatronics application which deals with the angular position controller for a DC servo system with actuator dead zone and control signal saturation. The results show the reduction of the LQR's objective function for a single input process application.

Keywords: implementation; Iterative Feedback Tuning; mechatronics; state feedback control; real-time experiments

1 Introduction

The improvement and optimisation of control system (CS) performance is normally obtained by minimizing objective functions (OFs) with several expressions [1]-[9] including integral quadratic performance indices. This also provides a convenient way to deal with the degrees of freedom associated with the pole placement design of Multi Input-Multi Output (MIMO) systems.

The Linear-Quadratic Regulator (LQR) approach, which is frequently used for the tuning of the optimal state feedback CSs, can actually be used only when linearized or linear models of the process and the knowledge on all state variables available for feedback are assumed. Alternatively, the Iterative Feedback Tuning (IFT) [10]-[12] offers a direct data-based offline-adaptive controller tuning strategy. IFT solves the problem by a gradient-based minimisation of the OF using data collected from the real-world CS. Attractive applications of IFT reported in the literature are chemical process control [13], servo drive control [14], [15], nonlinear process control [16]-[18], on-line IFT control of processes that vary over time [19] and IFT combined with fuzzy control [20].

We discussed in [21] the signal processing aspects of the IFT-based state feedback control for second-order positioning systems which have an integral component. A state-space formulation of IFT is analyzed in [22], and the solution converges to the analytical solutions for the state feedback gain matrix and to the Kalman gain. A Linear Quadratic Gaussian (LQG) formulation supported by the transfer function formulation, validated by digital simulation results for a first order process, is offered in [23]. Another LQG formulation dedicated to servo systems control with the Kalman filter state observer was validated in our recent paper [24].

This paper presents an extension of IFT for the optimal state feedback control techniques. Our state feedback CS estimates the OF gradients directly on the basis of measurements carried out during the CS operation. The accent is put on the interpretation of the results obtained in the particular case where a LQR-based tuning is attempted. An original IFT-based approach based on a data-based algorithm to improve the performance of state feedback control systems for single input processes is offered. A comparison between the model-based design for state feedback optimal control systems (the LQR problem) and the experimental-based design using IFT is carried out.

The LQR approach is applied in this paper to initially tune the parameters of the state feedback controller, and our approach ensures further improvement of the CS performance. This improvement is achieved by the alleviation of the OF using experiment-based information from the real-world CS. Our approach makes use of the LQR to guarantee that the initial controller is sufficiently close to the optimal one for the gradient scheme to converge. Our approach is appealing due to several situations that can occur in practice: differences between process models and reality, process changes in time and modifications of performance specifications.

This paper is structured as follows: the next section discusses the general framework to for tuning the state feedback CSs by means of IFT. Section 3 focuses on the new IFT algorithm. Section 4 is dedicated to the case study of an IFT-based angular position controller for a DC servo system with actuator dead zone and control signal saturation. Several practical recommendations for CS designers are also given. The conclusions are highlighted in Section 5. Appendix 1

shows the connection between the LQR OF, which drives the analytical solutions of the optimisation problem, and the IFT OF, which is subjected to practical evaluations in our data-based algorithm.

2 IFT of State Feedback Control Systems

Let us consider a process characterized by the single input discrete-time linear time-invariant (LTI) state-space model

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{A} \mathbf{x}(k) + \mathbf{B} u(k) + \bar{\mathbf{B}} \mathbf{w}(k), \\ \mathbf{y}(k) &= \mathbf{C} \mathbf{x}(k) + \bar{\mathbf{C}} \mathbf{v}(k),\end{aligned}\quad (1)$$

where $k \in \mathbf{N}$ is the discrete time argument, u is the control signal, $\mathbf{x} = [x_1 \ \dots \ x_n]^T \in \mathbf{R}^n$ is the state vector, n is the system order, $\mathbf{y} \in \mathbf{R}^{n_y}$ is the controlled output, $\mathbf{A} \in \mathbf{R}^{n \times n}$, $\mathbf{B} \in \mathbf{R}^{n \times 1}$, $\bar{\mathbf{B}} \in \mathbf{R}^{n \times n}$, $\mathbf{C} \in \mathbf{R}^{n_y \times n}$, $\bar{\mathbf{C}} \in \mathbf{R}^{n_y \times n}$ are constant matrices, and $\mathbf{w} \in \mathbf{R}^n$ and $\mathbf{v} \in \mathbf{R}^{n_y}$ are the uncorrelated process noise vector and measurement noise vector, respectively. All elements of the vectors \mathbf{w} and \mathbf{v} are normal independent identically distributed random variables with zero means and variances σ_w^2 and σ_v^2 , respectively. Zero initial conditions are assumed throughout the paper for the process dynamics without generality loss. The process is supposed to be controllable and observable.

The vector \mathbf{y} is the controlled position and speed in the cases of positioning systems and of servo systems in several applications [25]-[32], but our approach is not limited to positioning systems, servo systems or mechatronics. The transfer characteristics of the actuator and of the measurement instrumentation of the state variables x_i , $i = 1..n$, are both included in the process. The corresponding deterministic discrete-time LTI state-space model of the process is

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{A} \mathbf{x}(k) + \mathbf{B} u(k), \\ \mathbf{y}(k) &= \mathbf{C} \mathbf{x}(k).\end{aligned}\quad (2)$$

The following infinite horizon quadratic performance index can be imposed as performance specification of the CS such that its minimization can ensure very good CS performance:

$$I(\boldsymbol{\rho}) = \sum_{k=0}^{\infty} [\mathbf{x}^T(\boldsymbol{\rho}, k) \mathbf{Q} \mathbf{x}(\boldsymbol{\rho}, k) + \lambda u^2(\boldsymbol{\rho}, k)], \quad (3)$$

where $\boldsymbol{\rho} = [\rho_1 \ \dots \ \rho_n]^T$ is a parameter vector, the state vector and the control signal are parameterised by $\boldsymbol{\rho}$, and the weights are

$$\mathbf{Q} \geq 0, \mathbf{Q} = [q_{ij}]_{i,j=1..n}, q_{ij} = q_{ji}, i, j = 1..n, \lambda > 0. \quad (4)$$

The parametric optimisation of the state feedback control systems can be formulated as the following optimisation problem of finding the optimal parameter vector $\boldsymbol{\rho}^*$ which corresponds to the optimal gain matrix $(\boldsymbol{\rho}^*)^T$:

$$\boldsymbol{\rho}^* = \arg \min_{\boldsymbol{\rho}} I(\boldsymbol{\rho}). \quad (5)$$

The solution to the discrete-time infinite horizon optimisation problem defined in (5) is the control law $u(\boldsymbol{\rho}^*, k) = -(\boldsymbol{\rho}^*)^T \mathbf{x}(\boldsymbol{\rho}^*, k)$ which together with (2) drives the state vector to zero under the CS's spectrum characterized by the system matrix $\mathbf{A}^{cl} = \mathbf{A} - \mathbf{B}(\boldsymbol{\rho}^*)^T$.

The reference inputs are commonly introduced for each state variable when it is needed to drive the state vector to a different point in the state space. The resulting state feedback controller is defined in terms of the control law

$$u(\boldsymbol{\rho}, k) = \boldsymbol{\rho}^T \mathbf{e}(\boldsymbol{\rho}, k), \boldsymbol{\rho}^T = [\rho_1 \ \dots \ \rho_n], \quad (6)$$

$$\mathbf{e}(\boldsymbol{\rho}, k) = \mathbf{r}(k) - \mathbf{x}(\boldsymbol{\rho}, k),$$

where $\mathbf{r} = [r_1 \ \dots \ r_n]^T$ is the reference input vector, r_i are the reference inputs that correspond to the state variables x_i , $i = 1..n$, $\mathbf{e} = [e_1 = r_1 - x_1 \ \dots \ e_n = r_n - x_n]^T$ is the state control error vector that consists of the state variable errors e_i , $i = 1..n$, $\boldsymbol{\rho}^T$ is the state feedback gain matrix, referred to also as the gain matrix, $\boldsymbol{\rho}$ is the parameter vector, and T indicates the matrix transposition. The vector \mathbf{e} is applied as an input to the state feedback gain matrix $\boldsymbol{\rho}^T$ as shown in Figure 1, where \mathbf{P} is the process and \mathbf{C} is the controller, and the difference from the matrix \mathbf{C} in (1) will be pointed out in the sequel when necessary.

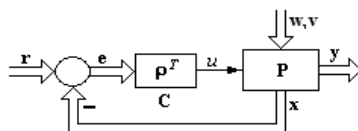


Figure 1

State feedback control system structure

Introducing reference inputs for the state variables, the optimisation problem defined in (5) makes use of the following modified OF:

$$I(\boldsymbol{\rho}) = \sum_{k=0}^{\infty} [\mathbf{e}^T(\boldsymbol{\rho}, k) \mathbf{Q} \mathbf{e}(\boldsymbol{\rho}, k) + \lambda e_u^2(\boldsymbol{\rho}, k)], \quad (7)$$

where the control signal error $e_u(\boldsymbol{\rho}, k)$ is defined as the difference between the control signal and its steady-state value $u(\boldsymbol{\rho}, \infty)$:

$$e_u(\boldsymbol{\rho}, k) = u(\boldsymbol{\rho}, k) - u(\boldsymbol{\rho}, \infty). \quad (8)$$

In order to apply the IFT to solve the optimisation problem (5), using the OF defined in (7), we will use a modified OF, referred to as J , defined as follows over the finite time horizon N for reasons of practical evaluations of the OF:

$$J(\boldsymbol{\rho}) = \sum_{k=0}^N [\mathbf{e}^T(\boldsymbol{\rho}, k) \mathbf{Q} \mathbf{e}(\boldsymbol{\rho}, k) + \lambda e_u^2(\boldsymbol{\rho}, k)]. \quad (9)$$

The OF (9) can be represented by the following approximation if N is sufficiently large to capture all transients in the CS response:

$$I(\boldsymbol{\rho}) \approx J(\boldsymbol{\rho}). \quad (10)$$

IFT algorithms can conveniently be employed to find a solution $\boldsymbol{\rho}^*$ to the optimisation problem

$$\boldsymbol{\rho}^* = \arg \min_{\boldsymbol{\rho} \in D_S} J(\boldsymbol{\rho}), \quad (11)$$

where D_S stands for the stability domain of all state feedback gain matrixes that ensure a stable CS. The two optimisation problems defined in (5) and respectively in (11) essentially are equivalent. However, differences may appear due to the infinite and respectively the finite time horizons in the OFs and to the more general stochastic framework that is necessary to be taken into consideration when the IFT problem is set.

The finite time optimal state feedback control problem is characterized by a time-varying gain matrix, while the infinite time state feedback optimal control problem is characterized by a steady-state gain matrix $\boldsymbol{\rho}^T$. The calculation of the matrices used in both cases requires process models that are affected by modelling and identification errors. In order to solve the optimisation problem (11), a parameter vector $\boldsymbol{\rho}$ has to be found such that

$$\frac{\partial J}{\partial \boldsymbol{\rho}} = \left[\frac{\partial J}{\partial \rho_1} \quad \dots \quad \frac{\partial J}{\partial \rho_n} \right]^T = [0 \quad \dots \quad 0]^T, \quad (12)$$

which, for an OF J defined in (9), becomes

$$\frac{\partial J}{\partial \rho_l} = 2 \sum_{k=0}^N \left\{ \left[\sum_{\substack{i,j=1 \\ i \geq j}}^n (q_{ij} e_i \frac{\partial e_j}{\partial \rho_l}) \right] + \lambda e_u \frac{\partial e_u}{\partial \rho_l} \right\} = 0, \quad l = 1 \dots n. \quad (13)$$

The cases of constrained optimisation problems use the Karush-Kuhn-Tucker optimality conditions instead of the null gradient given by (12).

Partial derivatives $\frac{\partial e_i}{\partial \rho_l}$ and $\frac{\partial e_u}{\partial \rho_l}$ need to be first calculated in order to obtain the

derivatives $\frac{\partial J}{\partial \rho_l}$, $l = 1 \dots n$, in the gradient of the OF. We will present in the next

section an experimental method developed to calculate these partial derivatives.

The IFT algorithms are presented as follows in the more general stochastic framework. Therefore the OF defined in (9) and evaluated on a finite-time horizon becomes a random variable and therefore should be defined as

$$J(\boldsymbol{\rho}) = E \left\{ \sum_{k=0}^N [\mathbf{e}^T(\boldsymbol{\rho}, k) \mathbf{Q} \mathbf{e}(\boldsymbol{\rho}, k) + \lambda \mathbf{e}_u^2(\boldsymbol{\rho}, k)] \right\}, \quad (14)$$

where $E\{ \}$ is the expectation with respect to the stochastic disturbances. However, the deterministic case results in the simplification of the IFT algorithms.

The IFT algorithms can solve the optimisation problem defined in (14) by using the Robbins-Monro stochastic approximation algorithm, which iteratively approaches a zero of a function without the need to know its complete expression. There is no need for evaluations of the OF, but its first and eventually second partial derivatives are important. This result holds not only for the tuning approach based on sensitivity functions, but also the stochastic convergence is ensured with useful consequences when dealing with real world processes. The parameter vector $\boldsymbol{\rho}$ values are iteratively updated according to the following equation:

$$\boldsymbol{\rho}^{i+1} = \boldsymbol{\rho}^i - \gamma^i (\mathbf{R}^i)^{-1} \text{est} \left[\frac{\partial J}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}^i) \right], \quad \mathbf{R}^i > 0, \quad (15)$$

where $i \in \mathbf{N}$ is the current iteration/experiment index, $\gamma^i > 0$ is the step size,

$\text{est} \left[\frac{\partial J}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}^i) \right]$ is the unbiased estimate of the gradient, and the regular matrix \mathbf{R}^i

can be the estimate of the Hessian matrix, the Gauss-Newton approximation of the Hessian, or the identity matrix in the case of less demanding and slower convergent computations. The step size sequence $\{\gamma^i\}_{i \in \mathbf{N}}$ should evolve in time such as to satisfy some bounds. With this regard the conditions to ensure the convergence of the stochastic algorithm are given in [10], [12], [22].

3 Description and Implementation of IFT Algorithm

LQR requires always a linearized model or a collection of local models of the process (e.g., in the gain scheduling approach) in order to calculate the optimal parameter vector $\boldsymbol{\rho}^*$ which corresponds to the optimal gain matrix $(\boldsymbol{\rho}^*)^T$. The identification problem itself is a rather complex undertaking in the case of MIMO systems, which requires a special design of the experiments.

On the other hand, the IFT-based approach does not need exact process models, and special gradient experiments can be conveniently designed to avoid abnormal operation regimes. The initial tuning of the gain matrix is not a problem in the case of the LQR-based approach. However, finding an initial stabilising controller without knowing the process is not a trivial task. Finally, the IFT can be used to fine tune controllers for nonlinear processes under constraints [16].

The IFT-based approach offers a notable degree of flexibility. The OF (11) is not only weighting the state variable errors and the control signal error associated with the LTI state-space model of the CS defined in (1), but it can weigh the reference model tracking error trajectories as well. As shown in [18], the IFT can be used as an alternative solution to the popular pole placement design of optimal state feedback controllers. However, the form in which it is used here is similar to the classical LQR optimisation problem.

As mentioned in the previous section, the main advantage of the IFT resides in its gradient computation algorithm together with the stochastic convergence result. The MIMO IFT-based approach is particularly well suited to solving the optimisation problem defined in (9). From (1) and (6), the LTI state feedback CS is characterized by

$$\begin{aligned} \mathbf{x}(\boldsymbol{\rho}, k) &= \mathbf{P}_{u\mathbf{x}}(q^{-1}) u(\boldsymbol{\rho}, k) + \mathbf{P}_{w\mathbf{x}}(q^{-1}) \mathbf{w}(k), \\ u(\boldsymbol{\rho}, k) &= \boldsymbol{\rho}^T \underbrace{[\mathbf{r}(k) - \mathbf{x}(\boldsymbol{\rho}, k)]}_{\mathbf{e}(\boldsymbol{\rho}, k)}, \end{aligned} \quad (17)$$

where $\mathbf{P}_{u\mathbf{x}}(q^{-1}) \in \mathbf{R}^{n \times 1}$ is the process pulse transfer matrix operator from the input u to the state vector \mathbf{x} , $\mathbf{P}_{w\mathbf{x}}(q^{-1}) \in \mathbf{R}^{n \times n}$ is the disturbance pulse transfer matrix operator from the process noise vector \mathbf{v} to the state vector, and \mathbf{w} , \mathbf{x} and u are defined in accordance with (1). The dependence of the variables involved in (17) on $\boldsymbol{\rho}$ is underlined accordingly.

As suggested in (13), we need to calculate the derivatives $\frac{\partial e_i}{\partial \rho_l}$. Taking into

account the state feedback control law defined in (6) and the fact that \mathbf{r} does not depend on $\boldsymbol{\rho}$, the partial derivatives obtain the expressions

$$\begin{aligned}\frac{\partial e_i(\boldsymbol{\rho}, k)}{\partial \rho_l} &= -\frac{\partial x_i(\boldsymbol{\rho}, k)}{\partial \rho_l}, \\ \frac{\partial e_u(\boldsymbol{\rho}, k)}{\partial \rho_l} &= \frac{\partial u(\boldsymbol{\rho}, k)}{\partial \rho_l} - \frac{\partial u(\boldsymbol{\rho}, \infty)}{\partial \rho_l}, \quad i, l = 1 \dots n.\end{aligned}\quad (18)$$

The derivative of the CS state vector with respect to a certain process parameter $\rho_l, l = 1 \dots n$, can be expressed as

$$\frac{\partial \mathbf{x}(\boldsymbol{\rho}, k)}{\partial \rho_l} = \mathbf{P}_{u \times} (q^{-1}) \frac{\partial u(\boldsymbol{\rho}, k)}{\partial \rho_l}. \quad (19)$$

Similarly, the derivative of the control signal in the state feedback control law expressed in (6) with respect to the same parameter $\rho_l, l = 1 \dots n$, is

$$\frac{\partial u(\boldsymbol{\rho}, k)}{\partial \rho_l} = \frac{\partial \boldsymbol{\rho}^T}{\partial \rho_l} \mathbf{e}(\boldsymbol{\rho}, k) - \boldsymbol{\rho}^T \frac{\partial \mathbf{x}(\boldsymbol{\rho}, k)}{\partial \rho_l}. \quad (20)$$

The derivative of the gain matrix $\boldsymbol{\rho}^T$ with respect to one parameter ρ_l is a row vector with the same dimension as $\boldsymbol{\rho}^T$, but with a single nonzero element that takes the value 1, and when multiplied by \mathbf{e} it keeps only the l^{th} state variable error. The derivative of the control signal is then

$$\begin{aligned}\frac{\partial u(\boldsymbol{\rho}, k)}{\partial \rho_l} &= [0 \quad \dots \quad \rho_l \quad \dots \quad 0] \begin{bmatrix} e_1(\boldsymbol{\rho}, k) \\ \vdots \\ e_l(\boldsymbol{\rho}, k) \\ \vdots \\ e_n(\boldsymbol{\rho}, k) \end{bmatrix} - \boldsymbol{\rho}^T \frac{\partial \mathbf{x}(\boldsymbol{\rho}, k)}{\partial \rho_l} \\ &= e_l(\boldsymbol{\rho}, k) + \boldsymbol{\rho}^T \left(\underset{=0}{\mathbf{r}} - \frac{\partial \mathbf{x}(\boldsymbol{\rho}, k)}{\partial \rho_l} \right),\end{aligned}\quad (21)$$

where e_l is the l^{th} state variable error. Equation (21) shows how to conduct the gradient experiments with the process: by injecting an additive term in the control signal of the state feedback CS and letting the reference input vector \mathbf{r} equal to zero, the derivatives of the state variables and of the control signal with respect to the parameter ρ_l in $\boldsymbol{\rho}^T$ are obtained. The injected term is e_l , i.e., the l^{th} element of the state control error vector obtained in a normal experiment. All specific experiments of IFT are described as follows.

An initial experiment, called the normal experiment, is carried out to record the evolution of the state variables and the corresponding state variable errors and control signal error respectively, in the state feedback CS shown in Figure 1.

Other n gradient experiments are then subsequently carried out in order to calculate estimates of the derivatives $\frac{\partial x_i}{\partial \rho_l}$ and $\frac{\partial u}{\partial \rho_l}$, and use is made of (17) and

(21). Let l denote as a superscript the l^{th} gradient experiment corresponding to $\rho_l, l=1\dots n$:

$$\begin{aligned} \mathbf{x}^l(\boldsymbol{\rho}, k) &= \mathbf{P}_{u\mathbf{x}}(q^{-1}) u^l(\boldsymbol{\rho}, k) + \mathbf{P}_{w\mathbf{x}}(q^{-1}) \mathbf{w}^l(k) \\ &= \mathbf{P}_{u\mathbf{x}}(q^{-1}) [e_l(\boldsymbol{\rho}, k) - \boldsymbol{\rho}^T \mathbf{x}^l(\boldsymbol{\rho}, k)] + \mathbf{P}_{w\mathbf{x}}(q^{-1}) \mathbf{w}^l(k). \end{aligned} \quad (22)$$

Equation (22) provides the basis for the experimental setup (illustrated in Figure 2) employed in the iterative calculation of the partial derivatives $\frac{\partial x_i}{\partial \rho_l}$ and $\frac{\partial u}{\partial \rho_l}$

needed in the minimization of the OF. We actually obtain at each gradient experiment the estimates of the gradient of the state variables with respect to the gain matrix parameters. In other words, the state variables of the gradient experiments are actually the gradient estimates. This is because at each experiment the process noise acts upon the CS. Equation (22) results in $E\{\mathbf{x}^l\} = \frac{\partial \mathbf{x}}{\partial \rho_l}$.

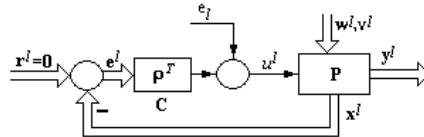


Figure 2

Experimental setup to compute $\frac{\partial x_i}{\partial \rho_l}$ and $\frac{\partial u}{\partial \rho_l}$

The IFT algorithm consists of the following steps:

- *Step 0.* Set the step size, the initial controller parameters $\boldsymbol{\rho}^0$ and the weights in the OF.
- *Step 1.* Conduct the initial (normal) experiment making use of the CS structure presented in Figure 1 and record the evolution of all state variables.

- *Step 2.* Conduct the n gradient experiments making use of the experimental setup presented in Figure 2 to obtain all partial derivatives $\frac{\partial x_i}{\partial \rho_l}$ and $\frac{\partial u}{\partial \rho_l}$.
- *Step 3.* Conduct the normal experiment again such that the states contain realizations of noise that differ from the noise at *step 2* to ensure the unbiased estimate of the gradient.
- *Step 4.* Calculate the estimates of the gradient of the OF according to (13).
- *Step 5.* Compute \mathbf{p}^{i+1} in terms of the update law (15).

Step 0 is done only once. *Steps 1* to *5* are repeated iteratively. *Step 0* requires an initial set of parameters that stabilise the state feedback CS to be obtained here by LQR. In the case of Single Input-Single Output (SISO) systems, we can use the Ziegler-Nichols tuning [33] or other techniques like the Virtual Reference Feedback Tuning [34], [35] in order to get these parameters.

There exists a difference between the deterministic case and the stochastic case in terms of the objective function and of the objectives that are targeted. Specifically, IFT is developed as an experimental-based technique in which the noise enters the CS and therefore the objective function also contains a factor that depends on the noise; therefore the minimization of the energy transfer between the noise and the state variables is also attempted, in addition to the minimization of the state control error and of the control signal energy that are objectives specific to the LQR deterministic problem. This aspect is illustrated in Appendix 1.

4 Case Study

The case study is a second-order positioning CS for a modular DC servo system with an integral component. The process is characterized by the single input discrete-time LTI state-space model defined in (4) with the matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 0.0487 \\ 0 & 0.9471 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0.1867 \\ 7.3993 \end{bmatrix}, \mathbf{C} = \mathbf{I}_2, \quad (23)$$

and with the angular position and the angular speed as state variables. The experimental setup is built around the INTECO DC servo system laboratory equipment. The control signal u for the accepted laboratory equipment is the PWM duty-cycle constrained to $-1 \leq u \leq 1$. The actuator exhibits a ± 0.15 width insensitivity zone applied to u and compensated by an inverse nonlinearity.

The simplified model (23) was obtained by the parameter identification of the first-principle model of the equipment resulting in the simplified process transfer function (considering u as the input and the angular position as the output)

$$P(s) = k_p / [s(1 + T_\Sigma s)], \quad (24)$$

where k_p is the process gain and T_Σ is the small time constant. The values of the process parameters were obtained as $k_p = 139.88$ and $T_\Sigma = 0.92$ s. Using the notation T_s for the sampling period, the sampling period of $T_s = 0.05$ s was set.

The detailed mathematical model of the process is time variant due to the interchanging modules (inertial load, encoder and eventually backlash). The re-identification is not used in our approach. An experimental scenario is presented as follows to illustrate the benefits of the IFT-based approach over the classical LQR-based approach.

The weights \mathbf{Q} and λ in the infinite horizon quadratic performance index defined in (5) are

$$\mathbf{Q} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}, \lambda = 400. \quad (25)$$

The results are presented for a step angular position reference input of 40 rad and zero angular speed reference input, i.e., $\mathbf{r} = [40 \ 0]^T$. A first order low-pass digital filter with a cut-off frequency of 20 rad/s is used in the experiments to reduce the errors and the noise that occurs during the measurement of the angular speed. This filter will change the process model, but IFT is independent with this regard. This choice also supports the idea that the tuning can be carried out whenever the process model changes in time, without the need of identification and optimal redesign via LQR.

The weights (25) do not cause the saturation of the actuator. Thus the undesired behaviour due to the nonlinearities is avoided. This undesired behaviour usually occurs in the LQR-based approach where the nonlinear actuator is not included in the process model.

For benchmarking purposes the control system performance indices that are used are the OF, the control signal energy defined as

$$E_u = \sum_{k=0}^N u^2(k), \quad (26)$$

the 10% to 90% rise time of the position response (t_r), and the maximum speed (ω_{\max}). The IFT-based approach is next used to further reduce the OF, taking advantage of the experiments conducted on the real-world experimental setup.

In order to provide a relevant improvement, we start with a process model that is very different from the identified model. This is the same as assuming that the process model is time variant or that the identification is not accurate. The starting model for the LQR design uses the process parameters $k_p = 180$ and $T_\Sigma = 1.2$ s in the transfer function (24). For the weights (25), the state feedback gain matrix is

$$(\mathbf{p}^T)_{LQR_1} = [\rho_1 = 0.020496 \quad \rho_2 = 0.021368], \quad (27)$$

The gain matrix $(\mathbf{p}^T)_{LQR_1}$ is further tuned using our IFT algorithm. The initial step size in the IFT algorithm employed to minimize the OF (9) is set to the initial value $\gamma^0 = 2 \cdot 10^{-8}$, the values of the consequent step sizes are set in terms of (17), and $\mathbf{R}^i = \mathbf{I}_2$ is used.

The reduction of the value of the OF is emphasized to illustrate that our IFT algorithm ensures the performance improvement of the state feedback CS. The following expression of the gain matrix is obtained after 15 iterations:

$$(\mathbf{p}^T)_{LQR_2} = [\rho_1 = 0.018900 \quad \rho_2 = 0.017355]. \quad (28)$$

The evolution of the OF with respect to the iteration number (i.e., during the tuning) is presented in Figure 3. The evolutions of the controller parameters (i.e., the elements of the gain matrix) versus the iteration number are presented in Figure 4. The time responses of the CS before and after the application of the IFT algorithm are presented in Figure 5.

Figure 4 illustrates that the OF is affected by random disturbances when it is evaluated on the real-world process. The values of the OF for the gain matrices defined in (27) and (28) are $J_{LQR_1} = 3821.89$ and $J_{LQR_2} = 3772.10$, respectively. The following performance indices were obtained:

- for the initial CS response (i.e., before IFT): $E_u = 2.5482$, $t_r = 2.94$ s, $\omega_{\max} = 27.0847$ rad/s,
- for the final CS response (i.e., after IFT): $E_u = 2.4654$, $t_r = 2.53$ s, $\omega_{\max} = 27.9519$ rad/s.

A discussion on these results follows. The relatively large number of iterations shown in this case indicates that the slow convergence is due to the fact that the steepest descent direction is used and due to the fact that we are close to the true local minimum, and therefore only modest but still quantifiable improvements are seen in the figures. In practical situations it suffices to do several experiments in order to improve the performance in terms of the OF.

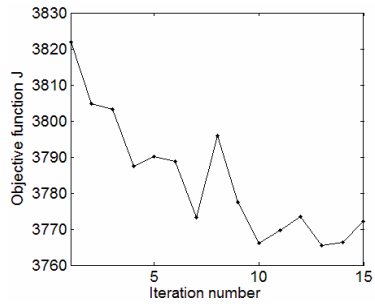


Figure 3

The evolution of the objective function versus the iteration number

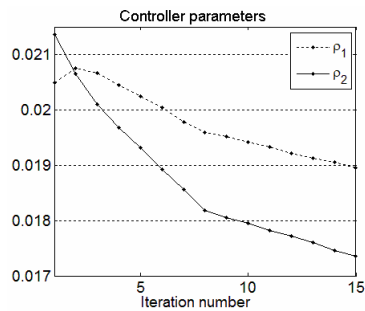


Figure 4

The evolution of the controller parameters versus the iteration number

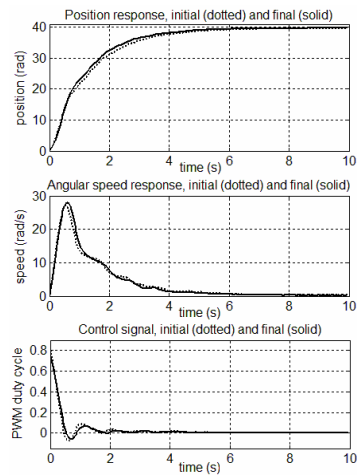


Figure 5

Control system responses of the CS before IFT and after IFT

The minimization of J is dedicated to reducing the energy transfer from the process noise to the state variables. In conclusion, nonzero reference inputs are reflected by targeting three objectives: the minimization of the tracking error energy, the minimization of the control effort, and the noise rejection problem [13]. The improvements via IFT shown in the previous section ensure the reduction of the OF value and of its variance, due to the lower sensitivity to noise. This idea is also backed up by Appendix 1.

The time responses of the experimental results shown are not very different, and this shows the robustness of state feedback CS with respect to the controller and process parametric variations. However, the solution is an evident improvement of the LQR design, and when the noise contribution in the OF is small, it is expected that the tuning procedure gets near the optimal gain matrix, which results in an optimal state feedback CS with robustness properties. When the noise contribution is important, the robustness properties of the optimal state feedback CS still hold, as suggested by the simulation scenarios with included process noise.

The weights in the optimisation problems were set so as to ensure the linear operation of the process and of the actuator, viz., without entering saturation. The experimental results illustrate that the steady-state error of the position response is improved in spite of the process nonlinearities.

IFT requires $1+n$ real-time experiments per iteration, n of them being successive gradient computation experiments. This number cannot be reduced using ideas similar to those presented in [36]-[38] because the number of gain matrix parameters is equal to the product between the numbers of process inputs and outputs.

Conclusions

This paper has presented an extension of the IFT approach to improve the performance of state feedback CSs where the performance specifications aim the minimization of OFs expressed as quadratic performance indices.

Our general IFT approach provides an efficient way to deal with some of the specific problems of ill-defined processes when the strongly model-dependent LQR design gives solutions that are far away from the optimal solution. In such cases, when the LQR approach cannot anymore allow finding the minimum of the OF, the IFT approach can be applied to further reduce the OF. The experimental results presented in Section 4 show that the IFT approach, which allows an estimation the OF gradients on the basis of sensitivity functions and of real-time measurements during the CS operation, can successfully be used.

A limitation of our IFT approach is that it actually ensures the strong improvement of the CS performance and the strong reduction of the OF only with respect to the considered particular reference input. Modifications of the reference input will yield different results with different dynamic characteristics. Our IFT approach does not use state estimators, being developed for a specific situation where all

states are measured. However, the introduction of state estimators in future research is not problematic because the estimator gain can also be included in the IFT algorithm.

Future research will deal with the extension of the proposed IFT approach to MIMO control systems in mechatronics applications and to the tuning of state feedback fuzzy control systems. Further study of the convergence of the IFT algorithms is needed for all these nonlinear applications. Similar model-free tuning methods will be implemented including extremum seeking with emphasis on the direct application to the tuning of fuzzy controllers [39]-[42].

Acknowledgement

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0109, and by a grant from the NSERC of Canada.

References

- [1] I. Škrjanc, S. Blažič, O. E. Agamennoni: Identification of Dynamical Systems with a Robust Interval Fuzzy Model, *Automatica*, Vol. 41, No. 2, 2005, pp. 327-332
- [2] M. A. Khanesar, M. Teshnehlab, O. Kaynak: Model Reference Fuzzy Control of Nonlinear Dynamical Systems Using an Optimal Observer, *Acta Polytechnica Hungarica*, Vol. 8, No. 4, 2011, pp. 35-54
- [3] L. Kovács, B. Benyó, J. Bokor, Z. Benyó: Induced L_2 -Norm Minimization of Glucose-Insulin System for Type I Diabetic Patients, *Computer Methods and Programs in Biomedicine*, Vol. 102, No. 2, 2011, pp. 105-118
- [4] E. S. Nicoară, F.-G. Filip, N. Paraschiv: Simulation-based Optimization Using Genetic Algorithms for Multi-objective Flexible JSSP, *Studies in Informatics and Control*, Vol. 20, No. 4, 2011, pp. 333-344
- [5] S. Aykut, A. Kentli, S. Gülmez, O. Yazicioglu: Robust Multiobjective Optimization of Cutting Parameters in Face Milling, *Acta Polytechnica Hungarica*, Vol. 9, No. 4, 2012, pp. 85-100
- [6] E. D. Niño: SAMODS and SAGAMODS: Novel Algorithms Based on the Automata Theory for the Multiobjective Optimization of Combinatorial Problems, *International Journal of Artificial Intelligence*, Vol. 8, No. S12, 2012, pp. 147-165
- [7] P. R. Srivastava, M. Chis, S. Deb, X.-S. Yang: An Efficient Optimization Algorithm for Structural Software Testing, *International Journal of Artificial Intelligence*, Vol. 8, No. S12, 2012, pp. 68-77
- [8] J. Vaščák, M. Paľa: Adaptation of Fuzzy Cognitive Maps for Navigation Purposes by Migration Algorithms, *International Journal of Artificial Intelligence*, Vol. 8, No. S12, 2012, pp. 20-37

- [9] F. Yao, Z. Dong, K. Meng, Z. Xu, H. Iu, K. Wong: Quantum-inspired Particle Swarm Optimization for Power System Operations Considering Wind Power Uncertainty and Carbon Tax in Australia, *IEEE Transactions on Industrial Informatics*, Vol. 8, No. 4, 2012, pp. 880-888
- [10] H. Hjalmarsson, M. Gevers, S. Gunnarsson: A Convergent Iterative Restricted Complexity Control Design Scheme, *Proceedings of 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, FL, USA, 1994, pp. 1735-1740
- [11] H. Hjalmarsson, M. Gevers, S. Gunnarsson, O. Lequin: Iterative Feedback Tuning: Theory and Applications, *IEEE Control Systems Magazine*, Vol. 18, No. 4, 1998, pp. 26-41
- [12] H. Hjalmarsson: Iterative Feedback Tuning – An Overview, *International Journal of Adaptive Control and Signal Processing*, Vol. 16, No. 5, 2002, pp. 373-395
- [13] J. K. Huusom, N. K. Poulsen, S. B. Jørgensen: Data Driven Tuning of State Space Control Loops with Unknown State Information and Model Uncertainty, *Computer Aided Chemical Engineering*, Vol. 26, 2009, pp. 441-446
- [14] R.-E. Precup, S. Preitl, I. J. Rudas, M. L. Tomescu, J. K. Tar: Design and Experiments for a Class of Fuzzy Controlled Servo Systems, *IEEE/ASME Transactions on Mechatronics*, Vol. 13, No. 1, 2008, pp. 22-35
- [15] S. Kissling, P. Blanc, P. Myszkowski, I. Vaclavik: Application of Iterative Feedback Tuning (IFT) to Speed and Position Control of a Servo Drive, *Control Engineering Practice*, Vol. 17, No. 7, 2009, pp. 834-840
- [16] H. Hjalmarsson: Control of Nonlinear Systems Using Iterative Feedback Tuning, *Proceedings of 1998 American Control Conference (ACC 1998)*, Philadelphia, PA, USA, 1998, Vol. 4, pp. 2083-2087
- [17] J. Sjöberg, F. De Bruyne, M. Agarwal, B. D. O. Anderson, M. Gevers, F. J. Kraus, N. Linard: Iterative Controller Optimization for Nonlinear Systems, *Control Engineering Practice*, Vol. 11, No. 9, 2003, pp. 1079-1086
- [18] A. J. McDaid, K. C. Aw, S. Q. Xie, E. Haemmerle: Gain Scheduled Control of IPMC Actuators With ‘Model-free’ Iterative Feedback Tuning, *Sensors and Actuators A: Physical*, Vol. 164, No. 1-2, 2010, pp. 137-147
- [19] A. J. McDaid, K. C. Aw, E. Haemmerle, S. Q. Xie: Control of IPMC Actuators for Microfluidics with Adaptive “Online” Iterative Feedback Tuning, *IEEE/ASME Transactions on Mechatronics*, Vol. 17, No. 4, 2012, pp. 789-797
- [20] R.-E. Precup, M.-B. Rădac, M. L. Tomescu, E. M. Petriu, S. Preitl: Stable and Convergent Iterative Feedback Tuning of Fuzzy Controllers for

- Discrete-Time SISO Systems, Expert Systems with Applications, Vol. 40, No. 1, 2013, pp. 188-199
- [21] M.-B. Rădac, R.-E. Precup, S. Preitl, E. M. Petriu, C.-A. Dragoș, A. S. Paul, S. Kilyeni: Signal Processing Aspects in State Feedback Control Based on Iterative Feedback Tuning, Proceedings of 2nd International Conference on Human System Interaction (HSI '09), Catania, Italy, 2009, pp. 40-45
- [22] J. K. Huusom, N. K. Poulsen, S. B. Jørgensen: Data Driven Tuning of State Space Controllers with Observers, Proceedings of European Control Conference 2009 (ECC '09), Budapest, Hungary, 2009, pp. 1961-1966
- [23] J. K. Huusom, N. K. Poulsen, S. B. Jørgensen: Iterative Feedback Tuning of Uncertain State Space Systems, Brazilian Journal of Chemical Engineering, Vol. 27, No. 3, 2010, pp. 461-472
- [24] M.-B. Rădac, R.-E. Precup, E. M. Petriu, S. Preitl: Application of IFT and SPSA to Servo System Control, IEEE Transactions on Neural Networks, Vol. 22, No. 12, 2011, pp. 2363-2375
- [25] L. Horváth, I. J. Rudas: Modeling and Problem Solving Methods for Engineers, Academic Press, Elsevier, Burlington, MA: 2004
- [26] J. A. Iglesias, P. Angelov, A. Ledezma, A. Sanchis: Evolving Classification of Agents' Behaviors: A General Approach, Evolving Systems, Vol. 1, No. 3, 2010, pp. 161-171
- [27] Zs. Cs. Johanyák: Survey on Five Fuzzy Inference-based Student Evaluation Methods, in: Computational Intelligence in Engineering, I. J. Rudas, J. Fodor, J. Kacprzyk (Eds.), Springer-Verlag, Berlin, Heidelberg, New York, 2010, pp. 219-228
- [28] D. Antić, S. Nikolić, M. Milojković, N. Danković, Z. Jovanović, S. Perić: Sensitivity Analysis of Imperfect Systems Using Almost Orthogonal Filters, Acta Polytechnica Hungarica, Vol. 8, No. 6, 2011, pp. 79-94
- [29] O. Linda, M. Manic: Uncertainty-Robust Design of Interval Type-2 Fuzzy Logic Controller for Delta Parallel Robot, IEEE Transactions on Industrial Informatics, Vol. 7, No. 11, 2011, pp. 661-670
- [30] P. Baranyi, Á. Csapó: Definition and Synergies of Cognitive Infocommunications, Acta Polytechnica Hungarica, Vol. 9, No. 1, 2012, pp. 67-83
- [31] H.-N. Teodorescu: Taylor and Bi-local Piecewise Approximations with Neuro-Fuzzy Systems, Studies in Informatics and Control, Vol. 21, No. 4, 2012, pp. 367-376
- [32] M. Iwasaki, K. Seki, Y. Maeda: High-Precision Motion Control Techniques: A Promising Approach to Improving Motion Performance, IEEE Industrial Electronics Magazine, Vol. 6, No. 1, 2012, pp. 32-40

- [33] S. Kissling, P. Blanc, P. Myszkowski, I. Vaclavik: Application of Iterative Feedback Tuning (IFT) to Speed and Position Control of a Servo Drive, *Control Engineering Practice*, Vol. 17, No. 7, 2009, pp. 834-840
- [34] F. Previdi, T. Schauer, S. M. Savaresi, K. J. Hunt: Data-driven Control Design for Neuroprotheses: A Virtual Reference Feedback Tuning (VRFT) Approach, *IEEE Transactions on Control Systems Technology*, Vol. 12, No. 1, 2004, pp. 176-182
- [35] M. C. Campi, S. M. Savaresi: Direct Nonlinear Control Design: The Virtual Reference Feedback Tuning (VRFT) Approach, *IEEE Transactions on Automatic Control*, Vol. 51, No. 1, 2006, pp. 14-27
- [36] H. Hjalmarsson: Iterative Feedback Tuning of Linear Time-invariant MIMO Systems, *Proceedings of 37th IEEE Conference on Decision and Control*, Tampa, FL, USA, 1998, pp. 3893-3898
- [37] H. Hjalmarsson: Efficient Tuning of Linear Multivariable Controllers Using Iterative Feedback Tuning, *International Journal of Adaptive Control and Signal Processing*, Vol. 13, No. 7, 1999, pp. 553-572
- [38] H. Jansson, H. Hjalmarsson: Gradient Approximations in Iterative Feedback Tuning for Multivariable Processes, *International Journal of Adaptive Control and Signal Processing*, Vol. 18, No. 8, 2004, pp. 665-681
- [39] R.-E. Precup, S. Preitl: *Fuzzy Controllers*, Editura Orizonturi Universitare Publishers, Timisoara: 1999
- [40] R. E. Precup, S. Doboli, S. Preitl: Stability Analysis and Development of a Class of Fuzzy Control Systems, *Engineering Applications of Artificial Intelligence*, Vol. 13, No. 3, 2000, pp. 237-247
- [41] R.-E. Precup, S. Preitl, G. Faur: PI Predictive Fuzzy Controllers for Electrical Drive Speed Control: Methods and Software for Stable Development, *Computers in Industry*, Vol. 52, No. 3, 2003, pp. 253-270
- [42] R.-E. Precup, S. Preitl, E. M. Petriu, J. K. Tar, M. L. Tomescu, C. Pozna: Generic Two-Degree-of-Freedom Linear and Fuzzy Controllers for Integral Processes, *Journal of The Franklin Institute*, Vol. 346, No. 10, 2009, pp. 980-1003
- [43] A. S. Bazanella, M. Gevers, L. Mišković, B. D. O. Anderson: Iterative Minimization of H_2 Control Performance Criteria, *Automatica*, Vol. 44, No. 10, 2008, pp. 2549-2559

Appendix 1. Connection between LQR and IFT objective functions

This Appendix illustrates the connection between the LQR OF which drives the analytical solutions of the optimisation problem, and the IFT OF which is subjected to practical evaluations in our data-based algorithm. We assume two

cases for the OF, defined in the deterministic case and in the stochastic case related to the state feedback CS. The dependence on the parameter vector $\boldsymbol{\rho}$ is omitted for the sake of simplicity. Our development follows a similar development to that presented in [43], and the two cases, a) and b), are presented as follows.

a) *The deterministic case.* We assume that the following operational relationships are valid:

$$\begin{aligned}\mathbf{x}(\boldsymbol{\rho}, k) &= \mathbf{P}_{\mathbf{r}\mathbf{x}}(\boldsymbol{\rho}, q^{-1}) \mathbf{r}(k), \quad u(\boldsymbol{\rho}, k) = \mathbf{P}_{\mathbf{r}u}(\boldsymbol{\rho}, q^{-1}) \mathbf{r}(k), \\ \mathbf{e}(\boldsymbol{\rho}, k) &= \mathbf{r}(k) - \mathbf{x}(\boldsymbol{\rho}, k) = \mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(\boldsymbol{\rho}, q^{-1}) \mathbf{r}(k),\end{aligned}\quad (29)$$

where $\mathbf{P}_{\mathbf{r}\mathbf{x}}(\boldsymbol{\rho}, q^{-1}) \in \mathbf{R}^{n \times n}$ is the process pulse transfer matrix operator from the reference input vector \mathbf{r} to the state vector \mathbf{x} and $\mathbf{P}_{\mathbf{r}u}(\boldsymbol{\rho}, q^{-1}) \in \mathbf{R}^{n \times 1}$ is the process pulse transfer matrix operator from \mathbf{r} to the control signal u . The dependence on $\boldsymbol{\rho}$ is assumed but not explicitly written as follows in order to simplify notation.

The infinite horizon OF specific to the formulation of the LQR problem corresponding to this case is

$$\begin{aligned}I(\boldsymbol{\rho}) &= \sum_{k=0}^{\infty} \{ \mathbf{e}(k)^T \mathbf{Q} \mathbf{e}(k) + \lambda u^2(k) \} = \sum_{k=0}^{\infty} \{ [\mathbf{r}(k) \\ &- \mathbf{Q} [\mathbf{r}(k) \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1}) \mathbf{r}(k)]^T - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1}) \mathbf{r}(k)] + \lambda [\mathbf{P}_{\mathbf{r}u}(q^{-1}) \mathbf{r}(k)]^2 \}.\end{aligned}\quad (30)$$

b) *The stochastic case.* The following relations hold:

$$\begin{aligned}\mathbf{x}(k) &= \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1}) \mathbf{r}(k) + \mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k), \\ u(k) &= \mathbf{P}_{\mathbf{r}u}(q^{-1}) \mathbf{r}(k) + \mathbf{P}_{\mathbf{w}u}(q^{-1}) \mathbf{w}(k), \\ \mathbf{e}(k) &= \mathbf{r}(k) - \mathbf{x}(k) = \mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1}) \mathbf{r}(k) - \mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k).\end{aligned}\quad (31)$$

The reference input vector and the process noise are assumed to be quasi-stationary and uncorrelated, i.e.,

$$E\{\mathbf{r}(k) \mathbf{w}^T(k)\} = \mathbf{0}.\quad (32)$$

The expression of the OF used in IFT in this case is

$$\begin{aligned}J(\boldsymbol{\rho}) &= E\left\{ \sum_{k=0}^{\infty} \mathbf{e}(k)^T \mathbf{Q} \mathbf{e}(k) + \lambda u^2(k) \right\} = E\left\{ \sum_{k=0}^{\infty} [\mathbf{r}(k) \\ &- \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1}) \mathbf{r}(k) - \mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k)]^T \mathbf{Q} [\mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1}) \mathbf{r}(k) \\ &- \mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k)] + \lambda [\mathbf{P}_{\mathbf{r}u}(q^{-1}) \mathbf{r}(k) + \mathbf{P}_{\mathbf{w}u}(q^{-1}) \mathbf{w}(k)]^2 \right\},\end{aligned}\quad (33)$$

$$\begin{aligned}
J(\boldsymbol{\rho}) &= \sum_{k=0}^{\infty} E\{[\mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1})\mathbf{r}(k)]^T \mathbf{Q} [\mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1})\mathbf{r}(k)]\} \\
&- \sum_{k=0}^{\infty} E\{[\mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1})\mathbf{r}(k)]^T \mathbf{Q} \mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k)\} \\
&- \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k)]^T \mathbf{Q} [\mathbf{r}(k) - \mathbf{P}_{\mathbf{r}\mathbf{x}}(q^{-1})\mathbf{r}(k)]\} \\
&+ \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k)]^T \mathbf{Q} [\mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1}) \mathbf{w}(k)]\} \\
&+ \lambda \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{r}u}(q^{-1})\mathbf{r}(k)]^2\} + \\
&+ 2\lambda \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{r}u}(q^{-1})\mathbf{r}(k)][\mathbf{P}_{\mathbf{w}u}(q^{-1}) \mathbf{w}(k)]\} \\
&+ \lambda \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{w}u}(q^{-1}) \mathbf{w}(k)]^2\}.
\end{aligned} \tag{34}$$

The second, the third and the sixth terms in (34) are zero due to the uncorrelation between \mathbf{r} and \mathbf{w} . Therefore the following expression of $J(\boldsymbol{\rho})$ is obtained:

$$\begin{aligned}
J(\boldsymbol{\rho}) &= I(\boldsymbol{\rho}) + \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1})\mathbf{w}(k)]^T \mathbf{Q} [\mathbf{P}_{\mathbf{w}\mathbf{x}}(q^{-1})\mathbf{w}(k)]\} \\
&+ \underbrace{\lambda \sum_{k=0}^{\infty} E\{[\mathbf{P}_{\mathbf{w}u}(q^{-1})\mathbf{w}(k)]^2\}}_{J_{\mathbf{w}}(\boldsymbol{\rho})}.
\end{aligned} \tag{35}$$

The term $J_{\mathbf{w}}(\boldsymbol{\rho})$ is dedicated to the minimization of the energy transfer from \mathbf{w} to \mathbf{x} and to u . Inherently, in experiment-based tuning via IFT, this objective is also targeted in addition to the objectives to minimize the state control error energy (set-point tracking) and E_u . If $\mathbf{r} = \mathbf{0}$ and $\lambda = 0$ are chosen, the OF $J(\boldsymbol{\rho})$ enables the minimization of the energy transfer from the process noise to the state variables, resulting in a non-robust structure.

Evaluation of the Quality of Experience for 3D Future Internet Multimedia

Ivett Kulik and Tuan Anh Trinh

Department of Telecommunications and Media Informatics, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary
kulik@tmit.bme.hu, trinh@tmit.bme.hu

Abstract: Provisioning 3D video stream-based services online in an acceptable quality, even in a wireless access environment, is a big challenge for Future Internet service providers. Characterizing the necessary Quality of Service requirements is hard, since only a few empirical results are known about the user perceived 3D quality. In this paper a statistical analysis of subjective perception of 3D stereoscopic video Quality of Experience (QoE) are investigated with respect to network level QoS. The network is configured to demonstrate a real environment; thus, GPON-based aggregation is used. Our results show characteristics of QoE-QoS relationship in the case of 3D video playback. We also tackle the challenge by carrying out GPON-based transport network with IEEE802.11n standard based WiFi access measurements focusing the QoE of 3D content. And according to our results we propose cubic fitting function for modeling QoE-QoS relationship in the case of throughput degradation.

Keywords: 3D stereoscopic video; Quality of Experience-QoE; Quality of Service-QoS; GPON-based network; WiFi network; Mean Opinion Score-MOS; subjective evaluation

1 Introduction

The Internet has approached an historic turning-point, when mobile platforms and applications are poised to replace the fixed-host/server model that has dominated since its inception. The existing Internet architecture has been designed for efficient communication but not for real-time data distribution. The exponential growth of smart mobile devices with Internet access, and the need of users to be “always connected” definitely indicate that the Internet has become the core mobile communication environment for business, entertainment, education, and for social and human interactions.

Over the past decades, new network architectures and protocols have been proposed that sketch the idea of the Future Internet. Paul et al. [1] presented a

comprehensive survey on the networking research on network architecture for future networks and the next generation Internet. The articular network neutrality aspect, where users are able to access any web content and to use any applications according to their choice without restrictions or limitations, is becoming the biggest challenge for Internet Service Providers (FISP).

FISP has to prepare for the capability to support multiple types of terminals, hosts and nodes, protocols and applications. The major design goals of FISP networks are: *mobility* as the norm with dynamic host and network mobility at scale; *robustness* with respect to intrinsic properties of wireless medium; *trustworthiness* in the form of enhanced security and privacy for both mobile networks and wired infrastructure; and *usability* features, such as support for context-aware pervasive mobile services, evolvable network services, manageability and economic viability.

The Future 3D Media Internet has generated a significant amount of research work recently, which should be designed to overcome current limitations of network architecture, involving content and service mobility, new forms of 3D content provisioning, etc. [15] [3]. A seamless delivery of 3D video streams means that the provider needs to be able to observe and react quickly to Quality of Service (QoS) problems in transport network, and the importance of Quality of Experience (QoE) appears as well. QoE are customer-centric metrics, while QoS is network-centric. Human perception of video streams is best characterized in term of QoE, which looks at the streaming content from the standpoint of end users. Today, in the era of increasing fast resolution, mobile-phone owners commonly watch movie trailers or whole films on their small favorite devices, while customer satisfaction will remain dominant criteria for future applications. Consequently, appropriate QoS support at the service providers side and satisfactory level of 3D video QoE at the client side provided through the wireless access for mobile handhelds remains a big challenge for Future Internet researchers, as well. Investigation of QoE characteristics based on QoS degradation for 3D multimedia contents delivery is in focus recently. The assessment of QoE in multimedia services can be performed either by subjective or objective methodologies [2].

More research subjects have brought into focus the QoE and QoS [3] [12] or evaluation of stereoscopic images [4] [11] [6] [10], but more investigations are needed for appropriate QoE provisioning in wireless network based networks. The Gigabit Passive Optical Network (GPON) GPON transport based test-bed with wireless client access is an appropriate representation of an environment for measurements, and recent research works have appeared for the evaluation of QoE for 3D multimedia delivery by means of QoS in Future Internet wireless access scenarios.

This contribution is publishing a few results of subjective tests carried out by participants focusing on describing the relationship between QoE and QoS for 3D

contents delivery in a real network environment. Obviously, network level QoS parameters such as throughput, delay, jitter and packet loss affect user level QoE parameters. First, we carried out experiments based on subjective testing of 3D video files, where 50 participants observed QoE changes due to the degradation of QoS parameters. The results of this experiment are published in [15]. We followed up on our experiments and this contribution shows the results of the QoE-QoS relationship investigation when one video file was observed in 3D and 2D types of visualization, as well. 40 users watched videos with QoS degradations, while jitter increased and throughput decreased. The second part of this paper describes a few results of an experiment where 36 participants assessed the quality of 3D video content when the network was a representative combination of GPON-based transport network and IEEE802.11n standard based WiFi access.

The paper is structured as follows. In Section 2 we explain the network environment. Section 3 describes the method of measurements. Section 4 discusses results in the case of a GPON environment. Section 5 shows a few results with a WiFi network, from the client side. Finally, this paper is concluded in Section 6.

2 The Network Environment

Based on the 3D multimedia transport requirements, the appropriate test network was planned and realized. Basically, the multimedia server is connected with a broadband and reliable connection, and 3D video contents were transferred through the network in unicast mode using TCP transport. Types of encoding and compression affect the demand of bandwidth in the case of multimedia content transport. The used average bandwidth can be between 10 Mbit/s and 20 Mbit/s via stream, or more, but in the case of higher motion level scenes, even 40 Mb/s throughput is needed. Videos were displayed by the Nvidia Vision Player v1.6.

The GPON-based transport network was efficient with 2.5 Gbit/s download speed and 1.5 Gbit/s upload speed [7] via broadband and responsible access to video server with 3D multimedia streams. The whole GPON-based network architecture with wireless sub-networks on the client side is shown in Figure 1.

The GPON-based transmission network consists of four components: Optical Line Terminal (OLT) on the provider side, Optical Network Terminal (ONT) on the customer side, optical cables for connecting, and passive splitters that can split optical signals in split ratios 1:2. The OLT and ONT devices are managed by the Siemens EM-PX manager client. The hardware configuration of the server and clients are shown in the Table 1.

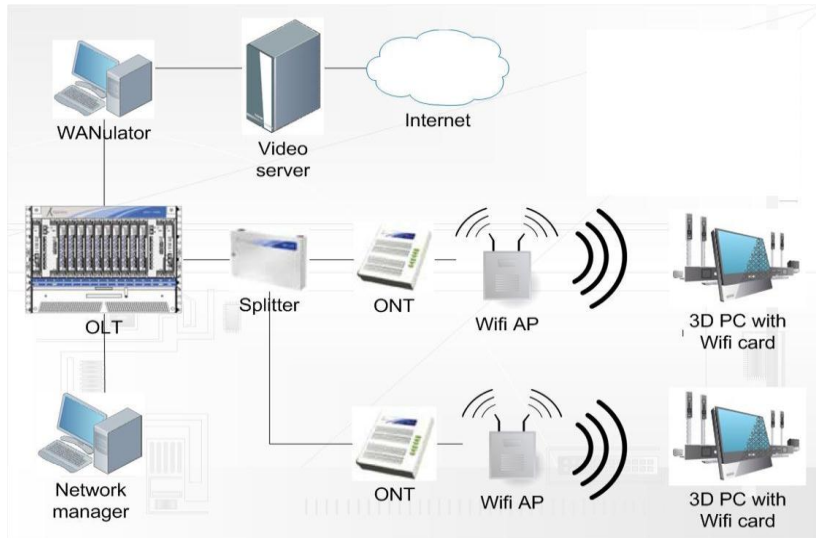


Figure 1

The GPON-based network with WiFi sub-networks for 3D video streams investigation

Table 1

Hardware configuration of the client and server

CLIENT	Components	Notes
Processor	Intel Core 2 Quad, Q8300, 2,5GHz	Needs: At least Intel Core 2 Duo, or AMD X2 Athlon
Video-card	NVIDIA GeForce GT 240	Needs: 8 series, 9 series or 200 series NVIDIA video-card
Memory	4GB RAM	
Spectacles	Nvidia 3D Vision	
SERVER	Components	
Motherboard	Asus P5B Deluxe	
Processor	Intel Core 2 Duo, 2,13GHz	
Memory	1 GB RAM	

The video server was responsible for the storage and sharing of the 3D and 2D video files, which was guaranteed by the VLC program. The WANulator software simulated different Internet conditions, such as delay, jitter or packet loss, providing the proper QoS degradation level in the transport network, and bandwidth limitation was set Netlimiter.

Figure 2 shows the network architectures of the experiment for both scenarios: firstly, when the 2D and 3D videos were delivered and watched on the PCs connected directly to the GPON; and secondly, when the 3D video was transferred through the GPON to clients with WiFi 802.11n access to the transport network.

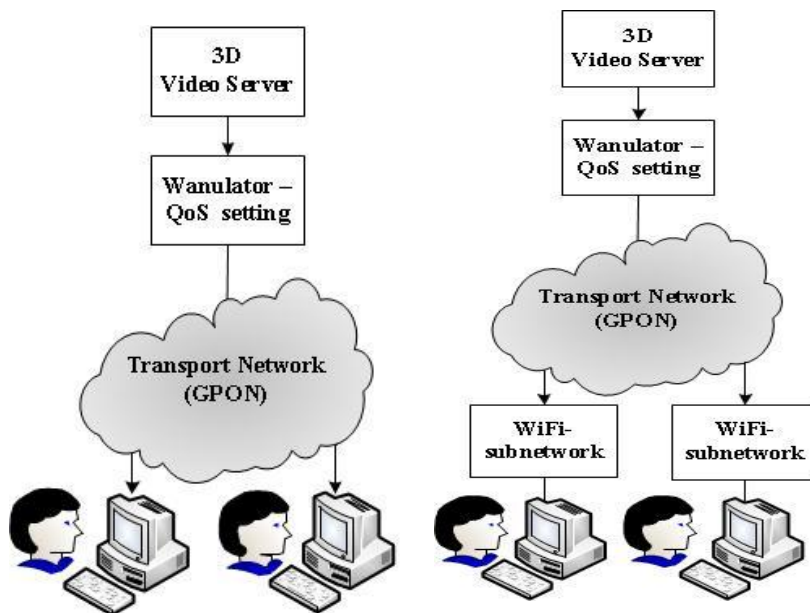


Figure 2

Network architectures of the experiment without and with the WiFi sub-networks

3 Method of Measurements

The common practice for estimating user perception from network-level performance criteria is to conduct large experiments in a controlled environment. The QoE can be affected by many factors: network features which refer to QoS metrics such as packet loss, delay, jitter, reordering, and bandwidth limitation; and also multimedia features, which include higher levels' specific parameters such as coding, quantization, bit-rate, frame-rate and motion level. All could have an effect on the QoE [12].

Multimedia sequences (undistorted and distorted contents as well) can be scored by the Mean Opinion Score (MOS) in the case of subjective evaluation, which is the core of our experiments. The Mean Opinion Score (MOS) [17] quality scale method is typically applied for voice and video traffic scale (shown in Table 2). Reference sequence quality can be also graded by MOS for more detailed results, but usually only the outcome needs to be done.

Based on the first-hand experience of our testing [15], we prepared an investigation regarding the QoE-QoS relation not only for 3D video streams but

also for 2D content as well. Our goal was to use statistical analysis to obtain more information on the relationship between the degradation of QoS parameters and QoE evaluations.

Table 2
MOS Quality Scale

Score	Sequence quality
5	Excellent
4	Good
3	Regular
2	Bad
1	Awful

In both cases (in the GPON environment and in the WiFi network topology) participants watched a short part of the 3D stereoscopic film *Avatar*, the features of which are shown in Table 3, and had to evaluate the following questions about quality during video watching focusing on the empirical quality of the video.

- 1) Rate continuity of the video content.
- 2) Rate the quality of picture. Did you notice disintegration of picture?
- 3) How did you assess the 3D experience on the whole?
- 4) How did you feel conformity between the picture and voice?
- 5) What was the quality like on the whole?

Table 3
Features of the investigated 3D video

Title	Video codec	Audio codec	Container format
Avatar	WMPv9 (VC-1 Simple/Main)	WMAv2	wmv
Length (mm:ss)	Resolution	Video bitrate (kb/s)	Audio bitrate (kb/s)
03:32	1280*720	9646	192

The order of these points was also essential. The first 4 points were about the QoE from various points of view. The last one was about QoE on the whole, which is usually much more complicated than only the recapitulation of the first 4 points. We also asked users to weight their answers for the correct statistical analysis. These weights helped us to calculate the weighted average for representation of the QoE-QoS relationship based on the subjective tests.

4 Test Results in GPON Environment

We gathered some basic demographic information. 40 users (37 men, 3 women, 16 wearing glasses, and with an average age of 22) took part in this experiment. They watched a trailer for the 3D stereoscopic film *Avatar* mentioned above and also the same part of the film in 2D. A short part was enough because the goal was the QoE estimation and not an assessment of the film content [14].

Two types of degradation were made on the 3D and 2D video file, as well. And the test users scored the videos in the case of the following scenarios via the MOS:

- 1) Reference undistorted video files
- 2) Videos disturbed only by jitter increase
- 3) Videos disturbed by bandwidth limitation and jitter increase

The value of bandwidth limitation was calculated based on the maximum bandwidth demand, which was around 40 Mb/s for the 3D content in the case of the highest motion level scenes. The mean value of the bandwidth used was around 32 Mb, so we set the bandwidth threshold to 32 Mb/s, which caused throughput limitation. This value was set by the Netlimiter software for each client.

Value settings of these scenarios are shown in Table 3 and Table 4.

Table 3
Parameters values for jitter degradation

QoS setting	Type of video	Values refer to every measuring	1. test	2. test	3. test	4. test
Jitter	2D	9400 packets + 470 burst for jitter; Bandwidth limit. none	Jitter: 100 ms	Jitter: 120 ms	Jitter: 140 ms	Jitter: 160 ms
	3D	9400 packets + 470 burst for jitter; Bandwidth limit. none	Jitter: 90 ms	Jitter: 100 ms	Jitter: 120 ms	Jitter: 160 ms

The results of the reference tests (watching the undistorted video file) showed that people who had watched 3D movies or videos before this experiment (36 persons) perceived the 3D content as lower quality than the rest of them (4 person). The average value of 3D experience (point 3 in the questionnaire) was 3.83 (almost 4, i.e. good quality) which was very good score on the whole.

After evaluation of the averages, we counted the weighted average based on weighted answers gathered from users, and we could assign one QoE value to every certain value of the QoS parameters. If an answer was given a larger weight by the user, this meant that this feature (one of points 1-5 above) was more important for the user. A summary of this information is shown in Table 5.

Table 4
Parameters values for throughput limitation + jitter

QoS setting	Type of video	Values refer to every measuring	1. test	2. test	3. test	4. test
Band-width limit. + Jitter	2D	9400 packets + 470 burst for jitter; Bandwidth 32 Mb/s	Jitter: 100 ms	Jitter: 120 ms	Jitter: 140 ms	Jitter: 160 ms
	3D	9400 packets + 470 burst for jitter; Bandwidth 32 Mb/s	Jitter: 90 ms	Jitter: 100 ms	Jitter: 120 ms	Jitter: 160 ms

Table 5
Summary of weighted values

3D QoS	reference	90 ms jitter	100 ms jitter	120 ms jitter	160 ms jitter
3D QoE	4,355	4,225	3,7775	2,955	2,2425
2D QoS	reference	100 ms jitter	120 ms jitter	140 ms jitter	160 ms jitter
2D QoE	4,8193	4,771	4,5199	4,143	3,1998

3D QoS	reference	90ms jitter + BW32Mb/s	100ms jitter + BW32Mb/s	120ms jitter + BW32Mb/s	160ms jitter + BW32Mb/s
3D QoE	4,355	3,625	3,205	2,395	1,8325
2D QoS	reference	100ms jitter + BW32Mb/s	120ms jitter + BW32Mb/s	140ms jitter + BW32Mb/s	160ms jitter + BW32Mb/s
2D QoE	4,8193	4,6951	4,0471	3,3898	2,7311

We can clearly recognize QoE deterioration based on an increase of QoS. Observers watched content on two PCs simultaneously and separately connected to GPON by two WiFi access points. The NVPv1.6 player was set up with 440 ms de-jittering buffer and it was not changed during the whole experiment.

Figure 3 shows QoE degradation based on jitter increase by using interpolation lines in the case of the 2D and 3D video.

Applying the method of least squares we got the next solutions:

$$\bullet \quad 2D: 1.41046 * 10^{-6} x^3 - 0.000558526 x^2 + 0.0398079 x + 4.88527 \quad (6)$$

$$\bullet \quad 3D: 4.19435 * 10^{-6} x^3 - 0.00116773 x^2 + 0.0644745 x + 4.22993 \quad (7)$$

The QoE-QoS relationship shows a cubic correlation, and the sensitivity is more pronounced in the case of 3D video.

Figure 3 shows the confidence interval (CI) of the QoE values, where the normal distribution is applied and a 90% confidence interval, and the critical value was calculated for this 90% CI. Lines of averages are plotted with bold lines and the margins of CI are plotted with dashed lines. In the case of the 3D video, the CI is more descending.

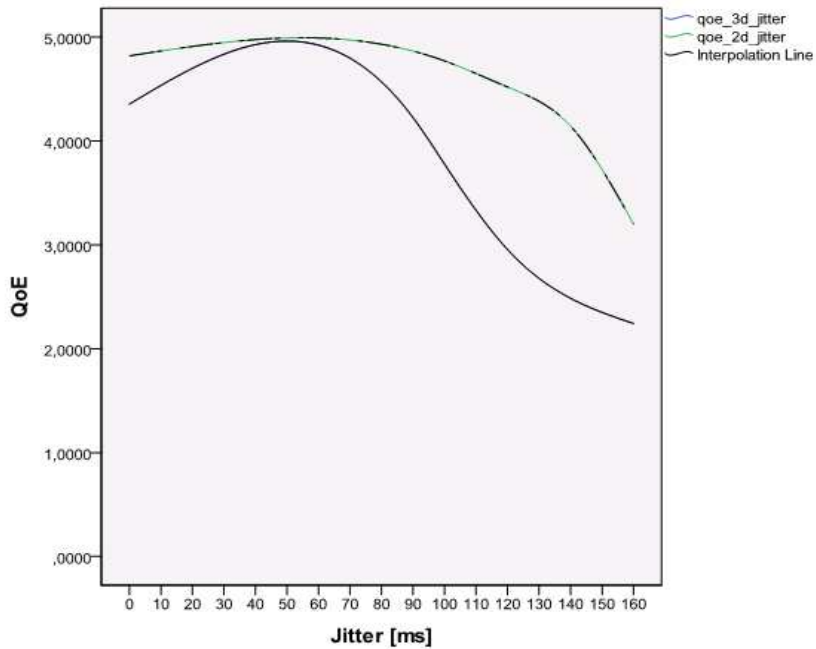


Figure 3
QoE based on jitter increase

A jitter value of 90 ms was the threshold for the 3D video, and a jitter value of 100ms was the threshold for the 2D video when the vision quality was still good, without jerkiness and freezing during the watching. The quality rapidly broke down from this point and participants were not satisfied with the quality due to jerkiness and, later, even a freezing picture. This method of evaluation was used in case of jitter increase and throughput limitation at the same time, when the threshold values were kept at 90 ms and 100 ms jitter value, but fell down rapidly from this point.

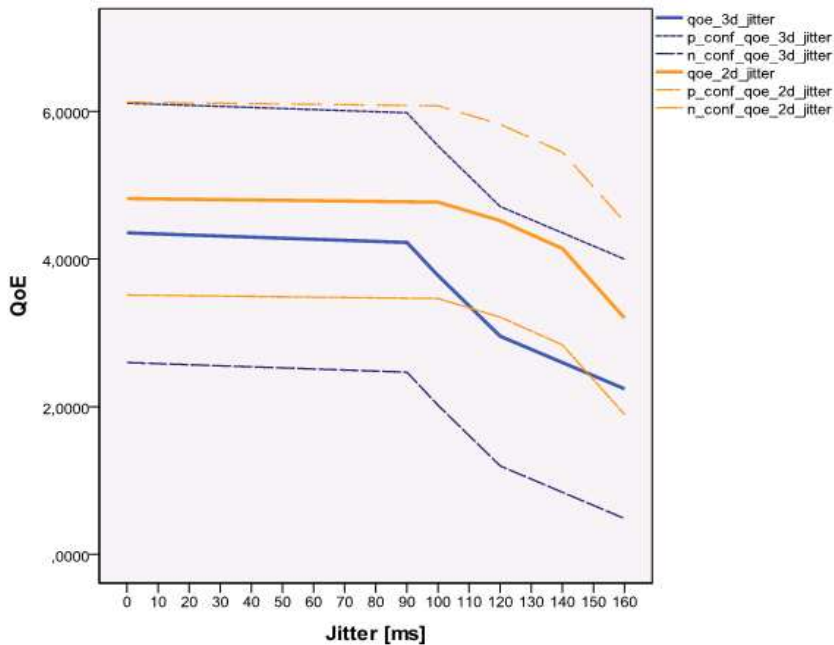


Figure 4
Confidence interval of QoE in case of jitter increase

5 Test Results in a WiFi Environment

In this experiment 36 participants attended (34 men and 2 women), who study at the Budapest University of Technology and Economics. 18 of them wore glasses, and their mean age was 22.14. The youngest student was 20 years old, while the oldest one was 27. 32 participants had watched 3D movies before the tests.

Observers watched content on two PCs simultaneously, separately connected to GPON by two WiFi access points. When people watched 3D stereoscopic content on two PCs simultaneously, playback was not fully fluent especially during higher motion level scenes, even in the case without any QoS parameter degradation in the transport. Simultaneously, two wireless configurations were investigated and loaded condition of them significantly affected our measurements, which could appear in real networks as well. Using WiFi channel-13 caused a medium load, while channel-3 showed an extremely crowded wireless condition.

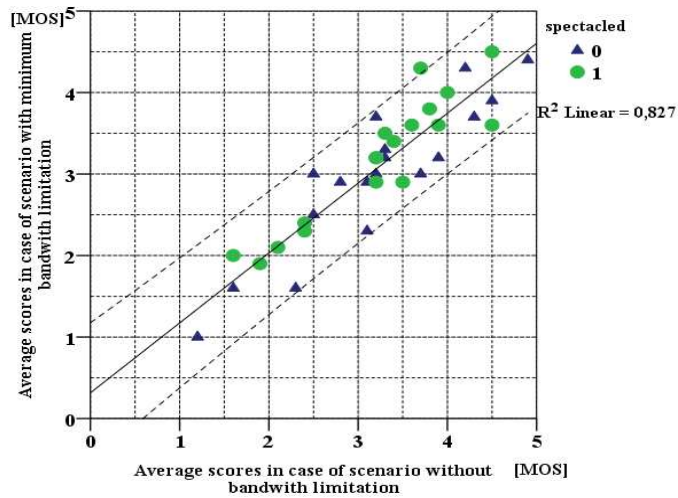


Figure 5

QoE scores comparison between scenarios with the moderate bandwidth limitation on channel-13: x-axis MOS values in case of bandwidth 40 Mb/s and y-axis MOS values in case of bandwidth value 36 Mb/s with bandwidth limitation 4 Mb/s

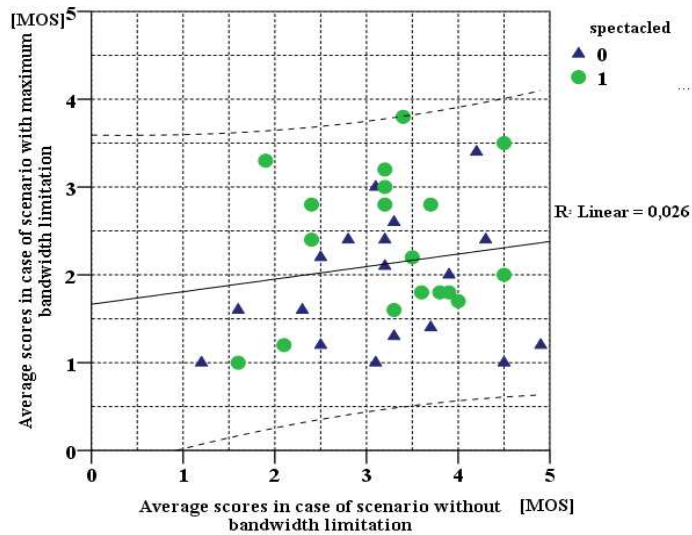


Figure 6

QoE scores comparison between scenarios with the high bandwidth limitation on channel-13: x-axis MOS values in case of bandwidth 40 Mb/s and y-axis MOS values in case of bandwidth 28 Mb/s with bandwidth limitation 12 Mb/s

Figure 5 shows linear regression with small deviation between average scores based on observation results in the case of small QoS degradation. This means that only small differences appeared in scoring between intact and moderately limited playback. In the second case, the video was played back with small QoS degradation, namely the bandwidth was limited with 4 Mb/s compared to the intact situation. Bandwidth limitation values were calculated on the average demand bandwidth value of the 3D stream, which during 95% of the playing time was 32 Mb/s, except in case of the highest motion level scenes when spine values appeared, exceeded this 32 Mb/s value up to 40 Mb/s. According to our experiments, with respect to the offered load, 40 Mb/s was considered as the highest load in the network, thus considered as an intact situation. During the tests, bandwidth in the transport was limited. The threshold was set to 36 Mb/s when the bandwidth limitation was 4 Mb/s, and so on, which caused network QoS degradation during our experiments.

Figure 6 shows a comparison of the results in the intact case and the highest bandwidth limitation setting when the bandwidth threshold value was 28 Mb/s with bandwidth limitation of 12 Mb/s. As we can see, the linearity disappeared in this case. When the quality of continuity became unacceptable because of jerkiness and freezing, some participants' average score still remained above 3 (regular quality). As can be seen in the figure, these participants were mostly with glasses, and they did not assess the poor quality so critically.

Also, it is observation that only spectacled people scored better the playback with higher bandwidth limitation in both cases depicted in Figure 5 and Figure 6.

In the article [12], the IQX hypothesis is presented, which is a natural and generic relationship between QoE and QoS. They demonstrated the feasibility of exponential relationship through a couple of case studies, for example measurements results for web browsing in a fast network taken from G.1030. Our experiments show correlation with quadratic and even with cubic model is much better than applying exponential model assuming the limitations of moderate and high crowded channel, channel 3 and channel 13, respectively. The applied models are shown in Figure 7 and Figure 8. This means that the QoE-QoS relationship for 3D stereoscopic video playback shows cubic correlation with R square of 0.964 on channel 3.

We can recognize a bigger contrast in the case of channel 13, as shown in Figure 8, where a higher QoE were evaluated with better scores at the beginning, but from the threshold bandwidth limitation value of 8Mb/s, a stronger QoE decrease appeared. The QoE-QoS relationship also shows a cubic correlation with R square 0.993. This difference from the logarithmic approaches found in [18] [19] and the correlation model proposed in [12] is caused by 3D video content specifics compared to data centric QoE observations.

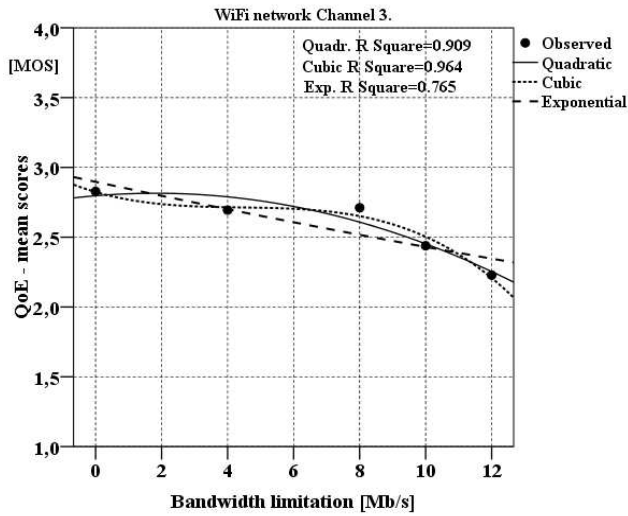


Figure 7

QoE mean scores results for 3D video watching carried on channel 3 and compared with quadratic model, cubic model, and exponential model

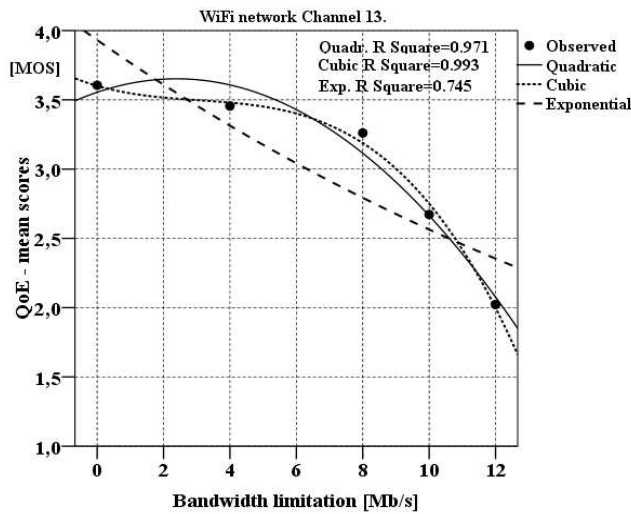


Figure 8

QoE mean scores results for 3D video watching carried on channel 13 and compared with quadratic model, cubic model, and exponential model

In the case of services such as video transport, continuity is the most significant factor in the case of QoE evaluation. We can recognize it from Figure 9, which shows in detail the evaluation of QoE degradation caused by bandwidth limitation for each question. From the boxplots the following observations can be made.

- 1) The rate of continuity was scored the most critically because the highest mean score was only 2.5 (between regular and bad), which was caused by data sequences stuck during high motion level parts of video. The threshold bandwidth limitation value was 8 Mb/s. In the case of 10 Mb/s and 12 Mb/s limitation values, the quality of continuity was unacceptable because of the jerkiness and freezing which occurred during playback.
- 2) The quality of picture was scored much better than continuity, usually between 4 and 3 (good and regular quality), because blurriness did not appear during the experiment, even in the worst case.
- 3) *The assessment of the 3D experience on the whole* was not so much criticized as the continuity, and the best mean score was 3 - good even for 4 Mb/s limitation. This point is interesting, because this means people are still accustomed to 2D screening, and they are more tolerant in the case of 3D quality impairment than in the case of video continuity stalling or short jerkiness. And the 3D QoE, such as the depth of picture, was not so sensitive to the QoS degradation than the screening continuity.
- 4) *Conformity between picture and voice* was scored with the biggest deviation and was acceptable, except in the last two scenarios with 10 Mb/s and 12 Mb/s bandwidth limitation values, when due to heavy continuity degradation, voice quality also rapidly fell off.
- 5) *The quality of 3D video watching, like as on the whole*, was scored with big deviation even in case of no bandwidth degradation. Some people scored it with 4 (good) but some even with 2 (bad); therefore, even the best mean score is only under three (less than regular quality), representing the most subjective part of the experiments.

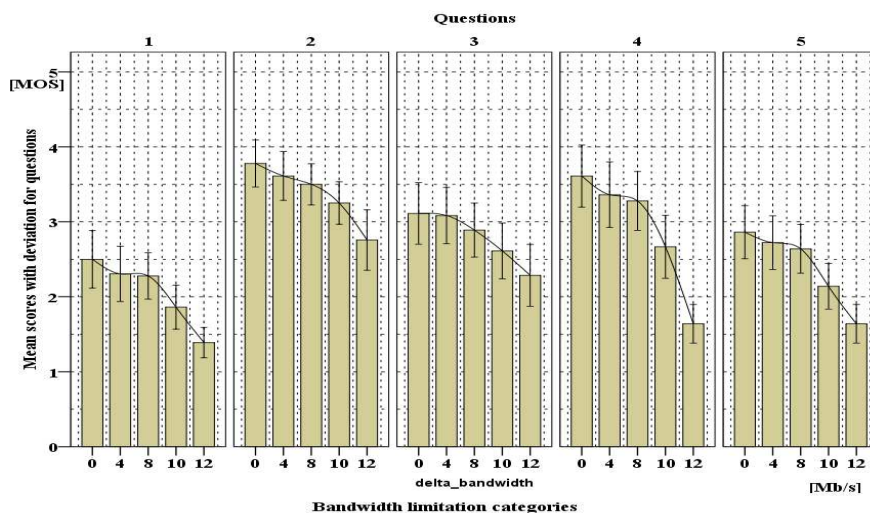


Figure 9

Boxplots mean scores with deviation for each bandwidth limitation values separately and clustered by questions: 1-continuity, 2-quality of picture, 3-3D experience, 4-conformity between picture and voice, 5-3D video vision quality on the whole

Consequently, participants were the most sensitive to the fluidness and continuity of scenes, and the 3D experience was less important when they evaluated the subjective video quality.

Conclusion

Within this paper a complex subjective test method of QoE investigation of 3D stereoscopic video files has been introduced. The GPON network, with its capacity, was suitable for the efficient transport of these contents even in unicast mode.

Firstly, the relationship between the QoE and QoS was shown based on the gathered results for 3D stereoscopic multimedia content, compared with results of the 2D implementation of the same content. The evaluation of data was carried out by IBM Statistics software. QoS metrics such as jitter and throughput limitation disturbance were demonstrated by tests results which showed cubic correlation in both cases. The quality of 3D presentation, such as depth impression, is influenced by multimedia features as well, and dynamic, high-movement sections in video are more sensitive to the QoS degradation.

In Future Internet research, one significant concept is to obtain network neutrality by extending of heterogeneity in the network architecture and service support. In the second part of this article are presented some results of subjective test results of the QoE-QoS relationship character with 36 participants in suitable

environment representing a common future environment, the GPON-based transport network + WiFi sub-network based on IEEE 802.11n in the 2.4 GHz band on the client side. Characteristics of QoE degradation were shown and analyzed on gathered MOS scores of participant experiments. The good quality guarantee is more complex in the case of WiFi access because the QoE is influenced by the nature of wireless technology (such as bandwidth limitation of multiple clients or channel interferences) and by the QoS level in transport network, as well. Robustness of 3D content, QoS degradation and limitation of WiFi network together cause stronger QoE deterioration on the client side.

The goal was to compare gathered experiments with exponential fitting function based on the IQX hypothesis [12] in the case of vision quality investigation of 3D stereoscopic video delivery through a WiFi network. Applying the cubic fitting function to measurement results leads to better correlation with *R Square* values 0.964 and 0.993 than exponential fitting function with *R Square* values 0.765 and 0.745 in the investigated bandwidth limitation interval. This different result was caused by 3D video content delivery service investigation and subjective QoE assessment by users.

Our results show that the fluidness and permanent continuity of video-streams is the most important aspect for good QoE. The primary importance of QoE investigation in wireless network environments has come to the forefront due to worldwide growth of video-stream presentation on smart small mobile devices, and results of this contribution could be helpful for ISPs in the case of 3D based multimedia services.

In the future, more measurements and investigation are needed with various QoS disturbances such as delay, jitter and packet loss in a wireless environment and with explicit channel parameters such as WiFi Access Category, Beacon time, Max. Agg. Frames as long as the resulting A-MPDU fits within the configured TXOP limit, etc. considered. The goal is the mathematical modeling of the functional relationship between QoE and QoS metrics, which is needed for an optimal solution of 3D stereoscopic video contents delivery with appropriate display quality.

Acknowledgement

This work was supported by NKTH-OTKA grant CNK77802.

References

- [1] Jianli Pan; Paul, S.; Jain, R.; "A Survey of the Research on Future Internet Architectures," *Communications Magazine*, IEEE, Vol. 49, No. 7, pp. 26-36, July 2011 doi: 10.1109/MCOM.2011.5936152
- [2] Zahariadis T., Daras P., Laso-Ballesteros I. „Towards Future 3D Media Internet” *Network & Electronic Media – Summit*, St. Malo France, October 2008

-
- [3] Casas P., Belzarena P., Vaton S. „End-2-End Evaluation of IP Multimedia Services, a User Perceived Quality of Service Approach” 18th ITC Specialist Seminar of Quality of Experience, Karlskrona, Sweden, May 2008, pp. 13-23
- [4] Mrak M., Grgic M., Kunt M. High-Quality of Visual Experience, Chapter 3, You J., Xing L., Perkis A. „Quality of Visual Experience for 3D Presentation – Stereoscopic Image” Signals and Communication technology, 2010, I, pp. 51-77
- [5] Kroeker L. Kirk „Looking beyond Stereoscopic 3D’s Revival” Communications of the ACM, Volume 53, Issue 8, August 2010, pp. 14-16
- [6] Xing L., You J., Ebrahimi T., Perkis A. „Estimating Quality of Experience on Stereoscopic Images” ISPACS 2010 – International Symposium on Intelligent Signal Processing and Communication Systems, Chengdu, December 2010
- [7] Cale I., Salihovic A., Ivekovic M. „Gigabit Passive Network – GPON” ITI 2007 – International Conference on Information Technology Interfaces, Cavtat, Croatia, June 2007, pp. 679-684
- [8] Zilly F., Müller M., Eisert P., Kauff P. „The Stereoscopic Analyzer – An Image-based Assistance Tool for Stereo Shooting and 3D Production” ICIP 2010 – IEEE International Conference, Hong Kong, September 2010
- [9] Häkkinen J., Kawai T., Takatalo J., Leisiti T., Radun J., Hirsaho A., Nyman G. „Measuring Stereoscopic Image Quality Experience with Interpretation Based Quality Methodology” IS&T/SPIE’s International Symposium on Electronic Imaging, San Jose, California USA, January 2008
- [10] Lambooij M., Ijsselsteijn W., Heynderickx I., „Visual Discomfort in Stereoscopic Displays: A Review” Journal of Imaging Science and Technology - May/June 2009, Volume 53, Issue 3, pp. 030201-(14)
- [11] Shibata T., Kurihara S., Kawai T., Takahashi T., Shimizu T., Kawada R., Ito A., Häkkinen J., Takatalo J., Nyman G. „Evaluation of Stereoscopic Image Quality for Mobile Devices Using Interpretation-based Quality Methodology” Proc. SPIE, Vol. 7237 (2009)
- [12] Fiedler M., Hossfeld T., Phuoc Tran-Gia „A Generic Quantitative Relationship between Quality of Experience and Quality of Service” IEEE Network, March/April 2010, Volume 24, Issue 2, pp. 36-41
- [13] Fort S. „2020 3D Media: New Directions in Immersive Entertainment” SIGGRAPH 2010 – International Conference and Exhibition on Computer Graphics and Interactive Techniques, Los Angeles, USA, July 2010
- [14] Quin Dai „A Survey of Quality of Experience” In. Ralf Lehnert (Ed.) EUNICE 2011, LNCS, Vol. 6955, pp. 146-156 Springer, Heidelberg (2011)

- [15] Kulik I., Trinh T. A. „Investigation of Quality of Experience for 3D Streams in GPON” In. Ralf Lehnert (Ed.) EUNICE 2011, LNCS, Vol. 6955, pp. 157-168 Springer, Heidelberg (2011)
- [16] Milton J. S., Arnold J. C. „Probability and Statistics in the Engineering and Computing Sciences” McGraw Hill International Editions (1986)
- [17] International Telecommunication Union (2003) ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology
- [18] ITU-T Rec. G.1030, “Estimating End-to-End Performance in IP Networks for Data Applications,” Nov. 2005
- [19] S. Khirman and P. Henriksen, “Relationship between Quality-of-Service and Quality-of-Experience for Public Internet Service,” 3rd Passive Active Measurement Wksp., Fort Collins, CO, March 2002
- [20] Rugel S., Knoll T. M., Eckert M., Bauschert T. „A Network-based Method for Measuring of Internet Video Streaming Quality” 1st European Teletraffic Seminar, Poznan, Poland, February 2011
- [21] International Telecommunication Union, “Subjective Video Quality Assessment Methods for Multimedia Applications”, rec. ITU-T P.910, 1999
- [22] Garroppo R., Giordano S., Oppedisano F., Procissi G., “A Receiver Side Approach for Real Time Monitoring of IP Performance Metrics”, Proc. of the EuroFGI Workshop on IP QoS and Traffic Control, pp. 169-176, 2007
- [23] ur Rehman Laghari K., Crespi N., Molina B., Palau C. E. , "QoE Aware Service Delivery in Distributed Environment", Advanced Information Networking and Applications (WAINA) 2011 IEEE Workshops of International Conference on, Vol., No., pp. 837-842, 22-25, March 2011
- [24] Zheng H., Boyce J., "An Improved UDP Protocol for Video Transmission over Internet-to-Wireless Networks", Multimedia, IEEE Transactions on, Vol. 3, No. 3, pp. 356-365, Sept. 2001

Colour Space Selection for Entropy-based Image Segmentation of Folded Substrate Images

Magdolna Apró¹, Dragoljub Novaković¹, Szabolcs Pál², Sandra Dedijer¹, Neda Milić¹

¹ University of Novi Sad, Faculty of Technical Sciences, Department of Graphic Engineering and Design, Novi Sad, Serbia, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, E-mail address: apro@uns.ac.rs, novakd@uns.ac.rs, dedijer@uns.ac.rs, milicn@uns.ac.rs

² University of Novi Sad, Faculty of Technical Sciences, Department of Computing and Control Engineering, Novi Sad, Serbia, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, E-mail address: sabolc.pal@rt-rk.com

Abstract: This paper is focused on analysing the effects of the chosen colour space on image segmentation accuracy for a permanent quality control of the folding process. The folding process is one of the basic operations in print finishing, but during this converting operation the printed or non-printed substrates are exposed to high tensile stresses. These stresses can cause coating cracks on the folding line, which decrease the expected aesthetic feature or even the functionality of the product. High production efficiency of the folding process could be provided by a control system for automated visual inspection. Such a quality control algorithm was proposed by the authors in previous papers. Since the proposed algorithm relies on qualitative image segmentation, it is very important to determine all the factors which influence the segmentation quality. This paper investigates the influence of colour spaces. The applied image segmentation algorithm (Maximum Entropy) works on grey-scale images, and therefore only the luminance components of the five selected colour spaces (HSI, HSL, HSV, CIE Lab and CIE xyY) were used. The segmentation quality was determined by using six different measures (quantitative and qualitative), which were combined in order to obtain a single performance measure for algorithm evaluation.

Keywords: colour space; segmentation; fold quality

1 Introduction

Folding a paper is one of the basic print finishing operations and its quality control is done by the machine operator, inspecting the folded paper mostly visually [2]. Besides such technical issues (non-precise register, double sheets folding, paper crinkling), the surface cracking along the folding line must be detected and

prevented or at least minimised during the folding process. There are different fold-ability or crack-resistance evaluation methods well known in the paper industry, for example: residual tensile strength, residual tensile stretch, residual bending stiffness, folding endurance, etc. [22]. Based on these methods under controlled conditions, a detailed investigation can be done for the fold-ability properties and fold line crack-resistances of the paper, but for real-time production quality control they cannot be applied, since they are time consuming and require special equipment. The visual control method which was used in [2], [12], [22] and [25] involves a human observer, and thus the obtained results are not repeatable, are highly dependable the observer's experience and are of a subjective nature. A simple analysis based on the white pixel analysis of the digitalised images of folding lines can be found in [3], [4], [6], [19] and [21]. The presented image analysis gave an objective quality grade for surface cracking using commercial image analysis software, but the evaluations were done separately from the production phase. Although the white pixel analysis solves the subjectivity issue, a more complex pattern analysis is needed to explore the true nature of the surface damage (quantity, distribution, size, length and width of the cracked lines, etc. [9]) and the influence of the printed colour on visual perception and therefore on the aesthetic feature.

With an objective folding quality estimation method implemented in the production process, the above described problems could be overcome. As in other fields of industry, computer vision based quality control can be applied.

A basic Objective Folding Quality Assessment (OFQA) was proposed by authors [15], which was based on a set of image analysis algorithms in combination with neural network (Figure 1).

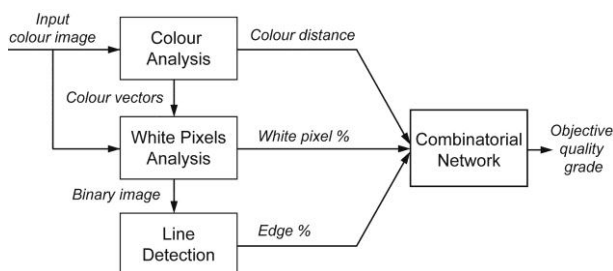


Figure 1

Block diagram of the proposed OFQA algorithm

Due the observed parameters (colour and white pixel analysis, line detection) the presented algorithm obtained objective quality measures of the SAPPI evaluation scale [2] in a good correlation with subjective grades. Although the initial results were promising, they also showed that for the industrial application, further development was necessary to improve the method's accuracy. The development and improvement process included a complete revision of some parts of the

algorithm and it revealed out that the choice of the proper image segmentation method is the most crucial step. On the observed images, in most of the cases, three different areas could be noticed: damaged area, printed area and shadow, and therefore lightness or intensity based monochrome/grey-scale auto-thresholding techniques were selected. Since the final algorithm should work autonomously with a wide range of samples, only algorithms with auto-detected thresholds were considered. In our previous paper [16], different types of grey-scale image thresholding algorithms were analysed, and the best performing algorithms were based on entropy thresholding (the so called Maximum Entropy and Renyi Entropy). As appropriate colour spaces, two colour spaces were selected: in the image processing, the widely used HSL and the perceptual uniform CIE Lab colour space. The obtained results indicated that further investigation was necessary to find the most suitable colour space for the chosen image segmentation method. The aim of this investigation was to evaluate the applicability of different colour spaces, in order to improve the image segmentation accuracy.

2 Methods

2.1 Image Segmentation

Image segmentation is an essential component of many image analysis and pattern recognition applications, conditioning the performance of subsequent analysis steps. It can be defined as the process of partitioning an image into a set of non-overlapping regions whose union is the entire image. Image segmentation techniques can vary widely according to the type of image (e.g., binary, grey, colour), the choice of mathematical framework (e.g., morphology, image statistics, graph theory), the type of features (e.g., intensity, colour, texture, motion) and the approach (e.g., top-down, bottom-up, graph-based). The simplest image segmentation technique is histogram thresholding, which assumes that the histogram of an image can be separated into as many peaks as there are different regions present in the image. The early segmentation algorithms were based on grey-level segmentation (monochrome). With the requirement changes (the emerging need of segmenting colour images) and growth of available computational power, the monochrome segmentation techniques have been extended to segment colour images. Some colour image thresholding approaches consider the 3D histograms that simultaneously contain all the colour information in the image, but since the storage and processing of multidimensional histograms is computationally expensive, most approaches consider 1D histograms computed for one or more colour components in some colour space [8, 11, 18].

In situations where the luminance (intensity) information on an image is discriminative enough, the 1D histogram approach can be used. In other words, if the intensity (grey) levels of pixels of the object of interest are substantially different from the intensity (grey) levels of the pixels belonging to the background, a 1D thresholding algorithm can be used to separate objects of interest from the background. A well-chosen colour space, whose luminance channel carries the most information about the analysed surface, could be much more effective than a segmentation based on all channels, especially if it is applied in real-time in industrial conditions [7, 11, 17].

Based on previous research of the authors [16] which was focused on the evaluation of image segmentation algorithms, Kapur et al.'s method was found to perform the best for the given class of images (images of folded substrates) and it was used further in the evaluation of colour spaces presented in this paper.

The algorithm evaluation was done with the ImageJ open source image analysing software. Since the selected method is implemented under the name "MaxEntropy" correlating to the basic concept of the thresholding method in the following, it will be referred to as Maximum Entropy [10, 23].

The Maximum Entropy thresholding method exploits the entropy distribution of the grey-levels in an image. The principle of entropy is based on uncertainty as a measure of the information contained in a source. The Maximum Entropy thresholding method considers the image foreground and background as two different signal sources, and the optimal image thresholding is achieved when the resulted image preserves as much information as possible, namely when the sum of the foreground and background entropies reaches its maximum [1, 5, 17].

Assuming that the $h(i)$ is the normalized histogram of the analysed image, the optimum threshold for the Maximum Entropy can be defined as following [23]:

$$T_{opt} = \underset{t=0..i_{max}}{ArgMax} [H_B(t) + H_W(t)] \quad (1)$$

Where $H_B(t)$, the entropy of black pixels and $H_W(t)$, the entropy of white pixels, are defined as [23]:

$$H_B(t) = - \sum_{i=0}^t \frac{h(i)}{\sum_{j=0}^t h(j)} \log \frac{h(i)}{\sum_{j=0}^t h(j)} \quad (2)$$

$$H_W(t) = - \sum_{i=t+1}^{i_{max}} \frac{h(i)}{\sum_{j=t+1}^{i_{max}} h(j)} \log \frac{h(i)}{\sum_{j=t+1}^{i_{max}} h(j)} \quad (3)$$

In order to improve the quality of the image segmentation, a smoothing filter can be integrated as a pre-filter step into the processing chain. By removing redundant

details and noise from the input image, the pre-filtering step can reduce the problem of complex textures (halftone and rosette pattern of the observed surfaces). In our previous paper [16], three different filters were tested and evaluated (the Gaussian, Mean and Kuwahara filters). Although the improvement was modest, based on the obtained results, the Mean filter was selected for further use as the smoothing pre-filter. The pre-filtering and the segmentations were done using the ImageJ software, which has a broad range of plug-ins, including the plug-in for the Maximum Entropy, named as MaxEntropy, and the Mean filter.

2.2 Colour Spaces

A colour space is a geometrical representation of colours in a space and allows for specifying colours by means of three components, whose numerical values define a specific colour. They can be distinguished according to their characteristics in the following four families [14].

Primary spaces based on the trichromatic theory, which states that any colour can be expressed as a mixture of three primaries. The primaries correspond to the three types of colour sensing elements (cones) found in the human eye. The primary spaces can be [13, 20]:

- real primary colour space with physically realizable primaries (RGB, rgb),
- and imaginary primary colour space like (XYZ and xyz), whose primaries physically do not exist.

Luminance-chrominance colour spaces represent colours in terms of luminosity (L) and two chromaticity components (Cr_1 and Cr_2). The luminance-chrominance components are derived from the RGB colour space by linear or nonlinear transformations. The luminance-chrominance colour spaces can be classified as [13, 14, 20]:

- perceptually uniform spaces (CIE $L^*u^*v^*$ and CIE $L^*a^*b^*$), which determine the correspondence between the colour distance measured in colour space and the colour difference perceived by a human observer,
- television spaces like (YIQ and YUV), where the luminosity and the chromaticity signals are separated for the signal transmission,
- antagonist (opponent) spaces (wb,rg,by and YC_1C_2) based on the opponent colour theory in order to model the human visual system,
- and other spaces (such as Irg and Yxy or CIE xyY), which cannot be directly classified in the above mentioned sub families but are applied in colour image analysis as well.

Perceptual spaces quantify the colour according to the subjective human colour perception by means of the intensity, the hue and the saturation of colour.

Perceptual spaces can be also considered as luminance-chrominance spaces, since they are consisted of a luminance and two chrominance components. The perceptual colour spaces have luminance – chrominance components expressed by polar coordinates (e.g. HSL, HSV, HSB) [13, 14, 20].

Independent axis (or statistically independent component) spaces result from other spaces by applying mathematical operations that aim at de-correlating individual components ($I_1 I_2 I_3$, $P_1 P_2 P_3$, IJK) [14].

For the needs of this investigation, five colour spaces were selected with luminosity/intensity component. Two colour spaces were selected from the luminance-chrominance group (CIE xyY, CIE Lab) and three forms the perceptual colour space group (HSL, HSV and HSI). The used image acquisition equipment for sample digitalisation obtained the RGB values (sRGB), and therefore for some of the selected colour spaces, the colour components from RGB had to be transformed into XYZ space first, using the transform matrix, and then into the target spaces, applying the adequate calculations/equations. Details about the basic characteristics and transformation equations for intensity channel of selected colour spaces are presented in Table 1 [20].






Table 1
Basic characteristics and transformation equations for selected colour spaces [20]

Colour spaces	Components	Conversion form sRGB to XYZ	Conversion the intensity component
CIE xyY	x, y – chromatic component Y – luminance	$X=m_{11}R+m_{12}G+m_{13}B$ $Y=m_{21}R+m_{22}G+m_{23}B$ $Z=m_{31}R+m_{32}G+m_{33}B$	$Y = Y$
CIE Lab	a – green-red axis b – blue-yellow axis L – luminance	where $m_{11}, m_{12} \dots m_{33}$ are transformation coefficients	$L^* = \begin{cases} 116 \times \sqrt[3]{\frac{Y}{Y^w}} - 16 & \text{if } \frac{Y}{Y^w} > 0,008856 \\ 903,3 \times \frac{Y}{Y^w} & \text{if } \frac{Y}{Y^w} \leq 0,008856 \end{cases}$ where X^w, Y^w and Z^w are tristimulus values of the reference white
HSL	H – hue S – saturation L – brightness	or	$L = \frac{1}{2}(M + m)$ where: $M = \max(R, G, B)$ $m = \min(R, G, B)$
HSV	H – hue S – saturation V – value	$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = [M] \begin{bmatrix} R \\ G \\ B \end{bmatrix}$	$V = M$ where: $M = \max(R, G, B)$
HSI	H – hue S – saturation I – intensity	where M is the transform matrix	$I = \frac{1}{3}(R + G + B)$

2.3 Samples Preparation

The proposed OFQA algorithm was developed for image analysis of the captured images of printed, folded and gathered substrates. In order to bring into account as many as possible different situations from real production process, a test form was prepared based on [12] and [25]. The test form consists of five different printed colour areas and a blank area for folding, as well as areas for print quality assurance. Table 2 shows the enlarged views of the printed areas for folding behaviour (surface damage) analysis, their CMYK notation of ink coverage and their target use case.

Table 2
The printed areas for folding behaviour analysis using CMYK notation

No.	Printed area	CMYK notation	Comment/explanation
1.		C 50%	damage visibility on light halftone pattern
2.		K 50%	damage visibility on dark halftone pattern
3.		K 100%	damage visibility on dark solid tone with total ink coverage of 100%
4.		C 40% + M 40% + Y 50% + K 20%	damage visibility on simulated colour image
5.		C 80% + M 80% + Y 80% + K 80%	folding behaviour at total ink coverage of 320%

The samples were made from uncoated, glossy- and matte-coated paper with basic weights of 100 g/m², 140/150 g/m² and 170 g/m². 50 samples of each paper grade were prepared in machine and cross grain direction 48 hours after printing at standard conditions (a temperature of 22°C, a relative humidity of 55%).

The sample-preparing process included the following operations and equipments:

- a) printing of the test forms was performed on KBA Performa 74 offset machine (process colours: Sun Chemical WORLD SERIES, plates: Agfa Azura TS CtP plates, dampening solution: 3% DS Acedin DH with 8% DS IPA, anti-set-off spray powder: DS 2020 B);
- b) cutting the printed test forms into suitable format for folding was done on a Perfecta 76 high-speed cutting machine (with a clamping pressure of 20 and 25 kN);
- c) folding the test forms in machine and cross direction was performed on a Horizon AFC546AKT folding machine (only one buckle folding used, standard fold rollers: combination of soft polyurethane foam rubber and steel roller, standard roller gap adjustment according to the manufacturer recommendation [24] working speed of 50 m/min).

2.4 Test Images

After the preparation, the samples were digitalised using a commercial digital camera, a flatbed scanner and a USB digital microscope. The basic settings of used equipment are presented in Table 3.

Table 3
Technical parameters and adjustments for used equipment

Used equipment	Canon A520	CanoScan 5600F	Veho VMS-001
Type	Commercial digital camera	Flatbed scanner	USB digital microscope
Colour mode	RGB	RGB	RGB
Embedded colour profile	sRGB	sRGB	-
Resolution	resolution 180 ppi	resolution 1200 ppi	resolution 300 ppi
Bit depth	8	8	8
Format	JPEG	BMP	BMP
Other	no flash, 100% digital zoom, auto white balance focal length of 5,8 mm	-	no light source, magnification of 200X, CMOS sensor

With the obtained digitalisation process, an extensive base image set was derived from the folded samples. A subset of 12 images has been selected from the base set, covering all three digitalisation methods and four different surface textures: halftone cyan, halftone black, solid-tone black and CMYK halftone pattern. The CMYK halftone pattern printed area, with 80% of each process colour was excluded from the evaluation set since the visual appearance of 320% total coverage was very similar to solid tone of black (K 100%). The selected halftone patterns present a particular challenge for automated image segmentation due to the complex texture – the halftone printed surface with damages (see Figure 2a, b and c).



Figure 2

Examples of selected images captured with digital microscope with (a) 50% cyan, (b) 50% black halftone printing and (c) rosette pattern of C 40% + M 40% + Y 50% + K 20%

To recognise the perfect folds without any surface damage is another demanding task for an image segmentation algorithm. In order to evaluate the segmentation quality for this class, 3 images were added to the selected subset, one image for each digitalisation method. Figure 3 presents an example of such a perfect folding.

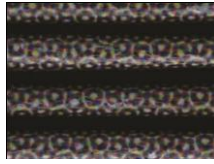


Figure 3

Example of folded paper without surface damages on the folding line

2.5 Quantitative Measures

To evaluate the segmentation performance of used colour representation, six performance measures have been used:

Misclassification error (ME) reflects the percentage of background pixels wrongly assigned to the foreground, and vice versa. For the two-class segmentation problem, ME can be simply expressed as:

$$ME = 1 - \frac{|B_o \cap B_T| + |F_o \cap F_T|}{|B_o| + |F_o|}, \quad (4)$$

where B_o and F_o denote the background and foreground of the ground truth image, B_T and F_T denote the background and foreground areas of the tested image, and $|\cdot|$ is the cardinality of the set. ME varies from 0 for a perfectly classified image to 1 for a total mismatch between reference and tested image [17].

The Hausdorff distance can be used to assess the shape similarity of the thresholded regions to the ground-truth shapes. Since the maximum distance is sensitive to outliers, Sezgin and Sankur [17] proposed a modification where the shape distortion is measured via the average of the Modified Hausdorff distances (MHD) over all objects. It can be defined as:

$$MHD(F_o, F_T) = \frac{1}{|F_o|} \sum_{f_o \in F_o} d(f_o, F_T), \quad (5)$$

where $d(f_o, F_T)$ denotes the minimal Euclidean distance of a pixel in the thresholded image from any pixel in the ground-truth image, and $|F_o|$ is the number of foreground pixels in the ground-truth image. Since an upper bound for the Hausdorff distance cannot be established, the normalization of the MHD metric can be performed by computing its reciprocal (with a small modification):

$$NMHD = 1 - \left(\frac{1}{1 + 0.2 \times (MHD - 1)} \right), \quad (6)$$

The measure derived by this formula has its optimal at 0 and its worst point at 1, as for the ME measure.

Positive false detection (PFD) is the proportion of background pixels wrongly assigned to the foreground object. Normalization (NPFDR) can be done using the overall number of pixels in the image. However, in order to maximize the covered range in $[0, 1]$, the normalization was done using the number of background pixels.

Negative false detection (NFD) is the proportion of foreground pixels wrongly assigned to the background. Normalization (NNFD), following a similar logic as for the PFD, was performed using the number of foreground pixels.

As a derived measure the positive false-negative false detection ratio, or shorten false detection ratio (FDR) is defined, too. It serves as an auxiliary measure to make the balancedness of false detection values easy to read. It is defined as:

$$FDR = \begin{cases} \frac{PFD}{NFD}, & \text{if } PFD \geq NFD \\ \frac{NFD}{PFD}, & \text{else} \end{cases} \quad (7)$$

This measure has a minimum in 1, which is also its optimum, desired value, whereas its maximum value cannot be analytically determined. For this reason, the following method is proposed for the normalization (NFDR):

$$NFDR = 1 - \left(\frac{1}{FDR} \right) \quad (8)$$

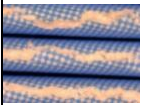
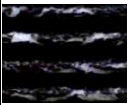

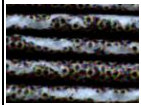
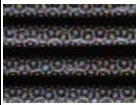

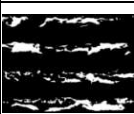
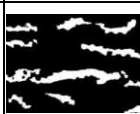
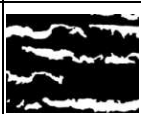

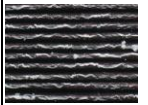




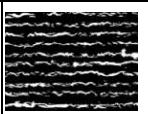
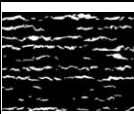
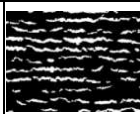
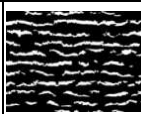



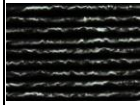


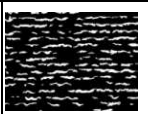
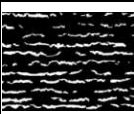
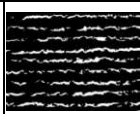
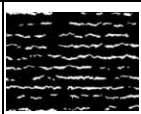

Relative foreground area error (RAE) was first proposed by [17] and is a modification of relative ultimate measurement accuracy (RUMA) measure used by Zhang (as cited in [17]). It is a comparison of object properties, more specifically the area of the detected and expected foreground. It is defined by the following formula:

$$RAE = \begin{cases} \frac{A_0 - A_T}{A_0}, & \text{if } A_T < A_0 \\ \frac{A_T - A_0}{A_T}, & \text{else} \end{cases} \quad (9)$$

where A_0 is the area of reference image and A_T is the area of the thresholded image. For a perfect match RAE is 0, while if there is zero overlap of the object areas, the RAE is 1.

All these measures require a reference or ground truth image, which was derived by hand, segmenting every sample image, marking just the cracked surfaces as foreground objects (see Table 4).

Table 4
Selected test images and their ground truth segmented pair

Type of images		Test images				
		1.	2.	3.	4.	5.
Microscope	Original					
	Ground truth					
Scan	Original					
	Ground truth					
Camera	Original					
	Ground truth					

The attempt to derive a combined measure was presented in [17] with the attempt to simplify segmentation quality evaluation. The authors proposed a simple averaging of the derived measures. However, since all of the measures carry different information about segmentation quality, simple arithmetic averaging might not give the best approximate of the overall quality measure. For this reason a further analysis of the measures were carried out during this research. As the result, the listed measures were divided into two groups: quantitative and qualitative measures. In the quantitative group there was just the ME measure. It defines the amount of misclassified pixels; hence, it is a direct measure of the overall error in detection. All the other measures belong to the second, qualitative group. In order to obtain a single, joint performance score (JPS), the following formula is proposed:

$$JPS = ME * \left(1 + \frac{NMHD + NFDR + RAE}{3} \right) \quad (10)$$

The NPDF and the NNFD were omitted from the JPS since their information is contained in the NFDR measure. However, they were used in further analysis to compare the behaviour of segmentation using different colour spaces. In order to crosscheck the derived results the evaluation was repeated by using just the ME measure. The two evaluation results were then compared.

3 Results and Discussion

A representative test image and its hand-segmented pair are shown in Figure 4a and 4b (as original and ground truth). In Figures 4c, d, e, f and g, the automatically segmented images are presented based on the luminance/intensity component from the HSI, HSL, HSV, CIE Lab and CIE xyY colour spaces, respectively.

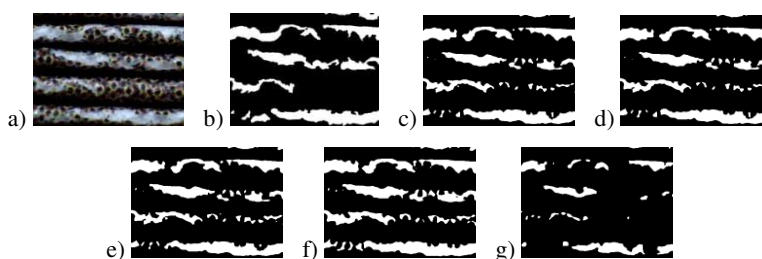


Figure 4

Example of a (a) test image, (b) its hand-segmented pair and the automatically segmented images based on (c) HIS, (d) HSL, (e) HSV, (f) CIE Lab and (g) CIE xyY colour spaces

All test images were processed by the segmentation algorithm using each of the colour spaces. The derived binary images were then pixel-wise compared to the appropriate ground truth image generating all six performance measures. Using (10), the measures were combined to form a single performance measure. These JPS values are presented in Table 5 for all test images and graphically shown in Figure 5.

Based on the average JPS values, the CIE xyY can be recognised as the best performing colour space, producing an average JPS value of 0.2293, which is significantly better than the second best performance (less than 60% of 0.3976). By further analysing each test image, it can be seen that segmentation accuracy using CIE xyY colour space results in an even performance throughout the whole test set (JPS value is around 0.2). There are three exceptions:

- *4_scan* and *4_pict* test images, which are presenting a cyan halftone substrate. Although, the CIE xyY based segmentation performed the best for these samples, significant over-detection could be observed (a detailed analysis on substrate type will be presented latter).

- 5_scan test image, which is showing a black halftone sample, was also poorly segmented with all colour spaces. As for the previous test images, significant over-detection could be observed by a slight advantage for the CIE xyY based segmentation.

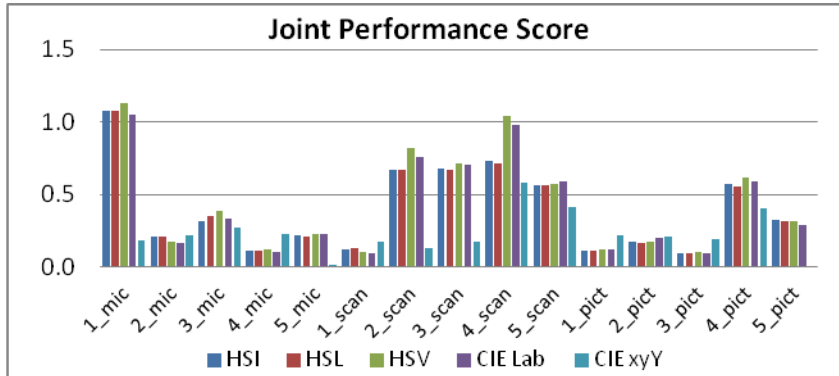


Figure 5

Joint performance score values for all test images and colour spaces

Table 5

Obtained joint performance score values for all test images

Test images	HSI	HSL	HSV	CIE Lab	CIE xyY
1_mic	1.072391	1.075942	1.127071	1.053293	0.185257
2_mic	0.208690	0.207112	0.172954	0.166763	0.218564
3_mic	0.317635	0.356105	0.386754	0.333201	0.276614
4_mic	0.112946	0.112013	0.120381	0.108577	0.228994
5_mic	0.220070	0.213945	0.229888	0.227978	0.019143
1_scan	0.124323	0.126958	0.105902	0.098954	0.178914
2_scan	0.670473	0.671548	0.821479	0.761589	0.133970
3_scan	0.674908	0.667136	0.713143	0.708855	0.174782
4_scan	0.732409	0.715503	1.044099	0.980283	0.586245
5_scan	0.566517	0.565210	0.575609	0.588636	0.415236
1_pict	0.114178	0.111927	0.123873	0.119140	0.217959
2_pict	0.171629	0.170618	0.177982	0.198934	0.211951
3_pict	0.095528	0.094056	0.103709	0.097141	0.190041
4_pict	0.570955	0.559490	0.613510	0.589825	0.402043
5_pict	0.326565	0.317093	0.317557	0.286431	0.000211
AVG	0.398614526	0.397643688	0.442260715	0.421306646	0.229328377

From the derived results (see Table 5), it can also be noticed that colour spaces HSI and HSL performed very similar for the entire set. Their average JPS measures differed by less than 0.001, whereas the highest discrete difference was 0.0385 for the 3_mic sample. Based on the JPS values, HSV can be announced as the worst performing colour space with the highest average JPS value. The CIE Lab colour space performed worse than HSI or HSL, which confirms the first results presented by the authors in [16].

In order to crosscheck the proposed joint measure, the same analysis was performed based only on the ME measure. These results can be seen in Figure 6. As can be noticed, the two charts are similar and the conclusions drawn from them are also very similar, with the differences just in magnitudes. However, it can be observed that the joint measure emphasises some differences according to favourable qualitative measures (see for example 4_mic or 2_scan samples).

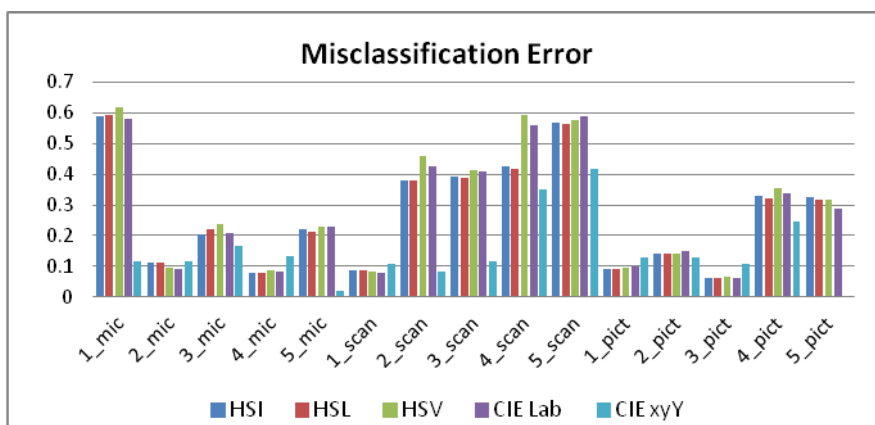


Figure 6

Obtained values for misclassification error for every test image

In order to analyse the performance of image segmentation for different digitalisation techniques and substrate types, the results of JPS were grouped accordingly. The results for image acquisition methods are shown in Figure 7. For this purpose, the results of the first 4 samples were averaged for each group. The fifth sample was omitted because it represents a perfect fold and it will be separately analysed. As can be seen, the best overall results (all colour spaces performed similarly well) were obtained for the acquisition by digital camera. This can be explained by the fact that this type of digitalisation reduces the most the effects of halftone and rosette patterns (because of the resolution and the closeness of the substrates). It is somewhat unexpected that the scanner based digitalisation resulted in the worst segmentations, since this method had ideal illumination, the best resolution and no effects of blurring or geometric distortions. Samples digitalised by scanner are over-detected, which could be explained by the method's high resolution (emphasized halftone and rosette patterns). The samples

derived by digital microscope are segmented somewhat better than those digitalised by the scanner. This is also an unexpected result, since one would expect that the halftone and rosette patterns are the most emphasized by this method. However, it seems that the high magnification helps the algorithm to correctly detect the base paper colour, hence helps avoiding misclassification of printed surfaces.

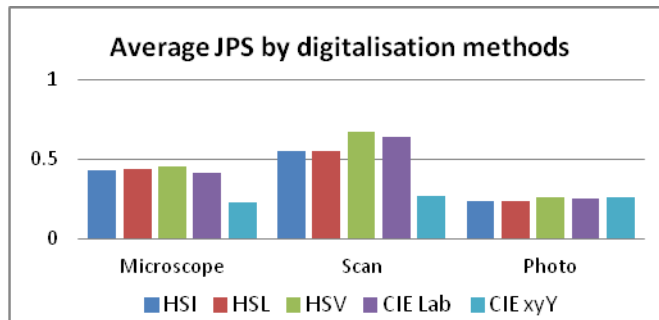


Figure 7

Average joint performance score values by digitalisation methods

Figure 8 presents performances by printed texture types. As was expected, the best results were derived for solid black printed areas (2_mic, 3_pict and 1_scan), where the difference between damages and printed surfaces is the biggest. Samples with rosette pattern are the second best segmented. At the first glance, this result is surprising because of the complex colour pattern (all four process colours are present). However, if we consider that the algorithm is working on grey-scale images, where these differences are less noticeable, then the results are less surprising. Halftone black and cyan samples were the worst segmented. This is especially true for the cyan halftone samples, where besides the halftone pattern, the relatively small difference between the grey-levels of cyan and paper colour renders the correct detection more difficult. It should also be noted that the CIE xyY based segmentation had balanced performance for all printed texture types.

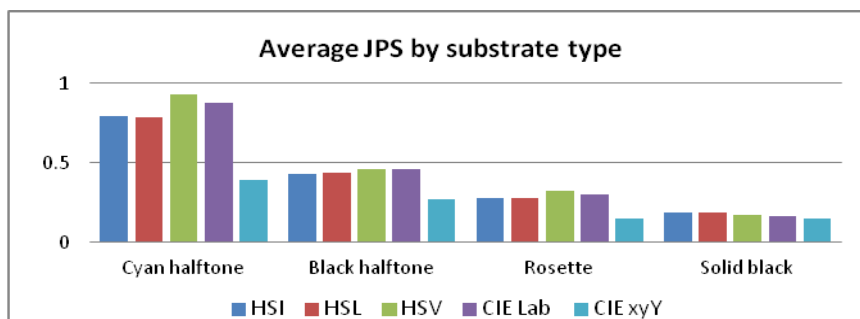


Figure 8

Average joint performance score values by printed texture type

The analysis of perfectly folded substrates (no damages visible) was set apart from the analysis of other substrates because of two reasons. On one hand, most of the measures could not be defined for these samples, making the joint performance score less relevant. On the other hand, these samples are special because there should be no foreground objects detected (there are no damages). This is a serious problem for most of the algorithms, since they are all configured to find a foreground object. Viewing this problem from the folding quality assessment point of view this would mean that there is a problem of detecting substrates, which does not require any action (the folding machine is configured well). The problem is even more serious if we consider that these substrates will be the most often presented to the OFQA. The results of this separate analysis are presented in Figure 9. As can be seen, the CIE xyY based segmentation is superior compared to the other three, having almost no over-detection for two (digitalised by microscope and digital camera) of the samples and 30% less for the scanned sample.

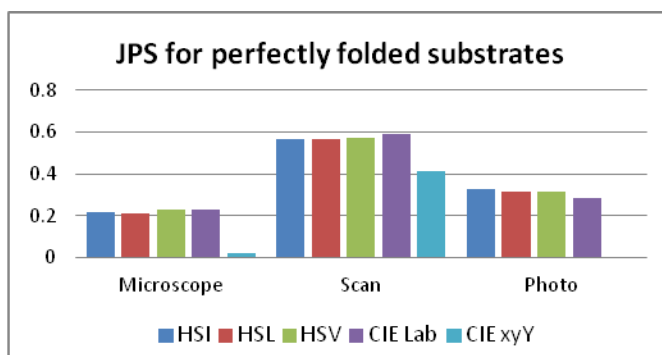


Figure 9

Joint performance score values for perfectly folded substrates

The presented results confirm the first conclusions derived by the average JPS, that the CIE xyY based segmentation performs the best from the chosen group of colour spaces. However, analysing the performances by positive and negative false detections, an oddity of the CIE xyY colour space is revealed. Namely, it tends to under-detect the damaged surfaces. This is visible in Figure 10, where the positive false detections are presented, and in Figure 11, where the negative false detections are presented.

It can be seen that the positive false detection of CIE xyY is the lowest for all samples, whereas for the negative false detection, the situation is the opposite, i.e. it produces significantly higher values. This behaviour is not always favorable, but could be successfully exploited, for example, in a two-stage segmentation process, where this step would be used to initially detect damages.

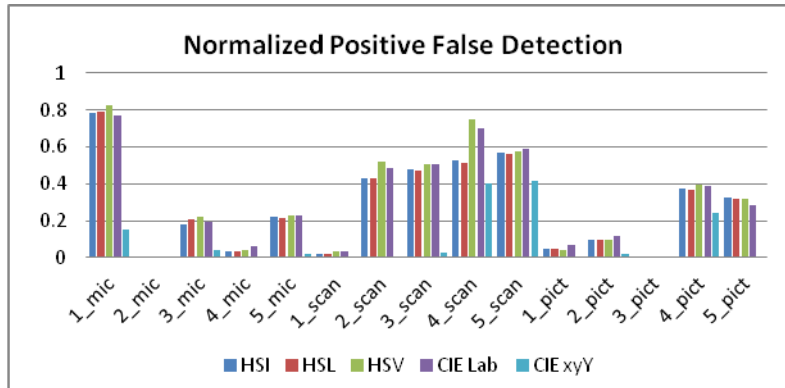


Figure 10

Normalized positive false detection values for all substrates and colour spaces

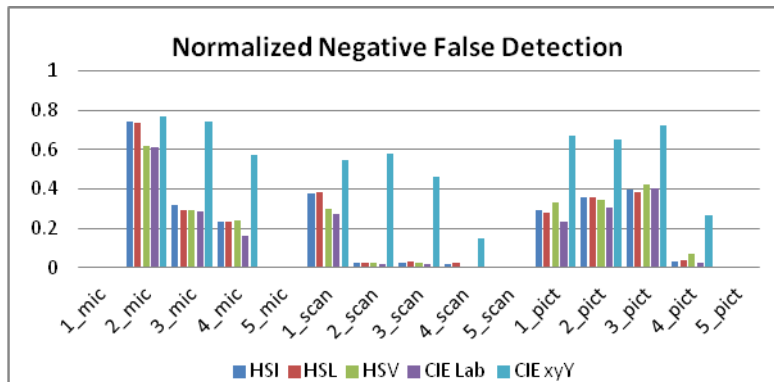


Figure 11

Normalized negative false detection values for all substrates and colour spaces

Conclusion

This paper presented a detailed evaluation of five colour spaces (HSI, HSL, HSV, CIE Lab and CIE xyY) in respect to their influence on image segmentation quality for a given set of test samples. The used segmentation algorithm was the Maximum Entropy algorithm, which is working on a 1D histogram. For this reason only the lightness (luminance) information of the colour spaces was used for evaluation. Six different measures have been derived to determine segmentation quality: misclassification error, modified Hausdorff distance, relative foreground area error, positive and negative false detection and their ratio. In an attempt to combine the measures into a unique grade, as a side effect of the research, a new combination method was proposed, which combines the metrics into a single measure. However, in order to crosscheck the results a separate verification, based only on misclassification error, was also performed. The two analyses gave similar results, showing that the segmentation shows best

performance by using CIE xyY colour space. However, it should be mentioned that segmentation based on this colour space tends to under-detect the damaged areas, while resulting in almost no false positive detections. This feature could be exploited in a two-step segmentation algorithm, where the 1D histogram analysis would be the first step. Also, the analysis of the results showed that the samples digitalised by a digital photo camera as having the lowest JPS, making this method the best choice for acquisition. As was expected, the solid black printed substrates were segmented with the least error, while the cyan halftone substrates were the hardest to segment. Based on the obtained results, further development and improvement of OFQA (Objective Folding Quality Assessment) could be achieved in refining the method's accuracy for real-time and in-line quality control on folding machines.

Acknowledgement

This work was supported by the Serbian Ministry of Science and Technological Development, Grant No.: 35027 "The development of software model for improvement of knowledge and production in graphic arts industry"

References

- [1] A. L. Barbieri, G. F. de Arruda, F. A. Rodrigues, O. M. Bruno, L. da F. Costa: An Entropy-based Approach to Automatic Image Segmentation of Satellite Images, *Physica A*, Vol. 390, 2011, pp. 512-518
- [2] Anon.: Folding and Creasing, Sappi's Technical Brochures, 2nd, revised edition, 2006 (Online) Available from: <http://www.sappi.com/NR/rdonlyres/F3F8F3B0-89B8-4528-9684-7D40C7A5817A/0/FoldingandCreasing.pdf> [Accessed 20 April 2008]
- [3] A. Yang, Y. Xie: From Theory to Practice: Improving the Foldcrack Resistance in Industrially Produced Triple Coated Paper, TappiPaperCon Conference Covington, Kentucky, USA, 2011, pp. 1845-1858 (Online) Available from: <http://www.tappi.org/Downloads/Conference-Papers/2011/2011-PaperCon-Conference/11PAP28.aspx> [Accessed 19 August 2011]
- [4] C. Barbier: On Folding of Coated Papers, Doctoral Thesis no. 56, Royal Institute of Technology, Department of Solid Mechanics, Stockholm, Sweden, 2004 (Online) Available from: http://www.t2f.nu/s2p2/S2P2_MS_9%20.pdf [Accessed 18 April 2008]
- [5] C.-I. Chang, Y. Du, J. Wang, S.-M. Guo, P. D Thouin: Survey and Comparative Analysis of Entropy and Relative Entropy Thresholding Techniques, *IEEE Proceedings of Vision, Image and Signal Processing*, Vol. 153, No. 6, 2006, pp. 837-850
- [6] C.-K. Kim, W.-S. Lim, Y. K. L: Studies on the Fold-Ability of Coated Paperboard (I): Influence of Latex on Fold-Ability during Creasing/Folding

- Coated Paperboard, Journal of Industrial and Engineering Chemistry, Vol. 16, No. 5, 2010, pp. 842-847
- [7] C. Zhang, P. Wang: A New Method of Color Image Segmentation Based on Intensity and Hue Clustering, in Proceedings of 15th International Conference on Pattern Recognition, Barcelona, Spain, Sept 3-7, Vol. 3, 2000, pp. 613-616
- [8] H. Cheng, X. Jiang, Y. Sun, J. Wang: Color Image Segmentation: Advances and Prospects, Pattern Recognition, Vol. 34, No. 12, 2001, pp. 2259-2281
- [9] International Standard: ISO 4628-4:2003, Paints and varnishes. Evaluation of degradation of coatings. Designation of quantity and size of defects, and of intensity of uniform changes in appearance. Assessment of degree of cracking
- [10] ImageJ (2011) Homepage: Image Processing and Analysis in Java, Plugins. (Online) Available from: <http://rsbweb.nih.gov/ij/plugins/index.html#segmentation> [Accessed 15 March 2011]
- [11] J. Delon, A. Desolneux, J. L. Lisani, A. B. Petro: Color Image Segmentation Using Acceptable Histogram Segmentation, Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, Vol. 3523-2005, 2005, pp. 239-246
- [12] J. Eklund, B. Österberg, L. Eriksson, L. Eindenvall: Finishing of Digital Prints – a Failure Mapping, in: Proceedings of the International Congress on Digital Printing Technologies, IS&T NIP 18, San Diego, California, USA, 2002, pp. 712-715 (Online) Available from: www.t2f.nu/t2frapp_f_56.pdf [Accessed 5 May 2008]
- [13] L. Busin, N. Vandebroucke, L. Macaire, J. G. Postaire: Color Space Selection for Unsupervised Color Image Segmentation by Histogram Multi-Thresholding, in: Proceedings of International Conference on Image Processing, ICIP '04, Vol. 1, 2004, pp. 203-206
- [14] L. Busin, N. Vandebroucke, L. Macaire: Color Spaces and Image Segmentation, Advances in Imaging and Electron Physics, Vol. 151, 2008, pp. 65-168
- [15] M. Apro, D. Novaković, Sz. Pal: Objective Fold Quality Evaluation, in: Proceedings of "BlažBaromić" International Conference on printing, design and graphic communications, Senj, Croatia, 2009, pp. 21-24

-
- [16] M. Apro, D. Novaković, Sz. Pal: Evaluation of Image Segmentation Algorithms for Folded Substrate Analysis, in: *Advances in Printing and Media Technology*, Vol. XXXVIII, Edited by: N. Enlund and M. Lovreček, Iarigai, 2011, pp. 209-217, ISBN 978-3-9812704-2-6
- [17] M. Sezgin, B. Sankur: Survey over Image Thresholding Techniques and Quantitative Performance Evaluation, *Journal of Electronic Imaging*, Vol. 13, No. 1, 2004, pp. 146-165
- [18] O. Marques: *Practical Image and Video Processing Using MATLAB*, John Wiley & Sons, Inc. Hoboken, New Jersey, 2011
- [19] P. Alam, M. Toivakka, R. Carlsson, P. Salminen, S. Sandås: Balancing between Fold-Crack Resistance and Stiffness, *Journal of Composite Materials*, Vol. 43, No. 11, 2009, pp. 1265-1283
- [20] P. Colantoni, Al: Color Space Transformations, Technical report, 2004, (Online) Available from: <http://colantoni.nerim.net/download/colorspacetransform98.pdf> [Accessed 5 September 2011]
- [21] P. Rättö, J. Hornatowska: The Influence of Coating Colour Composition on the Crack Area after Creasing, *Nordic Pulp and Paper Research Journal*, Vol. 25, No. 4, 2010, pp. 488-494
- [22] R. E. Popil: Prediction of Fold-Cracking Propensity through Physical Testing, TappiPaperCon Conference, Atlanta, GA, USA, 2010 (Online) Available from: <http://www.tappi.org/Downloads/Conference-Papers/2010/PaperCon-2010-Conference/10PAP101.aspx> [Accessed 19 August 2011]
- [23] S. Jarek: Maximum Entropy Thresholding, 2004, (Online) Available from: http://ij-plugins.sourceforge.net/plugins/segmentation/Maximum_Entropy_Thresholding.pdf [Accessed 3 September 2011]
- [24] User manual for Horizon AFC546AKT cross folding machine, n.d.
- [25] V. Gidlöf, J. Granås, M. Dahlström: Functionality in Digital Packaging Printing, in: *Proceedings of the TAGA conference*, San Antonio, Texas, USA, 2004 [Online] Available from: http://www.t2f.nu/t2frapp_f_140.pdf [Accessed 5 May 2008]

Construction of a Realistic Signal Model of Transients for a Ball Bearing with Inner Race Fault

Lajos Tóth

Department of Electrical and Electronic Engineering
University of Miskolc
H-3515 Miskolc-Egyetemváros, Hungary
e-mail: elkl11@uni-miskolc.hu

Tibor Tóth

Department of Information Engineering
University of Miskolc
H-3515 Miskolc-Egyetemváros, Hungary
e-mail: toth@ait.iit.uni-miskolc.hu

Abstract: This paper considers the creation of a transient vibration signal model established for signals generated in deep groove ball bearings with pitting (spalling) formulation on their inner race. The fault on the inner race was created artificially. The derivation of the signal model is based on data acquisition, signal filtering and parametric identification. A new filtering method is presented that is suitable to eliminate the effect of amplitude modulation and noise that usually arises in the case of bearing vibration measurements. We show that a three-parameter signal model is adequate to describe unmodulated transient pulses. Our signal model can be used in developing new bearing vibration analysis and condition monitoring methods.

Keywords: Condition monitoring; bearing vibration analysis; model identification

1 Introduction

Nowadays, bearings are commonly used components in machinery. This component plays a prominent role in the operation of devices. Failure can therefore not only cause enormous damage, but sometimes put human lives at risk. The failure of rolling element bearings during operation is indicated by the

unusual behaviour of the bearings. Improper operation may be indicated by a rising or increased vibration level in a bearing. The initial state of "unusual behaviour" is usually followed by sudden failure of the bearings. In this way bearing failures may have unforeseen consequences, and therefore the periodic inspection, preventive maintenance and replacement of defective parts is important.

In order to avoid unexpected failures, various procedures have been developed. These include condition monitoring and vibration analysis. All of these methods strongly rely upon the kinematical or dynamic model of the bearing. At the early stage of bearing failure, micro-cracks develop under the rolling surfaces due to the repetitive load on the contacting elements. This usually produces high-frequency (ultrasonic) vibrations that can be sensed by acoustic emission methods. In the next phase of bearing failure, these cracks reach the surface. When two surfaces contact each other in this situation, resonance is excited in the bearing.

The mathematical model of the vibration response of the bearing with a single point defect is investigated in [1, 2, and 3]. The authors in [1 and 2] consider the propagation path of the vibration and the effect of load distribution. This type of mathematical model is an exponentially damped sinusoid function. Later, in [4, 5], the signal model was extended to cases of multipoint defect.

Another approach is to model the bearing as many degrees of freedom (DOF) system. The authors in [6, 7 and 8] use a 2 DOF model, while others in [9] use a 3 DOF model. These models are based on the Hertzian contact theory, considering the centrifugal load effect and the radial clearance.

Our aim was to develop a realistic signal model of vibration response of a bearing with a single-point defect by creating an artificial fault. We set up test equipment to record the vibration response of this type of defect, and we also created a model with an appropriate number of unknowns and tried to find numerical values to fit these variables by parametric identification, where the measured and the theoretical vibration response are in good agreement.

2 Establishing a Vibration Signal Model

Vibrations generated by bearings appear at different frequency ranges. Periodically occurring transient pulses are produced at frequencies determined by bearing geometry and speed. The frequencies of transient pulses depend on the characteristic frequencies of the bearing. There are also low frequency vibrations originating from unbalancing. The subject of our examination is the model of these transient pulses.

2.1 Bearing Selection and Data Acquisition

The primary consideration at bearing selection was its applicability for testing. We had to choose a bearing that can be disassembled and assembled without destruction. Taking into account the available resources and above considerations, a 6204-type, plastic cage, single row, deep groove radial ball bearing was chosen (see Fig. 1).



Figure 1

An assembled and disassembled deep-groove ball-bearing, type 6204

In order to obtain a signal model, we had to form a point-wise fault on the surface of the inner ring of the bearing. Since bearing material is very hard (HRC 58-65), we experimented with applying sulphuric acid and nitric acid on the surface, but in neither case was the fault point wise. Finally, we were able to form a single-point fault (Fig. 2) with an electric arc engraver and with a laser engraver-cutter.

The shape of the failure created by electric arc differs from the ideal circle, while by using a laser beam a hole can be created that is very close to the ideal circular shape. The artificially created fault on the inner race of a deep groove ball bearing is shown in Fig. 3.

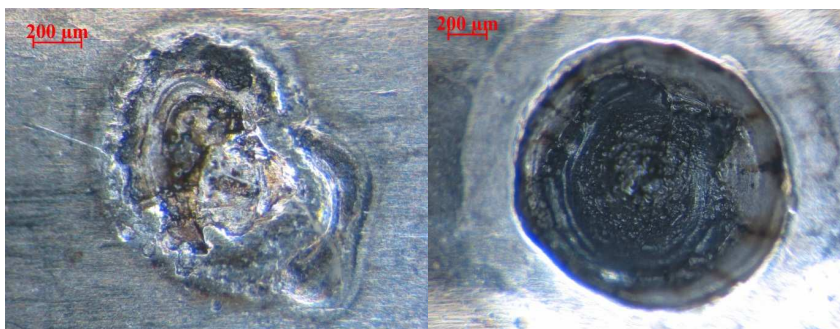


Figure 2

The "crater" formed on the inner race of the bearing by electric arc (left) and by laser beam (right)



Figure 3

An artificially created fault on the inner race of a deep groove ball bearing

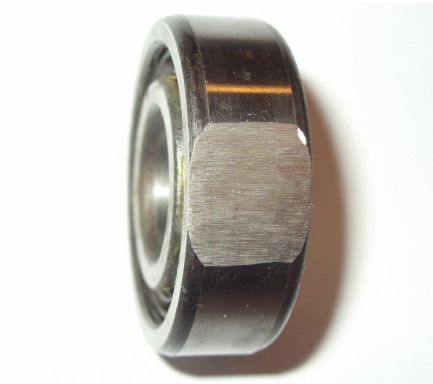


Figure 4

The machined outer ring

For the test rig we used a turning machine of E1N type. The bored and reassembled bearing was mounted on a shaft fixed in the chuck. We used a rod fixed in the tool post as a support. The rotating nature of the tool post made it possible to apply radial load on the outer ring of bearing, where the force was set to be perpendicular to the rotating shaft. Our primary goal was to minimize the force/vibration transmission path, since noise can come from a number of different sources. They can be mechanical or electrical noises. Electrical noise can be eliminated by properly set up measurement devices, while mechanical noise usually comes from the test rig. A portion of the outer ring of the bearing was machined by grinding (Fig. 4). An accelerometer of KISTLER 8702B50 type was attached to the flat area with beeswax.

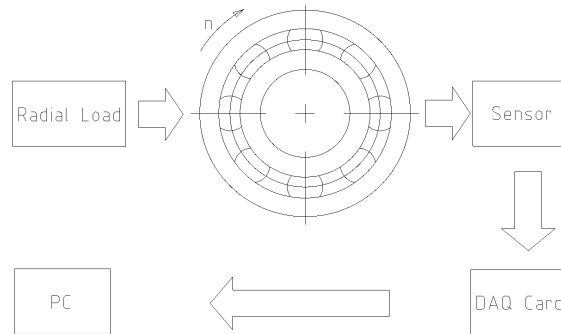


Figure 5
Measurement set up

For data acquisition we used the following devices:

- HAMEG, HM507 analogue-digital oscilloscope, 100 Ms/s real-time sampling rate,
- KISTLER accelerometer 8702B50,
- KISTLER 5108 charge amplifier,
- PCI 6063E PCMCIA DAQ card, 500 ks/s sampling rate.

The DAQ card was controlled by software developed under the NI LabWindows/CVI programming environment. Validation of our software was performed using a HITACHI VG-4429 function generator and digital oscilloscope.

Sampling was performed at a constant inner ring speed of 1812 min^{-1} . This value satisfies the specifications of American ANSI [10] and German DIN [11] standards ($1800 \text{ min}^{-1} \pm 2\%$) concerning bearing vibration measurements. The outer ring was stationary, as it delivered radial load. The sampling frequency and gain were set to be 30 kHz and unity, respectively.

When a ball rolls over the fatigue point, it excites resonance in the bearing at one of the natural frequencies. The amplitudes of excited impulses are proportional to the load and influenced by the load distribution factor. The closer a fault is located to the load zone, the higher the amplitude of excited impulse is. The repetition frequency of impulses is 30.2 Hz and the time between successive impulses is $T=33 \text{ ms}$. The resonance excited by an impact embedded in noise and the corresponding spectrum are plotted in Fig. 6. Since our test rig contains a gearbox to transmit power from the motor to the lathe spindle, the gear mesh frequencies should appear in the spectrum. These frequencies occupy the higher frequency ranges of the spectrum. Using a faultless test bearing of the same type as a reference gauge we were able to distinguish between gear mesh and bearing frequencies.

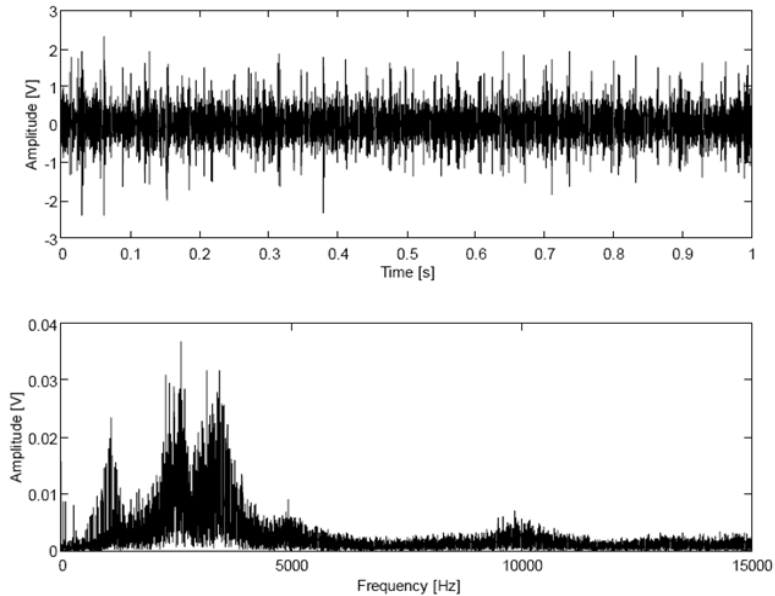


Figure 6

Time plot and corresponding amplitude spectra of the bearing with inner race fault

Bearing defect frequencies could also be calculated from the geometry, which is shown in Table 1.

Table 1

Bearing defect frequencies of 6204 bearing for $n=1812 \text{ min}^{-1}$, stationary outer ring

BPFI	149.339 Hz
BPFO	92.261 Hz
BSF	60.349 Hz
FTF _i	11.533 Hz

We used a 6th-order Butterworth 300-1800 Hz band-pass filter to remove the unwanted “noise”.

The signal at the output stage of filter still contains a considerable amount of noise. To establish the signal model of this kind of bearing failure we need a “noiseless” time signal. To resolve this problem we created a new filtering method that used *a priori* information about the signal.

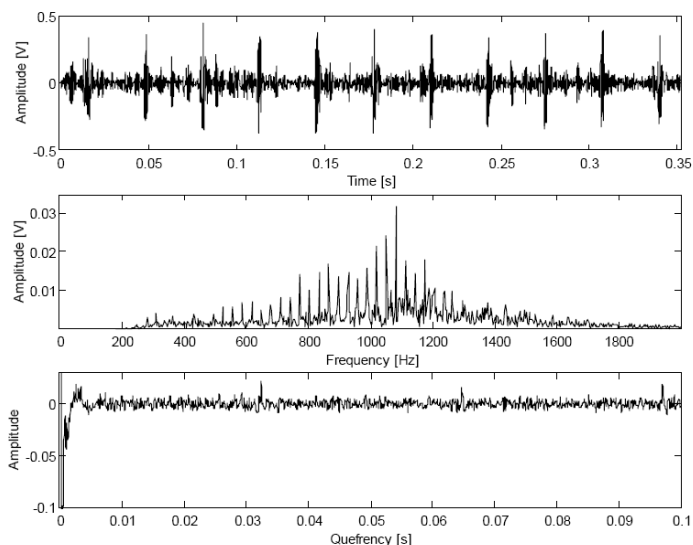


Figure 7

Time plot and corresponding amplitude spectra of the faulty bearing after filtering

2.2 Examination of Bearing Vibration Signals

Bearing vibration signals, especially those which result from pitting formulation, are a series of amplitude modulated transient pulses. The source of amplitude modulation is the load distribution, which is unequal along the inner or outer ring of bearing (Fig. 8). The load distribution (Equation (2)) is expressed in terms of the *load distribution factor* ε . The external radial force generates a number of reactive forces whose amplitude varies with the contact function in Eq. (1). That is, the closer the fault (pitting) is located to the load zone, the higher the amplitude of the transient vibration is.

In case of $\varepsilon < 1$ the load zone can be characterised by the *contact function* Ψ_e [12].

$$\Psi_e = \arccos(1 - 2\varepsilon) \quad (1)$$

The load on a rolling element at arbitrary Ψ_e angle is given as:

$$Q_\psi = Q_{\max} \cdot \left[1 - \frac{1}{2\varepsilon} \cdot (1 - \cos(\Psi)) \right]^n, \quad (2)$$

where Q_{\max} is the maximum load on a rolling element [12], $n = 3/2$ for bearings with point contact (ball bearings), and $n = 10/9$ for bearings with line contact (roller bearings).

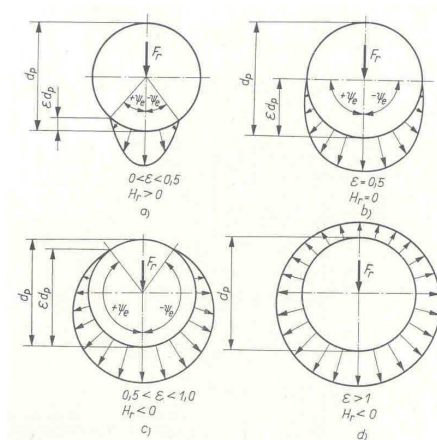


Figure 8

Interpretation of load distribution factor ε [12]

On the basis of static balance, the vertical components of rolling element load should be equal to the external radial load:

$$F_r = \sum_{\psi=0}^{\psi=\pm\Psi} Q_{\psi} \cos(\Psi). \quad (3)$$

Therefore,

$$F_r = Q_{max} \sum_{\psi=0}^{\psi=\pm\Psi} \left[1 - \frac{1}{2\varepsilon} \cdot (1 - \cos(\Psi)) \right]^n \cos(\Psi). \quad (4)$$

Applying *Newton's* second law on (4), i.e. that the acceleration is parallel and directly proportional to the net force, it is clear that the instantaneous amplitudes of vibration acceleration are determined by the radial load corresponding to *contact function*. That means amplitude modulation.

2.3 Filtering of Amplitude Modulated Transient Pulses

Transient signals are finite energy signals by definition (5)

$$E = \int_{-\infty}^{\infty} x^2(t) dt < \infty, \quad (5)$$

where E denotes the energy of signal $x(t)$.

For characterising those in the time domain, parameters such as rise-time, overshoot, settling-time or fall-time can be used. The shorter its time extent, the wider space it occupies from the frequency plane:

$$\mathcal{F}(\delta(t))=1, \quad (6)$$

where $\mathcal{F}(\)$ denotes the Fourier transform operator and $\delta(t)$ is the Dirac delta function.

Real world electrical signals are usually embedded in noise. In the case of vibration measurements we convert the mechanical displacement into voltage. Noise might come from the equipment where the investigated part is located. This usually happens, even if it operates under normal condition. Noise can also be measurement noise that arises during the converting process of mechanical quantities into electrical values. Or it might originate from EMC disturbances. But it can also be quantization error arising during A/D conversion.

Depending on the source of the noise, its spectrum may be located within a specific frequency area or it might occupy the whole frequency range. The filtering of transient signals embedded in noise by conventional methods is difficult, since it is hard to establish the cut-off frequencies of the filters accurately.

In certain engineering processes, periodically repeated transient impulses arise whose amplitude varies with time depending on some physical parameters. This means that the transient impulses amplitude is modulated. The amplitude modulation alters the original pulse spectrum. The location of sidebands depends on the modulation index and the shape of modulating signal. In addition, this spectrum is usually buried in one of the previously mentioned noises, depending on the signal to noise ratio (SNR).

Assuming that each transient impulse is just as likely to appear in the sampled data, and its time course – aside from the differences caused by amplitude modulation – is the same, we obtain the most accurate frequency domain representation if we take as many samples of the pulse as possible. If the transient pulses are generated by a deterministic process, the repetition rate is constant or well defined.

The Fourier transform of absolutely integrable finite-energy signals with only a finite number of local extremes is a continuous function. Amplitude spectra of periodic functions are line spectra, where the distance between individual spectral lines is equal to the frequency calculated from the periodicity. As a result, the Fourier transform of periodically recurring transient signals must be line spectra.

When a periodic signal is amplitude modulated, side bands appear in its spectra. The spectral line corresponds to the carrier signal and the sidebands to the modulating frequency, respectively.

One form of amplitude modulation (AM-DSB, or A3E) in time-domain can be written as:

$$x(t) = [A + m(t)] \cdot c(t), \quad (7)$$

where $x(t)$ is the amplitude modulated signal, $m(t)$ is the information (modulating signal, base band signal), $c(t)$ is the carrier and A is a constant.

The general form of the carrier signal

$$c(t) = C \cdot \sin(\omega_c t + \phi_c), \quad (8)$$

where C , ϕ_c and ω_c are the amplitude, phase and angular frequency of the carrier.

The time function of the modulating signal

$$m(t) = M \cdot \cos(\omega_m t + \phi), \quad (9)$$

where M is the amplitude maxima of the modulating signal, and ϕ and ω_m are the phase and angular frequency of the modulating signal.

The carrier and the modulating signal frequency can be calculated based on Equations (10) and (11).

$$f_c = \frac{\omega_c}{2\pi} \quad (10)$$

$$f_m = \frac{\omega_m}{2\pi} \quad (11)$$

The carrier frequency is always greater than the frequency of the modulating signal:

$$f_c \gg f_m. \quad (12)$$

If $A=0$ then we have double-sideband suppressed-carrier transmission (DSBSC).

The condition of double-sideband amplitude modulation: $A \geq M$.

The modulation depth can be calculated using

$$m = \frac{M}{A} \quad (13)$$

Assuming an additive, uniformly distributed "white noise" $e(t)$ with zero mean, the amplitude modulated signals embedded in noise can be written in the form:

$$y(t) = x(t) + e(t). \quad (14)$$

White noise is a signal where the consecutive samples do not correlate:

$$E\{e(t_1)e^*(t_2)\} = \begin{cases} \sigma^2 \delta_{t_1, t_2}, & t_1 = t_2 \\ 0, & t_1 \neq t_2 \end{cases}, \quad (15)$$

where $E\{e(t)\}$ is the expected value of random variable $e(t)$ (mean value, average value) and the asterisk stands for the complex conjugation

$\sigma^2 = E\{|e(t)|^2\}$ variance of random variable $e(t)$ (power).

The autocorrelation function of a white noise is a pulse at $t_1 = t_2$.

Since the Fourier transform of the autocorrelation function is the power spectral density (PSD), PSD of the white noise is constant throughout the entire frequency range. This means that all frequencies are present in the white noise:

$$\mathcal{F}\{e(t)\} = \sigma_e^2. \quad (16)$$

The Fourier transform of a noisy, amplitude modulated signal is

$$\mathcal{F}\{y(t)\} = A \cdot \mathcal{F}\{c(t)\} + \mathcal{F}\{m(t) \cdot c(t)\} + \mathcal{F}\{e(t)\}, \quad (17)$$

where the first part of the right side of the equation is the Fourier transform of the carrier, the second part is the Fourier transform of the modulated carrier, and the third part is the PSD of white noise. The Fourier transform of the second term in Equation (17) can be calculated as follows:

$$\mathcal{F}\{m(t) \cdot c(t)\} = \frac{1}{2\pi} \hat{m}(\omega) \cdot \hat{c}(\omega), \quad (18)$$

where the hat denotes Fourier transformation.

The harmonics in this term are symmetrical to the higher-frequency carrier. Those harmonics which correspond to the carrier are missing from the spectra.

Using Equation (18) we see that

$$\mathcal{F}\{m(t) \cdot c(t)\} = \frac{1}{2} \pi [\delta(\omega - 4) + \delta(\omega + 2) + \delta(\omega - 2) + \delta(\omega + 4)]. \quad (19)$$

The first part of the right side of Equation (17) represents transient pulses.

$$\hat{C}(\omega) = \mathcal{F}\{c(t)\} \quad (20)$$

This term gives line spectra, since we have periodic function. The distance between successive spectral lines is the aforementioned repetition frequency.

If the amplitude or the frequency of modulating signal changes with time, then the position of spectral lines representing carrier frequencies in Equation (17) does not

change, but the amplitude of side bands (second term) and their distance from the carrier also alter. The location of the sidebands is determined by the modulating signal.

$$\hat{M}(\omega) = \mathcal{F}\{m(t) \cdot c(t)\} \quad (21)$$

Using Equation (12, 18 and 19) Equation (17) can be rewritten as (22)

$$\hat{Y}(\omega) = A \cdot \hat{C}(\omega) + \hat{M}(\omega) + \sigma_e^2. \quad (22)$$

The second term containing modulation information from Equation (22) by spectral sampling can be removed, since it has no spectral component corresponding to the carrier frequency:

$$\hat{Y}_s(\omega) = \sum_{n=0}^N \delta(\omega - n \cdot \Delta f) \cdot \hat{Y}(\omega), n \in \mathbb{Z} \quad (23)$$

where $\hat{Y}_s(\omega)$ denotes the sampled spectra, \mathbb{Z} denotes the set of integer numbers, and Δf is the periodicity we get from Cepstral analysis:

$$c[\tau] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|\hat{Y}(e^{j\omega})| e^{j\omega\tau} d\omega, \quad (24)$$

$$\Delta f = \tau, \quad \text{where } c(\tau) = \max[c(\tau)]. \quad (25)$$

To determine the frequency domain form of transient pulses we can use:

$$\hat{C}(\omega) = \frac{1}{A} \hat{Y}_s - \sigma_e^2 \quad (26)$$

The value A is a constant determined by the amplitude of the modulating signal. When it is unknown, then let $A = 1$. The filtered transient pulses can be obtained by using inverse Fourier transformation

$$c(t) = \mathcal{F}^{-1}\{\hat{C}(\omega)\} \quad (27)$$

The proposed method consists of:

- sampling, considering Shannon's theorem.
- performing Cepstral analysis to determine periodicities in the spectra.
- estimating the noise level using signal spectra.
- sampling the amplitude spectra, according to the results of Cepstral analysis.
- decreasing the amplitude values by the estimated noise level
- applying inverse Fourier transformation.

2.3.1 Application of the Method for Filtering Bearing Vibration Signals

In our case transient pulses correspond to the carrier wave, while the effect of amplitude modulation caused by load distribution factor corresponds to the modulating wave (Fig. 9). Applying the filtering method on sampled data of vibration of laser drilled bearing, we were able to eliminate the effect of amplitude modulation. The time-domain behaviour of each transient in the filtered signal is the same, which makes it possible to establish a signal model of transient vibration for this type of bearing failure.

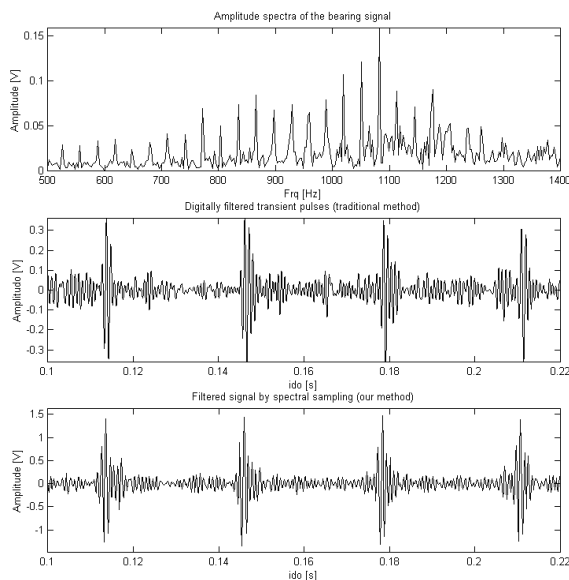


Figure 9

Filtered vibration acceleration signals of the laser drilled bearing

2.4 Model Identification

Model identification can be performed in various ways. Our main idea was signal enveloping followed by curve fitting. Since the transient signal of this type of defect does not contain a frequency modulated component (see Fig. 12), all we have to be concerned with is the amplitude modulating element.

In [3, 12] the signal model of transient pulse generated by a point-wise fault on the inner race of a deep groove ball bearing is defined as an exponentially damped sine function

$$x(t) = A \cdot e^{-Ct} \cdot \sin(\omega_n \cdot t), \quad (28)$$

where $\omega_n = 2\pi f_n$, f_n is the n^{th} natural frequency of bearing system, C is a damping factor, and A is the initial amplitude.

We created our signal model using *a priori* knowledge of the process. The idea is that the transient pulse can be decomposed into two parts: one that is responsible for the amplitude rise and decay of the transient in time, and one that corresponds to one of the natural frequencies of the bearing. Our signal model consists of product of these terms:

$$x(t) = m(t) \cdot n(t), \quad (29)$$

where $m(t)$ is the amplitude modulating part and $n(t)$ is the frequency modulating part.

To determine the envelope function $m(t)$ we calculated the magnitude of the Hilbert transform of the sampled and filtered transient:

$$m(t) = |\mathcal{H}\{x(t)\}|, \quad (30)$$

where $\mathcal{H}\{ \}$ denotes the Hilbert transform operator.

The time plot of the sampled and filtered transient and its envelope is shown in Fig. 10.

Equation (28) assumes a sudden rise in the transient pulse. This may be a valid way of describing a process in which the excitation of material with shock pulse and test of the response is within the same material. The transient pulse emitted by a ball rolling into the fault triggers a series of plastic deformation. Our signal model takes into account the transient formulation and decay process. We assumed that, if coefficients used in our signal model fit well with the envelope of the transient, the signal model can be considered appropriate.

We searched the equation of demodulated transient in the form of the following:

$$m(t) = A \cdot t^n \cdot e^{-C \cdot t}, \quad t \in [0, \infty), \quad A, C, n \in \mathfrak{R}, \quad (31)$$

where A, C, n are process parameters and \mathfrak{R} denotes the set of real numbers.

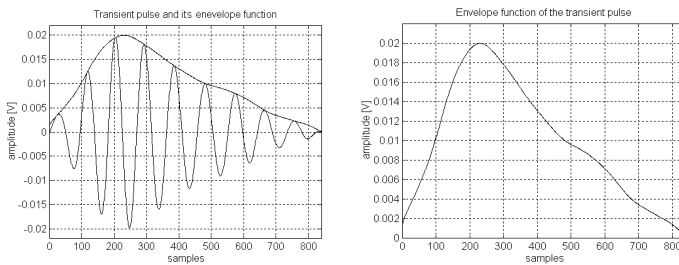


Figure 10

Time plot of sampled and filtered transient pulse and its envelope

We used the *Nelder-Mead* non-linear simplex method to find the best-fit parameters A , n and C that minimize the Mean Squared Error (MSE). The calculated best-fit parameters are shown in Table 3.

Table 3
Best-fit parameters for the demodulated transient in Equation (31)

A	5.5749e-6
n	1.8334
C	0.008
MSE	1.8729e-5

The result of the curve fitting process can be seen in Figure 11.

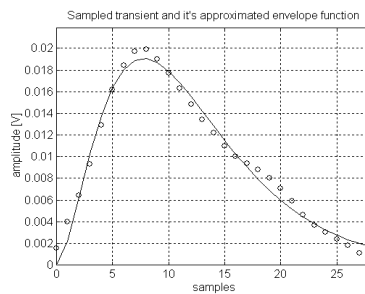


Figure 11

The envelope plot of sampled and approximated data

The curve obtained by approximation displays an adequate fit with the envelope of the sampled and filtered data. Because we were able to find parameters with which our equation well approximates the envelope of the sampled and filtered transient, our signal model can be considered appropriate. We performed the same fitting process with four variables. The result showed that value of the fourth variable was very close to one. Thus we considered the three-parameter model appropriate.

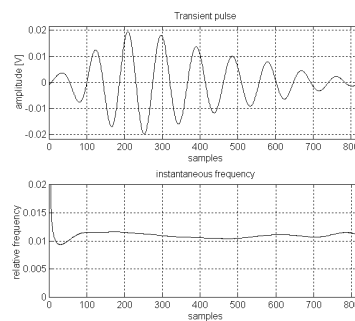


Figure 12

Instantaneous-frequency plot of the transient

To examine the behaviour of the transient in the frequency domain, we calculated the instantaneous-frequency values.

It can be seen in Figure 12 that the instantaneous frequency hardly changes with time. This means that the transient is not modulated in frequency. This gives the second part of (29) in the following form:

$$n(t) = \sin(\omega_n \cdot t), \quad (32)$$

where $\omega_n = 2\pi f_n$, f_n and is one of the natural frequencies of the bearing.

2.5 Model Verification

We used the vibration samples of a faulty bearing of type 6209 to validate our model. After filtering and demodulating the transient we performed the same fitting process as before. We were able to find coefficients which gave reasonable error (Table 4).

Table 4
The best-fit parameters

A	1.0065e-10
N	4.583
C	0.0114
MSE	0.2765

The time plot of the sampled transient pulse of vibration signal of the laser-drilled bearing and its approximated envelope function are shown in Fig. 13.

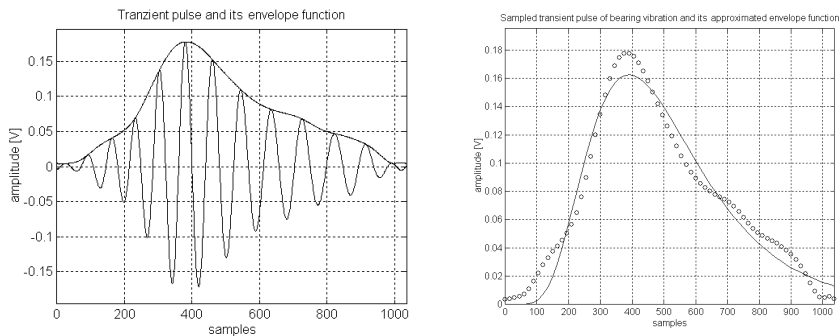


Figure 13

Time plot of sampled signal and simulated transient signal model

3 The Proposed Signal Model

Our studies have shown that our signal model (Eq. (33)) can be used to describe the signal caused by a point-like defect on the inner race of a deep groove ball bearing.

$$x(t) = A \cdot t^n \cdot e^{-Ct} \cdot \sin(\omega_n \cdot t), \quad t \in [0, \infty), \quad A, C, n \in \mathfrak{R} \quad (33)$$

where $\omega_n = 2\pi f_n$, and f_n is one of the natural frequencies of the bearing.

Table 5
Model parameters of Equation (35)

A	5.5749e-6
n	1.8334
C	0.008

For simulation purpose the values shown in Table 5 can be used. But care should be taken as these values change with load.

Conclusions

In this paper we developed a new signal model for a point-wise fault on the inner race of a deep groove ball bearing. A new filtering method was presented which is suitable for eliminating the effect of amplitude modulation and noise that usually arise in the case of bearing vibration measurements. We showed that a three-parameter model is adequate to describe the transient pulses. This model, however, does not take into account the magnitude of the load. By doing so it would be possible to give quantitative values of the model parameters. Our signal model can be used in developing new analysis methods.

Acknowledgement

This research was carried out as part of the TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

References

- [1] McFadden PD, Smith JD., Model for the Vibration Produced by a Single Point Defect in a Rolling Element Bearing, *Journal of Sound and Vibration* 96 (1984), pp. 69-82
- [2] N. Tandon, A. Choudhury, An Analytical Model for the Prediction of the Vibration Response of Rolling Element Bearings due to a Localized Defect, *Journal of Sound and Vibration*, Vol. 205 (1997), pp. 275-292
- [3] S. Ericsson, N. Grip, E. Johansson, L.-E. Persson, R. Sjöberg, J.-O. Strömberg: Towards Automatic Detection of Local Bearing Defects in

- Rotating Machines, Mechanical Systems and Signal Processing 19 (2005) pp. 509-535
- [4] P. D. McFadden, J. D. Smith, The Vibration Produced by Multiple Point Defects in a Rolling Element Bearing, *Journal of Sound and Vibration*, Vol. 98 (1985) pp. 263-273
- [5] A. Choudhury, N. Tandon, A Theoretical Model to Predict Vibration Response of Rolling Bearings to Distributed Defects under Radial Load, *ASME Transactions* Vol. 120 (1998) pp. 214-220
- [6] C. S. Sunnersjo. Varying Compliance Vibrations of Rolling Element Bearings. *Journal of Sound and Vibration*, Vol. 3 (1978) pp. 363-373
- [7] A. Rafsanjani, S. Abbasian, A. Farshidianfar, and H. Moeenfar. Nonlinear Dynamic Modeling of Surface Defects in Rolling Element Bearing Systems. *Journal of Sound and Vibration*, Vol. 319 (2009) pp. 1150-1174
- [8] M. Tadina, M. Boltezar, Improved Model of a Ball Bearing for the Simulation of Vibration Signals due to Faults during Run-up, *Journal of Sound and Vibration*, Vol. 330 (2011) pp. 4287-4301
- [9] H. Arslan and N. Akturk. An Investigation of Rolling Element Vibrations Caused by Local Defects. *Journal of Tribology*, Vol. 130 (2008), pp. 1-12
- [10] American National Standard ANSI/AFBMA Std 13-1970, ANSI B3.13-1970, Rolling Bearing Vibration and Noise (Methods of Measuring)
- [11] Deutsches Institut für Normung DIN 5426, Laufgeräusche von Wälzlagern, Prüfverfahren
- [12] L. Molnár and L. Varga: Gördülőcsapágyazások tervezése (Roller Bearings Design), Műszaki Könyvkiadó, Budapest, 1977 (in Hungarian)

Hybrid Feedback Linearization Slip Control for Anti-lock Braking System

Samuel John¹, Jimoh O. Pedro²

¹ Department of Mechanical Engineering
The Polytechnic of Namibia
P. Bag 13388, Windhoek Namibia
Email: sjohn@polytechnic.edu.na

² School of Mechanical, Industrial and Aeronautical Engineering
The University of the Witwatersrand
P. Bag 3, Johannesburg, South Africa
Email: jimoh.pedro@wits.ac.za

Abstract: The Anti-lock braking system (ABS) is an active safety device in road vehicles, which during hard braking maximizes the braking force between the tyre and the road irrespective of the road conditions. This is accomplished by regulating the wheel slip around its optimum value. Due to the high non-linearity of the tyre and road interaction, and uncertainties from vehicle dynamics, a standard PID controller will not suffice. This paper therefore proposes a non-linear control design using input-output feedback linearization approach. To enhance the robustness of the non-linear controller, an integral feedback method was employed. The stability of the controller is analysed in the Lyapunov sense. To demonstrate the robustness of the proposed controller, simulations were conducted on two different road conditions. The results from the proposed method exhibited a more superior performance and reduced the chattering effect on the braking torque compared to the performance of the standard feedback linearization method.

Keywords: anti-lock braking system; wheel slip; friction model; hybrid systems; feedback linearization; PID

1 Introduction

The anti-lock braking system is a device that senses when the wheels of a vehicle are about to lock during hard-braking and it releases the brakes, so that locking of wheels does not occur. This operation results in the improvement of the longitudinal stability and hence the driver's steering control, thereby improving the driver's ability to avoid obstacles. In addition, the action of the ABS results in maximizing the frictional forces between the tyres and the road, consequently

minimizing the braking distance. Most commercial ABSs have a design objective of maximising the friction force between the tyres and the road surface to achieve shorter braking distance and better steering control [1, 2]. They are implemented using an algorithm that is based on complicated logic rules (table rules) that attempt to capture all possible operating scenarios. These rules are executed by a control computer that switches on and off solenoid valves to ensure the right pressures are delivered to the wheels while avoiding slippage [3]. Current ABS research is based on slip control. The goal of the slip control is to track pre-determined slip trajectory optimally in the face of un-modeled dynamics and external disturbances. To achieve this, correct slip estimation, which is crucial in the performance of the controller, is necessary [4].

The major challenge in controlling the wheel slip is the fact that the tyre and road interaction is highly non-linear. In addition, there are uncertainties from the vehicle dynamics as well as from the road conditions. These challenges therefore require a more robust non-linear control scheme. However, the *Proportional Integral and Derivative* (PID) controller, which is basically a linear controller, have been applied to ABS [5, 1, 6, 7]. This is due to the fact that the PID controller has been a success story in industrial applications [8, 9]. Solyom S. [1] proposed a slip tracking approach in which the design objective is for each wheel to follow a reference trajectory for the longitudinal wheel slip. A quarter-car model is used for the analysis. A gain scheduled PI(D) controller was implemented for the design. Braking commenced from an initial speed of $30m/s$, and the vehicle achieved stopping distances of between $36m$ and $41m$, which is a considerable improvement to currently available ABS. One of the major advantages in [1] is exploring the accuracy of the PID controller and ease of tuning; however, the model did not consider a number of system dynamics, such as the suspension dynamics, braking actuator, and the pitching effect. The PID control method has been known to behave poorly when systems are highly non-linear, and hence Jiang and Gao [7] have proposed a *nonlinear PID* (NPID) controller. The NPID controller incorporates a nonlinear function to the linear PID as the major modification to the linear PID. The method of gain scheduling implemented for the NPID is same for the linear PID. A comparison between the two control methods revealed that the NPID has a better robustness than the linear PID when tested on ABS stopping distance, road conditions and tyre conditions. The NPID performance shows an average of 25% improvement over the linear PID.

The non-linearity of the ABS makes it difficult to capture all the dynamics in a mathematical model, hence the motivation for the introduction of the sliding mode control (SMC). The SMC consists of a robust controller, an equivalent controller and a sliding surface estimator. The robust controller compensates for a broad range of uncertainties, while the equivalent controller tracks the desired slip. However, the major drawback with the SMC is the chattering caused by the non-linearity in the ABS model, which could affect the life-span of the ABS elements.

According to a study by Austin and Morrey [10], it is reported that some researchers have tried to solve the chattering problem by introducing a saturation function in place of the switching sign function for different road conditions. The introduction of the saturation function eliminates the chattering; however, it introduces a steady state error [10, 11]. Some authors have proposed other solutions to the chattering effect of the SMC. Jing et al [12] proposed a moving sliding surface for the slip control, based on global sliding mode control strategy. In this method, unlike in the conventional SMC, the sliding surface moves from an initial condition to the desired sliding surface, thus achieving fast tracking of the desired slip, concluded Jiang et al [12]. This strategy is aimed at eliminating the reaching phase that causes chattering in the conventional SMC method. In addition, the radial basis function of the neural network is used for the sliding mode controller in this work. Simulation results on a quarter-car model comparing the proposed method and the conventional method indicate that the proposed method reduced the chattering.

Another new SMC methodology called the grey sliding mode control method (GSMC), provides robustness to partially unknown parameters and alleviates the chattering in the conventional SMC. This control method is becoming popular as an improvement to the standard SMC. It has been applied to various control problems [13, 14], including the ABS [15, 16]. In their work, Kayancan and Kaynak [15] proposed a grey sliding mode controller for the regulation of the wheel slip, on the basis of the vehicle longitudinal speed. The slip controller anticipates the slip value as a means of control input. Simulations and experimental validations on a quarter-car laboratory ABS test equipment were carried out. Simulations and experimentation on sudden changes in road conditions were conducted to evaluate the robustness of the controller. The proposed controller achieved faster convergence and better noise attenuation than the conventional SMC. It was concluded that the GSMC, with its predictive capabilities, can be a viable alternative approach when the conventional SMC cannot meet the desired performance specifications.

An investigation conducted by comparing the performances of four different control methods (threshold control, PID control, variable structure control and fuzzy logic control [17]) concluded that it is difficult for any single ABS control method to provide optimal control, accuracy and robustness under all possible braking conditions. To compensate for the shortcomings of various ABS controller designs, hybrid controllers have been proposed in the literature. For example, Assadian [18] investigated a mixed H_∞ and fuzzy logic controller. In this study, simulation results showed that using fuzzy logic mapping to vary the commanded slip value based on the vehicle deceleration input provides optimal results with H_∞ as the main controller or regulator of the torque. Park and Lim [19] presented simulation results of a hybrid wheel slip control, employing feedback linearisation control method with an adaptive sliding mode control. The novelty of this work is the introduction of a time delay to the input. It is claimed

that the time delay is necessary to compensate for the actuator's time delay. To compensate for the time delay, the sliding mode controller is incorporated to bound the uncertainties, using a method proposed by Shin *et al* [20]. From the simulation results it was concluded that the system with the time delay exhibits better performance compared with the system without a time delay.

Chattering of the braking torque is observed when using the feedback linearization control method with pole placement. The current work therefore proposes a combination of the feedback linearization method with a PID controller. The goal of this combination is to reduce the chattering effect on the braking torque. The performance of the proposed controller is tested in simulations on two extreme road conditions. The results from the proposed method exhibited a superior performance and reduced the chattering effect on the braking torque compared to the performance of the standard feedback linearization method.

2 Model Dynamics

2.1 Quarter-Car Model

A quarter-car model is used to develop the longitudinal braking dynamics. It consists of a single wheel carrying a quarter mass m of the vehicle, and at any given time t , the vehicle is moving with a longitudinal velocity $v(t)$. Before brakes are applied, the wheel moves with an angular velocity of $\omega(t)$, driven by the mass m in the direction of the longitudinal motion. Due to the friction between the tyre and the road surface, a tractive force F_x is generated. When the driver applies the braking torque, it will cause the wheel to decelerate until it comes to a stop. A two degree of freedom quarter-car model is shown in Figure 1.

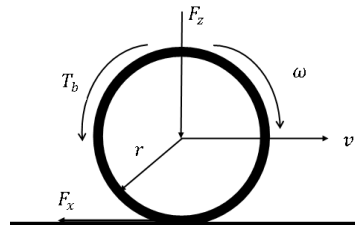


Figure 1
Quarter-car model

Applying Newton's second law of motion, the equations describing the vehicle, tyre and road interaction dynamics during braking are given by Equations (1) and (2).

The equation describing the wheel rotational dynamics is given by:

$$\dot{\omega} = \frac{1}{J} [r\mu(\lambda)F_z - B\omega - T_b(\text{sign}(\omega))] \quad (1)$$

where ω is the angular velocity of the wheel, J is the rotational inertia of the wheel, r is the radius of the tyre, B is the viscous friction coefficient of the wheel bearings and T_b is the effective braking torque, which is dependent on the direction of the angular velocity.

The equation describing the vehicle longitudinal dynamics is given by:

$$\dot{v} = -\frac{1}{m} [\mu(\lambda)F_z + Cv^2] \quad (2)$$

where v is the longitudinal velocity of the vehicle, C is the vehicle's aerodynamic friction coefficient, μ is the longitudinal friction coefficient between the tyre and the road surface λ is the longitudinal tyre slip and F_z is the normal force exerted on the wheel.

The hydraulic brake actuator dynamics is modelled as a first-order system given by:

$$\dot{T}_b = \frac{1}{\tau} (-T_b + k_b P_b) \quad (3)$$

where k_b is the braking gain, which is a function of the brake radius, brake pad friction coefficient, brake temperature and the number of pads [21], and P_b is the braking pressure from the action of the brake pedal which is converted to torque by the gain k_b . The hydraulic time constant τ accounts for the brake cylinder's filling and dumping of the brake fluid [21].

2.2 Friction Model

The friction coefficient between the road and the tyre has a significant influence on the braking or traction of the vehicle. The wheel slip results in the deformation and sliding of tread elements in the tyre/road patch. A simple definition of the longitudinal slip (λ) is given by:

$$\lambda = \frac{v - r\omega}{v} \quad (4)$$

The frictional forces developed between the tyre and the road surface is a complex non-linear problem, which attracted a lot of research work in the eighties to nineties [22, 23, 24, 25]. When the rotation of the wheel around its axle is free, partly or fully locked, three phenomena are likely to take place; these are free

rolling, skidding and full locking. The available maximum acceleration / deceleration of the vehicle body is determined by the maximum friction coefficient describing the contact of the road and the wheels. For this reason, the behaviours of various tyres under various environmental conditions are extensively studied [22, 25, 26, 27, 28]. The physics involved in the modeling of the rolling phenomenon is complex. Bakker et al. [22] in the late eighties and Zanten et al. [29] in the early nineties concentrated on clarifying the role of the so called “*wheel slip*” parameter, which is defined by Equation (4). The *wheel slip* was found to be a critical parameter on which the available maximal friction coefficient depends [30], also shown in Figure 2. A parameter essentially identical to this *wheel slip* is found to be crucial by measurements [31]. However, a practically useful model need not be so complex. The development of a practical friction model may enhance the performance of the ABS controller.

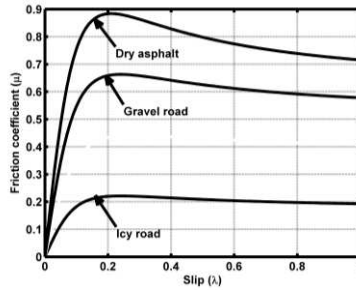


Figure 2
 $\mu - \lambda$ Curves for different road conditions

Equation (4) has a physical deficiency because of the normalization with v . In the definition of λ , no dependence on the *absolute value* of $|v|$ is taken into account, though the relative velocity of the skidding surfaces is determined by $v - r\omega$. Most friction models like the magic formulae [22] are affected by this physical deficiency. The Burckhardt's model given by Equation (5) on the other hand contains the relative velocity of the car body to road, v .

$$\mu_x(\lambda, v) = \left[C_1 \left(1 - e^{-C_2 \lambda} \right) - C_3 \lambda \right] e^{-C_4 \lambda v} \quad (5)$$

where:

C_1 is the maximum value of the friction curve

C_2 is the friction curve shape

C_3 is the friction curve difference between the maximum value and the value at $\lambda = 1$

C_4 is the wetness characteristics, which are in the range $0.02 \leq C_4 \leq 0.04$ s/m

This formulae was evidently developed for $v > 0$ (λ cannot be defined for $v = 0$) and for $r\omega \in [0, v]$. Since ω and v are physically independent variables, different possible ω/v ratios may occur in practice. For example, the $v \rightarrow 0$ situation may happen while $\omega \neq 0$ that makes λ indefinite. Besides this problem, it can be observed that the indefinite quantities may appear in the exponent of the approximating functions in the model. In order to eliminate these numerical difficulties, which could make the results of numerical situations unreliable, the present work has adopted a static tyre-road friction model proposed by [32] given as:

$$\mu(\lambda) = 2\mu_0 \frac{\lambda_0 \lambda}{\lambda_0^2 - \lambda^2} \quad (6)$$

in which λ_0 can be interpreted as the “optimal slip ratio” and μ_0 denotes the available maximum of the friction coefficient. This model has the advantage that it gives a good result when $\lambda \rightarrow \pm\infty$, and its sign modification can also be physically interpreted as turning from braking to accelerating phase, and vice versa. However, it also has the special property that for $v - r\omega = 0$ (i.e. for rolling without skidding) it yields $\mu = 0$. Therefore, it can describe the case of locked wheels when $v = 0$.

3 Controller Design

3.1 Input-Output Feedback Linearization

In the input-output feedback linearization method, the output y is differentiated r times to generate an explicit relationship between the output and the input. For any controllable system of order n , it will require at most n differentiations of the output for the control input to appear. This implies that $r \leq n$, where r is referred to as the *relative degree* of the system. If the control input never appears, the system is uncontrollable or undefined. If on the other hand $r < n$, there will be *internal dynamics*, which cannot be seen from the external input-output relationship. Depending on whether the internal dynamics are stable or unstable, further transformation of the states and analysis is required. The internal dynamics are usually difficult to analyse, and hence the *zero dynamics* are often analysed instead [33, 34]. For a well-defined system (i.e. $r \leq n$), a controller is designed to cancel the non-linearity. For an undefined or uncontrollable system, it is not possible to linearise the system. The current work is based on a well-defined relative degree as will be shown later.

In addition, the feedback linearization control approach is applicable to a class of non-linear systems described by the canonical form given by Equation (7) [35].

$$\dot{\mathbf{x}} = f(\mathbf{x}) + g(\mathbf{x})u \quad (7)$$

$$y = h(\mathbf{x}) \quad (8)$$

where the state variables $\mathbf{x} = [x_1, x_2]^T$ are the wheel angular velocity ω and the vehicle longitudinal velocity v respectively, $f, g : \mathbf{R} \otimes \mathbf{R}$ are smooth functions and y is the output slip function. It will be shown that the ABS problem is a single input single output (SISO) affine non-linear system of the form represented by Equation (7).

The wheel slip dynamics is obtained by taking the derivative of the longitudinal wheel slip (Equation 4) with respect to time, assuming that the radius of the tyre remains constant.

$$\frac{d\lambda}{dt} = \frac{\partial\lambda}{\partial v} \frac{dv}{dt} + \frac{\partial\lambda}{\partial\omega} \frac{d\omega}{dt} + \frac{\partial\lambda}{\partial r} \frac{dr}{dt} \quad (9)$$

$$\dot{\lambda} = \frac{\omega r}{v^2} \dot{v} - \frac{r}{v} \dot{\omega} \quad (10)$$

Substituting (1) and (2) into (10) yields the following:

$$\dot{\lambda} = -\frac{r}{v} \left(\frac{rF_x - T_b}{J} \right) - \frac{\omega r}{v^2} \left(\frac{F_x}{m} \right) \quad (11)$$

Rearranging (11) and knowing that $F_x = \mu F_z(\lambda, \mu_0)$ yields the slip dynamics as

$$\dot{\lambda} = -\frac{1}{v} \left(\frac{\omega}{mv} + \frac{r^2}{J} \right) \mu F_z(\lambda, \mu_0) + \frac{r}{Jv} T_b \quad (12)$$

The slip dynamics represented by Equation (12) resembles Equation (7) where

$$f(x) = -\frac{1}{v} \left(\frac{\omega}{mv} + \frac{r^2}{J} \right) \mu F_z(\lambda, \mu_0), \quad g(x) = \frac{r}{Jv} \quad \text{and} \quad u = T_b. \quad \text{In this case, } f(\mathbb{L}) \text{ and}$$

$g(\mathbb{L})$ are nonlinear dynamic functions. The goal of the ABS is to track a predetermined slip set-point (λ_d). At this operating point, it is safely assumed that $g(x) \neq 0$, and hence the control input can be chosen as:

$$u = \frac{1}{g(x)} [v - f(x)] \quad (13)$$

where v is a virtual input.

The nonlinearity in (12) is therefore cancelled and a simplified relationship between the integral of the output $\dot{\lambda}$ and the new input ν can be presented as:

$$\dot{\lambda} = \nu \quad (14)$$

The wheel slip control is an output tracking problem. The objective is to find a control action $u(t)$ that will ensure the plant follows the desired slip trajectory within acceptable boundaries, keeping all the states variables and controllers bounded. On this basis, the following assumptions are necessary [36]:

Assumption 1

The vehicle velocity v and wheel speed ω are measurable or observable.

Assumption 2

The desired trajectory vector defined within a compact subset of \mathbf{R}^1 , $l_d(t) \hat{\in} U_d$, is assumed to be continuous, available for measurement, and $\|l_d(t)\| \leq W_x$ with W_x as a known bound.

Let the tracking error (e) be given as:

$$e = \lambda(t) - \lambda_d(t) \quad (15)$$

and let the new input be chosen as:

$$\nu = \dot{\lambda}_d - \kappa e \quad (16)$$

where κ is a positive constant. From (13) and (14), the closed-loop tracking error dynamics will be:

$$\dot{e} + \kappa e = 0 \quad (17)$$

3.2 Stability Analysis

Even though the control input (16) is shown to provide perfect tracking, as shown by (17), it is desirable to show that the system is stable in the Lyapunov sense.

If the new input ν is defined as:

$$\nu = \kappa_v \rho + \Lambda_1 e^{(n-1)} + \dots + \Lambda_{(n-1)} e + \dots + \dot{x}_{nd} \quad (18)$$

where the design parameters κ_v and Λ_s are chosen heuristically [37], and ρ is the filtered error signal given by:

$$\rho = \frac{d^{n-1} e}{dt^{n-1}} + \Lambda_1 \frac{d^{n-2} e}{dt^{n-2}} + \dots + \Lambda_{n-1} e \quad (19)$$

Taking the derivative of ρ will yield:

$$\dot{\rho} = \kappa_v \rho \quad (20)$$

Considering the following Lyapunov function:

$$V = \frac{1}{2} \rho^2 \quad (21)$$

the derivative of (21) will yield:

$$\dot{V} = \rho = \frac{d^{(n-1)}e}{dt^{(n-1)}} + \Lambda_1 \frac{d^{(n-2)}e}{dt^{(n-2)}} + \dots + \Lambda_{(n-1)} e \quad (22)$$

It can be seen that $e \rightarrow 0$ exponentially as $\rho \rightarrow 0$ over time, while the design parameters $\Lambda_1 \dots \Lambda_{n-1}$ are chosen so that the system is stable. For example, if before braking $e(0) = \dot{e}(0) = 0$, then $e(t) \equiv 0 \forall t \geq 0$; this signifies perfect tracking of the slip.

4 Proposed Wheel Slip Controller Scheme

The current work proposes a hybrid system that combines feedback linearization (FBL) and PID controllers to realise the hybrid FBLPID controller. The FBLPID hybrid solution for the ABS takes advantage of the FBL approach, in which a non-linear system is transformed into a linear system [38, 34]. This makes it possible to apply a linear PID controller instead of the traditional pole placement controller. The problem identified with the FBL method incorporating pole placement approach for the ABS is that it inherently chatters. Since the linear controller realised from the input-output feedback linearization scheme already contains *proportional* and *derivative* terms of the error signal, the proposed scheme therefore adds an integral term to handle the chattering of the braking torque. The arrangement of the proposed hybrid system is shown in Figure 3, and the governing equation for the PID controller used in the hybrid system is given by Equation 23.

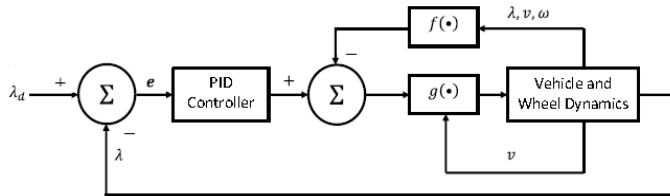


Figure 3
Hybrid FBLPID slip controller structure

$$U(s) = \left[K_p \left(\frac{T_i s + 1}{T_i s} \frac{T_d s + 1}{\psi T_d s + 1} \right) \right] E(s) \quad (23)$$

The Simulink[®] optimization toolbox incorporating a gradient descent search method is used to choose the gains for the PID controller. The gains are presented in Table 1

Table 1
PID gains for hybrid system

Parameter	Gain
K_p	1487
T_i	0.1255
T_d	0.25
ψ	0.025

5 Simulation Results and Discussions

In order to evaluate the performance of the proposed controllers, three performance indices are adopted. These are: the integral squared error [ISE] of the

slip $\int_0^{t_f} (\lambda - \lambda_d)^2 dt$, the integral squared control input $\int_0^{t_f} T_b^2 dt$, and the stopping

distance $\int_0^{t_f} v dt$ [40]. The desired performance will therefore be: small variations

from the desired slip, less effective braking torque, to achieve a shorter stopping distance.

Simulations are conducted on a straight-line braking operation, braking commenced at an initial longitudinal velocity of 80 km/h , and the braking torque was limited to 1200 Nm . In order to impose a desired slip trajectory, the following reference model is adopted [41].

$$\dot{\lambda}_d(t) + 10\lambda_d(t) = 10\lambda_c(t) \quad (24)$$

where the slip command is chosen to be $\lambda_c = 0.18$.

The parameters and numerical values used are presented in Table 2.

Table 2
System parameters and numerical values

Symbol	Description	Value	Unit
m	Quarter-car mass	395	kg
J	Moment of Inertia	1.6	Nms^2 / rad
r	Radius of wheel	0.3	m
C	Vehicle viscous friction	0.856	kg / m
B	Wheel viscous friction	0.08	$N kg m^2 / s$
τ	Hydraulic time constant	0.3	s
κ_b	Hydraulic gain	0.8	Constant
g	Gravitational acceleration	9.81	m / s^2
λ_d	Desired slip	0.18	Ratio

Simulations are conducted for high and low friction surfaces with friction coefficients of $\mu = 0.85$ and $\mu = 0.2$, respectively. These friction coefficients correspond to dry asphalt and icy road conditions, respectively [42]. The simulations are terminated at speeds of about $4km/h$. This is because as the speed of the wheel approaches zero, the slip becomes unstable, therefore the ABS should disengage at low speeds to allow the vehicle to come to a stop.

Simulation results for the vehicle and wheel deceleration for both FBL and FBLPID controllers are shown in Figures 4 to 7 for high and low friction road conditions. The vehicle deceleration is represented by dash-lines while the wheel angular deceleration is represented by a continuous line. Figures 8 to 11 gives the slip tracking plots the desired slip is shown in dash-lines and the tracking slip is shown as continuous line. It can be observed that in some cases the difference between the slip tracking and the desired slip is not obvious; this is a case of perfect tracking. The braking torque simulation results are presented in Figures 12 to 15 for both the standard FBL controller and the proposed hybrid controller, respectively. The summarised performance results are presented in Tables 3 and 4.

The primary objective for developing the hybrid system is to solve the chattering effect observed in the FBL controller performance by enhancing the FBL controller with a PID controller to smooth-out the control action. The proposed hybrid controller achieves stopping distances of $29m$ and $103m$ on high and low friction road conditions, respectively. This yields a 6% improvement on the high friction road and 19% improvement on the low friction road when compared to the standard FBL controller.

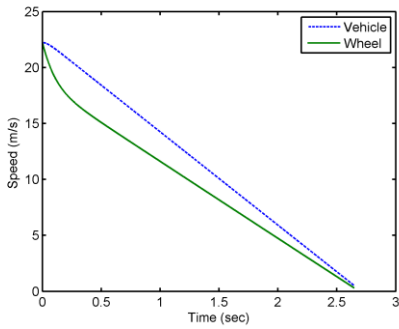


Figure 4
Vehicle & wheel deceleration using FBL controller with $\mu = 0.85$

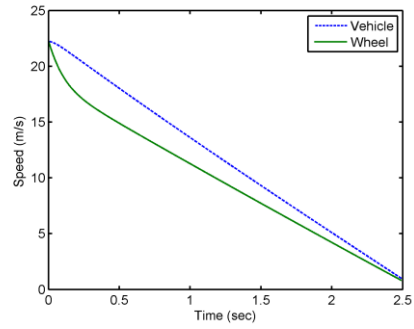


Figure 5
Vehicle & wheel deceleration using FBLPID controller with $\mu = 0.85$

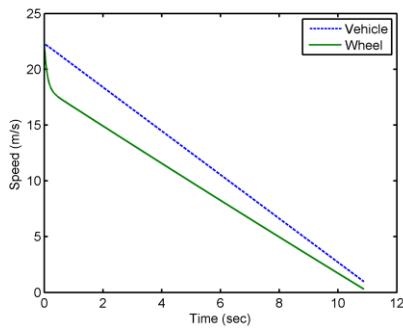


Figure 6
Vehicle & wheel deceleration using FBL controller with $\mu = 0.2$

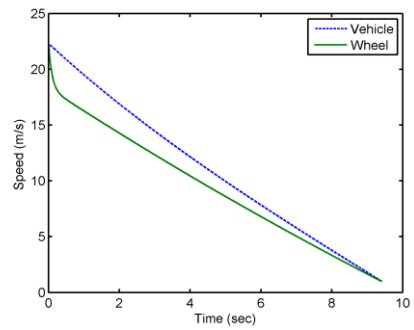


Figure 7
Vehicle & wheel deceleration using FBLPID controller with $\mu = 0.2$

Table 3
Simulation results for $\mu = 0.85$ road condition

Performance index	Specs.	FBL	Proposed-FBL
Integral square error	min	65.99	0.2113
Integral square input $(Nm)^2 \cdot 10^5$	min	2.411	2.083
Stopping distance (m)	≤ 50	31	29

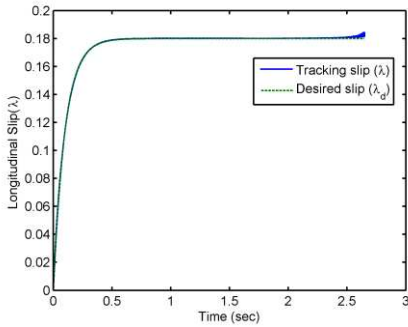


Figure 8
Slip tracking using FBL with $\mu = 0.85$

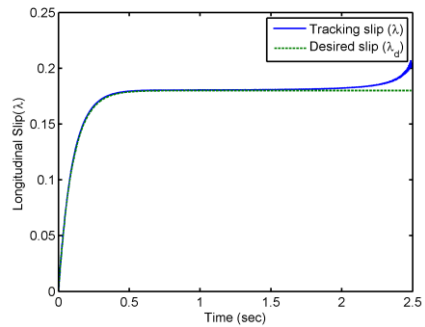


Figure 9
Slip tracking using FBLPID with $\mu = 0.85$

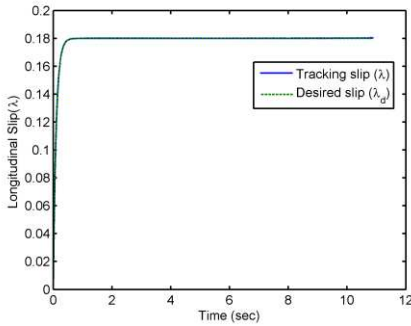


Figure 10
Slip tracking using FBL with $\mu = 0.2$

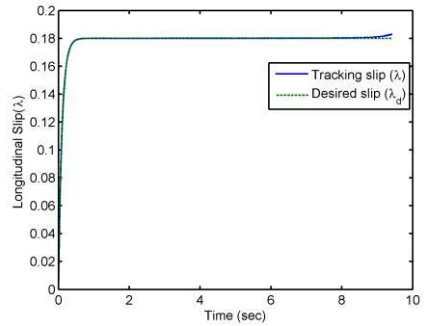


Figure 11
Slip tracking using FBLPID with $\mu = 0.2$

Table 4
Simulation results for $\mu = 0.2$ road condition

Performance index	Specs.	FBL	Proposed-FBL
Integral square error	min	1.412	4.148
Integral square input $(Nm)^2 \cdot 10^5$	min	0.645	0.481
Stopping distance (m)	≤ 50	127	103

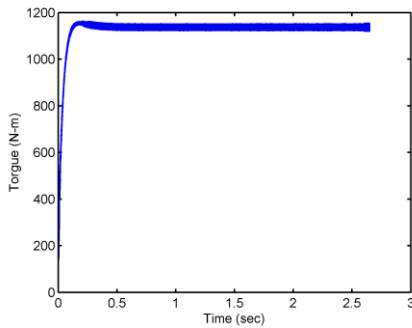


Figure 12

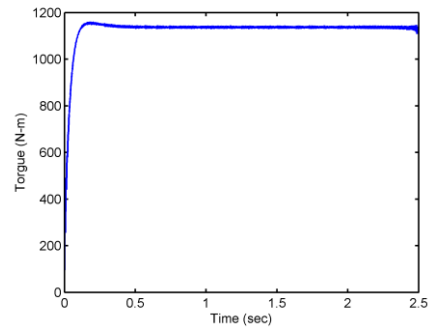
Braking torque using FBL with $\mu = 0.85$ 

Figure 13

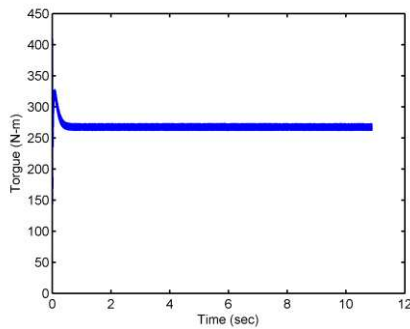
Braking torque using FBLPID with $\mu = 0.85$ 

Figure 14

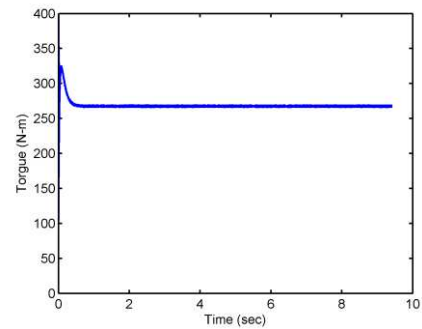
Braking torque using FBL with $\mu = 0.2$ 

Figure 15

Braking torque using FBLPID with $\mu = 0.2$

The hybrid system equally utilises less effective braking torques when compared to the maximum allowable torque. It recorded an effective braking torque of 1135 Nm on the high friction road, which is about 1% higher than the standard FBL, and 267 Nm on the low friction road, about 1.5% above that of the standard FBL. Comparing the plots of the hybrid system in Figures 13 and 15 to those of the standard FBL in Figures 12 and 14, chattering has been reduced considerably. The slightly higher braking torques utilised by the hybrid system is compensated for by the reduced chattering. The effective braking torques plots showed a high initial torque values at the on-set of the braking on all road conditions, but this phenomenon is more pronounced on the low friction road condition. Taking this situation into consideration, therefore, the hybrid controller out-performs the standard FBL as revealed on the effective braking torque performance index in Tables 3 and 4.

The slip tracking plots for the hybrid system are shown in Figures 9 and 11. The hybrid system exhibits a slight unstable slip situation towards the end of the simulation for the high friction road condition. It gives, however, almost perfect slip tracking on the low friction road condition. The hybrid system records lower performance index values than the FBL with respect to slip tracking, as indicated in Tables 3 and 4.

Conclusion and Future Work

This paper proposes a hybrid feedback linearization method in combination with a PID controller for solving the chattering problem observed in the application of the standard feedback linearization to the ABS control problem. The over-all performance of the proposed method demonstrates a superior performance over the standard FBL controller for the ABS. Correlating the plots with the performance results presented in Tables 3 and 4 confirms this assertion. The robustness of both controllers, however, can be seen in their performances at low friction road conditions, where both the transient and steady state conditions of the slip behaved quite well, thereby avoiding excessive slippage.

The scope of this paper covers vehicle dynamics modeling, controller design and analysis and implementation in simulations using the Matlab[®] / Simulink[®] simulation environment. Future work will investigate the application of intelligent-based FBL control scheme.

References

- [1] Solyom S.: Control of Systems with Limited Capacity, PhD Thesis, Lund Institute of Technology, 2004
- [2] Petersen I.: Wheel Slip Control in ABS Brakes Using Gain Scheduled Optimal Control with Constraints, PhD Thesis, Norwegian University of Science and Technology, Norway, 2003
- [3] Wellstead P., Pettit N.: Analysis and Re-Design of an Antilock Brake System Controller, IEE Proceedings on Control Theory and Applications, Vol. 144, No. 5, pp. 413-426
- [4] Aly A. A., Zeidan E. S., Hamed A., Salem F.: An Antilock-Braking System (ABS) Control: A Technical Review, Intelligent Control and Automation, August 2011, Vol. 2, pp. 186-195
- [5] Yoo D.: Model-based Slip Control via Constrained Optimal Algorithm, Master's thesis, RMIT University, 2006
- [6] John S., Pedro J., Pozna C.: Enhanced Slip Control Performance Using Nonlinear Passive Suspension System, Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Hungary, 2011, pp. 277-282

-
- [7] Jiang F., Gao Z.: An Application of Nonlinear PID Control to a Class of Truck ABS Problems, Proceedings of the 40th IEEE Conference on Decision and Control Vol. 1, 2001, pp. 516-521
- [8] O'Dwyer A. Handbook of PI and PID Controller Tuning Rules, 3rd Edition, Imperial College Press, 2009
- [9] Panagopoulos H, Astrom K. J.: PID Control Design and H_∞ loop Shapping, International Journal of Robust and Nonlinear Control, 2000, Vol. 10, No. 15, pp. 1249-1261
- [10] Austin L., Morrey D.: Recent Advances in Antilock Braking Systems and Traction Control Systems, Proceedings of the Institute of Mechanical Engineers, Part D: Journal of Automobile Engineering, January 2000, Vol. 214, No. 6, pp. 625-638
- [11] Buckholtz K. R. Reference Input Wheel Slip Tracking Using Sliding Mode Control, SAE Technical Series, 2002
- [12] Jing Y., Mao Y., Dimirovski G, Zheng S.: Adaptive Global Sliding Mode Control Strategy for the Vehicle Antilock Braking Systems, Proceedings of the American Control Conference, 2009, pp. 769-773
- [13] Lu H. C.: Grey Prediction Approach for Designing Grey Sliding Mode Controller, Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Vol. 1, 2004, pp. 403-408
- [14] Li J., Yibo Z., Haipeng P., Chattering-Free Grey Sliding Mode Control for Discrete Time-Delay Systems with Unmatched Uncertainty, Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, Vol. 1, 2008, pp. 983-986
- [15] Kayacan E., Oniz Y., Kaynak O. A., A Grey System Modeling Approach for Sliding-Mode Control of Antilock Braking System, IEEE Transactions on Industrial Electronics, August 2009, Vol. 56, No. 8, pp. 3244-3252
- [16] Oniz Y., Kayacan E., Kaynak O., Simulated and Experiental Study of Antilock Braking System Using Grey Sliding Mode Control, IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 90-95
- [17] Jun C.: The Study of ABS Control System with Different Control Methods, Proceedings of the 4th International Symposium on Adavced Vehicle Control, Nagoya Japan, 1998
- [18] Assdian F.: Mixed H_∞ and Fuzzy Logic Controllers for the Automobile ABS, SAE Technical Series, March 2001
- [19] Park K. S., Lim J. T., Wheel Slip Control for ABS with Time Delay Input Using Feedback Linearization and Adaptive Sliding Mode Control, Proceedings of the International Conference on Control, Automation and Systems, 2008, pp. 290-295

- [20] Shin H. S., Choi H. L., Lim J. T.: Feedback Linearization of Uncertain Nonlinear Systems with Time Delay, IEE Proceedings on Control Theory and Applications November 2006, Vol. 153, No. 6, pp. 732-736
- [21] Alleyne A.: Improved Vehicle Performance Using Combined Suspension and Braking Forces, Vehicle Systems Dynamics, International Journal of Vehicle Mechanics and Mobility, 1997, Vol. 27, No. 4, pp. 235-265
- [22] Bakker E., Pacejka H., Lidner L.: A New Tire Model with an Application in Vehicle Dynamics Studies, SAE Technical Series, 1998, pp. 83-95
- [23] Pacejka H., Bakker E.: The Magic Formula Tyre Model, Proceedings of the 1st International Colloquium on Tyre Models for Vehicle Dynamics Analysis, 1993, Vol. 21, pp. 1-18
- [24] Oosten J. V., Bakker E.: Determination of Magic Tyre Model Parameters, Proceedings of 1st International Colloquium on Tyre Models for Vehicle Dynamics Analysis, 1993, Vol. 21, pp. 19-29
- [25] Linder L.: Experience with the Magic Formula Tyre Model, 1st International Colloquium on Tyre Models for Vehicle Dynamics Analysis, 1993, Vol. 21, pp. 30-46
- [26] Canudas-de Wit C., Tsiotras P.: Dynamic Tire Friction Models for Vehicle Traction Control, Proceedings of the 38th IEEE Conference on Decision and Control, Vol. 4, 1999, pp. 3746-3751
- [27] Lacombe J.: Tire Model for Simulations of Vehicle Motion on High and Low Friction Road Surfaces, Simulation Conference Proceedings, Vol. 1, 2000, pp. 1025-1034
- [28] Olson B. J.: Nonlinear Dynamics of Longitudinal Ground Vehicle Traction, Master's Thesis, Michigan State University, 2001
- [29] Zantn A, Erhardt R., Lutz A.: Measurement and Simulation of Transients in Longitudinal and Lateral Tire Forces, SAE Technical Series, 1990, pp. 300-318
- [30] Corno M., Savaresi S. M., Balas G. J.: On Linear-Parameter-Varying (LPV) Slip-Control Design for Two-wheeled Vehicles, International Journal of Robust and Nonlinear Control, 2009, Vol. 19, No. 12, pp. 1313-1336
- [31] Yagi K., Kyogoku K., Nakahara T.: Relationship between Temperature Distribution in EHL Film and Dimple Formation, Journal of Tribology, 2005, Vol. 127, No. 3, pp. 658-665
- [32] Lin J. S., Ting W. E.: Nonlinear Control Design of Anti-Lock Braking Systems with Assistance of Active Suspension, Control Theory Applications IET, January 2007, Vol. 1, No. 1, pp. 343-348

-
- [33] Slotine J. J., Li W.: Applied Nonlinear Control, Printice-Hall, New Jersey, USA, 1991
- [34] Hedrick J, Girard A, Feedback linearization, 2005
- [35] Ball S., Barany E., Schaffer S., Wedeward K.: Nonlinear Control of Power Network Models Using Feedback Linearization, Proceedings of the Circuits, Signals and Systems, Oklobdzija, 2005, pp. 493-800
- [36] Yesildirek A., Lewis F.L.: Adaptive Feedback Linearization Using Efficient Neural Networks, Journal of Intelligent and Robotics Systems, May 2001, Vol. 31, pp. 253-281
- [37] Behera L., Kar I.: Intelligent Systems and Control: Principles and Applications, Oxford University Press, 2009
- [38] Henson M. A, Seborg D.E.: Nonlinear Process Control, Prentice-Hall, 1997
- [39] Zhang D., Chen Y., Xie J., Ai W., Yuan C.: A Hybrid Control Method of Sliding Mode and PID Controllers Based on Adaptive Controlled Switching Portion, Proceedings of the 29th Control Conference 2010, pp. 439-445
- [40] Mirzaeinejad H., Mirzaei M.: Anovel Method for Non-Linear Control of Wheel Slip in Anti-Lock Braking Systems, Control Engineering Practice, 2010, Vol. 18, No. 8, pp. 918-926
- [41] Poursamad A.: Adaptive Feedback Linearization Control of Antilock Braking Systems Using Neural Networks, Mechatronics, 2009, Vol. 19, No. 5, pp. 767-773
- [42] MacIsaac Jr J. D., Garrot W. R.: Preliminary Findings of the Effect of Tire Inflation Pressure on the Peak and Slide Coefficients of Friction, Technical Report, National Highway Traffic Safety Administration, Washington D.C., 20590, USA, 2002

Modularized Constraint Management in Model Transformation Frameworks

László Lengyel

Department of Automation and Applied Informatics
Budapest University of Technology and Economics
Magyar tudósok körútja 2, H-1117 Budapest, Hungary
lengyel@aut.bme.com

Abstract: Model-based development methods are increasingly being applied in the production of software artifacts. The processing of visual models, within these frameworks, is an essential issue that can be addressed using graph rewriting techniques. The precise definition of graph rewriting-based model transformation requires that beyond the topology of the rules, further textual constraints be added. These constraints often appear repetitively in a transformation; therefore, constraint concerns crosscut the transformation. It is useful to define often applied constraints as physically separated modules and indicate the places where to use them. This effort provides solutions to structuring, modularizing and propagating repetitively occurring and crosscutting constraints. We propose an aspect-oriented approach that allows for consistent constraint management, in which repetitive and crosscutting constraints can be semi-automatically identified.

Keywords: Aspect-oriented constraints; Constraint aspects; Constraint modularization; Graph rewriting

1 Introduction

Model-based software development [13] [18] applies different software models during system development. Model-based approaches highlight the relevance of model-driven methods in the software industry. They facilitate defining the applications with software models and automatically transform them into executable artifacts.

Model transformations appear in various situations in application development [2]. Graph rewriting is a widely utilized technique for model transformation [8] [9] [19]. Model transformations, like all software, must be validated to ensure their usefulness for each intended application. In [10] [11], an approach has been introduced for validating model transformation that applies Object Constraint Language (OCL) [14]. Constraints are the pre- and post-conditions of

transformation rules. OCL as a constraint and query language in software modeling is an effective way to define textual constraints [3] [5]. We have already demonstrated that it can also be utilized in model transformation definitions [15].

Often we require the validation of several rules or whole transformations, which may cause the same constraint concerns to appear numerous times in a transformation. Regarding this recurrence of constraint concerns, it is beneficial to distinguish between the classical constraint repetition and the crosscutting constraints. According to [17], the definition for the term *concern* is "any matter of interest in a software system".

The classical constraint repetition is similar to the frequently appearing lines of program code in a source file (also known as code clones). In the source code domain, this problem is handled with program segmentation. In most cases, it is implemented with functions; the recurring lines of source code are placed into a function and the function is then called from the appropriate position. This method can be applied to model transformation constraints as well. This can be achieved by extracting the repetitive constraints into separated components and, similarly to function calls, manually designating the points in the model transformation in which they will be applied.

Regarding crosscutting concerns, the situation is significantly different. As opposed to repetitions, crosscutting concerns of a design cannot be modularly separated. If a concern attempts to decompose, according to a specified design principle, other concerns will crosscut this decomposition. This implies that crosscutting is relative to each particular decomposition.

To summarize crosscutting concerns, there is no way to achieve a modular design. In the case of repetitive constraints, consistent constraint management is difficult. In order to mitigate these issues, our aim is to physically separate the different concerns, namely the structure of the transformation rules and constraints, and design them separately. Next, using a weaving mechanism, we generate the executable artifact that combines the two concerns. This generated representation, containing both repetitive and crosscutting constraints, is similar to a binary file compiled from source code and is not edited by the transformation engineer. Therefore, no problems arise, despite the generated artifact concerns not being separated.

The approach presented in this paper provides solutions for both repetitive and crosscutting constraints. Our previous works [10] [12] have already introduced the problem of crosscutting constraints in model transformations. In order for this paper to be self-contained, we briefly summarize the constructs and methods we have developed for crosscutting constraint management in model transformations. The novel results provided by this paper are: (i) the distinction of repetitive and crosscutting constraints in model transformations, (ii) the mechanism that handles the repetitive constraints and (iii) a generalized, semi-automatic identification of repetitive and crosscutting constraints.

With the help of a case study, the next section introduces the problem of repetitive and crosscutting constraints in model transformations. Section 3 gives background information about our model transformation framework and introduces the approach developed for managing crosscutting constraints. In Section 4, we identify the difference between repetitive and crosscutting constraints and discuss the handling of repetitive constraints in model transformations. Section 5 provides a generalized method for semi-automatic detection of repetitive and crosscutting constraints. Finally, concluding remarks are elaborated.

2 Constraint Management Problems in Model Transformations

Graph rewriting [16] is a widely applied technique for graph transformation. The basic elements of graph transformations are graph rewriting rules. Each rule consists of a left-hand side graph (LHS) and right-hand side graph (RHS). Initially, performing a rule requires locating an occurrence (match) in which the rule is applied on the LHS of a graph and replacing this pattern with the RHS. In most model transformation tools, the LHS and RHS of the rules are defined via pattern language [1] [8] [9]. In this case, the structure defined by the pattern language must be found, not an isomorphic occurrence.

A *precondition* assigned to a transformation rule is a Boolean expression that must hold at the moment the rule is fired. A *postcondition* assigned to a transformation rule is a Boolean expression that must hold after the completion of the rule. If a *precondition* of a transformation rule is invalid, then the rule fails without being fired. If a *postcondition* of a transformation rule is invalid after the execution of the rule, then the transformation rule fails. OCL expressions in model transformation rules correlate with it: in the LHS of a transformation rule they represent preconditions, and in the RHS, OCL expressions are postconditions [10].

The dominant decomposition of model transformations provides the functional behavior. The additional constraints ensure the correctness of certain transformation properties. These constraints are responsible for correctness, but often they are treated with secondary importance. They are applied repetitively and in several cases crosscut the transformation. Therefore, it is difficult for the designer to perform the intuitive activities required to verify the transformation.

In order to illustrate the issue of repetitive and crosscutting constraints, a case study is introduced. In [12], a variation of the "class model to relational database management system (RDBMS)" model transformation (also referred to as object-relational mapping) [19] is presented. In Figure 1, using the concrete syntax of our model transformation environment (VMTS, Section 3.1), the control flow model of the transformation is presented.

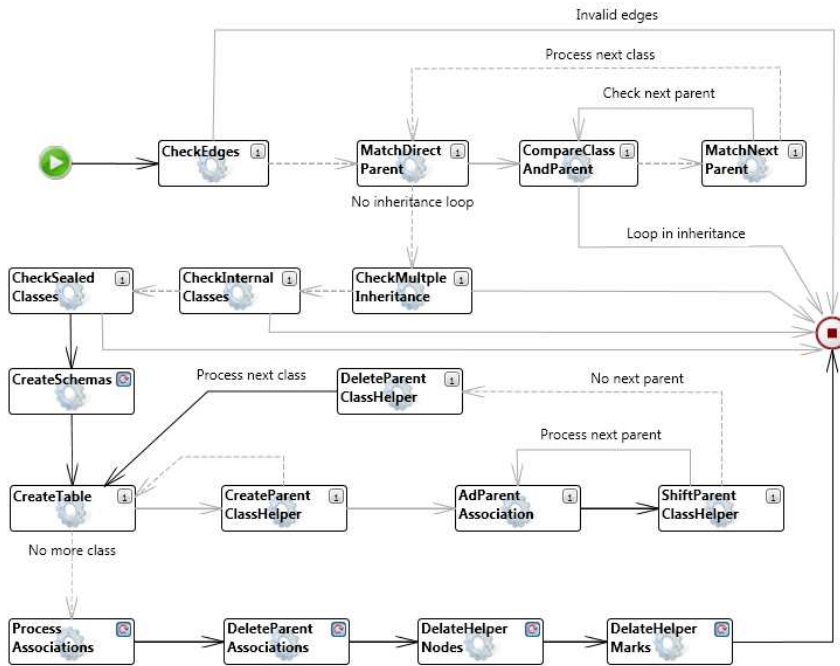


Figure 1

The control flow model of the transformation *ClassToRDBMS*

This model is a stereotypical activity diagram, in which each activity represents a rule. According to the goal of the units, the model can be divided into four parts: (i) The rules *CheckEdges*, *MatchDirectParent*, *CompareClassAndParent*, *MatchNextParent*, *CheckMultipleInheritance*, *CheckInternalClasses*, *CheckSealedClasses* verify the input model. (ii) The rule *CreateSchemas* and the substantial loop in the middle (*CreateTable*, *CreateParentClassHelper*, *AddParentAssociation*, *ShiftParentClassHelper*, *DeleteParentClassHelper*) are responsible for the schema and table creation as well as inheritance-related issues. (iii) The rule *ProcessAssociations* processes the associations. (iv) Finally, the last three rules remove the helper nodes and temporary associations.

In the control flow model, some rules have two outgoing edges. If a rule is successful, then the control is passed via the solid line; otherwise, the dashed line is used. For example, the first rule (*CheckEdges*) is successful if there is at least one dangling edge in the input model. Therefore, the solid outgoing line goes to the end node, because dangling edges are not permitted. If the rule *CheckEdges* was unsuccessful, then the control is passed to rule *MatchDirectParent*.

The first seven transformation rules verify five class diagram-related conditions. We differentiate between class diagram-related conditions that are general language-independent conditions (Conditions 1 and 2), and specific programming

language (Conditions 3, 4 and 5). These condition groups form our well-formedness concerns.

Condition 1. Each association and inheritance edge should connect two nodes because no dangling edges are allowed in class diagrams. This condition is checked by rule *CheckEdges*. The constraints related to this condition are as follows:

```
context Association inv DanglingEdges1:
self.LeftNodeID is NULL or self.RightNodeID is NULL
```

```
context Inheritance inv DanglingEdges2:
self.LeftNodeID is NULL or self.RightNodeID is NULL
```

Condition 2. The '*no directed inheritance loop is allowed*' condition is checked by rules *MatchDirectParent*, *CompareClassAndParent*, and *MatchNextParent*. The rule *MatchDirectParent* selects a class yet to be processed, marks it, then matches its direct parent class. Rule *CompareClassAndParent* verifies that the class marked by a previous rule and actual parent class are not the same. The rule *MatchNextParent* matches the direct parent of the actual class. If the rule has successfully found the next parent, the control is passed to the rule *CompareClassAndParent*, where the originally marked class, the recently found parent, and the actual parent are compared. Otherwise, if there is no next parent, then the transformation continues with rule *MatchDirectParent* in conjunction with the next unprocessed class. If all of the classes have been checked, then the control is passed to rule *CheckMultipleInheritance*. If rule *CompareClassAndParent* finds that a class and its parent (direct indirect) are the same, then the transformation ends with error. The related constraint:

```
context Class inv ClassAndItsParentAreTheSame:
self = self.parentHelper.parent
```

Condition 3. No multiple direct parents are allowed. The condition is checked by rule *CheckMultipleInheritance*. If the rule finds a match where a class has more than one direct parent, then the transformation terminates with error.

Condition 4. The building blocks of software applications are components. They form the fundamental unit of deployment, version control, reuse, activation scoping and security permissions. A component is a collection of types and resources that are built to work in unison to form a logical unit of functionality. If the visibility of the class is set to '*internal*', the type it defines is accessible only to types within the same component. The condition is checked by the rule *CheckInternalClasses*. The constraint related to this rule is:

```
context Class inv CheckInternalCondition:
self.Internal = true and self.neighborClasses->
exists(neighborClass | neighborClass.package <> self.package)
```

Condition 5. If the *'sealed'* attribute of a class is set to true, then other classes cannot be inherited from it. The condition is checked by rule *CheckSealedClasses*. The constraint related to this rule is:

```
context Class inv CheckSealedCondition:
self.Sealed = true and self.childClasses->size() > 0
```

As was mentioned earlier, these conditions are aggregated into condition groups. The groups representing the well-formedness concerns are the *syntactic well-formedness* and the *semantic well-formedness* groups. Syntactic well-formedness conditions represent the general class diagram-related conditions. However, semantic well-formedness conditions are related to a specific programming language. Unfortunately, these concerns are logically scattered across several transformation rules. The syntactic well-formedness concern affects the rules *CheckEdges* and *CompareClassAndParent*. Furthermore, the semantic well-formedness concern affects the rules *CheckMultipleInheritance*, *CheckInternalClasses* and *CheckSealedClasses*. In the current case, these rules are developed based on their functional requirement, meaning they are designed around the functional concern. We could have designed the transformation around the well-formedness concerns, but in that case the rules would have crosscut the well-formedness conditions. In order to achieve the same functionality, transformation rules within a loop should be combined (e.g., with Concurrency Theorem [6]), and other rules should be designed in an unintuitive way.

In conclusion, the transformation cannot be refactored into a modular design in which both transformation rules and well-formedness conditions are elegantly expressed. Therefore, within these rules we can observe valid crosscutting.

The transformation rule *CreateTable* is shown in Figure 2. *CreateTable* works on the non-abstract classes and, based on them, defines tables for the resulting software model. The created table gets the same name based on the class. The table has an additional primary key column and a separate column for each class attribute. The rule matches the package of the class and the schema created for that package. Thus, the rule ensures that the table is created and inserted into the corresponding schema. In addition to these, *CreateTable* creates an edge between the class and its table. With the help of this edge, the subsequent rules can reach the right table from the class.

In order to ensure certain properties and provide validation for the rule *CreateTable*, six different constraints are propagated to it. Because we cannot discuss all transformation rules, we provide statistical data. The transformation *ClassToRDBMS* contains seventeen rules. The constraint *NonAbstract* appears 30 times and the constraint *Abstract* appears 16 times. Furthermore, the constraints *PrimaryKey* and *PrimaryAndForeignKey* are utilized 6 times. The constraints responsible for processing the associations between classes (*OneToOneOrOneToMany* and *ManyToMany*) are used 4 times. Gathering from this, the actual open issue is the repetitively appearing constraints. Further details of the transformation can be found in [12].

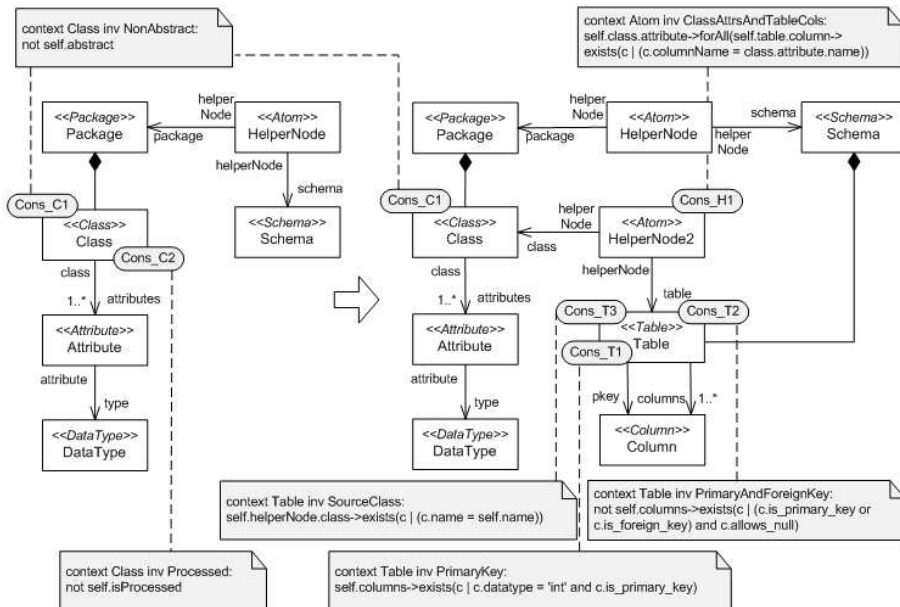


Figure 2
Transformation rule *CreateTable*

The problems of crosscutting and repetitive constraints make understanding both the constraints and model transformation more difficult. Therefore, our goal is to achieve a consistent constraint management by separating constraints and weaving them automatically.

3 Backgrounds

This section introduces the Visual Modeling and Transformation System (VMTS) [20], which is our modeling and model transformation framework. The aspect-oriented constructs provided by the VMTS are also discussed. These aspect-oriented constructs are used in later sections, during the discussion of the novel constraint identification and weaving algorithms.

3.1 The Visual Modeling and Transformation System

Visual Modeling and Transformation System (VMTS) supports domain-specific modeling via metamodeling. Visual metamodel definitions can be extended through textual constraints, defined in OCL.

Furthermore, VMTS is a model transformation system which applies template-based text generation and graph rewriting-based [16] model transformation. Templates are used to produce textual output from model definitions in an efficient way, while graph transformation describes transformations in a visual way. A set of rewriting rules define a graph transformation system. The applications of these rules are the elementary operations of graphs. In our framework, a *model transformation* defines the algorithm of a model processing. We use graph rewriting rules and a control flow graph, which specifies the execution order of the rules. Furthermore, VMTS makes possible the verification/validation of the constraints of the transformation rules.

The results discussed in this paper, handling repetitive constraints (Section 4) and semi-automatic modularization of transformation constraints (Section 5), have been validated in VMTS as a proof-of-concept implementation.

3.2 Managing Constraints in an Aspect-oriented Way

This section provides an overview of aspect-oriented constraint management that was developed to address the problem of the crosscutting constraints in graph rewriting-based model transformations. Depending on the parameterization settings, VMTS provides certain aspect-oriented constraint notions: *aspect-oriented constraints* and *constraint aspects*. In order to turn crosscutting constraints into a coherent module, they are separated from the transformation rules. If a separated constraint can be parameterized by types only in the constraint expression, it is called an aspect-oriented constraint. If a separated constraint is parameterized by a model structure, it is referred to as a constraint aspect. Subsequent sections introduce the concept of aspect-oriented constraints and discuss the advantages of their use in visual model transformations.

The approach presented highlights the different role of the transformation rule constraints and the model constraints. Model constraints, defined in metamodels, should always hold for each instance of a certain metatype. However, in model transformation, preconditions should hold only at the beginning of the rule execution and postconditions at the end of the rule execution. Of course, metamodel constraints hold because the input and output models should be valid instances of the input and output metamodels; this is ensured by the tool during the modeling and can also be checked by the transformation.

3.2.1 Aspect-oriented Constraints

In VMTS, aspect-oriented constraints are OCL constraints; we separate them physically from transformation rules. Weaver algorithms weave them into the rules. The context information of the aspect-oriented constraints is used as a type-based pointcut. This pointcut, based on the metatype information, selects the appropriate rule nodes. This weaving process is referred to as *type-based weaving* [12].

In order to further develop the weaving procedure, we apply weaving constraints. A weaving constraint is similar to a property-based pointcut [7]. This is also an OCL constraint, which restricts the type-based weaving. Obviously, weaving constraint is not added to. Weaving constraints allow for the verification of optional conditions during the weaving process. We refer to it as *constraint-based weaving* [12].

The physically separated constraints require a weaver that applies type-based and constraint-based weaving mechanisms and facilitates the assignment of constraints to transformation rules. Our approach addresses the challenge of aspect-oriented constraint propagation with the Global Constraint Weaver (GCW) algorithm. The GCW algorithm is presented in Section 3.2.3.

3.2.2 Constraint Aspects

In order to make both the constraint weaving process and the constraint evaluation more efficient, we have developed the concept of *Constraint Aspects*. A constraint aspect is a model structure (pattern) to which we assign textual OCL constraints. This means that a constraint aspect, besides the textual conditions, also contains structure information, metatype, and multiplicity conditions, as well as weaving constraints. The structure, metatype conditions and weaving constraints are checked at propagation time, while the OCL constraints are validated during the model transformation.

During the constraint aspect propagation, we search for topological matches throughout transformation rules. These matches must satisfy metatype requirements. Next, the weaving constraints are verified.

In comparison, constraint aspects and aspect-oriented constraints can express the same conditions, but the structure of the constraint aspects makes their propagation to transformation rules more efficient.

3.2.3 Constraint Weaving

The constraint weaving is an offline method that is performed once for a constraint set and once for a transformation. Because of the two different notations of the aspectified constraints, there are also two weaver algorithms in VMTS: the Global Constraint Weaver (GCW) and the Constraint Aspect Weaver (CAW). The GCW algorithm receives the transformation rule, the aspect-oriented constraints and the weaving constraints as input parameters. The CAW receives the transformation rule and the constraint aspects as input parameters. The output of both weavers is the transformation rule with the propagated constraints.

The GCW algorithm, using type-based weaving and applying weaving constraints, weaves the aspect-oriented constraints to the appropriate rule nodes of the transformation rules. The CAW algorithm, using similar methods to GCW, weaves constraint aspects into model transformations.

4 Managing Repetitive Constraints

In our approach, model transformation-related problems concerning validation constraint management are separated into two groups: namely, the management of repetitively appearing constraints and the management of crosscutting constraints. This section clarifies the differences between these two types of constraints and discusses the methods applied for the handling of repetitive constraints.

In software engineering, it is advisable to follow the separation of concerns [4] (SoC) principle. In essence, this indicates that, in dealing with complex problems, the only possible solution is to divide the problem into sub-problems, and then to solve them separately. Next, combine the partial solutions to create a complete solution. One type of concerns, such as rewriting rules, may smoothly be encapsulated within building blocks by means of conventional techniques of modularization and decomposition, whereas the same is not possible for other types. More specifically, these types crosscut the design and are therefore called crosscutting concerns. Because of their specialty, crosscutting concerns raise two significant problems:

- The *scattering problem*: the design of certain concerns is scattered over many building blocks.
- The *tangling problem*: a building block can include the design of more than one concern.

Recall that in the validation of model transformations there are two concerns: the functionality of the transformation and the constraints ensuring the validation. Sometimes modularizing one of the two concerns implies that the other concern will crosscut the transformation, and vice versa.

Both scattering and tangling have several negative consequences for the transformations they affect. However, the aim of aspect-oriented methods is to alleviate these problems by modularizing crosscutting concerns. Therefore, in the case of crosscutting constraints, aspect-oriented methods should be applied in order to achieve consistent constraint management. Both logically coherent constraints (crosscutting constraints) and repetitively appearing constraints should be physically maintained in a modularized manner.

For the problem of crosscutting constraint management, a solution has been provided in [10] and this solution has been summarized in Section 3. Current section provides a novel approach for handling repetitive constraints in model transformations.

4.1 The Constraint Management Process

As we previously stated, consistent constraint management requires a mechanism that supports the handling of repetitive constraints. Our approach provides the following methods for their management:

- Constraints are defined independently from transformation rules. This allows us to maintain the constraints in a physically separated place.
- Constraint calls are defined, along with the designation where the constraints should be applied. Using the generalized version of the Global Constraint Weaver, the approach automatically assigns the constraints to the indicated points of the transformation.

In this approach, the selection of the rules, where the aspect should be propagated (constraint calls), is performed manually by the transformation designer. This method is supported by the weaver tool: the potential transformation rule nodes are offered for the transformation designer, who can manually select those which are required.

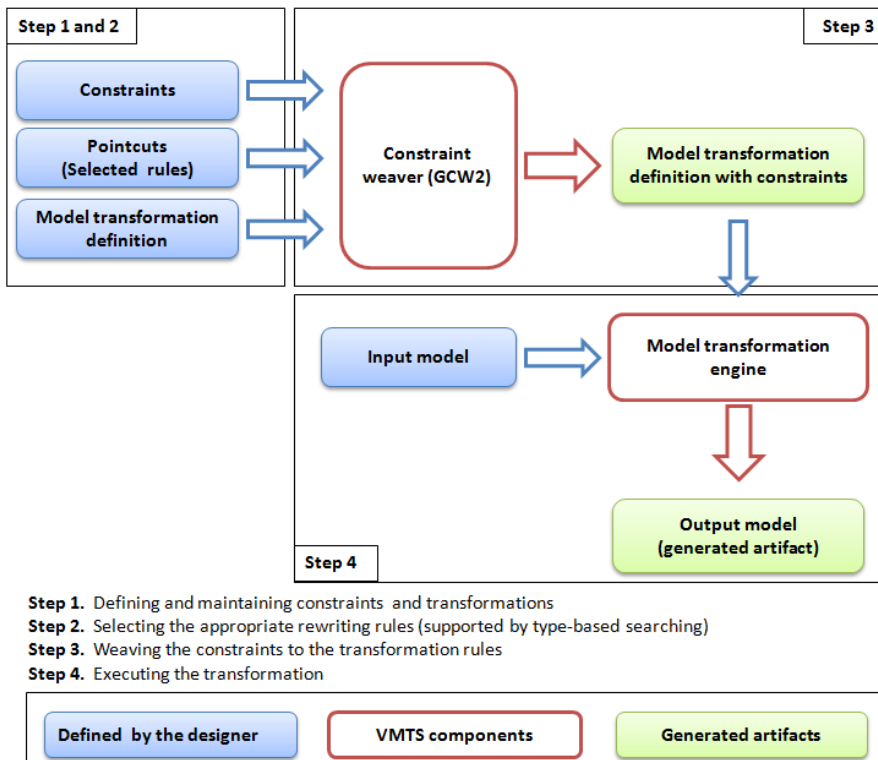


Figure 3

The process of repetitive constraint handling

The whole process of repetitive constraint handling, and its role in the model transformation, is illustrated in Figure 3. Related to this process, we have identified four steps:

- 1 *Defining and maintaining constraints and transformations.* This step is performed by the transformation designer.
- 2 *Selecting the appropriate rewriting rules.* This step is also completed by the transformation designer. The result of this step is the constraint calls that designate the rewriting rules where to propagate the constraints (from where the constraints should be called during the transformation).
- 3 *Propagating the constraints to the rules.* This step is executed by the weaver component. The weaving method receives the transformation, the constraints, and the constraint calls. The result of the weaving process is the transformation definition with the assigned constraints.
- 4 *Executing the transformation.* This step is performed by the model transformation engine. The inputs are the transformation definition that contains the constraints and the input model. The output of the model transformation is the generated artifact that can also be a model or optional text, e.g. source code.

4.2 Generalizing the Constraint Weaving

Based on the Global Constraint Weaver (GCW), presented in Section 3, a generalized constraint weaving mechanism has been developed. This Generalized GCW (GCW2) method supports the weaving of the following constraint constructs:

- Aspect-oriented constraints driven by weaving constraints (introduced in Section 3).
- Repetitive constraints driven by their constraint calls.

In this approach, aspect-oriented constraints and repetitive constraints both represent constraints which are defined separately from model transformations. They are handled separately, because their weaving is driven by different constructs. The weaving of aspect-oriented constraints is supported by the weaving constraints, and the weaving of repetitive constraints is driven by constraint calls. Therefore, these two types of constraints are not mixed.

The inputs of the GCW2 algorithm include the transformation definition, the aspect-oriented constraints with their weaving constraints, and the repetitive constraints with their constraint calls. The output of the weaver is the constrained transformation. Algorithm 1 depicts the pseudo code of the GCW2 algorithm.

Algorithm 1. Pseudo code of the GLOBALCONSTRAINTWEAVER2 algorithm

```

1: GLOBALCONSTRAINTWEAVER2 (Transformation T, ConstraintList AOCs, ConstraintList
   weavingCs, ConstraintList repetitiveCs, ConstraintCallList constraintCalls)
2: for all Constraint AOC in AOCs do
3:   for all TransformationRule R in T do
4:     nodesWithProperMetaT type = GETNODESBYMETATYPE (context type of AOC, R)
5:     nodesWithProperStructure = CHECKSTRUCTURE (nodesWithProperMetaT type, R,
   AOC)
6:     checkedNodes = CHECKWEAVINGCONSTRAINTS (nodesWithProperStructure,
   weavingCs)
7:     WEAVECONSTRAINT (AOC, checkedNodes)
8:   end for
9: end for
10: for all Constraint RC in repetitiveCs do
11:   for all ConstraintCall CC in constraintCalls do
12:     nodesToWeave = EVALUATECONSTRAINTCALL (CC, RC)
13:     WEAVECONSTRAINT (RC, nodesToWeave)
14:   end for
15: end for

```

The GCW2 algorithm is passed through a model transformation, a list of aspect-oriented constraints, a list of weaving constraints, a list of repetitive constraints and a list of constraint calls. The algorithm, using type-based weaving and applying weaving constraints, weaves the aspect-oriented constraints to the appropriate nodes of the rules. Furthermore, the algorithm weaves the repetitive constraints to the rules designated by the constraint calls.

The GCW2 algorithm uses a different block to manage the aspect-oriented constraint weaving (line 1-8) and the repetitive constraint weaving (line 9-14). In the first block, for each aspect-oriented constraint and transformation rule pair, the algorithm identifies the possible places where the constraint can be woven. It then checks the surrounding structures of these locations and evaluates the weaving constraint for the appropriate places. Finally, the constraint is woven to the correct rules. In the second block, for each repetitive constraint and constraint call pair, the algorithm decides where to weave the actual repetitive constraint, then performs the weaving.

An example of a constraint that repetitively occurs in transformation *ClassToRDBMS* is the *PrimaryKey* constraint:

```

context Table inv PrimaryKey:
self.columns->exists(c | c.datatype = 'int' and c.is_primary_key)

```

The constraint call used to propagate the constraint *PrimaryKey* is the following:

```
ConstraintCall_PrimaryKey {constraint: PrimaryKey, rules:
  CreateTable (Table), CreateParentClassHelper (Table), AddParentAssociation (Table),
  ProcessAssociations (Table1, Table2)}
```

The constraint call definition is named and contains a constraint reference (*PrimaryKey*) and an optional number of rule references. The enlisted rule names indicate from where the repetitive constraint should be called. The node names, following the rule names, are the parameters of the constraint calls. They designate where the exact rule nodes call the constraints.

The proposed method for handling repetitive constraints facilitates the definition of constraints independent of transformation rules and designates the rewriting rules, i.e., where to apply them. The approach automatically weaves the constraints to the designated points in the transformation. The benefit of this approach is that the constraints are maintained in one place and in one copy. Furthermore, our method supports a better understanding of both the transformations and constraints.

This section introduced the GCW2 algorithm, which facilitates the constraint weaving driven by both weaving constraints and manually defined constraint calls. The next section discusses the method to modularize transformation constraints if they already exist in model transformations.

5 Semi-Automatic Modularization of Transformation Constraints

In [12], a mechanism is introduced for systematically identifying crosscutting constraints. This section provides a generalized, semi-automatic method for modularizing both repetitive and crosscutting constraints.

In model transformations, some validation or other concerns can be expressed by several constraints. These concerns (expressed by more than one constraint) are the source of the crosscutting. In our approach, transformation designers can aggregate constraints into groups, in which each group represents a concern. The examples provided are the *syntactic well-formedness* and the *semantic well-formedness* groups.

```
Group_SyntacticWellFormedness {DanglingEdges1, DanglingEdges2,
  ClassAndItsParentAreTheSame}
Group_SemanticWellFormedness {MultipleInheritance, CheckInternalCondition,
  CheckSealedCondition}
```

The inputs of the modularization method are the transformation itself and the grouping definitions. The expected outputs are the modularized constraints and the constraint calls that support the weaving process. The tasks required by the modularization method are as follows:

1. Collect the constraints from the transformation.
2. Identify the crosscutting constraints.
3. Identify the repetitive constraints.
4. Extract the crosscutting constraints as aspects, and generate the constraint calls to support their weaving.
5. Extract the repetitive constraints as aspects, and generate the constraint calls to support their weaving.

In Step 2, the identification of crosscutting concerns is supported by the grouping definition. The algorithm checks whether the semantically coherent concerns are, physically, in the same rule or scattered across several rules. Concerns represented by single constraints cannot crosscut the transformations, but if they appear several times they are classified as repetitive constraints.

The crosscutting constraint identification method, presented in [12], provides the coloring and extracting algorithms. These algorithms have been updated to support both the repetitive and crosscutting constraint modularization in a general way. Based on the groups and the identified concerns, the reworked coloring algorithm assigns different colors to the constraints of the transformation. The automatic concern identification also accounts for the constraints not appearing in any of the user defined groups. In the output of the coloring algorithm, each color represents a concern. These concerns should be modularized. After the coloring, the extracting algorithm creates aspects from crosscutting and repetitive constraints, as well as generates the constraint call definitions.

The subsequent sections elaborate upon the algorithms, and their operation is illustrated with the help of our case study.

5.1 Generalized Coloring Algorithm

The algorithm receives the transformation with its constraints and the grouping definitions. The expected result is a concern list and a coloring table which provides the transformation rule and affected concern relations.

A concern is represented by a color and can be an optional condition or property expressed by one (simple) or several constraints. Examples of this include: the well-formedness concerns of our case study, as well as more simplified versions, namely, those including an attribute value or the existence of adjacent nodes of a specific type.

Algorithm 2 shows the pseudo code of the COLORING algorithm. The model transformation T and its corresponding groups are passed to the algorithm. The algorithm creates a list of rule-constraint pairs. These contain each transformation rule-constraint pair assignment, defined in transformation T . Based on their rule constraint pair assignment, the algorithm identifies the crosscutting concerns for each group (line 4). Next, the coloring table is updated with the actual group information, even if there exists no crosscutting related to the actual group. Then, the algorithm creates a concern (constraint) list (line 7) in which each member of the list represents a separated concern. This means that, if a constraint in the transformation contains more than one concern, the constraint is decomposed into several constraints. Therefore, more than one list member is created from such constraints. Groups are also added to the constraint list. Based on the list of constraint, the algorithm identifies repetitive constraints (line 9) and updates the coloring table accordingly.

Algorithm 2. Pseudo code of the COLORING algorithm

```

1: COLORING (Transformation  $T$ , GroupList  $groups$ )
2: ColoringTable  $coloringTable$  = new ColoringTable();
3: RuleConstraintPairList  $ruleConstraintPairs$  = COLLECTRULECONSTRAINTPAIRS ( $T$ )
4: for all Group  $G$  in  $groups$  do
5:   CrosscuttingList  $crosscuttings$  = IDENTIFYCROSSCUTTING ( $G$ ,  $ruleConstraintPairs$ )
6:   UPDATECOLORINGTABLE ( $G$ ,  $ruleConstraintPairs$ ,  $crosscuttings$ )
7: end for
8: ConcernList  $concerns$  = COLLECTSEPARATEDCONCERNCONSTRAINTS ( $T$ )
9: for all Constraint  $C$  in  $concerns$  do
10:  ConstraintList  $repetitives$  = IDENTIFYREPETITIVECONSTRAINTS ( $C$ ,  $concerns$ )
11:  UPDATECOLORINGTABLE ( $C$ ,  $constraints$ ,  $repetitives$ )
12: end for

```

5.2 Generalized Constraint Extracting Algorithm

The algorithm receives the model transformation and the results of the coloring algorithm. The results of the algorithm are the modularized constraints and the constraint calls supporting the weaving.

The algorithm creates the modularized constraints based on the provided concern list. The group concerns are handled in a different way from simple constraints or constraint part concerns: each member of the group concern is modularized into a different constraint. The second part of the extracting algorithm creates the constraint calls both for crosscutting and repetitive constraints. These constraint calls contain the exact list of the rules from which they should be called. In

general, for modularized crosscutting constraints (aspects) we prefer to use weaving constraints instead of the constraint calls. This is because with weaving constraints more complex conditions can be defined, and this type of weaving definition is used when defining these artifacts manually. In the current case, the artifacts are created by the extracting algorithm. Our aim is to provide a simple method that modularizes the concerns and creates such weaving artifacts that can reproduce the original transformation, exactly. Therefore, creating constraint calls for crosscutting constraints is the correct decision.

Algorithm 3. Pseudo code of the EXTRACTING algorithm

```

1: EXTRACTING (Transformation T, ConcernList concerns, ColoringTable coloringTable)
2: ConstraintList modularizedConstraints = new ConstraintList ()
3: for all Concern groupConcern in concerns.GroupConcerns do
4:   for all Constraint C in groupConcern do
5:     modularizedConstraints.Add (C)
6:   end for
7: end for
8: for all Concern nonGroupConcern in concerns.NonGroupConcerns do
9:   modularizedConstraints.Add (nonGroupConcern.Constraint)
10: end for
11: ConstraintCallList constraintCalls = new ConstraintCallList()
12: for all ColoringItem coloringItem in coloringTable do
13:   ConstraintCall constraintCall = CreateConstraintCall(coloringItem, T)
14:   constraintCalls.Add (constraintCall)
15: end for

```

The EXTRACTING algorithm receives transformation *T*, the concerns identified by the COLORING algorithm and the *coloringTable*. The algorithm processes the concerns in two blocks. In the first block, the group concerns (*SyntacticWellFormedness* and *SemanticWellFormedness*) are processed; each constraint, although related to the group, is independent and is added to the modularized constraint list (line 2-6). In the second block, the constraints of the non-group concerns are processed: simple constraints and constraint parts (line 7-9). Next, using the coloring table (transformation rule - constraint mappings), the algorithm creates the constraint calls for each constraint.

Conclusions

We have discussed that in graph rewriting-based model transformations, the two main concerns are functionality, defined by the transformation rules, and the validation properties, expressed through constraints. Concerning model transformations, we have introduced the problem of repetitive and crosscutting

constraints. We have identified the difference between repetitive and crosscutting constraints. We have shown that, in certain cases, crosscutting cannot be eliminated, but it can be solved by applying aspect-oriented mechanisms. We have briefly summarized our previous results related to aspect-oriented constraint management in model transformations. As a novel contribution, we have provided a mechanism for handling repetitive constraints. Unifying their treatment, we have developed a generalized method with its algorithms for semi-automatic modularization of repetitive and crosscutting constraints in model transformations.

The disadvantage of the repetitive constraint management approach regard is its being based on manual decisions: the transformation designer should designate the points where a constraint call should be applied. Therefore, the designer can miss some constraint call definitions, which would result in unexpected behavior of the transformation execution, especially in the case of complex transformations. The introduced approach has an UML-compliant notation that is easy to use and simple to understand.

Acknowledgement

The author would like to thank Tihamér Levendovszky for his valuable comments and support. This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

- [1] AGG, The Attributed Graph Grammar System Website, <http://tfs.cs.tu-berlin.de/agg>
- [2] Assmann U., Ludwig, A.: *Aspect Weaving by Graph Rewriting, Generative Component-based Software Engineering*, Springer (2000)
- [3] Bottoni, P., Koch, M., Parisi-Presicce, F., Taentzer, G.: *Consistency Checking and Visualization of OCL Constraints*, 294-308 (2000)
- [4] Dijkstra, E. W.: *A Discipline of Programming*. Prentice Hall, Englewood Cliffs, NJ (1976)
- [5] Dresden OCL Toolkit Website, <http://dresden-ocl.sourceforge.net>
- [6] Ehrig, H., Ehrig, K., Prange, U., Taenzer, G.: *Fundamentals of Algebraic Graph Transformation, Monographs in Theo. Comp. Sci.*, Springer (2006)
- [7] Filman, R. E., Elrad, T., Clarke, S., Aksit, M.: *Aspect-Oriented Software Development*, Addison-Wesley (2004)
- [8] Karsai, G., Agrawal, A., Shi, F., Sprinkle, J.: *On the Use of Graph Transformation in the Formal Specification of Model Interpreters*, *Journal of Universal Comp. Science*, Special issue on Formal Spec. of CBS (2003)

-
- [9] Lara, J., Vangheluwe, H., Alfonseca, M.: Meta-Modelling and Graph Grammars for Multi-Paradigm Modelling in AToM, Software and Systems Modeling (SoSyM), Vol. 3(3), 194-209 (2004)
- [10] Lengyel, L.: Online Validation of Visual Model Transformations, PhD thesis, Budapest University of Technology and Economics, Department of Automation and Applied Informatics (2006)
- [11] Lengyel, L., Levendovszky, T., Charaf, H.: Validated Model Transformation-Driven Software Development, International Journal of Computer Applications in Technology, Vol. 31(1), 106-119 (2008)
- [12] Lengyel, L., Levendovszky, T., Angyal, L.: Identification of Crosscutting Constraints in Metamodel-Based Model Transformations, IEEE Eurocon 2009, St. Petersburg, Russia, 359-364 (2009)
- [13] OMG MDA Specification, MOMG document ormsc/01-07-01, 2001, <http://www.omg.org/>
- [14] OMG OCL Specification, Version 2.2, OMG document formal/2010-02-01, 2010, <http://www.omg.org/>
- [15] Pollet, D., Vojtisek, D., Jezequel, J. M.: OCL as a Core UML Transformation Language, WITUML: Workshop on Integration and Transformation of UML models, ECOOP 2002, Malaga, Spain (2002)
- [16] Rozenberg, G. (ed.): Handbook on Graph Grammars and Computing by Graph Transformation: Foundations, Vol. 1, World Sci., Singapore (1997)
- [17] Sutton, S. M., Rouvellou, I.: Modeling of Software Concerns in Cosmos. In Proceedings of the 1st International Conference on Aspect-Oriented Software Development, ACM Press, 127-133 (2002)
- [18] Sztipanovits, J., Karsai, G.: Generative Programming for Embedded Systems, In GPCE '02: ACM SIGPLAN/SIGSOFT Conf. on Generative Programming and Component Eng., Springer, London, UK, 32-49 (2002)
- [19] Taentzer, G., Ehrig, K., Guerra, E., de Lara, J., Lengyel, L., Levendovszky, T., Prange, U., Varro D., Varro-Gyapay, Sz.: Model Transformation by Graph Transformation: A Comparative Study, ACM/IEEE 8th Int. Conf. on Model Driven Engineering Languages and Systems, Jamaica (2005)
- [20] VMTS Website, <http://www.aut.bme.hu/vmts>

Inverse Problem of Failure Mechanics for a Drawing Die Strengthened with a Holder

Vagif M. Mirsalimov

Azerbaijan Technical University
Baku, Azerbaijan
E-mail: mir-vagif@mail.ru

Farid E. Veliyev

Institute of Mathematics and Mechanics of NAS of Azerbaijan
Baku, Azerbaijan
E-mail: iske@mail.ru

Abstract: A plane problem of failure mechanics is considered for concentrically integrated cylinders. It is assumed that the drawing die (internal cylinder) is negative allowance strengthened by means of external cylinder (holder), and near the surface of the drawing die there are N arbitrarily located rectilinear cracks of length $2l_k$ ($k=1,2,\dots,N$). Theoretical analysis on definition the negative allowance providing minimization of failure parameters (stress intensity factors) of drawing die was carried out on minimax criterion. A simplified method for minimization the failure parameters of a hard alloy drawing die was separately considered.

Keywords: hard-alloy drawing die; reinforcing cylinder; negative allowance; cracks; stress intensity factors; minimization of drawing die failure parameters

1 Introduction

Experience shows [1] the great reliability and durability of multicomponent constructions compared to homogeneous ones. At present, sandwich constructions are widely used in industry and engineering. While designing high pressure apparatus, a circuit of negative allowance connected multicomponent ring under internal pressure is often used. A similar circuit is implemented in draw-making while drawing the wires and rods of annular cross section. The drawing is a process when a wire, a rod or a pipe is given a draft through the hole of a special instrument (drawing die) that has some less section than the initial work piece.

The drawing dies are manufactured from hard alloys, industrial diamonds (to make thin rods) or tool steel (to draw rods and large section pipes). The hard alloys and diamond are embedded so that it could freely go in a draw hole and go out from the opposite side. The end is caught by a tractive mechanism [2] of a drawbench that gives the rod a draft through a drawing block and subjects it to deformation, i.e. to reduction and drawing.

The experience of the drawing industry shows that [2] the failure of hard alloy drawing dies with reinforcing rings (holder) occurs because of crack propagation arising on the boundary of the working and calibrating zones of the drawing die. In this connection, at the stage design of new constructions of drawing dies, it is necessary to perform limit analysis of the drawing die in order to determine that the would-be initial cracks arranged unfavorably will not grow to disastrous sizes and cause failure in the course of rated life. The size of the initial minimal crack should be considered as a design characteristic of the material.

At the current stage of development of engineering, the optimal design of the machine parts provided in order to increase their serviceability is of great importance. Therefore, the optimal design of composite (multicomponent) constructions increases in importance. An increase in the drawing die's serviceability may be substantially controlled by using design-technological methods, in particular by geometry negative allowance of the connection of a drawing die and a holder. The solution of a problem of mechanics on the determination of such negative allowance of a drawing die and reinforcing ring under which the stress field created by this tension could slow down the crack propagation in the drawing die, which is of particular interest.

2 Formulation of the Problem

Let us consider a stress-strain state in a hard alloy drawing die reinforced with a holder under the action of loads normal and tangential to the inner contour. It is accepted that the inner contour of the drawing die orifice is close to annular one. As is known, the real surface of the tool is never absolutely smooth and always has micro or macroscopic irregularities of a technological character. In spite of exceptionally small sizes of the unevenness that generate roughness, it has an essential effect on various operational properties of tools [3-6].

It is assumed that a hard alloy drawing die is negative allowance reinforced with the help of an annular ring (holder) made of mean carbon steel. The allowance function is not known beforehand and should be defined. Let a negative allowance reinforced elastic drawing die with an outer cylinder (ring) have N rectilinear cracks of length $2l_k$ ($k=1,2,\dots,N$). At the center of the cracks, locate the origin of local coordinate systems $x_kO_ky_k$ whose axis x_k coincides with the lines of cracks

and makes the angle α_k with the axis x (Fig. 1). It is assumed that the cracks' lips are free from external loads. Refer the two-component ring to the polar coordinate $r\theta$ system of having chosen the origin at the center of concentric circles L_0, L, L_1 with radii R_0, R, R_1 (Fig. 1), respectively.

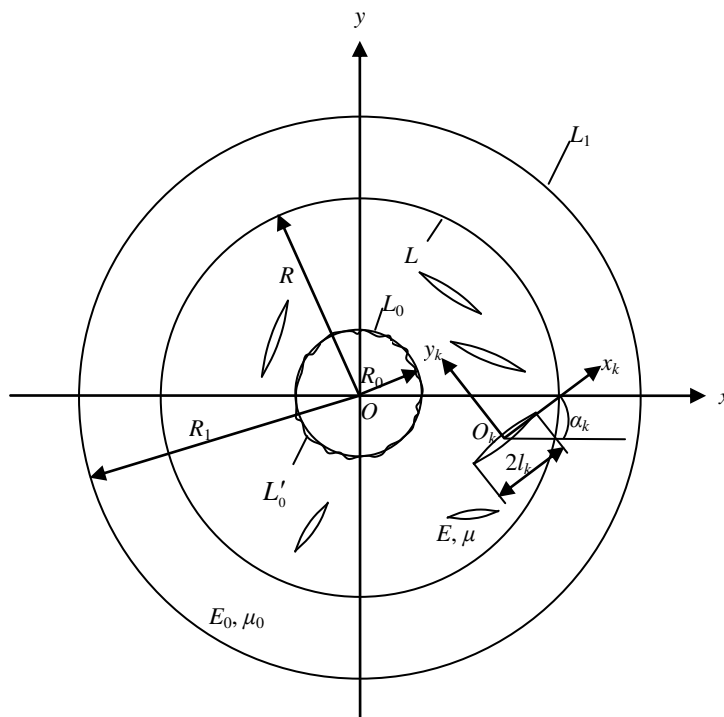


Figure 1

Calculation scheme of inverse problem of failure mechanics for a drawing die with reinforcing cylinder

Consider some realization of the rough inner surface of the drawing die. We will assume that the plane stress state condition is fulfilled. In the area occupied by the two-component ring (drawing die and holder), the stress tensor components $\sigma_r, \sigma_\theta, \sigma_{r\theta}$ should satisfy the differential equations of plane theory of elasticity [7].

Denote by E, μ and E_0, μ_0 the Young modulus and Poisson ratio of the drawing die and reinforcing ring, respectively. The boundary of the inner contour L_0 is presented in the form:

$$r = \rho(\theta) = R_0 + \delta(\theta) \quad (1)$$

Let the outer contour L_1 be free from loads. The boundary conditions of the considered problem are of the form:

$$\sigma_n = -p, \quad \tau_{nt} = -fp \quad \text{for } r = \rho(\theta) \quad (2)$$

$$\sigma_r = 0, \quad \tau_{r\theta} = 0 \quad \text{for } r = R_1 \quad (3)$$

$$\sigma_r^0 - i r_{r\theta}^0 = \sigma_r^d - i r_{r\theta}^d, \quad v_r^0 - i v_{r\theta}^0 = v_r^d - i v_{r\theta}^d - g(\theta) \quad \text{for } r = R \quad (4)$$

$$\sigma_n^d = 0, \quad \tau_{nt}^d = 0 \quad \text{on the crack's faces}$$

Here v_r, v_θ are radial and tangential constituents of the vectors of the displacement points of the contour L ; $g(\theta)$ is the desired allowance function; f is the friction factor of the “drawing die-wire rod” pair; $i = \sqrt{-1}$; p is the pressure on the inner surface of the drawing die.

The temperature of the surface layers of the drawing die increases under drawing under the action of contact friction. By drawing on the inner surface of the drawing die, on the area of contact friction with wire (wire rod), there acts a surface heat source heat caused by the outer friction. Tangential forces $\tau=fp$ promote the release of heat in the contact area of the tool and the wire rod in the drawing process. The general amount of heat in a time unit is proportional to the power of the friction forces, and the amount of the heat released at the point in the contact zone with coordinate θ will be equal to

$$Q(\theta) = Vf p,$$

where V is the mean displacement velocity of the wire rod with respect to the drawing die (drawing velocity).

The total amount of heat $Q(\theta)$ will be consumed as follows: heat flow in the drawing die $Q_d(\theta)$ and similar heat flow $Q_1(\theta)$ for increasing the wire rod heat.

In the case of steady heat exchange, the definition of the temperature field in the drawing die and annular holder may be reduced to the solution of boundary value problem of heat-conductivity

$$\text{in the drawing die } \Delta T = 0 \quad (5)$$

$$\text{in the reinforcing ring } \Delta T_0(r, \theta) = 0$$

$$\text{for } r = R: T = T_0, \quad \lambda \frac{\partial T}{\partial n} = \lambda_0 \frac{\partial T_0}{\partial n} \quad (6)$$

$$\text{for } r = \rho(\theta): \lambda \frac{\partial T}{\partial n} = -Q_d(\theta) \quad \text{on the contact area}$$

$$\text{for } r = R_1: \lambda_0 \frac{\partial T_0}{\partial r} + \alpha_2(T_0 - T_2) = 0$$

Here T is the temperature in an elastic isotropic drawing die; T_0 is the temperature in the reinforcing cylinder; λ, λ_0 are the thermal conductivity coefficients of the drawing die and holder, respectively; Δ is Laplace's operator; T_2 is the temperature of environment on the external surface of the holder; α_2 is the heat exchange from the outer cylindrical surface of the holder with external medium; $Q_d(\theta) = \alpha_d f p V$ is the intensity of the surface heat source for a drawing die; α_d is a coefficient of heat flow separation for a drawing die.

For finding the allowance function, the statement of the problem should be complemented by a condition (criterion) that allows us to determine the desired negative allowance.

According to the Irvin-Orovan theory [8] of quasibrittle fracture, the stress intensity factor is a parameter characterizing the stress state in the vicinity of the crack end. Consequently, the maximal value quantity of the stress intensity coefficient near the crack tip is responsible for the failure of the drawing die's material. Investigating the basic failure parameters and the influence of allowance of the drawing die's junction and reinforcing ring, the material properties and other factors on them, we can substantially control the failure by design-technological methods, in particular by varying the negative allowance (the function $g(\theta)$). Further, we accept minimization of quantity of maximal stress intensity factors on the vicinity of the crack tips in the drawing die. The minimization of the maximal value of the stress intensity coefficient will promote an increase in the serviceability of the drawing die of the drawing tool.

Thus, it is required to determine the junction negative allowance $g(\theta)$ such that the stress field created by it in the loading process prevents the crack from propagating.

Not losing the generality of the stated problem, it is accepted that the desired allowance function $g(\theta)$ may be represented as a Fourier series. Consequently, the coefficients $A_k^d = \alpha_k + i\beta_k$ in the expansion of the desired allowance function should be managed so that the minimization of the maximal stress intensity factors are provided. This additional condition allows to determine the desired function $g(\theta)$.

3 The Case of a Single Crack

In order to solve the stated inverse problem, it is necessary to solve a problem of failure mechanics for the "drawing die and reinforcing holder" pair. Represent the boundary of the internal contour L_0^1 of the drawing die in the form:

$$r = \rho(\theta) = R_0 + \varepsilon H(\theta)$$

$$H(\theta) = \sum_{k=0}^n (a_k^0 \cos k\theta + b_k^0 \sin k\theta)$$

where $\varepsilon = R_{\max} / R_0$ is a small parameter, R_{\max} is the greatest height of the bulge of irregularity of the surface friction, and $H(\theta)$ is a function independent of a low parameter.

Using a profilometer, the measurements have been made for a treated surface of the drawing die, and the approximate values Fourier coefficients for the function $H(\theta)$ describing each inner profile of the treated drawing die surface have been calculated for the function $H(\theta)$.

We look for temperatures, the stress tensor components and the displacements in the drawing die and holder in the form of expansions in small parameter ε :

$$T = T^{(0)} + \varepsilon T^{(1)} + \dots, \quad T_0 = T_0^{(0)} + \varepsilon T_0^{(1)} + \dots, \quad (7)$$

$$\sigma_r = \sigma_r^{(0)} + \varepsilon \sigma_r^{(1)} + \dots, \quad \sigma_\theta = \sigma_\theta^{(0)} + \varepsilon \sigma_\theta^{(1)} + \dots, \quad \tau_{r\theta} = \tau_{r\theta}^{(0)} + \varepsilon \tau_{r\theta}^{(1)} + \dots,$$

$$v_r = v_r^{(0)} + \varepsilon v_r^{(1)} + \dots, \quad v_\theta = v_\theta^{(0)} + \varepsilon v_\theta^{(1)} + \dots, \quad g(\theta) = g^{(0)}(\theta) + \varepsilon g^{(1)}(\theta) + \dots$$

where the terms with ε of higher order are neglected for simplification. Here $T^{(0)}, T_0^{(0)}$ are zero approximation temperatures; $T^{(1)}, T_0^{(1)}$ are first approximation temperatures, respectively; $\sigma_r^{(0)}, \sigma_\theta^{(0)}, \tau_{r\theta}^{(0)}$ are zero approximation stresses; $\sigma_r^{(1)}, \sigma_\theta^{(1)}$ and $\tau_{r\theta}^{(1)}$ are first approximation stresses; $v_r^{(0)}, v_\theta^{(0)}$ are radial and tangential displacements at a zero approximation; and $v_r^{(1)}, v_\theta^{(1)}$ are first approximation displacements. Each of the above approximations satisfies the system of differential equations of the plane theory of elasticity [7]. Expanding in series the expressions for temperature, stresses and displacements in the vicinity $r=R_0$ we obtain the values of constituents of temperature, stress sensor and displacement components for $r=\rho(\theta)$.

Using the perturbations method, with regard to what has been said, we arrive at the sequence of boundary conditions for the boundary value problems of fracture mechanics for a drawing die and reinforcing cylinder

$$\text{at a zero approximation for } r = R_0 \quad \lambda \frac{\partial t^{(0)}}{\partial r} = -Q$$

$$\text{for } r = R \quad t^{(0)} = t_0^{(0)}; \quad \lambda \frac{\partial t^{(0)}}{\partial r} = \lambda_0 \frac{\partial t_0^{(0)}}{\partial r} \quad (8)$$

$$\text{for } r = R_1 \quad \lambda_0 \frac{\partial t_0^{(0)}}{\partial r} + \lambda_2 t_0^{(0)} = 0$$

$$\text{for } r = R_0 \quad \sigma_r^{d(0)} = -p; \quad \tau_{r\theta}^{d(0)} = -fp$$

$$\text{for } r = R \quad \sigma_r^{0(0)} - i\tau_{r\theta}^{0(0)} = \sigma_r^{d(0)} - i\tau_{r\theta}^{d(0)} \quad (9)$$

$$v_r^{0(0)} - iv_\theta^{0(0)} = v_r^{d(0)} - iv_\theta^{d(0)} - g^{(0)}(\theta)$$

$$\text{for } r = R_1 \quad \sigma_r^{0(0)} = 0; \quad \tau_{r\theta}^{0(0)} = 0$$

$$\text{on the crack faces } \sigma_{y_1}^{d(0)} = 0; \quad \tau_{x_1y_1}^{d(0)} = 0 \text{ for } y_1 = 0, \quad |x_1| \leq l_1$$

$$\text{in the first approximation for } r = R_0 \quad \frac{\partial t^{(1)}}{\partial r} = -\frac{\partial^2 t^{(0)}}{\partial r^2} H(\theta),$$

$$\text{for } r = R \quad t^{(1)} = t_0^{(1)}; \quad \lambda \frac{\partial t^{(1)}}{\partial r} = \lambda \frac{\partial^2 t_0^{(1)}}{\partial r^2}, \quad (10)$$

$$\text{for } r = R_1 \quad \lambda_0 \frac{\partial t_0^{(0)}}{\partial r} + \alpha_2 t_0^{(1)} = 0,$$

$$\text{for } r = R_0 \quad \sigma_r^{d(1)} = N_0; \quad \tau_{r\theta}^{d(1)} = T_0$$

$$\text{for } r = R \quad \sigma_r^{0(1)} - i\tau_{r\theta}^{0(1)} = \sigma_r^{d(1)} - i\tau_{r\theta}^{d(1)}, \quad (11)$$

$$v_r^{0(1)} - iv_\theta^{0(1)} = v_r^{d(1)} - iv_\theta^{d(1)} - g^{(1)}(\theta)$$

$$\text{for } r = R_1 \quad \sigma_r^{0(1)} = 0; \quad \tau_{r\theta}^{0(1)} = 0$$

$$\text{on the crack faces } \sigma_{y_1}^{d(1)} = 0; \quad \tau_{x_1y_1}^{d(1)} = 0 \text{ for } y_1 = 0, \quad |x_1| \leq l_1$$

Here $t = T - T_c$; $t_0 = T_0 - T_c$ are excessive temperatures; T_c is temperature of environment

$$\text{for } r = R_0 \quad N_0 = -H(\theta) \frac{\partial \sigma_r^{d(0)}}{\partial r} + 2\tau_{r\theta}^{d(0)} \frac{1}{R_0} \frac{dH(\theta)}{d\theta} \quad (12)$$

$$T_0 = (\sigma_\theta^{d(0)} - \sigma_r^{d(0)}) \frac{1}{R_0} \frac{dH(\theta)}{d\theta} - H(\theta) \frac{\partial \tau_{r\theta}^{d(0)}}{\partial r}.$$

At each approximation, the solution of the boundary value problem of heat conductivity theory is sought by the method of separation of variables. We find temperatures t for a drawing die and t_0 for a reinforcing cylinder in the form

$$t^{(0)} = C_1^0 + C_2^0 \ln \rho, \quad \rho = r/R, \quad (13)$$

$$t_0^{(0)} = C_3^0 + C_4^0 \ln \rho_1, \quad \rho_1 = r/R_1,$$

$$t^{(1)} = C_1 + C_2 \ln \rho + \sum_{k=1}^{\infty} (C_1^{(k)} \rho^k + C_2^{(k)} \rho^{-k}) \cos k\theta + \sum_{k=1}^{\infty} (A_1^{(k)} \rho^k + A_2^{(k)} \rho^{-k}) \sin k\theta,$$

$$t_0^{(i)} = C_3 + C_4 \ln \rho_1 + \sum_{k=1}^{\infty} (C_3^{(k)} \rho_1^k + C_4^{(k)} \rho_1^{-k}) \cos k\theta + \sum_{k=1}^{\infty} (A_3^{(k)} \rho_1^k + A_4^{(k)} \rho_1^{-k}) \sin k\theta.$$

The constants $C_1^0, C_2^0, C_3^0, C_4^0, C_1, C_2, C_3, C_4, C_1^{(k)}, C_2^{(k)}, C_3^{(k)}, C_4^{(k)}, A_1^{(k)}, A_2^{(k)}, A_3^{(k)}, A_4^{(k)}$ are determined from the boundary conditions of the thermal conductivity theory problem (8), (10). Because of their length, the corresponding formulae are not presented here. To solve the thermoelasticity problem, we will use the thermoelastic displacement potential [9].

In the considered problem, the thermoelastic displacement potential for a drawing die F and reinforcing cylinder F_0 is determined at each approximation by the solution of the following differential equations

$$\Delta F^{(j)} = \frac{1+\mu}{1-\mu} \alpha t^{(j)}, \quad \Delta F_0^{(j)} = \frac{1+\mu_0}{1-\mu_0} \alpha_0 t_0^{(j)} \quad (j=0,1) \quad (14)$$

Here α, α_0 are the coefficients of linear temperature expansion for a drawing die and reinforcing holder, respectively, and μ, μ_0 are the Poisson ratio of a drawing die and reinforcing cylinder material. We will seek a solution of equations (14) in the form:

$$F = \sum_{n=0}^{\infty} (f_n \cos n\theta + f_n^* \sin n\theta), \quad F_0 = \sum_{n=0}^{\infty} (f_{n0} \cos n\theta + f_{n0}^* \sin n\theta) \quad (15)$$

At each approximation, for the functions $f_n(r), f_n^*(r), f_{n0}(r), f_{n0}^*(r)$, we obtain ordinary differential equations whose solutions are found by the method of variation of the constants. After determining the thermoelastic displacement potentials for a drawing die and reinforcing cylinder using well known formulae [9], we calculate the stresses $\bar{\sigma}_r^{d(j)}, \bar{\sigma}_\theta^{d(j)}, \bar{\sigma}_{r\theta}^{d(j)}$ and displacements $\bar{v}_r^{d(j)}, \bar{v}_\theta^{d(j)}$ for a drawing die, and also $\bar{\sigma}_r^{o(j)}, \bar{\sigma}_\theta^{o(j)}, \bar{\tau}_{r\theta}^{o(j)}$ and $\bar{v}_r^{o(j)}, \bar{v}_\theta^{o(j)}$ for a reinforcing cylinder that correspond to the thermoelastic displacement potentials at each approximation. The found stresses and displacements for a drawing die and reinforcing cylinder will not satisfy boundary conditions (9), (10), respectively. At each approximation, it is necessary to find the second stress strain state: $\bar{\sigma}_r^{=d(j)}, \bar{\sigma}_\theta^{=d(j)}, \bar{\tau}_{r\theta}^{=d(j)}, v_r^{=d(j)}, v_\theta^{=d(j)}$ for a drawing die, and $\bar{\sigma}_r^{=o(j)}, \bar{\sigma}_\theta^{=o(j)}, \bar{\tau}_{r\theta}^{=o(j)}, v_r^{=o(j)}, v_\theta^{=o(j)}$ for a reinforcing cylinder so that the boundary conditions (9), (11) be fulfilled.

Consequently, for determining the second stress-strain state at a zero approximation, for a drawing die and reinforcing cylinder we have the following boundary conditions:

$$\text{for } r = R_0 \quad \bar{\sigma}_r^{=d(o)} = -p - \bar{\sigma}_\theta^{=d(o)}, \quad \bar{\tau}_{r\theta}^{=d(o)} = -fp - \bar{\tau}_{r\theta}^{=d(o)} \quad (16)$$

$$\text{for } r = R \quad \bar{\sigma}_r^{=o(o)} - i\bar{\tau}_{r\theta}^{=o(o)} = \bar{\sigma}_r^{=d(o)} - i\bar{\tau}_{r\theta}^{=d(o)} + f_1(\theta)$$

$$v_r = v_r^{(o)}, \quad -iv_\theta = -iv_\theta^{(o)} = v_r^{(o)} - iv_\theta^{(o)} - g^{(0)}(\theta) + f_2(\theta)$$

$$\text{for } r = R_1 \quad \sigma_r = \sigma_r^{(o)}, \quad \tau_{r\theta} = \tau_{r\theta}^{(o)}, \quad \bar{\sigma}_r = -\bar{\sigma}_r^{(o)}, \quad \bar{\tau}_{r\theta} = -\bar{\tau}_{r\theta}^{(o)} \quad (17)$$

$$\text{for } y_1 = 0, |x_1| \leq l_1 \quad \sigma_{y_1} = \sigma_{y_1}^{(o)}, \quad \tau_{x_1 y_1} = \tau_{x_1 y_1}^{(o)}, \quad \bar{\sigma}_{y_1} = -\bar{\sigma}_{y_1}^{(o)}, \quad \bar{\tau}_{x_1 y_1} = -i\bar{\tau}_{x_1 y_1}^{(o)},$$

$$\text{where } f_1(\theta) = \bar{\sigma}_r^{(o)} - i\bar{\tau}_{r\theta}^{(o)} - (\bar{\tau}_r^{(o)} - i\bar{\tau}_{r\theta}^{(o)}), \quad f_2(\theta) = \bar{v}_r^{(o)} - i\bar{v}_\theta^{(o)} - (\bar{v}_r^{(o)} - i\bar{v}_\theta^{(o)}).$$

We can write the boundary conditions of problem (16)-(17) by means of the Kolosov Muskhelishvili formulae [7] in the form of a boundary value problem for finding two pairs of complex potentials: $\Phi^{(0)}(z)$, $\Psi^{(0)}(z)$ for the drawing die, $\Phi_o^{(0)}(z)$, $\Psi_o^{(0)}(z)$ for the reinforcing cylinder.

We will seek the complex potentials in the form [7, 10]

$$\Phi^o(z) = \Phi_1^{(0)}(z) + \Phi_2^{(0)}(z) + \Phi_3^{(0)}(z), \quad \Psi^o(z) = \Psi_1^{(0)}(z) + \Psi_2^{(0)}(z) + \Psi_3^{(0)}(z), \quad (18)$$

$$\Phi_1^{(0)}(z) = \sum_{k=-\infty}^{\infty} d_k z^k, \quad \Psi_1^{(0)}(z) = \sum_{k=-\infty}^{\infty} c_k z^k \quad (19)$$

$$\Phi_2^{(0)}(z) = \frac{1}{2\pi} \int_{-l_1}^{l_1} \frac{g_1^{(0)}(t) dt}{t - z_1}, \quad \Psi_2^{(0)}(z) = \frac{1}{2\pi} e^{-2i\alpha_1} \int_{-l_1}^{l_1} \left[\frac{\overline{g_1^{(0)}(t)}}{t - z_1} - \frac{\bar{T}_1 e^{i\alpha_1}}{(t - z_1)^2} g_1^{(0)}(t) \right] dt$$

$$\Phi_3^{(0)}(z) = \frac{1}{2\pi} \int_{-l_1}^{l_1} \left[-\frac{1}{z} - \frac{\bar{T}_1}{1 - z\bar{T}_1} e^{i\alpha_1} g_1^{(0)}(t) + \overline{g_1^{(0)}(t)} e^{-i\alpha_1} \frac{1 - T_1 \bar{T}_1}{\bar{T}_1 (1 - z\bar{T}_1)^2} \right] dt \quad (20)$$

$$\Psi_3^{(0)}(z) = \frac{1}{2\pi} \int_{-l_1}^{l_1} \left\{ g_1^{(0)}(t) e^{i\alpha_1} \left[\frac{1}{zT_1} - \frac{2}{z^2} - \frac{\bar{T}_1}{z(1 - z\bar{T}_1)} + \frac{\bar{T}_1}{(1 - z\bar{T}_1)^2} \right] + \overline{g_1^{(0)}(t)} e^{-i\alpha_1} \left[\frac{1 - T_1 \bar{T}_1}{z\bar{T}_1 (1 - z\bar{T}_1)^2} - \frac{1}{1 - zT_1} - \frac{2(1 - T_1 \bar{T}_1)}{(1 - z\bar{T}_1)^3} \right] \right\} dt$$

$$\Phi_0^{(0)}(z) = \sum_{k=-\infty}^{\infty} a_k z^k, \quad \Psi_0^{(0)}(z) = \sum_{k=-\infty}^{\infty} b_k z^k \quad (21)$$

Here $T_1 = te^{i\alpha_1} + z_1^o$, $z_1 = e^{-i\alpha_1}(z - z_1^o)$; $g_k(x_k)$ are the desired functions, characterizing the displacement discontinuity across the crack line

$$g_k^{(0)}(x) = \frac{2G}{i(1 + \kappa)} \frac{\partial}{\partial x} [u_k^+(x, 0) - u_k^-(x, 0) + i(v_k^+(x, 0) - v_k^-(x, 0))], \quad (22)$$

$\kappa = 3 - 4\mu$, in the considered case $k = 1$.

Using (18)-(21) for finding complex potentials $\Phi_1^{(0)}(z), \Psi_1^{(0)}(z)$ and $\Phi_0^{(0)}(z), \Psi_0^{(0)}(z)$ we represent the boundary conditions in the form:

$$\Phi_1^{(0)}(\tau_0) + \overline{\Phi_1^{(0)}(\tau_0)} - e^{2i\theta} \left[\bar{\tau}_0 \Phi_1^{(0)'}(\tau_1) + \Psi_1^{(0)}(\tau) \right] = -p(1-if) - (\bar{\sigma}_r^{d(0)} - i\bar{\tau}_{r\theta}^{d(0)}) \quad (23)$$

$$\Phi_1^{(0)}(\tau) + \overline{\Phi_1^{(0)}(\tau)} - e^{2i\theta} \left[\bar{\tau} \Phi_1^{(0)'}(\tau) + \Psi_1^{(0)}(\tau) \right] = \quad (24)$$

$$= \Phi_0^{(0)}(\tau) + \overline{\Phi_0^{(0)}(\tau)} - e^{2i\theta} \left[\bar{\tau} \Phi_0^{(0)'}(\tau) + \Psi_0^{(0)}(\tau) \right] - f_1(\theta) - (f_3 - if_4)$$

$$r = \rho(\theta) = R_0 + \delta(\theta) \quad \kappa \overline{\Phi_1^{(0)}(\tau)} - \Phi_1^{(0)}(\tau) + e^{2i\theta} \left[\bar{\tau} \Phi_1^{(0)'}(\tau) + \Psi_1^{(0)}(\tau) \right] = \quad (25)$$

$$= \frac{G}{G_0} \left\{ \kappa \overline{\Phi_0^{(0)}(\tau)} - \Phi_0^{(0)}(\tau) + \left[\bar{\tau} \Phi_0^{(0)'}(\tau) + \Psi_0^{(0)}(\tau) \right] e^{2i\theta} \right\} + 2Gg^{(0)'}(\tau) - f_2(\theta) - (f_5 - if_6)$$

$$f_3 - if_4 = \Phi_*^{(0)}(\tau) + \overline{\Phi_*^{(0)}(\tau)} - e^{2i\theta} \left[\bar{\tau} \Phi_*^{(0)'}(\tau) + \Psi_*^{(0)}(\tau) \right]$$

$$f_5 - if_6 = \alpha \overline{\Phi_*^{(0)}(\tau)} - \Phi_*^{(0)}(\tau) + e^{2i\theta} \left[\bar{\tau} \Phi_*^{(0)'}(\tau) + \Psi_*^{(0)}(\tau) \right]$$

$$\Phi_*^{(0)}(\tau) = \Phi_2^{(0)}(\tau) + \Phi_3^{(0)}(\tau); \quad \Psi_*^{(0)}(\tau) = \Psi_2^{(0)}(\tau) + \Psi_3^{(0)}(\tau)$$

$$\tau_1 = R_0 \exp(i\theta), \quad \tau = \text{Re}xp(i\theta).$$

We denote the left-hand side of the boundary condition (24) by the function $\sigma - i\tau$, then we have

$$\Phi_0^{(0)}(\tau) + \overline{\Phi_0^{(0)}(\tau)} - e^{2i\theta} \left[\bar{\tau} \Phi_0^{(0)'}(\tau) + \Psi_0^{(0)}(\tau) \right] - f_1(\theta) - (f_3 - if_4) = \sigma - i\tau \quad (26)$$

We assume that the function $\sigma - i\tau$, which is a self-balanced system of forces acting on the reinforcing cylinder as viewed from the drawing die, can be expanded on the circular contour L ($\tau = R \exp(i\theta)$) in a complex Fourier series

$$\sigma - i\tau = \sum_{k=-\infty}^{\infty} A_k e^{ik\theta} \quad (27)$$

For determining the complex potentials $\Phi_0^{(0)}(z)$ and $\Psi_0^{(0)}(z)$ we have condition (26) on the contour L , and the condition

$$\Phi_0^{(0)}(\tau_1) + \overline{\Phi_0^{(0)}(\tau_1)} - e^{2i\theta} \left[\bar{\tau}_1 \Phi_0^{(0)'}(\tau_1) + \Psi_0^{(0)}(\tau_1) \right] = -(\bar{\sigma}_r^{0(0)} - i\bar{\tau}_{r\theta}^{0(0)}) \quad (28)$$

on the contour L_1 ($\tau_1 = R_1 \exp(i\theta)$).

The functions $\Phi_0^{(0)}(z)$ and $\Psi_0^{(0)}(z)$ are analytical in the interior of the transverse cross section of the reinforcing cylinder $R \leq |z| \leq R_1$ and may be represented [7] by the series (21). We use the power series method [7] to find the coefficients a_k, b_k of the potentials $\Phi_0^{(0)}(z)$ and $\Psi_0^{(0)}(z)$.

For determining the still unknown quantities A_k , we consider the solution of the problem for a drawing die $R_0 \leq |z| \leq R$. After some transformations of the complex potentials, $\Phi_0^{(0)}(z)$ and $\Psi_0^{(0)}(z)$ permit representing the boundary conditions for the functions $\Phi_1^{(0)}(z)$ and $\Psi_1^{(0)}(z)$ in the form (23) and

$$\Phi_0^{(0)}(\tau) + \overline{\Phi_1^{(0)}(\tau)} - e^{2i\theta} \left[\overline{\tau} \Phi_1^{(0)'}(\tau) + \Psi_1^{(0)}(\tau) \right] = \sum_{k=-\infty}^{\infty} A_k e^{ik\theta} \quad (29)$$

$$\begin{aligned} & \kappa \overline{\Phi_1^{(0)}(\tau)} - \Phi_1^{(0)}(\tau) + \left[\overline{\tau} \Phi_1^{(0)'}(\tau) + \Psi_1^{(0)}(\tau) \right] e^{2i\theta} = \\ & = \sum_{k=-\infty}^{\infty} A_k^* e^{ik\theta} + 2Gg^{(0)'}(\tau) - f_2(\theta) - (f_5 - if_6) \end{aligned} \quad (30)$$

$$A_{-k}^* = \frac{G}{G_0} \left[\kappa_0 \bar{a}_k R^k - a_{-k} R^{-k} (1+k) + b_{-k-2} R^{-k-2} \right]$$

$$A_k^* = \frac{G}{G_0} \left[\kappa_0 \bar{a}_{-k} R^{-k} + (k-1) a_k R^k + b_{k-2} R^{k-2} \right].$$

For the functions $\bar{\sigma}_r^{d(0)} - i\bar{\tau}_{r\theta}^{d(0)}$, $g'(\tau)$, $f_2(\theta)$, $(f_3 - if_4)$, $(f_5 - if_6)$ we will assume that they can be expanded in Fourier series

$$\begin{aligned} -(\bar{\sigma}_r^{d(0)} - i\bar{\tau}_{r\theta}^{d(0)}) &= \sum_{k=-\infty}^{\infty} A_k' e^{ik\theta}, \quad g^{(0)'}(\tau) = \sum_{k=-\infty}^{\infty} A_k'' e^{ik\theta}, \\ -(f_5 - if_6) &= \sum_{k=-\infty}^{\infty} D_k e^{ik\theta}, \quad (f_3 - if_4) = \sum_{k=-\infty}^{\infty} B_k e^{ik\theta}, \quad f_2(\theta) = \sum_{k=-\infty}^{\infty} A_k''' e^{ik\theta}. \end{aligned}$$

Here, the coefficients D_k and B_k depend on the desired function $g_1^{(0)}(t)$ and are determined by residue theory.

The boundary conditions (23), (29) are used to determine the coefficients d_k, c_k and the boundary condition (30) is used to determine the quantities D_k . As a result, we find:

$$d_0 = \frac{A_0 R^2 - (A_0' - p(1-if)) R_0^2}{2(R^2 - R_0^2)}, \quad d_1 = \frac{\bar{A}_1 R_0}{1+\kappa}, \quad c_{-1} = -\kappa \frac{A_1' R_0}{1+\kappa} \quad (31)$$

$$d_1 = \frac{\bar{M}_{-1}}{R^4 - R_0^4} - \frac{2A_1 R_0}{(1 + \kappa)(R^2 + R_0^2)},$$

$$d_k = \frac{(1+k)(R^2 - R_0^2)M_k - \bar{M}_{-k}(R^{-2k+2} - R_0^{-2k+2})}{(1-k^2)(R^2 - R_0^2)^2 - (R^{2k+2} - R_0^{2k+2})(R^{-2k+2} - R_0^{-2k+2})} \quad (k=\pm 2, \pm 3, \dots),$$

$$c_{-2}R_0^{-2} = 2d_0 - (A_0' - p(1-if)),$$

$$c_{k-2}R^{k-2} = (1-k)d_k R^k + \bar{d}_{-k}R^{-k} - A_k,$$

$$M_k = A_k R^{-k+2} - A_k' R_0^{-k+2},$$

$$(1 + \kappa)\bar{d}_0 = A_0 + A_0^* + 2GA_0' - A_0'' + D_0,$$

$$(1 + \kappa)\bar{d}_k R_k = A_{-k} + A_{-k}^* + 2GA_{-k}' - A_{-k}'' + D_{-k},$$

$$(1 + \kappa)\bar{d}_{-k} R^{-k} = A_k + A_k^* + 2GA_k' - A_k'' + D_k.$$

The right-hand sides of the formulas for determining the coefficients a_k, b_k, d_k, c_k , and A_k contain the coefficients of expansions of the allowance function $g^{(0)}(\theta)$ and also the integrals of the desired function $g_1^{(0)}(t)$.

Satisfying by the functions (18) the boundary condition (17) on the crack faces, we obtain a complex singular integral equation with respect to the unknown function $g_1^{(0)}(x)$

$$\int_{-l_1}^{l_1} [R_{11}(t, x_1)g_1^{(0)}(t) + S_{11}(t, x_1)\overline{g_1^{(0)}(t)}] dt = \pi f(x_1), \quad |x_1| \leq l_1 \tag{32}$$

$$f_0(x_1) = -\left[\Phi_1^{(0)}(x_1) + \overline{\Phi_1^{(0)}(x_1)} + x_1 \overline{\Phi_1^{(0)}(x_1)} + \overline{\Psi_1^{(0)}(x_1)} \right] - \left(\overline{\sigma_{y_1}^{d(0)}} - i \overline{\tau_{x_1 y_1}^{d(0)}} \right).$$

Here, the variables x_1, t, l_1, z_1^0 are dimensionless quantities referred to $R_0, R_{nk}, S_{nk} (n = k = 1)$ are determined [10] by the formulae

$$R_{nk} = \frac{e^{i\alpha_k}}{2} \left\{ \frac{1}{T_k - X_n} + \frac{e^{-2i\alpha_n}}{\bar{T}_k - \bar{X}_n} + \frac{1}{X_n(X_n \bar{T}_k - 1)} + \frac{1 - T_k \bar{T}_k}{T_k(1 - \bar{X}_n T_k)^2} + \frac{(T_k \bar{T}_k - 1)(2X_n T_k \bar{X}_n^2 - 3\bar{X}_n T_k + 1) + T_k \bar{X}_n(1 - \bar{X}_n T_k)^2}{T_k \bar{X}_n^2 (T_k \bar{X}_n - 1)^3} e^{-2i\alpha_n} \right\}$$

$$S_{nk}(t, x) = \frac{e^{-i\alpha_k}}{2} \left\{ \frac{1}{\bar{T}_k - \bar{X}_n} - \frac{T_k - X_n}{(\bar{T}_k - \bar{X}_n)^2} e^{-2i\alpha_n} + \frac{1}{\bar{X}_n(T_k \bar{X}_n - 1)} + \right.$$

$$+ \frac{1 - T_k \bar{T}_k}{\bar{T}_k (1 - X_n \bar{T}_k)^2} + e^{-2i\alpha_n} \left\langle \frac{1}{\bar{T}_k \bar{X}_n^2} + \frac{X_n \bar{X}_n (1 - 2\bar{X}_n T_k) + 3\bar{X}_n T_k - 2}{\bar{X}_n^3 (1 - \bar{X}_n T_k)^2} \right\rangle$$

$$T_k = t e^{i\alpha_k} + z_k^0; \quad z_k = e^{-i\alpha_k} (z - z_k^0); \quad X_n = x e^{i\alpha_n} + z_n^0$$

For the inner crack, to the singular integral equation we should add additional equality expressing the displacement uniqueness condition in tracing the crack contour

$$\int_{-l_1}^{l_1} g_1^{(0)}(t) dt = 0 \quad (33)$$

Under the additional condition (33), the singular integral equation (32) by means of algebraization procedure (see the Appendix in [11]) is reduced to the system of M algebraic equations for determining M unknowns $g_1^{(0)}(t_m)$ ($m = 1, 2, \dots, M$):

$$\frac{1}{M} \sum_{m=1}^M l_1 \left[g_1^{(0)}(t_m) R_{11}(l_1 t_m, l_1 x_r) + \overline{g_1^{(0)}(t_m)} S_{11}(l_1 t_m, l_1 x_r) \right] = f_0(x_r) \quad (r = 1, 2, \dots, M-1) \quad (34)$$

$$\sum_{m=1}^M g_1^{(0)}(t_m) = 0,$$

$$t_m = \cos \frac{2m-1}{2M} \pi \quad (m = 1, 2, \dots, M), \quad x_r = \cos \frac{\pi r}{M} \quad (r = 1, 2, \dots, M-1).$$

If in (34) we pass to the complex conjugate values, we will get M additional algebraic equations. The obtained systems of equations with respect to $a_k, b_k, d_k, c_k, A_k, g_1^{(0)}(t_m)$ ($m = 1, 2, \dots, M$) permit for a given allowance $g(\theta)$ to find the stress-strain state of a drawing die and reinforcing cylinder in the presence of a crack in the drawing die at a zero approximation. In the stated optimal design problem, the coefficients $A_k^i = \alpha_k + i\beta_k$ ($k = 0, \pm 1, \pm 2, \dots$) are to be determined. Consequently, the obtained united algebraic system is not still closed. For the stress intensity factors near the crack tips at a zero approximation we have:

$$K_I^{(0)} - iK_{II}^{(0)} = \sqrt{\pi l_1} \sum_{m=1}^M (-1)^m g_1^{(0)}(t_m) \cot \frac{2m-1}{4M} \pi \quad \text{near the right vertex} \quad (35)$$

$$K_I^{(0)} - iK_{II}^{(0)} = \sqrt{\pi l_1} \sum_{m=1}^M (-1)^{m+M} g_1^{(0)}(t_m) \tan \frac{2m-1}{4M} \pi \quad \text{near the left vertex}$$

For constructing the missing equations, we require the minimization of maximal value of the stress intensity factor

$$K_{p \max}^{(0)} \rightarrow \min \quad (36)$$

with regard to restrictions connected with carrying capacity, heat stability of a drawing die, unavailability of plastic deformations, and also the fact that

$$K_{p \max}^{(0)} \leq K_{th},$$

where K_{th} is the characteristic of the threshold value of the drawing die material fracture toughness which is determined experimentally.

At a zero approximation, the optimization problem is reduced to the definition of the coefficients (the control parameters) of the expansion of the allowance function $g^{(0)}(\theta)$ in the Fourier series. The quantities $g_1^{(0)}(t_m)$ linearly depend on the coefficients $A_0^{(0)}$ of the Fourier series of the allowance function $g^{(0)}(\theta)$. Consequently, the quantity of the stress intensity coefficient (35) (the objective function) also linearly depends on the control parameters (control variables). Thus, using the minimax criterion, at a zero approximation, the considered problem is reduced to a linear programming problem.

Numerical calculations are performed by the simplex algorithm. In expansion of the allowance function $g^{(0)}(\theta)$ we were confined to seven terms. The calculations were conducted in conformity to the form of the drawing die N_013 [12]: $R_1=65\text{mm}$; $2R=29,5 \text{ mm}$; $2R_0=5,7 \text{ mm}$. The mechanical characteristics of the drawing die material (hard alloy BK6;) and holders (mean carbon steel) were accepted to be equal to $E=6,28 \cdot 10^5 \text{ MPa}$; $\nu=0,22$ and $E_0=2,06 \cdot 10^5 \text{ MPa}$; $\nu_0=0,28$. The internal pressure p changed within 391-1960 MPa.

After defining desired quantities of the zero approximation, we can go on to construct the solution of the problem at a first approximation. The functions N_0 and T_0 are determined on the base of the obtained solution for $r=R_0$. The boundary conditions (11) may be written in the form of a boundary value problem for finding complex potentials $\Phi^{(1)}(z)$, $\Psi^{(1)}(z)$ and $\Phi_0^{(1)}(z)$, $\Psi_0^{(1)}(z)$, which are sought in the form similar to (18)-(23) with obvious changes. The further course of the solution is the same as at a zero approximation. The obtained singular integral equation with respect to $g_1^{(1)}(t)$, $\overline{g_1^{(1)}(t)}$ under additional condition of type (33) by means of the algebraization method is reduced to the system of M algebraic equations for determining M unknowns $g_1^{(1)}(t_m)$ ($m=1,2,\dots,M$). The desired coefficients $a_k^{(1)}, b_k^{(1)}, d_k^{(1)}, c_k^{(1)}, A_k^{(1)}$ are contained in the right-hand side of this system.

In the stated optimal design problem, the coefficients $A_k^{(1)} = \alpha_k^{(1)} + i\beta_k^{(1)}$ ($k=0, \pm 1, \pm 2, \dots$) should be defined. Consequently, the obtained united algebraic system is still closed. For stress intensity coefficients in the vicinity of the crack tips, at a first approximation we have:

$$K_I^{(1)} - iK_{II}^{(1)} = \sqrt{\pi d_1} \sum_{m=1}^M (-1)^m g_1^{(1)}(t_m) \cot \frac{2m-1}{4M} \pi \quad \text{near the right vertex} \quad (37)$$

$$K_I^{(i)} - iK_{II}^{(i)} = \sqrt{\pi l_1} \sum_{m=1}^M (-1)^{m+M} g_1^{(i)}(t_m) \tan \frac{2m-1}{4M} \pi \quad \text{near the left vertex}$$

For constructing the missing equations, we use the minimax criterion

$$K_{\rho \max}^{(i)} \rightarrow \min$$

allowing for the above-mentioned restrictions.

The quantities $g_1^{(i)}(t_m)$ linearly depend on the coefficients $A_k^{(i)}$ of the Fourier series of the allowance function $g^{(i)}(\theta)$.

Consequently, the quantity of stress intensity factor (37) (the objective function) also linearly depends on control parameters $A_k^{(i)}$ (control variables). Thus, the optimization problem at a first approximation may also be reduced to a linear programming problem.

Numerical calculation was carried out by a simplex algorithm. The calculation results of the allowance function (the coefficients are given in mm) are given in Table 1 for the case $\alpha_1 = 30^\circ$; $l_1/(R-R_0) = 0,1$; $z_1^0 = (R_0 + 0,1(R-R_0))e^{i\pi\theta/12}$; $p=1200$ MPa.

Table 1
The values of Fourier coefficients of optimal allowance

α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7
0,1103	0,0792	0,0714	0,0642	0,0518	0,0489	0,0238	0,0157
	β_1	β_2	β_3	β_4	β_5	β_6	β_7
	0,0718	0,0452	0,0423	0,376	0,0249	0,0202	0,0105

If one of the crack ends the internal surface of the drawing die, then at each approximation, the equality (33) is replaced by an additional condition that expresses finiteness of stresses at the crack edge for $r=R_0$.

4 The Case of an Arbitrary Number of Cracks

Now we assume that in the elastic drawing die near the friction surface are N rectilinear cracks of length $2l_k$ ($k=1,2,\dots,N$) (Fig. 1). We consider the optimal design problem, or more exactly, a problem of the definition of such an allowance function for the junction of the drawing die and the reinforcing cylinder that minimization of maximal value of the stress intensity factors near the crack tips.

In this case, the problem is solved by analogy with the problem for a single crack. The complex potentials $\Phi_2^{(j)}$, $\Psi_2^{(j)}$ and $\Phi_3^{(j)}$, $\Psi_3^{(j)}$ ($j=0,1$) are generalized to the

case of arbitrary many cracks. By satisfying the boundary conditions on the crack edges, at each approximation we obtain the system of N singular integral equations of the functions $g_k^{(j)}(x_k)$ ($k=1,2,\dots,N$). At each approximation, to the system of singular integral equations for the inner cracks we should add the additional conditions

$$\int_{-l_k}^{l_k} g_k^{(j)}(t)dt = 0 \quad (j=0,1) \tag{38}$$

At each approximation, under the above mentioned conditions by means of the algebraization process, the system of singular integral equations is reduced to the system of $N \times M$ algebraic equations for determining $N \times M$ unknowns $g_k^{(j)}(t_m)$ ($j=0,1; n=1,2,\dots,N; m=1,2,\dots,M$)

$$\frac{1}{M} \sum_{m=1}^M \sum_{k=1}^N l_k \left[g_k^{(j)}(t_m) R_{nk}(l_k t_m, l_n x_r) + \overline{g_k^{(j)}(t_m)} S_{nk}(l_k t_m, l_n x_r) \right] = f_n^{(j)}(x_r) \tag{39}$$

$$\sum_{m=1}^M g_n^{(j)}(t_m) = 0.$$

Construction of the missing equations is realized at each approximation similar to the case for one crack. For the stress intensity factors in the vicinity of the crack end at each approximation we get

$$K_{In}^{(1)} - iK_{In}^{(1)} = \sqrt{\pi d_n} \sum_{m=1}^M (-1)^m g_n^{(1)}(t_m) \cot \frac{2m-1}{4M} \pi \quad \text{near the right vertex} \tag{40}$$

$$K_{In}^{(j)} - iK_{In}^{(j)} = \sqrt{\pi d_n} \sum_{m=1}^M (-1)^{m+M} g_n^{(j)}(t_m) \tan \frac{2m-1}{4M} \pi \quad \text{near the left vertex}$$

$$(j=0,1; n=1,2,\dots,N; m=1,2,\dots,M)$$

Using the minimax criterion, in the case of an arbitrary number of cracks, the considered problem is reduced to a linear programming problem allowing for the mentioned restrictions.

The numerical calculation was carried out with a symplex algorithm for the same form of the drawing die. It is assumed that the drawing die is provided with three cracks:

$$\alpha_1 = 45^\circ, l_1 / (R - R_0) = 0,05, z_1^0 = (R_0 + 0,1(R - R_0))e^{i\pi\theta/18};$$

$$\alpha_2 = 30^\circ, l_2 / (R - R_0) = 0,10, z_2^0 = (R_0 + 0,15(R - R_0))e^{i\pi\theta/12};$$

$$\alpha_3 = 15^\circ, l_3 / (R - R_0) = 0,075, z_3^0 = (R_0 + 0,05(R - R_0))e^{i\pi\theta/10}.$$

The results of calculations of the allowance function (the coefficients are given in mm) are cited in Table 2.

Table 2

The values of Fourier coefficients of optimal interference for the case of three cracks

α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7
0,1291	0,0874	0,0719	0,0648	0,0566	0,0493	0,0341	0,0207
	β_1	β_2	β_3	β_4	β_5	β_6	β_7
	0,0754	0,0687	0,0546	0,0481	0,0360	0,0204	0,0127

The optimal solution, i.e. the found coefficients α_k and β_k , promote an increase the carrying capacity of a drawing die (drawing tool).

5 A Simplified Method for Solving the Inverse Problem

In the case of several cracks, the amount of calculations increases. We consider a simplified method for solving the problem of the definition of the optimal negative allowance for the junction of the drawing die and the reinforcing cylinder. Expand the desired allowance function in the Fourier series with that amount of terms equal to the number of cracks tips. In the case of N internal cracks in the drawing die, we will be restricted by $2N$ coefficients of the allowance function expansion in the Fourier series. We require that at each approximation the stress intensity factors in the vicinity of cracks ends be equal zero. Adding $2N$ complex linear algebraic equations to the main resolving equations (they were discussed above) we get a closed algebraic system for defining all the unknowns, including the coefficients $\alpha_k^{(j)}$, $\beta_k^{(j)}$ ($j=0,1$) of the expansion of the allowance function into the Fourier series.

We assume that N_1 cracks have one end of a part on the internal surface of the drawing die. Then, the number of the cracks vertices will equal $(2N-N_1)$. In this case, when a part of the cracks is a surface crack, then in expansion of the desired allowance function in the Fourier series we will use $(2N-N_1)$ coefficients.

Require that the stress intensity factors near vertices be equal zero. Adding $(2N-N_1)$ linear algebraic equations to the main resolving equations, we get in this case a closed algebraic system of equations for determining all the unknowns. It is appropriate to apply the simplified method for solving a problem on the minimization of fracture parameters of a drawing die involving a great amount of cracks, when the use of the simplex method causes a great volume of calculations. For the numerical solution of the obtained system of equations, we use the Gauss method with a choice of the principal element. Thus, the suggested methods of the minimization of failure parameters complete each other.

Conclusions

The main equations obtained in the paper allow for the given negative allowance by numerical calculations, with the help of the definition of the stress intensity factors, to predict the growth of cracks existing in the drawing die and to establish the admissible level of deficiency and the maximal values of operation loads providing a sufficient reliability margin. At the design stage, the solution of the optimal design problem of the definition of the negative allowance in the junction of the drawing die and the reinforcing cylinder permits finding the optimal geometric parameters of the drawing die and the reinforcing cylinder, which ensures an increase in the carrying capacity. It should be noted that the obtained results are applicable in the case of a brittle fracture.

References

- [1] Reshetov D. N.: 'State and Tendency of Machine Parts Development', *Vestnik Mashinostroyeniya*, No. 10, pp. 11-15, 2000
- [2] Perlin I. L., Ermanok M. Z.: 'Drawing Theory' *Metallurgiya*, Moscow, 1971
- [3] Zolgharnein E., Mirsalimov V. M.: 'Nucleation of a Crack under Inner Compression of Cylindrical Bodies', *Acta Polytechnica Hungarica*, Vol. 9, No. 2, pp. 169-183, 2012
- [4] Thomas T. R.: 'Rough Surfaces' Longman, London, 1982
- [5] Aykut Ş.: 'Surface Roughness Prediction in Machining Castamide Material Using ANN', *Acta Polytechnica Hungarica*, Vol. 8, No. 2, pp. 21-32, 2011
- [6] Rusinko A., Rusinko K.: 'Plasticity and Creep of Metals' Springer, Verlag Berlin Heidelberg, 2011
- [7] Mushelishvili N. I.: 'Some Basic Problems of Mathematical Theory Elasticity' Amsterdam, Kluwer, 1977
- [8] Cherepanov G. P.: 'Mechanics of Brittle Fracture' Mc Graw Hill, New York, 1979
- [9] Parkus H.: 'Instationare Warmes-Pannungen' Springer, Wien, 1959
- [10] Panasyuk V. V., Savruk M. P., Datsyshyn A. P.: 'A General Method of Solution of Two-Dimensional Problems in the Theory of Cracks', *Eng. Fract. Mech.*, No. 2, pp. 481-497, 1977
- [11] Mirsalimov V. M.: 'Non-One-Dimensional Elastoplastic Problems' Nauka, Moscow, 1987
- [12] Samoilov V. S., Eichmans E. F., Falkovsky V. A. and others.: 'Metal-Working Hard – Alloy Tool' Reference book, *Mashinostroyeniya*, Moscow, 1958

Representing the Model of Impedance Controlled Robot Interaction with Feedback Delay in Polytopic LPV Form: TP Model Transformation based Approach

Péter Galambos, Péter Baranyi

Computer and Automation Research Institute, Hungarian Academy of Sciences
Kende u. 13-17. H-1111 Budapest, Hungary
{galambos,baranyi}@sztaki.hu

Abstract: The aim of this paper is to transform the model of the impedance controlled robot interaction with feedback delay to a Tensor Product (TP) type polytopic LPV model whereupon Linear Matrix Inequality (LMI) based control design can be immediately executed. The paper proves that the impedance model can be exactly represented by a finite element TP type polytopic model under certain constraints. The paper also determines various further TP models with different advantages for control design. First, it derives the exact Higher Order Singular Value Decomposition (HOSVD) based canonical form, then it performs complexity trade-off to yield a model with less number of components but rather effective for LMI design. Then the paper presents various different types of convex TP model representations based on the non-exact model in order to investigate how convex hull manipulation can be performed on the model. Finally the presented models are analyzed to validate the accuracy of the transformation and the resulting TP type polytopic LPV models. The paper concludes that these prepared models are ready for convex hull manipulation and LMI based control design.

Keywords: LPV/qLPV modeling; impedance control; compliance control; time delay; telemanipulation; haptics

1 Introduction

The literature of modern control theory shows that the representation of a given plant has considerable effect on the usability of the proper controller design method and on the achievable control performance. For instance, in the case of the qLPV state-space model given in a polytop representation and the LMI (Linear Matrix Inequality) based design techniques, we can observe that the disposition layout of the system matrix elements at the very beginning modeling phase already determines

the set of achievable control performance. Furthermore, the resulting controller really depends on the applied LMIs, which is why the majority of the related literature discusses how to manipulate LMIs in order to optimize for multi-objective control performance. At the same time, one of the key trends in modern control - \mathcal{H}_∞ based methodologies - bases the optimization of the required control constraints on integrating weighting functions into the system model before determining the polytopic representation and constructing the LMI-based synthesis.

Nevertheless, it was not emphasized as much that the LMIs are very sensitive for the polytop structure. Since the LMI in that sense can be considered as a non-linear transformation, a little modification of the convex hull may lead to considerable deviation of the resulting controller. Therefore, one may raise the question whether the convex hull manipulation plays an important role in the optimization of the control performance. Actually, this was one of the key motivations to develop the TP model transformation, which is readily capable of manipulating the convex hull of the convex polytop representation of a given model. The TP model transformation is actually a numerical representation of the HOSVD of given functions. It becomes a control design tool when it is executed on matrix functions, where the matrix function actually represents the non linear system matrix of a given LPV/qLPV model. In this general case the TP model transformation can be viewed as a TP type polytop decomposition technique having various advantages for complexity trade-off and convex hull manipulation, all relying on the power of the HOSVD. The key idea and further investigations about the utilization of TP model transformation is presented in papers [1, 2, 3]. Some further examples for control design oriented utilization of TP model transformation are in [4, 5, 6, 7, 8, 9, 10]. Authors Chumalee *et al.*, Rangajeeva *et al.*, Gai *et al.*, Sun *et al.* and Qin *et al.* introduce TP model transformation based novel approaches in avionics related control problems [11, 12, 13, 14, 15], thus leading to pioneering conceptual frameworks. In [16] Precup *et al.* introduced novel application-oriented TP models for the automatic transmission system of vehicles.

Paper [17] discusses the importance of the convex hull manipulation in the polytop representation based control design and how further improvements on the resulting control performance can be achieved. The paper also concludes that this manipulation results in a kind of relaxation of the conservativeness of the design. Based on this paper, Gróf *et al.* [18] deeply investigate an example how the convex hull manipulation influences the effectiveness of the LMIs, or even more how the improper selection of the convex hull may lead to infeasible LMIs. With this investigation, she has shown that the manipulation of the convex hull is as important as the selection of the LMIs to reach the best control performance.

In this paper, we examine two types of manipulation techniques. One type performs complexity trade-off on the number of the LTI vertex models, while the other focuses on the manipulation of the convex hull. Thus, we create the HOSVD based canonical form of the impedance model with approximation trade-off and generate different convex hulls using different convex transformation satisfying various constraints.

The paper is organized as follows: Section 2 introduces definitions related to TP model transformation. Section 3 defines the equation of motion of the investigated delayed dynamical system and specifies the required properties of the expected model form. In section 4, the properties of the resulted HOSVD-based canonical form are discussed and a trade-off is performed between the complexity and the accuracy of the TP model. Section 5 introduces the model with different types of convex hulls, and section 6 investigates the accuracy of the resulted convex TP model considering constant and varying time delay. The last section concludes the paper.

2 Basic concepts

The mathematical background of the TP model transformation and TP model transformation based LMI controller design was introduced and elaborated in [1, 2, 3]. Let us recall some of the related theorems and definitions:

Definition 1 (*qLPV model*): Consider the Linear Parameter Varying State Space model:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\mathbf{p}(t))\mathbf{x}(t) + \mathbf{B}(\mathbf{p}(t))\mathbf{u}(t) \quad (1)$$

$$\mathbf{y}(t) = \mathbf{C}(\mathbf{p}(t))\mathbf{x}(t) + \mathbf{D}(\mathbf{p}(t))\mathbf{u}(t),$$

with input $\mathbf{u}(t) \in \mathbb{R}^m$, output $\mathbf{y}(t) \in \mathbb{R}^l$ and state vector $\mathbf{x}(t) \in \mathbb{R}^k$. The system matrix

$$\mathbf{S}(\mathbf{p}(t)) = \begin{pmatrix} \mathbf{A}(\mathbf{p}(t)) & \mathbf{B}(\mathbf{p}(t)) \\ \mathbf{C}(\mathbf{p}(t)) & \mathbf{D}(\mathbf{p}(t)) \end{pmatrix} \quad (2)$$

is a parameter-varying object, where $\mathbf{p}(t) \in \Omega$ is a time varying N -dimensional parameter vector, and $\Omega = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_N, b_N] \in \mathbb{R}^N$ is a closed hypercube. $\mathbf{p}(t)$ can also include some elements of $\mathbf{x}(t)$. In this case, (2) is referred to as a quasi LPV (qLPV) model. This type of model is considered to belong to the class of non-linear models. The size of the system matrix $\mathbf{S}(\mathbf{p}(t))$ is $(k+l) \times (k+m)$.

A wide class of LMI based control design techniques are available for convex polytopic model representations; thus the finite element convex polytopic form of (1) is defined as:

Definition 2 (*Finite element polytopic model*):

$$\mathbf{S}(\mathbf{p}(t)) = \sum_{r=1}^R w_r(\mathbf{p}(t))\mathbf{S}_r. \quad (3)$$

where $\mathbf{p}(t) \in \Omega$. $\mathbf{S}(\mathbf{p}(t))$ is given for any parameter vector $\mathbf{p}(t)$ as the parameter varying combinations of LTI system matrices $\mathbf{S}_r \in \mathbb{R}^{(k+l) \times (k+m)}$ called LTI vertex

systems. The combination is defined by weighting functions $w_r(\mathbf{p}(t)) \in [0, 1]$. The term finite means that R is bounded.

Definition 3 (Finite element TP type polytopic model): $\mathbf{S}(\mathbf{p}(t))$ in (3) is given for any parameter as the parameter-varying combination of LTI system matrices $\mathbf{S}_r \in \mathbb{R}^{(k+l) \times (k+m)}$.

$$\mathbf{S}(\mathbf{p}(t)) = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \dots \sum_{i_N=1}^{I_N} \prod_{n=1}^N w_{n,i_n}(p_n(t)) \mathbf{S}_{i_1,i_2,\dots,i_N}, \quad (4)$$

applying the compact notation based on tensor algebra (Lathauwer's work [19]) one has:

$$\mathbf{S}(\mathbf{p}(t)) = S \underset{n=1}{\boxtimes}^N \mathbf{w}_n(p_n(t)) \quad (5)$$

where the $(N+2)$ dimensional coefficient tensor $\mathbf{S} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N \times (m+k) \times (m+k)}$ is constructed from the LTI vertex systems $\mathbf{S}_{i_1,i_2,\dots,i_N}$ (5) and the row vector $\mathbf{w}_n(p_n(t))$ contains univariate and continuous weighting functions $w_{n,i_n}(p_n(t))$, ($i_n = 1 \dots I_N$).

Remark 1 : TP model (5) is a special class of polytopic models (3), where the weighting functions are decomposed to a Tensor Product of univariate functions.

Definition 4 (TP model transformation): TP model transformation is a numerical method that transforms qLPV models given in the form of (1) to the form of (5), so that a large class of LMI based control design techniques can be applied to the resulting model. Detailed description of TP model transformation and application examples can be found in [1]. The TP model transformation gives a trade-off between the accuracy of the resulting model and the number of required vertexes for the LMI control design. The TP model transformation is also capable of providing a convex hull manipulation tool during execution. For further details please read papers [17, 18].

Definition 5 (HOSVD-based canonical form of qLPV models): The direct result of the TP model transformation when neither complexity trade-off nor convex hull manipulation is done is the numerical reconstruction of the HOSVD of a given function. It is like the HOSVD of tensors, but for functions where instead of singular matrices we have singular functions in an orthonormal structure, and the core tensor contains the higher order singular values. In the case of a system where matrix functions are used, the HOSVD canonical form has the same structure; the only difference is that the core tensor contains the system vertices assigned to the higher order singular values. For further details please be referred to papers [20, 21].

Definition 6 (Convex TP model): The TP model is convex if the weighting functions satisfy the following criteria:

$$\forall n, i, p_n(t) : w_{n,i}(p_n(t)) \in [0, 1]; \quad (6)$$

$$\forall n, p_n(t) : \sum_{i=1}^{I_n} w_{n,i}(p_n(t)) = 1. \quad (7)$$

Different convex hulls for TP type polytopic qLPV models can be defined. Some of the basic types are defined as follows:

Definition 7 (*SN type TP function*): The convex TP function is SN (Sum Normalized) if the sum of the weighting functions for all $x \in \Omega$ is 1.

Definition 8 (*NN type TP function*): The convex TP function is NN (Non-Negative) if the values of the weighting functions for all $x \in \Omega$ are non-negative.

Definition 9 (*NO/CNO, Normal type TP function*): The convex TP function is a NO (Normal) type model if its $w(p)$ weighting functions are Normal, that is, if it satisfies (6) and (7), and the largest value of all weighting functions is 1. Also, it is CNO (close to normal), if it satisfies (6) and (7) and the largest value of all weighting functions is 1 or close to 1.

Definition 10 (*IRNO, Inverted and Relaxed Normal type TP function*): The TP function is IRNO type if the smallest values of all weighting functions are 0, and the largest values of all weighting functions are the same.

3 Specification of the modeling problem

Since the extensive work of Hogan [22, 23, 24], wherein the concept of impedance control and its application was formulated, this control strategy has become one of the key technologies of modern robot control. Haptic rendering is a special area of robotics where the haptic device and the virtual environment together forms an impedance controlled interaction structure where time delays have an unfavorable effect on the stability of the system. A series of papers from DLR's researchers investigate the stability of haptic rendering from various aspects [25, 26, 27]. The present study focuses on the time delay that occurs in the control loop of the impedance controlled interaction.

In this paper, impedance model is understood as the dynamic relationship between the force and the resulted displacement. The impedance model is typically given by a virtual mass-spring-damper system. In the general case, a task-space impedance model can be described as

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{B}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} + \mathbf{C}(\dot{\mathbf{x}}, \mathbf{x}) = \mathbf{F}, \quad (8)$$

where \mathbf{x} denotes the Cartesian task space coordinates, \mathbf{M}, \mathbf{B} and \mathbf{K} are symmetric, positive-definite matrices describing the inertial, damping and stiffness parameters

respectively, and \mathbf{C} contains other non-linear terms of the impedance model, while \mathbf{F} denotes the external forces.

In many cases, the end effector path is prescribed and the displacement results from the impedance model is added to the predefined path. In this way the robot motion becomes compliant.

3.1 Equation of the impedance controlled actuation with feedback delay

Since this paper deals with a generic abstraction, a single degree of freedom model will be discussed, but the results can be extended to multidimensional cases. Consider the mechanical system depicted by Figure 1(a) as a simplified model of the impedance controlled robot interaction. Mass m and viscous damping b are virtual properties defining the desired dynamics of the manipulator, while k denotes the stiffness of the robot's environment.

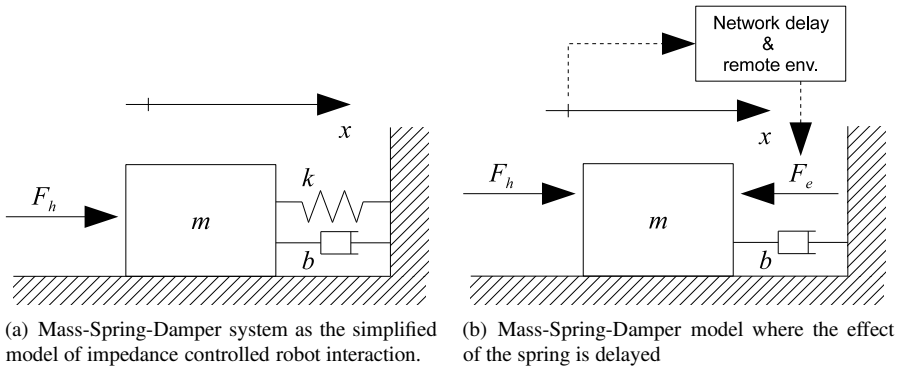


Figure 1

Introducing the time-delay in the measurement of the interaction force between the robot and its environment.

Virtual parameters have to be chosen according to the accuracy \leftrightarrow robustness trade-off [28]: The lower mass and damping result in faster and more accurate tracking with less robustness against feedback delay, and vice versa.

The equation of motion of this system is as follows:

$$\ddot{x}(t) = \frac{F_h(t)}{m} - \frac{b}{m}\dot{x}(t) - \frac{F_e(t)}{m} \quad (9)$$

Introducing the time-delay τ in the interaction as the overall delay of the force

monitoring due to the lag of the signal processing and/or network delays:

$$\ddot{x}(t) = \frac{F_h(t)}{m} - \frac{b}{m}\dot{x}(t) - \frac{F_e(t - \tau(t))}{m} \quad (10)$$

substituting the interaction force (F_e) by the elastic force (kx) in the formula as the simplest model of the environment, we get the following equation:

$$\ddot{x}(t) = \frac{F_h(t)}{m} - \frac{b}{m}\dot{x}(t) - \frac{k}{m}x(t - \tau(t)). \quad (11)$$

One can see that the resulted equation represents a mass-spring-damper system where the elastic effect is delayed by $\tau(t)$. Figure 1(b) illustrates the resulted model.

3.2 Specification of the expected qLPV representation

In this paper, we search for a representation of the investigated delayed dynamical system in TP type polytopic form (5), wherein the time delay τ becomes a parameter of the model and meets the following requirements:

- i* Fulfills the specifications of HOSVD based canonical form by finding the minimum number of LTI components that represent the original system in polytopic structure (4).
- ii* Complexity trade-off capability by means of approximation.
- iii* Eligibility for LMI based multi-objective control design.
- iv* The generated convex hull indirectly supports the feasibility of optimal control performance under the LMI based design concept.

4 The HOSVD based canonical form

In this section we utilize TP model transformation to determine the so called HOSVD based canonical form (Theorem 5) of the investigated model (11), which is a minimum and unique TP type polytopic representation. Since the investigated delayed model cannot be discretized by sampling in the first step of the TP model transformation, the discretized system tensor was determined based on the identification (see the paper [29]) of the model with multiple delay values. As the TP model transformation is a fully numerical method, the paper discusses a typical numerical example, wherein the following model parameters are considered: Mass $m = 1kg$, viscous damping $b = 100Ns/m$, Stiffness of the environment $k = 2000N/m$, Delay interval $\tau = 0..0.07s$. It is important to note that the main properties of the polytop structure are not influenced significantly by the model parameters in a wide range with practical relevancy.

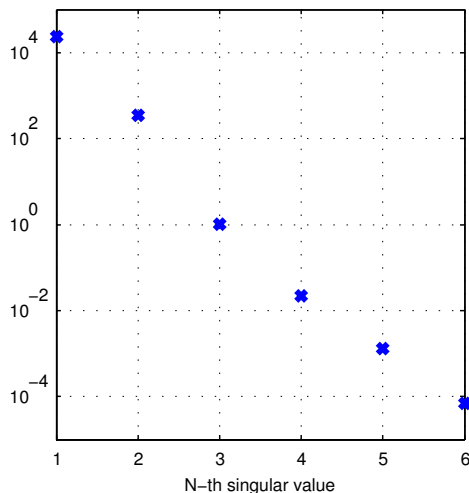


Figure 2
Singular values of the HOSVD based canonical form

4.1 Components and structure of the exact HOSVD based canonical form

After the execution of the TP model transformation on the impedance model, we get the minimum size qLPV representation composed of 6 LTI vertex models since the HOSVD leads to 6 non-zero singular values in the second step of the TP model transformation. Singular values are as follow: $\sigma_1 = 2.3414 \times 10^4$, $\sigma_2 = 3.5305 \times 10^2$, $\sigma_3 = 1.0331$, $\sigma_4 = 2.2164 \times 10^{-2}$, $\sigma_5 = 1.2964 \times 10^{-3}$, $\sigma_6 = 6.9808 \times 10^{-5}$. Note again that the different model parameters have no substantial effect on the resulted singular values, on the rank of the model or on the underlying polytopic structure, so this example properly shows the uniqueness of the representation. The consecutive singular values decrease exponentially by a factor of two orders of magnitude, which suggests a balanced contribution of vertices. Figure 2 displays the formation of the 6 singular values.

In the following, the components of the HOSVD based canonical form of the impedance model are introduced. The system is represented by the vertex models and the weighting functions. Let us partition the LTI vertices (\mathbf{S}_r^{can}) as follows:

$$\mathbf{S}_r^{can} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}. \quad (12)$$

As the transformation results in $\mathbf{C} = [1 \ 0]^T$ and $\mathbf{D} = 0$ for all $\tau \in \Omega$, only \mathbf{A} and \mathbf{B} are written in the list below:

$$\begin{aligned}
[\mathbf{AB}]_1^{can} &= \begin{bmatrix} -1.1515 \times 10^4 & -1.1896 \times 10^4 & -6.3944 \\ 1.1521 \times 10^4 & 1.1890 \times 10^4 & 6.3906 \end{bmatrix} \\
[\mathbf{AB}]_2^{can} &= \begin{bmatrix} -1.8082 \times 10^2 & 1.7522 \times 10^2 & 3.0496 \\ 1.7781 \times 10^2 & -1.7209 \times 10^2 & -3.0457 \end{bmatrix} \\
[\mathbf{AB}]_3^{can} &= \begin{bmatrix} 2.8404 \times 10^{-1} & -2.7537 \times 10^{-1} & -6.0693 \times 10^{-1} \\ 2.9928 \times 10^{-1} & -2.9109 \times 10^{-1} & 6.0668 \times 10^{-1} \end{bmatrix} \\
[\mathbf{AB}]_4^{can} &= \begin{bmatrix} 8.7316 \times 10^{-3} & -9.6607 \times 10^{-3} & 8.6616 \times 10^{-3} \\ 8.7412 \times 10^{-3} & -9.6704 \times 10^{-3} & -8.7600 \times 10^{-3} \end{bmatrix} \\
[\mathbf{AB}]_5^{can} &= \begin{bmatrix} 6.6570 \times 10^{-4} & 6.2594 \times 10^{-4} & 9.5164 \times 10^{-5} \\ 6.6535 \times 10^{-4} & 6.2629 \times 10^{-4} & 3.9904 \times 10^{-5} \end{bmatrix} \\
[\mathbf{AB}]_6^{can} &= \begin{bmatrix} -2.5581 \times 10^{-6} & -2.5893 \times 10^{-6} & 4.9197 \times 10^{-5} \\ -2.5570 \times 10^{-6} & -2.5904 \times 10^{-6} & 4.9258 \times 10^{-5} \end{bmatrix}
\end{aligned}$$

Figure 3 shows the weighting functions $w(\tau)$ over the range of Ω . The smoothness of the weighting functions shows that the applied reidentification method is stable along the investigated range of τ . This means that the applied identification algorithm does not alternate between different local solutions (local minimums). It is worth mentioning that if the identification method is switching between different solutions, additional ranks could appear in the HOSVD canonical form. By neglecting the extra singular values, HOSVD is able to (smoothly) approximate the ruggedness in least-square sense in a way similar to how SVD can be used for noise filtering in digital signal processing [30]. However, if the fluctuation is large, such approximation should be applied with circumspection.

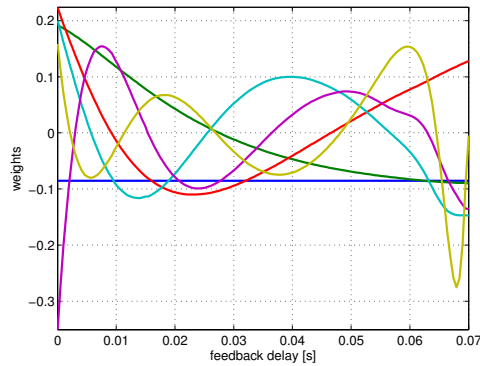


Figure 3
Weighting functions of the HOSVD based canonical form

4.2 Executing trade-off by TP model transformation

As was mentioned before, a trade-off can be determined between the complexity and the accuracy of the TP type polytopic model. The goal of this subsection is

to reveal the correlation between the accuracy and the number of utilized vertices. Considering that the LMI based design process is very sensitive to the complexity of the polytopic model, it is very important to find the minimum complexity that reaches the accuracy threshold of the given engineering problem.

Even for the computational solutions of LMI toolbox for MATLAB that introduces high quality LMI solvers [31], the computational requirements explodes exponentially by the number of vertex models. Over a certain number of vertices, the LMI solvers may not be able to provide the solution. Considering that the significance of the vertex models decreases uniformly (Figure 2), there is no theoretically appealing point from where to cut the less significant vertices to reduce the complexity of the model. However, a systematically executed trade-off could help to find the reasonable complexity. The HOSVD based canonical form readily supports a kind of principal component analysis of the investigated dynamical system model. In this analysis, the model accuracy is measured by the modeling error ϵ_r defined as:

$$\epsilon_r = \left\| \mathcal{S}^{D(\Omega, M)} - \mathcal{S}_{Approx_r}^{D(\Omega, M)} \right\|_{\mathcal{L}_2}, \quad (13)$$

where $\mathcal{S}^{D(\Omega, M)}$ can be computed using CHOSVD in the second step of the TP model transformation. $\mathcal{S}_{Approx_r}^{D(\Omega, M)}$ is computed analogously but considers only the vertex models according to the first r singular values. Modeling errors result as follows:

$$\begin{aligned} \epsilon_1 &= 3.53 \times 10^2 \leq 3.541 \times 10^2 \\ \epsilon_2 &= 1.0333 \leq 1.0566 \\ \epsilon_3 &= 2.22 \times 10^{-2} \leq 2.35 \times 10^{-2} \\ \epsilon_4 &= 1.3 \times 10^{-3} \leq 1.4 \times 10^{-3} \\ \epsilon_5 &= 6.9808 \times 10^{-5} \leq 6.9808 \times 10^{-5} \\ \epsilon_6 &= 1.1272 \times 10^{-11} \approx 0 \text{ (numerically zero)} \end{aligned}$$

As matrix $[\mathbf{AB}]_r$ contains element in the order of magnitude 10^3 , due to the definition of ϵ_r , ϵ_6 is much larger than 10^{-15} , which is typically considered as numerically zero if all the matrix elements are in the range of 10^1 . ϵ_r is upper bounded by the sum of the singular values of the neglected vertices. The upper bounds are also displayed in the above list.

Figure 4(a) displays the modeling errors. One can see that the modeling error decreases between ϵ_6 and ϵ_5 much larger than in the case of the other reduction steps.

This measure describes the model accuracy only over the discrete delay values defined by M and does not give information about the correctness between the discrete points that have been used in the first step of the TP model transformation. To follow a more extensive investigation and to ensure that the resulted TP model is not under-sampled, let us define the following measure:

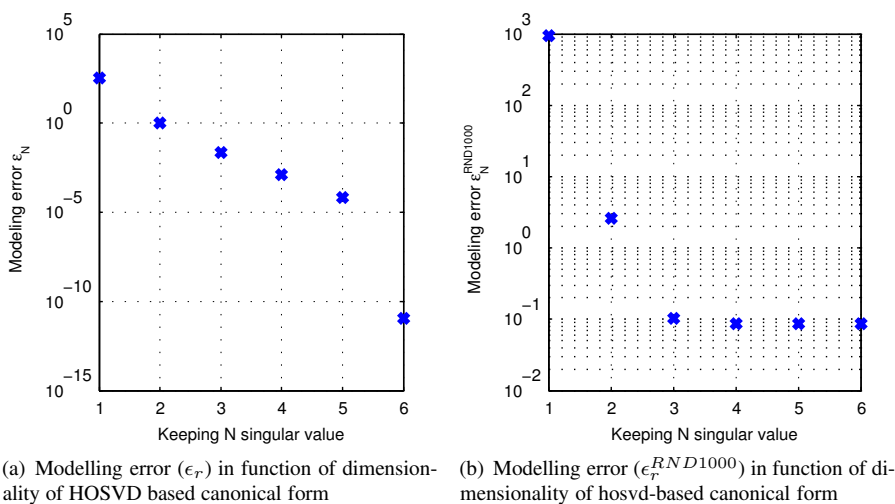


Figure 4
Accuracy-complexity trade-off

Definition 11 ($\epsilon_r^{RND1000}$)

$$\epsilon_r^{RND1000} = \left\| \mathcal{S}^{D(\Omega, M')} - \mathcal{S}_{Approx_r}^{D(\Omega, M')} \right\|_{\mathcal{L}_2}, \quad (14)$$

where M' denotes a discretization grid with 1000 randomly generated grid points over Ω . Grid M' is not equidistant and $M' \cap M = \emptyset$.

The measure $\epsilon_r^{RND1000}$ compares the reidentified and the approximated systems in 1000 randomly generated points considering the first r vertices of the HOSVD based model. $\epsilon_r^{RND1000}$ shows the model accuracy better in real situations where arbitrary varying delays occur. The resulted $\epsilon_r^{RND1000}$ s are listed below:

$$\begin{aligned} \epsilon_1^{RND1000} &= 9.6325 \times 10^2 \\ \epsilon_2^{RND1000} &= 2.7698 \\ \epsilon_3^{RND1000} &= 1.2099 \times 10^{-1} \\ \epsilon_4^{RND1000} &= 1.0519 \times 10^{-1} \\ \epsilon_5^{RND1000} &= 1.0514 \times 10^{-1} \\ \epsilon_6^{RND1000} &= 1.0514 \times 10^{-1} \end{aligned}$$

The values for $r = 4$, $r = 5$, $r = 6$ are almost the same and start to increase only at

$r = 3$. Figure 4(b) displays the $\epsilon_r^{\text{RND}1000}$ data in logarithmic scale and the big step by $r = 3$ is evident.

These results support the hypothesis that the number of non-zero singular values do not increase, even when the density of M is increased without bounds. Thus, results show that the representation is minimal and exact. It can also be concluded that the applied discretization is not under-sampled, hence complexity reduction by neglecting the less significant vertices is well established. Beyond this pure numerical comparison, the dynamic accuracy of the TP model is also investigated in section 6.

5 Different convex TP model representations

As was already emphasized, LMIs are very sensitive to the shape of the convex hull that defines the polytopic qLPV representation together with the weighting functions. Different types of convex hulls of the delayed impedance model can be generated utilizing the hull manipulation capabilities of TP model transformation. For the sake of brevity, only the non-exact, CNO type convex model with 3 vertices is detailed here. Even using the most developed optimization strategies, it is not possible to generate NO type convex hulls. From the engineering aspect, this hypothesis can be accepted since the CNO type convex hulls fulfill the requirements of control synthesis.

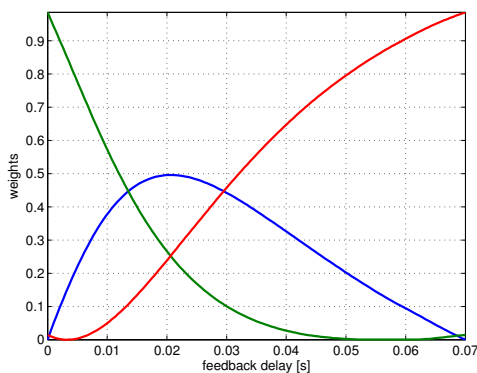


Figure 5

Weighting functions of CNO type convex hull of the reduced TP model with 3 vertices

$$\begin{aligned}
 [\mathbf{AB}]_1^{\text{cno}3} &= \begin{bmatrix} 9.7873 \times 10^2 & 1.0209 \times 10^3 & 8.6644 \times 10^{-1} \\ -9.7957 \times 10^2 & -1.0200 \times 10^3 & -8.6591 \times 10^{-1} \end{bmatrix} \\
 [\mathbf{AB}]_2^{\text{cno}3} &= \begin{bmatrix} 9.4943 \times 10^2 & 1.0519 \times 10^3 & 1.0101 \\ -9.5043 \times 10^2 & -1.0509 \times 10^3 & -1.0091 \end{bmatrix} \\
 [\mathbf{AB}]_3^{\text{cno}3} &= \begin{bmatrix} 1.0005 \times 10^3 & 9.9947 \times 10^2 & 1.8247 \times 10^{-1} \\ -1.0007 \times 10^3 & -9.9929 \times 10^2 & -1.8254 \times 10^{-1} \end{bmatrix}
 \end{aligned}$$

6 Analysis of the convex representation

The goal of this section is to illustrate the dynamical accuracy of the different TP models. Here the model accuracy is investigated by means of the difference between the step response of the polytopic model and the original delayed model. Due to the limited extent of this paper, only a set of practically interesting validation cases have been included. The comparison is broken into two parts according to the constant delay and varying delay cases.

6.1 Constant time-delay

Firstly, the dynamic accuracy of the HOSVD based canonical forms of the delayed impedance model with different complexity is examined. Figure 6 shows the step responses of the compared models at an arbitrarily chosen constant delay value ($\tau = 0.05567$). As input signal, a $1N$ force step was used at $0.1s$ in the simulation. Subfigures 6/a-f shows the step responses of TP models with a different number of neglected less significant vertices. The time plots confirm the result of the modeling error analysis done in 4.2. As the values of $\epsilon_r^{\text{RND}1000}$ suggested, the TP models show similarly good accuracy with 6, 5 and 4 vertices and the model accuracy begins to relapse with 3 vertices. A TP model with 2 and only 1 vertices cannot describe the dynamics of the original delayed system properly.

For the purpose of confidence, the same simulations have been executed on CNO type TP models with 5, 4 and 3 vertices (Figure 7). As was expected, the time plots show the same result as the HOSVD based canonical type with equivalent complexities.

We can conclude that the CNO type TP models with 5,4 and 3 vertices give very similar responses independently from the complexity. The convex hull of the investigated polytopic model cannot be formed with less than three vertices because the resulted domain of LPV models are not on the hyperline that can be defined by the convex combinations of two vertex systems. The results suggest that the CNO type convex TP model with 3 vertices provides sufficient accuracy for controller design purposes.

For the quantitative comparison, the \mathcal{L}_2 norm of the position error (the square root of sum of squares) and the maximum error is computed at four arbitrarily chosen τ values (neither of them are on the grid M) considering the $1s$ long execution of the previously discussed simulation scenario. Results are displayed in Table 1.

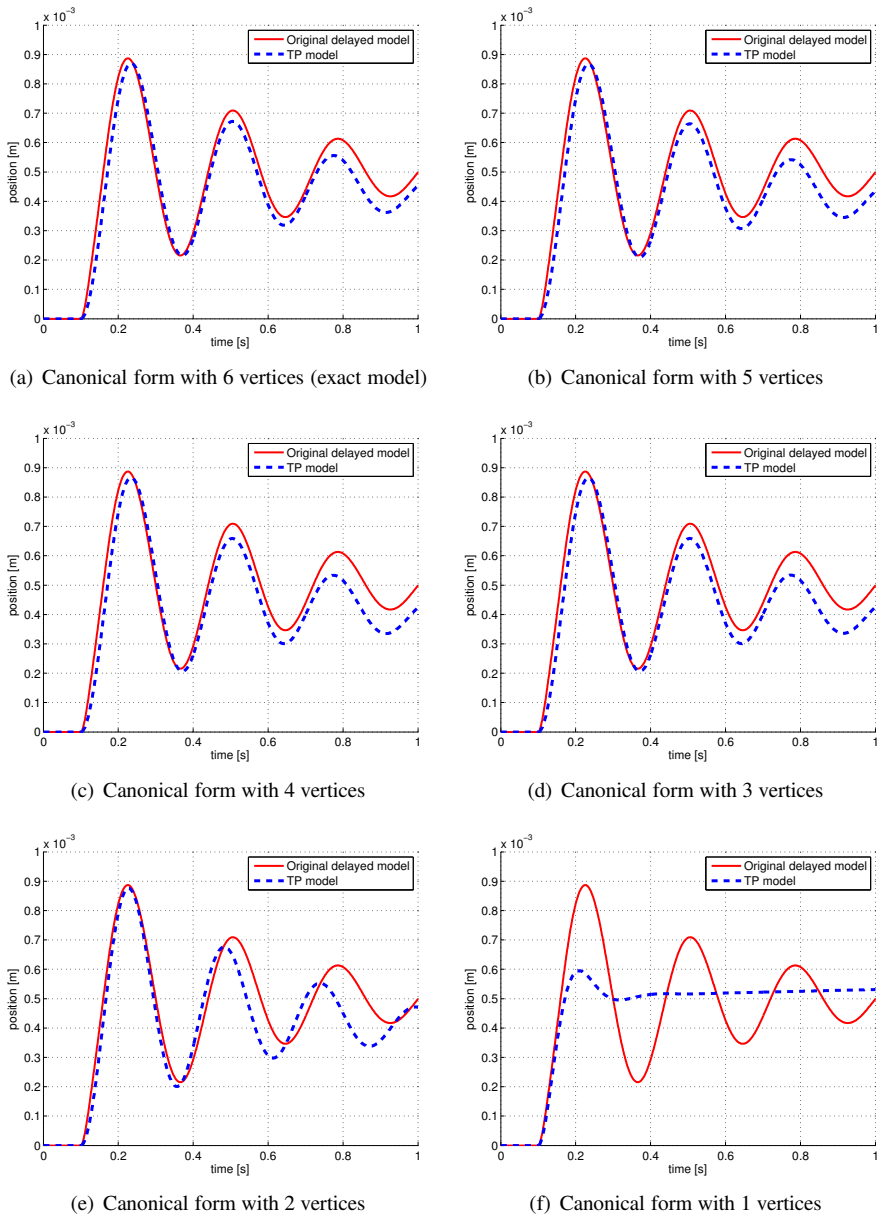


Figure 6
 Comparison of the original delayed model and the HOSVD-based canonical form of the TP model with different complexity

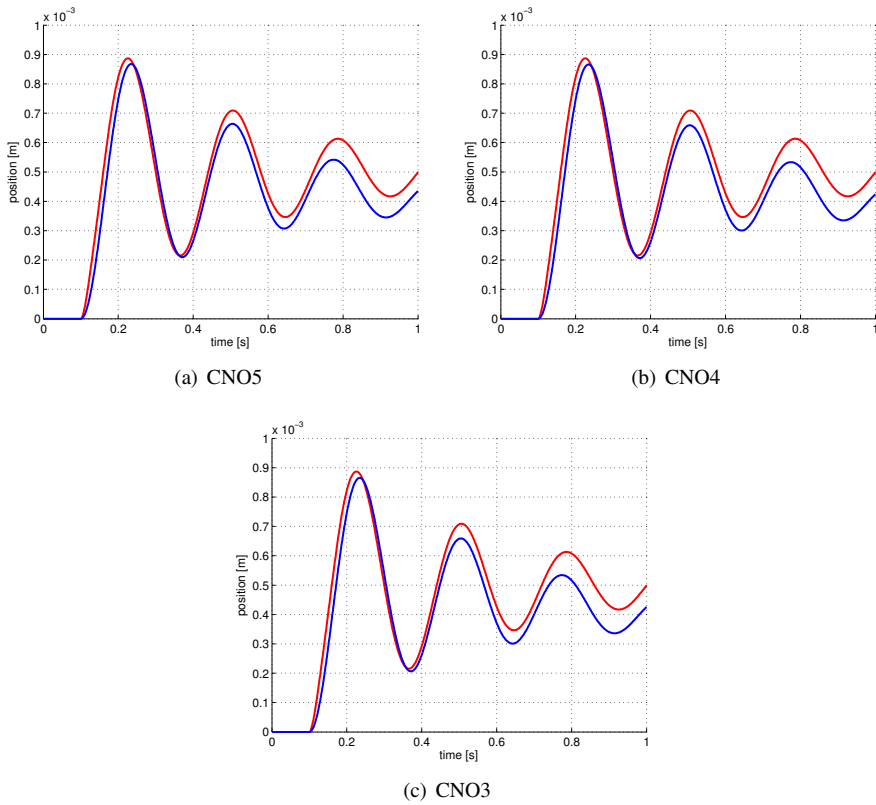


Figure 7

Comparison of the original delayed model and the CNO type TP model with different complexity

Table 1
Quantitative comparison of the original delayed model and the CNO type TP model with 3 vertices

	\mathcal{L}_2 error	Max error
$\tau = 0.01375s$	2.6279×10^{-5}	9.8521×10^{-7}
$\tau = 0.02941s$	4.0380×10^{-5}	5.9765×10^{-6}
$\tau = 0.04752s$	4.3281×10^{-5}	1.0500×10^{-5}
$\tau = 0.06393s$	1.0851×10^{-4}	1.3048×10^{-5}

6.2 Varying time-delay

The models have been compared under varying delay as well. The value of $\tau(t)$ was varied as a sine function of time $\tau(t) = 0.03 + \sin(t\pi)0.025$. The input signal was a square wave with the frequency of $2Hz$ and amplitude of $1N$. Figure 8 shows the result of the simulation.

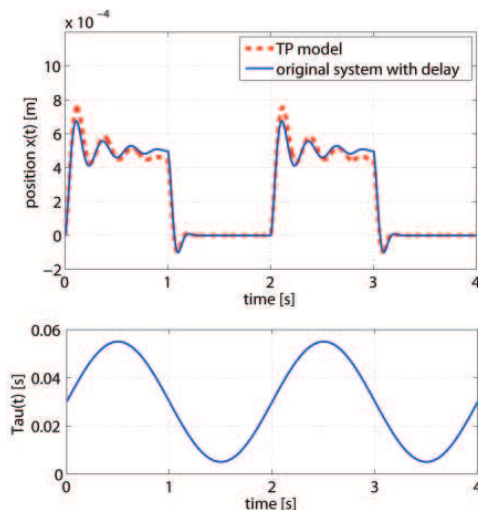


Figure 8
Comparison under varying delay

Conclusion

In this paper the HOSVD based canonical form of the model of the generic impedance controlled actuation with feedback delay was determined via a TP model transformation. A complexity trade off was also performed to determine non-exact TP models neglecting vertex systems with the less contribution. Via this investigation it has been proved that the TP model transformation is capable of manipulating the convex hull of the model wherein time delay τ appears as an external parameter. It has been shown that a convex polytop structure requires 6 vertex models for exact representation of the investigated model for any $\tau \in \Omega$. We presented the correlation between the number of vertex models and the number of singular values of the HOSVD based canonical form and the L_2 norm based error of the polytopic structure over the transformation space Ω . In order to satisfy the basic requirements of LMI based design and further convex hull manipulation based optimization of the control design SNNN, IRNO and CNO type convex TP models was generated for the reduced 3 vertex model based convex hull.

Acknowledgement

The research was supported by the Hungarian National Development Agency, (ERC-HU-09-1-2009-0004 MTASZTAK) (OMFB-01677/2009).

References

- [1] P. Baranyi, "TP model transformation as a way to LMI based controller design," *IEEE Transaction on Industrial Electronics*, vol. 51, pp. 387–400, April 2004.
- [2] P. Baranyi, "Tensor-product model-based control of two-dimensional aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 29, pp. 391–400, May-June 2005.
- [3] P. Baranyi, "Output feedback control of two-dimensional aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 29, pp. 762–767, May-June 2005.
- [4] F. Kolonic, A. Poljugan, and I. Petrovic, "Tensor product model transformation-based controller design for gantry crane control system - an application approach," *Acta Polytechnica Hungarica*, vol. 3, no. 4, pp. 95–112, 2006.
- [5] R. Precup, L. Dioanca, E. M. Petriu, M. Radac, S. Preitl, and C. Dragos, "Tensor product-based real-time control of the liquid levels in a three tank system," in *2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, (Montreal, QC, Canada), pp. 768–773, July 2010.
- [6] Z. Szabó, P. Gáspár, and J. Bokor, "A novel control-oriented multi-affine qLPV modeling framework," in *Control Automation (MED), 2010 18th Mediterranean Conference on*, pp. 1019–1024, June 2010.
- [7] S. Ilea, J. Matusko, and F. Kolonic, "Tensor product transformation based speed control of permanent magnet synchronous motor drives," in *17th International Conference on Electrical Drives and Power Electronics, EDPE 2011 (5th Joint Slovak-Croatian Conference)*, 2011.
- [8] C. Sun, Y. Huang, C. Qian, and L. Wang, "On modeling and control of a flexible air-breathing hypersonic vehicle based on LPV method," *Frontiers of Electrical and Electronic Engineering*, vol. 7, no. 1, pp. 56–68, 2012.
- [9] T. Luspay, T. Péni, and B. Kulcsar, "Constrained freeway traffic control via linear parameter varying paradigms," *Control of Linear Parameter Varying Systems With Applications*, p. 461, 2012.
- [10] B. Takarics and P. Baranyi, "Friction compensation in TP model form - aeroelastic wing as an example system," *Acta Polytechnica Hungarica*, (Submitted).
- [11] S. Chumalee and J. Whidborne, *LPV Autopilot Design of a Jindivik UAV*. American Institute of Aeronautics and Astronautics, 1801 Alexander Bell Dr., Suite 500 Reston VA 20191-4344 USA., 2009.

- [12] S. Chumalee, *Robust gain-scheduled \mathcal{H}_∞ control for unmanned aerial vehicles*. PhD thesis, Cranfield University, Cranfield, UK, 2010.
- [13] W. Qin, Z. Zheng, G. Liu, J. Ma, and W. Li, “Robust variable gain control for hypersonic vehicles based on LPV,” *Systems Engineering and Electronics*, vol. 33, no. 6, pp. 1327–1331, 2011.
- [14] S. Rangajeeva and J. Whidborne, “Linear parameter varying control of a quadrotor,” in *Industrial and Information Systems (ICIIS), 2011 6th IEEE International Conference on*, pp. 483–488, Aug. 2011.
- [15] W. Gai, H. Wang, T. Guo, and D. Li, “Modeling and LPV flight control of the canard rotor/ wing unmanned aerial vehicle,” in *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*, pp. 2187–2191, Aug. 2011.
- [16] R. Precup, C. Dragos, S. Preitl, M. Radac, and E. M. Petriu, “Novel tensor product models for automatic transmission system control,” *IEEE Systems Journal*, p. In print., 2012.
- [17] P. Baranyi, “Convex hull generation methods for polytopic representations of LPV models,” in *Applied Machine Intelligence and Informatics, 2009. SAMI 2009. 7th International Symposium on*, pp. 69–74, IEEE, 2009.
- [18] P. Gróf, P. Baranyi, and P. Korondi, “Convex hull manipulation based control performance optimisation,” *WSEAS Transactions on Systems and Control*, vol. 5, pp. 691–700, August 2010.
- [19] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, “A multilinear singular value decomposition,” *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.
- [20] P. Baranyi, L. Szeidl, P. Várlaki, and Y. Yam, “Definition of the HOSVD-based canonical form of polytopic dynamic models,” in *3rd International Conference on Mechatronics (ICM 2006)*, (Budapest, Hungary), pp. 660–665, July 3-5 2006.
- [21] L. Szeidl and P. Várlaki, “HOSVD based canonical form for polytopic models of dynamic systems,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, no. 1, pp. 52–60, 2009.
- [22] N. Hogan, “Impedance control: An approach to manipulation: Part I—Theory,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 107, pp. 1–7, Mar. 1985.
- [23] N. Hogan, “Impedance control: An approach to manipulation: Part II—Implementation,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 107, pp. 8–16, Mar. 1985.

- [24] N. Hogan, “Impedance control: An approach to manipulation: Part III—Applications,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 107, pp. 17–24, Mar. 1985.
- [25] T. Hulin, C. Preusche, and G. Hirzinger, “Stability boundary for haptic rendering: Influence of physical damping,” in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pp. 1570–1575, 2006.
- [26] T. Hulin, C. Preusche, and G. Hirzinger, “Stability boundary for haptic rendering: Influence of human operator,” in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pp. 3483–3488, 2008.
- [27] J. J. Gil, E. Sanchez, T. Hulin, C. Preusche, and G. Hirzinger, “Stability boundary for haptic rendering: Influence of damping and delay,” *Journal of Computing and Information Science in Engineering*, vol. 9, pp. 011005–8, Mar. 2009.
- [28] S. H. Kang, M. Jin, and P. H. Chang, “A solution to the Accuracy/Robustness dilemma in impedance control,” *Mechatronics, IEEE/ASME Transactions on*, vol. 14, no. 3, pp. 282–294, 2009.
- [29] P. Galambos, P. Baranyi, and P. Korondi, “Extended TP model transformation for polytopic representation of impedance model with feedback delay,” *WSEAS Transactions on Systems and Control*, vol. 5, no. 9, pp. 701–710, 2010.
- [30] E. Biglieri and K. Yao, “Some properties of singular value decomposition and their applications to digital signal processing,” *Signal Processing*, vol. 18, no. 3, pp. 277 – 289, 1989.
- [31] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Programming*. Philadelphia: SIAM, 1994.

Understanding Student Preferences for Postpaid Mobile Services using Conjoint Analysis

Marija Kuzmanovic, Marko Radosavljevic, Mirko Vujosevic

University of Belgrade, Faculty of Organizational Sciences

Jove Ilica 154, Belgrade, Serbia

marija.kuzmanovic@fon.bg.ac.rs, mari@fon.bg.ac.rs, mirkov@fon.bg.ac.rs

Abstract: In this paper, conjoint analysis is used to gain insights into how university students value various aspects of the postpaid mobile phone service. Preference-based segmentation is performed on the output from the conjoint analysis to isolate homogeneous consumer segments that possess similar preferences for mobile phone service. Based on the results the study suggests a marketing strategy for mobile phone operators.

Keywords: Conjoint analysis; students' preferences; mobile phone service; postpaid plan; preference-based segmentation

1 Introduction

In recent years, the mobile communication market has developed rapidly. At the end of 2011, there were 6 billion mobile subscriptions, which is equivalent to 87 percent of the world population [1]. It has been estimated that by the end of 2013 there will be 6.9 billion mobile phone subscriptions worldwide [1]. Global mobile service revenue in 2009 was 840 billion US\$, while the projected revenue for 2013 is 1038.6 billion US\$ [2]. In Serbia, there were 10.2 million subscriptions in 2011 (corresponding to a 142.99% penetration rate), while the mobile service revenue for 2011 reached 846.7 million euros [3].

Despite its continued global expansion, mobile user growth is slowing. This decline is acute in developed markets, such as Serbia, and reflects saturating market conditions [1]. As mobile service providers seek to counter the slowing growth, the youth have emerged as an important segment. University students have been labeled as one of the most important target markets [4] as well as the largest consumer group for mobile phone services [5]. Recently, there have been several studies concerning the interaction of young people with mobile phone technology, e.g. their attitude [6], motivation [7], psychological effects [8, 9], the

impact on their social life [10], and their use of mobile phone services [5, 11]. However, there have been only a small number of attempts to explore student preferences towards certain features of a mobile phone service offering [12, 13].

The mobile phone markets show some changes from one country to the next [14]. There are three mobile phone operators in Serbia. All of them offer both prepaid and postpaid plans. University students are among the users that widely use postpaid services.

The objective of this paper is to determine the factors affecting the preferences of university students for postpaid service plans in Serbia, and to provide insights on how mobile phone operators can attract as many subscribers as possible. The results of our research are expected to inform mobile phone operators about student perceptions regarding various aspects of mobile phone services, and to help them design business models and perform successful marketing strategy based on the students' needs.

In order to measure student preferences, this paper used conjoint analysis. Conjoint analysis is a multivariate technique that can be used to understand how an individual's preferences are developed. Specifically, the technique is used to gain insights into how consumers value various product attributes based on their evaluation of the complete product. Conjoint analysis has been widely used in marketing literature to evaluate consumer preferences for hypothetical products and services [15, 16, 17], as well as for pricing research [18]. The method has been applied to understanding the preferences in various markets including retail [19, 20], transportation [21], education [22], the labor market in the context of personnel selection decisions [23], telecommunications [24, 25] and health care services [26]. However, few studies have used conjoint analysis within the mobile industry [27, 28, 29].

This paper is organized as follows. The research design is covered in the following section. The type of data and how the data was collected, are also explained in that section. The main part of the paper is devoted to an explanation of the empirical results. Finally, the main conclusions are summarized.

2 Method

2.1 Conjoint Analysis

Conjoint analysis is an experimental approach used for measuring customer preferences regarding the attributes of a product or service. Originally developed in the field of mathematical psychology, conjoint analysis has attracted

considerable attention, especially in marketing research, as a method that portrays consumer decisions.

Conjoint analysis, sometimes called ‘trade-off analysis’, reveals how people make complex judgments. The technique assumes that complex decisions involve not only one factor or criterion, but rather several factors ‘considered jointly’. It is based on the simple premise that consumers evaluate the value of a product or service by combining the separate amounts of value provided by each attribute. Accordingly, conjoint analysis enables the investigator to better understand the interrelationship of multiple factors as they contribute to the preferences.

Conjoint experiments involve individuals being asked to express their preferences for various experimentally designed, real or hypothetical alternatives. These hypothetical alternatives are descriptions of potential real-world alternatives, in terms of their most relevant features or attributes (both quantitative and qualitative); hence, they are multi-attribute alternatives. Lists of attributes describing single alternatives are called profiles or concepts. Typically, the set of relevant attributes is generated by expert opinions, reviewing the research literature and performing pilot research with techniques such as focus groups, factor listings, or repertory grids. Two or more fixed values, or “levels”, are defined for each attribute, and these are then combined to create different profiles.

Moreover, the number of product attributes selected must be reconciled with the characteristic of the given conjoint method: the traditional approach is ideal in the case of a maximum of six attributes, but if more than six attributes must be included, then the adaptive conjoint analysis is the appropriate method [30]. Though nowadays adaptive conjoint analysis and choice-based conjoint methods are very popular, sometimes it is more convenient to use the traditional approach. Adaptive conjoint analysis must be computer-administered. The interview adapts to respondents’ previous answers, which cannot be done via the "paper and pencil" method. On the other hand, the choice-based conjoint method can be administered by PC or via paper and pencil, but results have traditionally been analysed at the aggregate, or group, level. Aggregate-level analysis is useful for detecting and modeling subtle interactions that may not always be revealed with individual-level models. While these advantages seem to favor aggregate analysis from choice data, academics and practitioners have argued that consumers have unique preferences, and that aggregate-level models which assume homogeneity cannot be as accurate as individual-level models [31].

Thus, the traditional approach proved the better choice in this study, because it calculates a set of utilities for each individual. The experimental procedure involves profiles being presented to respondents who are asked to express their preferences by rating or ranking real or hypothetical profiles. Preference functions are estimated from this data, using ordinary least square regression for rating the data, as well as non-metric techniques when the rankings are obtained.

2.2 Postpaid Attributes and Their Levels

The first stage in the design of a conjoint analysis study is the selection of the attributes. The selection of key attributes in this study has been carried out through a preliminary survey. The survey was conducted using the traditional “paper and pencil” method. The respondents were asked to evaluate the significance of each of the 10 offered characteristics of post-paid mobile phone packages. The grading ranged from 1 to 10, where a score of 10 indicated the most important criterion, while the score of 1 indicated the least important criterion. The survey was completed by 28 respondents, members of the student population. The average score of each of the criteria as well as their ranking are presented in Table 1. A high value of standard deviation, especially in the case of the Mobile Phone Operator criterion, indicates heterogeneity of preferences of the student population.

A subset of 7 attributes that stood out by their average ranking was selected for the conjoint analysis. Although the Mobile Phone Operator criterion was ranked very low (last, ninth place) according to the results of pre-research, it has been added to the selected set of attributes in order to determine the student preferences towards the existing operators in Serbia, as well as their level of satisfaction with their currently selected operator. Two criteria have been omitted from further analysis: the number of branches of the network operator, and the existence of tariff add-ons at extra cost.

Table 1
The results of the preliminary survey

Rang	Criteria	Avg. Rate	St. Dev.
1	Possibility of transferring unused traffic to the next month	8.04	2.25
2	Conversation billing interval (1s; 60s+1s; 60s+60s)	7.21	2.67
3	Free internet within package	6.86	2.46
4	Account balance check (prompt or delayed update)	6.61	2.25
5	Promotions (discount) after expiration of the signed contract	5.43	2.67
6	Level of availability and quality of technical support	4.82	2.47
7	Possibility of choosing preferred phone number	4.57	2.78
8	Number of branch offices	4.04	2.28
9	Mobile phone operator	3.82	3.03
10	The existence of tariff add-ons at extra cost	3.61	1.93

Having chosen the attributes, levels must be assign to them. These should be realistic, plausible and capable of being traded. The attributes and levels chosen for this study are shown in Table 2.

Table 2
Attributes and their levels

No.	Attribute	Attribute description	Attribute levels
1.	Operator	Mobile phone operator	MTS Telenor VIP
2.	Transfer	Possibility of transferring unused free traffic to the next month	Yes No
3.	Support	Level of availability and quality of technical support	High Low
4.	Internet	Free internet within package	Yes No
5.	Interval	Conversation billing interval	1s 60s+1s 60s+60s
6.	Number	Possibility of choosing preferred phone number	Yes No
7.	Promotions	Promotions following the expiration of the contract	Yes No
8.	Checking	Account balance check	Prompt (update) Delayed (update)

Three mobile operators are currently operating in Serbia: MTS, owned by the company “Telekom Serbia”, based in Belgrade, Serbia; Telenor, a member of the company “Telenor group” which is based in Oslo, Norway; and VIP, a part of the company “Telekom Austria”, with headquarters in Vienna, Austria. Accordingly, the attribute Operator belongs to the category of nominal attributes, and the existing three operators are the levels that are assigned to it. The next attribute, Transfer, refers to the possibility of transferring unused free minutes during one month (minutes of conversation, SMS, MMS, GPRS ...) to the free minutes intended for the next month. The practice of operators in Serbia is that if the option is available, the transferred traffic must be used within a certain period of time. This attribute is of the ordinal type, with levels where an option Exists (Yes) or Does not exist (No).

Support is an attribute that refers to the availability of technical support in terms of possibilities of establishing contact with call center operators. This attribute has been included in the analysis because practice has shown that it is frequently almost impossible to contact a call center, and the idea was to determine whether and to what extent this factor affects overall student preferences. The attribute is ordinal, with the levels of High and Low as the levels of availability and quality. Internet is an attribute that describes whether the free traffic within the post-paid packages includes a certain extent of Internet access. Considering that students are the population that has the highest percentage of Internet and modern technologies users in Serbia, the assumption is that the existence of this option is an important criterion for choosing a particular mobile phone package among the student

population. The attribute is ordinal, with the levels Exists (Yes) and Does not exist (No).

Interval is an attribute that shows the manner in which time consumption is billed during conversations. The levels are: 1s - there is no rounding-off of the duration of a conversation, the exact number of seconds of a conversation shall be deducted from the remaining free minutes, or additionally charged if the free minutes have been used up; 60s +1s – as soon as a connection is established, the conversation is rounded off to 60 seconds, and after the first minute the billing is performed per second of conversation; 60s +60s - each initiated minute is billed as a minute spent. With certain postpaid packages, operators offer users the possibility to select a new phone number according to their wishes. Therefore, the analysis also includes the attribute Number, and it has been assigned the levels Exists (Yes) and Does not exist (No).

Promotions is an attribute that refers to the existence of promotions following the expiration of a time related contract between the user and operator (e.g., a cheaper phone if the user decides to renew a contract, a discount on a subscription for several months, etc.). The attribute is ordinal, and the levels are Exists (Yes) and Does not Exist (No). Checking is an attribute that refers to the promptness of updates of the remaining free traffic, or new billing after the use of a service by the user. In Serbia, it often happens that status updates are late by more than a week. The attribute is ordinal as the previous one, with the levels Prompt updates and Delayed updates.

2.3 Conjoint Experimental Design

Once attributes and attribute levels are selected, they must be combined to form different hypothetical services for survey respondents to assign preference ratings. In this study, a full profile approach was used to design the product profiles. The attributes and levels in Table 2 gave rise to 576 possible profiles ($3^2 \times 2^6$). Since it is difficult, from a customer's perspective, to evaluate a large number of service profiles, it is necessary to select fewer of them. Therefore in this study the fractional factorial experimental design was used. A component of the statistical package SPSS 16.0 (Orthoplan) was used to reduce the possible number of profiles to a manageable level, while still allowing the preferences to be inferred for all of the combinations of levels and attributes. The use of Orthoplan results in an orthogonal main effects design, thus ensuring the absence of multicollinearity between the attributes. Through the use of this design, the 576 possible profiles were reduced to 16. Two control profiles (holdout tasks) were added to the given design. These 2 profiles were not used by the conjoint procedure for estimating the utilities. Instead, the conjoint procedure calculates correlations between the observed and predicted rank orders for these profiles, as a check of the validity of the utilities. The 18 hypothetical service profiles considered are shown in Table 3.

Table 3
Hypothetical mobile service profiles

Profile	Operator	Transfer	Support	Internet	Interval	Number	Promotions	Checking
1	MTS	Yes	High	Yes	60s+60s	No	Yes	Prompt
2	VIP	Yes	Low	No	1s	Yes	Yes	Delayed
3	Telenor	Yes	Low	No	60s+60s	No	Yes	Delayed
4	Telenor	No	High	Yes	1s	No	No	Delayed
5	Telenor	Yes	Low	Yes	60s+1s	Yes	No	Prompt
6	MTS	Yes	High	Yes	1s	Yes	Yes	Prompt
7	MTS	No	Low	No	60s+60s	Yes	No	Prompt
8	MTS	No	Low	Yes	60s+1s	No	Yes	Delayed
9	VIP	Yes	Low	Yes	1s	No	No	Prompt
10 ^h	MTS	No	High	Yes	1s	Yes	No	Delayed
11	MTS	Yes	High	No	60s+1s	Yes	No	Delayed
12	MTS	Yes	High	No	1s	No	No	Delayed
13 ^h	MTS	No	High	Yes	60s+1s	No	Yes	Prompt
14	MTS	No	Low	Yes	1s	Yes	Yes	Delayed
15	VIP	No	High	No	60s+1s	No	Yes	Prompt
16	VIP	No	High	Yes	60s+60s	Yes	No	Delayed
17	MTS	No	Low	No	1s	No	No	Prompt
18	Telenor	No	High	No	1s	Yes	Yes	Prompt

^h holdout profiles

2.4 Survey

The survey was conducted in Belgrade, Serbia, in February 2011. Data collection was conducted online through a web-based questionnaire. This method of data collection was chosen for several reasons:

- Online surveys are less expensive than the traditional “paper and pencil”. In this study specifically, free web hosting and a free domain were used.
- An online survey can be filled out simultaneously by a greater number of people. The number is practically unlimited.
- The collected data is very easily exported into SPSS or Excel format
- The questionnaire is available to a greater number of people.

The questionnaire included: (1) Instructions for completion, (2) Demographic questions, and (3) Conjoint questions from an effective experiment plan with two control (holdout) tasks.

The instructions for completion explain to the respondents how the questioning is performed. The method of evaluation of whole profiles has been chosen as the method of evaluation by the respondents. The respondents expressed their preferences for a particular service, or the real or hypothetical combination of attributes of the mobile telephony, on a scale of 1 to 9, where 1 stands for absolutely undesirable, and 9 stands for absolutely desirable. The questionnaire

also included some basic demographic questions, but also questions related to the current habits of the students in relation to the services of the mobile operators. The aim was to determine whether there is a difference in preferences among students of different demographic characteristics.

Given the subject matter and objective of the research, the respondents were exclusively members of the student population and were selected randomly. Students were invited via email to complete a questionnaire which, as noted, was available online. A list of students' email addresses was drawn both from the some student forums and the official faculty mailing lists. We sent an invitation to a total of 700 addresses, and 146 students answered the survey (approximately a 21% response rate). After the elimination of incomplete surveys and ineligible participants, 134 eligible surveys were collected. The demographic information is summarized in Table 4.

Table 4
Demographics of respondents

Variable	Description	Count (<i>n</i> =134)	Percent (%)
Gender	Male	74	55.2%
	Female	60	44.8%
Monthly income	/	88	65.7%
	Occasional income	30	22.4%
	Regular income	16	11.9%
Residence	With parents	67	50.0%
	In rented apartment	47	35.1%
	On the Campus	20	14.9%
Current mobile phone operator	MTS	73	54.5%
	Telenor	44	32.8%
	VIP	17	12.7%
Current tariff plan	Prepaid	60	44.8%
	Postpaid	74	55.2%
Average monthly traffic consumption	0-500 RSD	27	20.1%
	501-1000 RSD	57	42.5%
	1001-1500 RSD	28	20.9%
	1501-2000 RSD	14	10.4%
	More than 2000 RSD	8	6.0%

2.5 Conjoint Model Specification

Having collected the information on individual preferences, the responses needed to be analysed. To determine the relative importance of different attributes to respondents, the trade-offs that individuals make between these attributes, as well as the overall benefit taking into account these trade-offs, a relationship must be specified between the attributes' utility and the rated responses. The simplest and most commonly used model is the linear additive model. This model assumes that the overall utility derived from any combination of attributes of a given good or service is obtained as the sum of the separate part-worths of the attributes. Thus, respondent *i*'s predicted conjoint utility for profile *j* can be specified as follows:

$$U_{ij} = \sum_{k=1}^K \sum_{l=1}^{L_k} \beta_{ikl} x_{jkl} + \varepsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, J \quad (1)$$

where K is the number of attributes, L_k is the number of levels of attribute k , and β_{ikl} is respondent i 's utility with respect to level l of the attribute k . x_{jkl} is such a $\{0,1\}$ variable that equals 1 if profile j has attribute k at level l , otherwise it equals 0. ε_{ij} is a stochastic error term.

The parameters β_{ikl} , also known as part-worth utilities, can be used to establish a number of things. Firstly, the value of these parameters indicates the amount of any effect that an attribute has on overall utility – the larger the coefficient, the greater the impact. Secondly, part-worths can be used for preference-based segmentation. Namely, given that part-worth utilities are calculated at the individual level, if preference heterogeneity is present, the researcher can find it. Respondents who place a similar value on the various attribute levels will be grouped together into a segment. Thirdly, part-worths can be used to calculate the relative importance of each attribute, which is known as an importance score or value (FI_{ik}). These values are calculated by taking the utility range for each attribute separately, and then dividing it by the sum of the utility ranges for all of the factors. Calculations are done separately for each respondent:

$$FI_{ik} = \frac{\max\{\beta_{ik1}, \beta_{ik2}, \dots, \beta_{ikL_k}\} - \min\{\beta_{ik1}, \beta_{ik2}, \dots, \beta_{ikL_k}\}}{\sum_{k=1}^K (\max\{\beta_{ik1}, \beta_{ik2}, \dots, \beta_{ikL_k}\} - \min\{\beta_{ik1}, \beta_{ik2}, \dots, \beta_{ikL_k}\})}, \quad i = 1, \dots, I, \quad k = 1, \dots, K \quad (2)$$

and the results are then averaged to include all of the respondents:

$$FI_k = \frac{1}{I} \sum_{i=1}^I FI_{ik}, \quad k = 1, \dots, K \quad (3)$$

To estimate the parameters of the model, this paper used the statistical package SPSS 16.0 (Conjoint procedure). The parameters were estimated for each respondent in the sample individually, as well as for the entire sample.

3 Analysis and Results

3.1 Results at the Aggregate Level (Averaged Preferences)

Results from the analysis are shown in Table 5 and Figure 1. Table 5 presents the (averaged) part-worth of each level of the attributes, while Figure 1 is the graph

description of the attributes importance. The goodness of fit statistics for the estimated models is reported also in Table 5.

Table 5
Averaged part-worth utilities

Attribute levels	Part-worth utilities (β)	
Mobile phone operator		
MTS	0.188	
Telenor	0.062	
VIP	-0.250	
Possibility of transferring unused free traffic to the next month		
Yes	0.531	
No	-0.531	
Level of availability and quality of technical support		
High	0.401	
Low	-0.401	
Free internet within package		
Yes	0.508	
No	-0.508	
Conversation billing interval		
1s	0.159	
60s+1s	0.047	
60s+60s	-0.206	
Possibility of choosing preferred phone number		
Yes	0.302	
No	-0.302	
Promotions following the expiration of the contract		
Yes	0.321	
No	-0.321	
Account balance check		
Prompt	0.343	
Delayed	-0.343	
Constant	4.622	
Correlation between the observed and estimated preferences		
	Value	Significance
Person's R	0.984	0.000
Kendall's tau	0.908	0.000
Kendall's tau for Holdouts	1.000	

A high value of the Pearson coefficient, 0.984, confirms the high level of significance of the obtained results. Similarly, a high value of the Kendall correlation coefficient, 0.908, indicates a high level of correlation between the observed and estimated preferences. The Kendall coefficient for two holdout profiles has a value of 1.000, which is an additional indicator of the high quality of the obtained data.

As we can see in Table 5, when it comes to the only nominal attribute, the **Operator**, the highest average utility is held by the level **MTS** (0.188), followed by **Telenor** (0.062). The operator **VIP** was identified by the respondents as

undesirable, giving it a negative utility value of (-0.250). All of the other attributes are of the ordinal type, and the respondents displayed the expected behavior towards them, i.e. the levels that were presumed to have greater utility did indeed have it. For example, when it comes to the attribute **Interval**, the level “1s”, as expected, showed a greater utility (0.159) when compared to the intermediate level “60s+1s” (0.047) and the least desirable level “60s+60s” (-0.206).

The constant whose value is 4.622 represents a stochastic error obtained through regression analysis, and it is used to calculate the total utility of each profile.

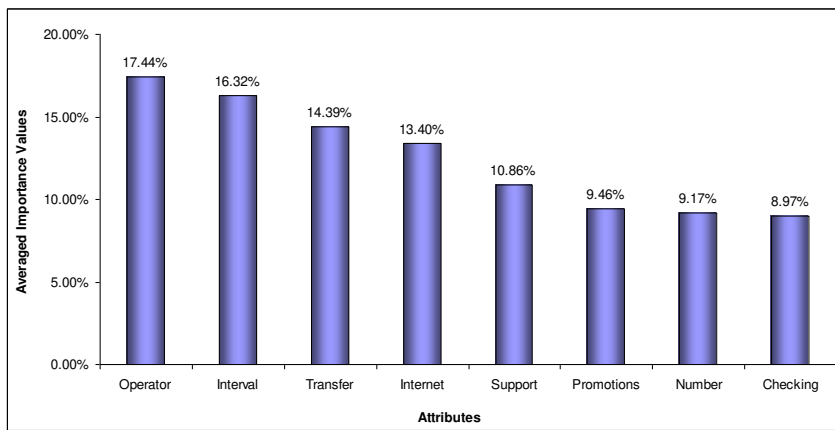


Figure 1

Averaged importance values

It can be seen in Figure 1 that the most important attribute to users is the attribute **Operator**, and its average importance value at the aggregate level is 17.44%. This result is particularly interesting due to the fact that during pre-research, during direct method surveying, that same attribute was positioned in the penultimate place. Still, the question remains whether such a high importance of the attribute **Operator** is a result of the averaging of attribute importance values at the sample level, or the fact that the conjoint analysis revealed hidden respondent preferences.

The attribute **Interval** has shown to be second by importance (16.32%). Such a high ranking of this attribute is not surprising, because most conversations among the student population last for less than a minute. With this in mind, the respondents are fully aware of the fact that the package with fewer minutes and rounding off to 1 second is better for them than the package with a greater number of minutes and rounding off to 60s+1s or 60s+60s.

Next in line according to importance is the attribute **Transfer** (14.39%), which leads to the conclusion that a great number of respondents do not spend all of their free traffic within a month, and therefore they find it important to be able to

transfer the traffic onto the next month when it might be used, and therefore avoid additional billing.

High positioning of the attribute **Internet** (13.40%) is a result of the fact that the student population greatly uses the Internet, while mobile phones have become devices from which the internet is increasingly being accessed. In addition, there are a growing number of mobile phone services that require constant Internet access.

The poor positioning of the remaining attributes can be interpreted as the fact that students mostly think about current monthly spending and internet access, while the quality of service and potential future promotions have currently no great importance among the student population.

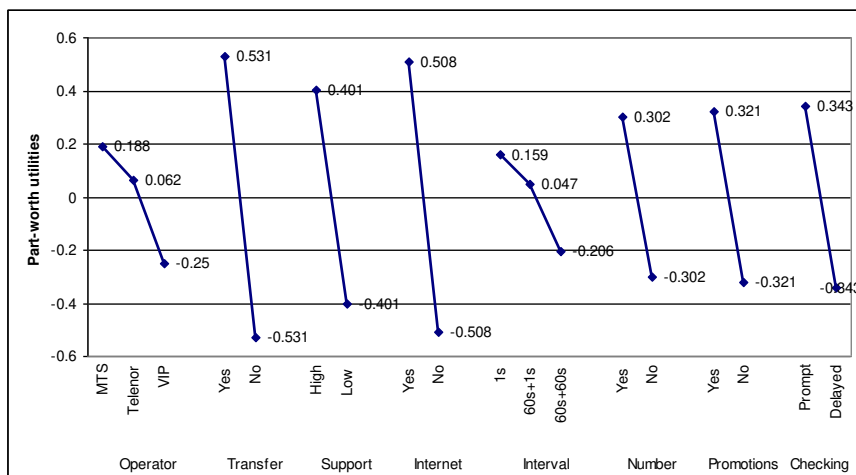


Figure 2
Part-worth utility functions

Figure 2 shows the part-worth utility functions for all of the attributes included in the study. It may be noted that all of them are extremely sensitive to level changes, but for the attribute **Interval** this sensitivity varies depending on the interval. Namely, the preferences decline much faster in the interval 60s+1s to 60s+60s than in the interval 1s to 60s+60s. Nevertheless, only the best levels of each attribute increase the overall respondent preferences, while the worst decrease them (negative sign for part-worths).

3.2 Preference-based Segmentation

A more detailed analysis of part-worths at the individual level revealed wide heterogeneity in consumer preferences. Therefore, a cluster analysis was

performed to classify respondents into more homogeneous preference groups. These part-worths are then used as input for cluster analysis. This approach has been conducted by various researchers across industries, in order to determine customer segments based on distinct preference profiles [19, 32, 33, 34].

The k-means cluster procedure in SPSS 16.0 was used to perform the segmentation. Based on the sample size, the solutions were searched in two and three clusters. The 3-cluster solution resulted in one segment that was very small in size and could not be statistically reliable ($n < 15$). A 2-cluster solution was chosen due to the size of the segments and statistical significance. An analysis of variance revealed that the segments in the 2-cluster solution differed significantly from each other, with respect to their part-worths generated by the conjoint analysis.

The mean part-worths for each of the levels of the attributes of the two segments are given in Table 6, while the importance scores are shown on Figure 3.

Table 6
Cluster analysis results of mean part-worths

Attribute and levels	Segment I <i>n</i> = 42 (31.34%)	Segment II <i>n</i> = 92 (68.66%)
Mobile phone operator		
MTS	-0.4	0.45
Telenor	0.57	-0.17
VIP	-0.18	-0.28
Possibility of transferring unused free traffic to the next month		
Yes	0.36	0.61
No	-0.36	-0.61
Level of availability and quality of technical support		
High	0.39	0.4
Low	-0.39	-0.4
Free internet within package		
Yes	0.35	0.58
No	-0.35	-0.58
Conversation billing interval		
1s	0.73	-0.1
60s+1s	0.15	0
60s+60s	-0.88	0.1
Possibility of choosing preferred phone number		
Yes	0.24	0.33
No	-0.24	-0.33
Promotions following the expiration of the contract		
Yes	0.24	0.36
No	-0.24	-0.36
Account balance check		
Prompt	0.45	0.29
Delayed	-0.45	-0.29

3.2.1 Characteristics of Segment I

The first and smaller segment consists of 42 respondents (31.34%). The most important attribute to them by far is the Interval (importance value = 24.25%), while the most preferred is level “1s”. Next by importance is the attribute “Operator” with an importance value of 14.61%, while the most preferred operator is Telenor (part-worth utility = 0.57). Among the more important attributes in this group is also Checking (importance value = 13.55%).

If we observe the demographic data of the respondents that belong to this segment, it can be noted that the majority of them do not live with their parents (64.3%), and most of them are already using postpaid services (66.7%). It is also interesting that half of them are already using the services of Telenor. Based on this data, we can conclude that this segment mostly includes students who do not live with their parents, so they find it easier to have their parents pay the monthly phone bill instead of having to set aside money for credit several times a month. Considering that they also find the manner in which their conversations are billed to be important, it can be concluded that they all very careful not to exceed their subscriptions.

An operator who wishes to win this segment over should offer such a service where the emphasis would be on rounding off conversations according to the “1s” method, with instant balance updates. The possibility of transferring unused traffic and free internet within the package would only further attract new users, as well as keep the old ones.

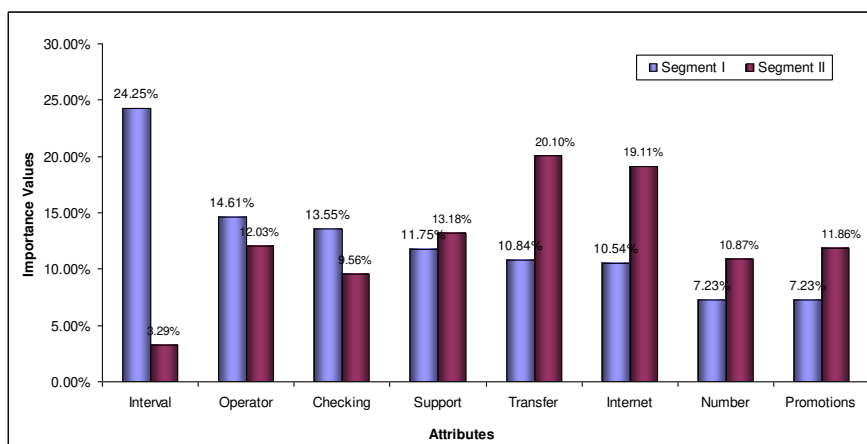


Figure 3
Importance values of attributes by segments

3.2.2 Characteristics of Segment II

The second, larger segment consists of 92 respondents (68.66%). The attribute with the greatest importance in this segment is the attribute Transfer (importance value = 20.1%), and right behind it is the attribute Internet with an importance of 19.11%. The third attribute by importance is the Support (13.18%), while fourth is the attribute Operator (12.03%) with an emphasis on the operator MTS. It is interesting that the attribute with the highest importance value in the first segment, the Interval, is by far in last place (importance value = 3.29%) in the second segment.

The demographic data of this segment shows that the majority of the respondents still live with their parents (56.5%), which is in sharp contrast with the first segment. Most of them use the services of MTS (62.0%), which indicates that they are satisfied with the current service.

The offer for this segment should emphasize the transfer of unused traffic, and free internet access within the package. Considering the fact that they do not find the billing interval to be very important, the model “60s+60s” could be left in this case, which allows for higher profits.

Conclusion

For mobile phone operators who operate in a highly competitive environment, it is very important to investigate the preferences of the segment of young people, who make up a significant base of future users. Meeting the needs and desires of this category of users can have an outcome of long-term loyalty to a particular company and its products or services.

The purpose of this paper was to use the conjoint analysis method to investigate how students from Serbia think when choosing a mobile postpaid package, i.e. what is it precisely that makes them choose a package of a specific operator, and not the offered services of a competing company.

The findings of the study are significant to marketers on both the theoretical and practical level. On the theoretical level, they add to our knowledge of the relative importance of the various mobile phone service factors that influence young consumer decisions. On the practical level, the results provide information to mobile phone operators which could help them provide appropriate customer service levels more effectively. Namely, based on the results showing the level of perception that university students have regarding postpaid mobile services, this study suggested a marketing strategy for mobile operators. Moreover, the study identified two segments that differed according to preferences, and thus suggested two different marketing strategies for each of them.

The implementation of a conjoint analysis should be repeated after a certain period of time because user preferences change over time as well, and this is especially present in the high tech sector, which also includes mobile telephony.

Acknowledgement

This research was partially supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Project numbers: III44007 and TR33044.

References

- [1] ITU: Key Global Telecom Indicators for the World Telecommunication Service Sector, 2011. Last retrieved October 15, 2012, from <http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf>
- [2] Plunkett's Wireless, Wi-Fi, RFID & Cellular Industry Almanac 2013. Last retrieved October 15, 2012, from Plunkett Research: <http://www.plunkettresearch.com/wireless-cellphone-rfid-market-research/industry-statistics>
- [3] RATEL: An Overview of Telecom Market in the Republic of Serbia in 2011, Last retrieved October 15, 2012, from http://www.ratel.rs/market/overviews_of_telecom_market.129.html
- [4] Totten, J. W., Lipscomb, T. J., Cook, R. A., Lesch, W.: General Patterns of Cell Phone Usage among College Students. *Services Marketing Quarterly*, 26(3) (2005) 13-39
- [5] McClatchey, S.: The Consumption of Mobile Services by Australian University Students. *International Journal of Mobile Marketing*, 1(1) (2006) 1-9
- [6] Aoki, K., Downes, E.: An Analysis of Young Peoples Use of and Attitudes toward Cell Phones. *Telematics and Informatics*, 20(4) (2003) 349-364
- [7] Leung, L.: Unwillingness-to-Communicate and College Students' Motives in SMS Mobile Messaging. *Telematics and Informatics*, 24 (2007) 115-129
- [8] Ha, J. H., Chin, B., Park, D.-H., Ryu, S.-H., Yu, J.: Characteristics of Excessive Cellular Phone Use in Korean Adolescents. *CyberPsychology & Behavior*, 11(6) (2008), 783-784. doi:10.1089/cpb.2008.0096
- [9] Walsh, S. P., White, K. M., Cox, S., McD. Young R.: Keeping in Constant Touch: The Predictors of Young Australians' Mobile Phone Involvement. *Computers in Human Behavior*, 27(1) (2011) 333-342
- [10] Kennan, W., Hazleton, V., Janoske, M., Short, M.: The Influence of New Communication Technologies on Undergraduate Preferences for Social Capital Formation, Maintenance, and Expenditure. *Public Relations Journal*, 2(2) (2008)
- [11] Kennedy, G., Krause, K., Judd, T., Churchward, A., Gray, K., Unit, B.: First Year Students' Experiences with Technology: Are They Really Digital Natives. *Educational Technology*, 24(1) (2008) 108-122

-
- [12] Akin, M.: Predicting Preferences of University Students for Prepaid vs Post Paid Cell Phone Service Plans. *Expert Systems with Applications*, 38(8) (2011) 9207-9210, doi:10.1016/j.eswa.2011.01.122
- [13] Kohne, F., Totz, C., Wehmeyer, K.: Consumer Preferences for Location-based Service Attributes: A Conjoint Analysis. *International Journal of Management and Decision Making*, 6(1) (2005) 16-32
- [14] Zhang, X., Prybutok, V.: How the Mobile Communication Markets Differ in China, the U.S., and Europe. *Communications of the ACM*, 48(3) (2005) 111-114
- [15] Hair, J. F., Anderson, R. E., Tathan, R. L., and Black, W. C.: *Multivariate Data Analysis*, Englewood Cliffs, NJ: Prentice Hall, 1995
- [16] Kuzmanovic, M., Martic, M.: An Approach to Competitive Product Line Design Using Conjoint Data. *Expert Systems with Application*, 39(8) (2012) 7262-7269. doi: 10.1016/j.eswa.2012.01.097
- [17] Kuzmanovic, M., Martic, M.: Using Conjoint Analysis to Create Superior Value to Customers, *Metalurgia International*, 17(2) (2012) 93-99
- [18] Kuzmanović, M., Obradović, T.: The Role of Conjoint Analysis in the New Product Price Sensibility Research. *Management – Journal for Theory and Practice Management*, 15(54) (2010) 51-58
- [19] Kuzmanovic, M., Panic, B., Martic, M.: Identification of Key Positioning Factors in the Retail Sector: A Conjoint Analysis Approach, *African Journal of Business Management*, 5(26) (2011) 10376-10386
- [20] Wilson-Jeanselme, M., Reynolds, J.: The Advantages of Preference-based Segmentation: An Investigation of Online Grocery Retailing. *Journal of Targeting, Measurement and Analysis for Marketing*, 14(4) (2006) 297-308
- [21] Hensher, D.: The Valuation of Commuter Travel Time Savings for Car Drivers: Evaluating Alternative Model Specifications. *Transportation*, 28 (2001) 101-118
- [22] Sohn, S. Y., Ju, Y. H.: Conjoint Analysis for Recruiting High Quality Students for College Education. *Expert Systems with Applications*, 37 (2010) 3777-3783
- [23] Popović M., Kuzmanović M., Martić, M.: Using Conjoint Analysis to Elicit Employers' Preferences toward Key Competencies for a Business Manager Position. *Management – Journal for Theory and Practice Management*, 17(63) (2012) 17-26, doi: 10.7595/management.fon.2012.0011
- [24] Kim, Y.: Estimation of Consumer Preferences on New Telecommunications Services: IMT-2000 Service in Korea. *Information Economics and Policy*, 17(1) (2004) 73-84

- [25] Sobolewski, M., Czajkowski, M.: Network Effects and Preference Heterogeneity in the Case of Mobile Telecommunications Markets. *Telecommunications Policy*, 36(3) (2012) 197-211
- [26] Kuzmanovic, M., Vujošević, M., Martić, M.: Using Conjoint Analysis to Elicit Patients' Preferences for Public Primary Care Service in Serbia, *HealthMED*, 6(2) (2012) 496-504
- [27] Head, M., Ziolkowski, N.: Understanding Student Attitudes of Mobile Phone Applications and Tools: A Study Using Conjoint, Cluster and SEM Analyses. *Proceedings of the 18th European Conference on Information Systems (ECIS 2010) Pretoria, South Africa, 2010*
- [28] Kim, C., Choe, S., Choi, C., Park, Y.: A Systematic Approach to New Mobile Service Creation. *Expert Systems with Applications*, 35 (2008) 762-771
- [29] Nakamura, A.: Estimating Switching Costs Involved in Changing Mobile Phone Carriers in Japan: Evaluation of Lock-In Factors Related to Japan's SIM Card Locks. *Telecommunications Policy*, 34(11) (2010) 736-746, doi: 10.1016/j.telpol.2010.10.003
- [30] Majláth, M.: Evaluation of Environmentally Friendly Product Attribute – Results of an Empirical Research. *Proceedings of the MEB 7th International Conference on Management, Enterprise and Benchmarking, Budapest, June 5-6, 2009*, 201-212
- [31] Orme, B.: *Which Conjoint Method Should I Use?* Research Paper Series, Sawtooth Software, Inc., 1996
- [32] Baker, G. A., Burnham, T. A.: The Market for Genetically Modified Foods: Consumer Characteristics and Policy Implications. *International Food and Agribusiness*, 4 (2002) 351-360
- [33] Haddad, Y., Haddad, J., Olabi, A., Shuayto, N., Haddad, T., & Toufeili, I.: Mapping Determinants of Purchase Intent of Concentrated Yogurt (Labneh) by Conjoint Analysis. *Food Quality and Preference*, 18 (2007) 795-802
- [34] Makila, M.: Retaining Students in Retail Banking through Price Bundling: Evidence from the Swedish Market. *European Journal of Operational Research*, 155 (2004) 299-316

Towards Convergence of Accounting for Emission Rights

Éva Karai and Mónika Bárány

Budapest University of Technology and Economics (BME), Department of Finance, Magyar tudósok körútja 2, H-1117 Budapest, Hungary, karai@finance.bme.hu; barany@finance.bme.hu

Abstract: Already from the start of the EU ETS, several investigations and analyses have shown that the involved actors are far from treating the emission rights uniformly in their accounting. There is no accepted and uniformly applicable method to determine the exact category and value of these new asset items and to identify how the obligations arising due to the reimbursement of emission rights are to be assessed. At the same time, the various measurement methods may cause significant differences in the profits reported by companies. Therefore, the companies and professional organisations involved in emissions trading are indeed entitled to demand clear guidelines about the accounting treatment of emission rights. The main problem arising in practice is that it is not clarified nor even considered from a theoretical aspect how far the various presentation and measurement methods contribute to the original objective of the emissions trading system, and hence which procedure would represent the most advantageous approach from accounting and social perspectives. The purpose of this research is, through the critical evaluation of the contents of national guidelines issued by professional bodies of European countries and through the review of the impacts of these specifications, to contribute to the creation of a clear and uniformly applicable method in the field of accounting for emission rights. A convergence in accounting for emission rights would be beneficial not only for companies, but also for professional bodies and legislators, independent of which member state are they from.

Keywords: emission rights; EU ETS; accounting; IFRS

1 Introduction

Today it seems to be a more accepted view that the human factor is decisive in the currently experienced change of climate, although in many cases there are scientific statements to the contrary as well. The natural greenhouse effect is a precondition of life on Earth, because it is indispensable for providing a tolerable temperature. With the progress of industrialisation, the ratio of greenhouse gases has continuously grown in the atmosphere, and this – according to the dominant

scientific view – contributes to the global climate change. Each country has various means to reduce the anthropogenic factors of global climate change, such as decreasing greenhouse gases, and these means include the fostering of environmentally conscious thinking, providing precise information to consumers about the impact of their consumption decisions on the emission of carbon dioxide, and supporting energy efficient solutions and a number of economic and financial incentives, of which only one is the setting up of the quota trading market on which this paper focuses. [1]

According to the theoretical model, the emission rights applying to the relevant period are distributed among the actors of this market, keeping in mind that the permitted emission level should be gradually lowered from period to period by each actor. In case an actor (organisation or individual) exceeds the emission level permitted for it, it can purchase the required quotas from an actor of the market who has surplus emission rights. The market mechanism ensures in this way the reduction of total emissions, because first those actors will curb their emissions which are able to do so by spending a limited amount, and then they are followed by those for whom reduction is much more expensive. In the United States, the system set up for sulphur dioxide emissions is based on this model, and the European Union bases its scheme introduced for carbon dioxide emissions also on this system.

1.1 The European Union Emissions Trading Scheme (EU ETS)

The Community and its Member States agreed that they would jointly meet their obligations to curb the climate change caused by anthropogenic factors, and to establish a European market which ensures the efficient trade of the emission allowances¹ of greenhouse gases. The related guidelines were accepted in 2003. The system covers all those sectors which are responsible for most of the EU's total greenhouse gas emissions. The experimental period (2005 to 2007) of introducing the trading system was followed by the first five-year trading period between 2008 and 2012, which coincided with the obligation period of the Kyoto Protocol. The third period of the system will run from 2013 to 2020.

For each period, the Member States elaborate their own national plans, in which they determine how many allowances will be distributed in the given period, by which method and for which facilities. This plan must be approved by the European Commission. The competent authority credits the relevant annual emission allowances every year by 28 February to the operator's account. The

¹ One emission allowance gives eligibility to emit one tonne of carbon dioxide equivalent in a specified period. Tonne of carbon dioxide equivalent: one metric tonne of carbon dioxide (CO₂) or such a quantity of any other greenhouse gas with an equivalent global-warming potential.[8]

allowances can be transferred within the Community between entities, and between entities within the Community and entities in third countries, if the latter recognise the allowances without limitations. The emission allowances are generally received free of charge by the operators involved, but depending on the Member State's decision, one part of the total quantity – up to 5% in the first three-year period and up to 10% between 2008 and 2012 – can be purchased at an auction. The allowances only apply to the emissions which were made in the period for which they were issued. [8]

Every year, up to 30 April at the latest, the emission allowances corresponding to the total controlled emission of the relevant facilities must be surrendered by the operator of the facilities to the state, and then the emission allowances handed over are cancelled. An operator which does not submit by the deadline the allowances of an appropriate quantity covering the previous year's emissions must pay a fine. The excess emission penalty is 100 Euros on each tonne of carbon dioxide equivalent emitted by the facilities, but uncovered by surrendered allowances (in the first three-year period the penalty was lower, only 40 Euros). Paying the fine does not relieve the operator from handing over in the following year the emission allowance of a quantity corresponding to the excess emission. [8]

According to the analyses carried out so far, the EU ETS can be considered to be a successful scheme, because it has obviously contributed to the member countries meeting their obligations undertaken in the Kyoto Protocol. However, the experience gathered in recent years has highlighted several problems, on the basis of which the European Commission identified many modification proposals. For example, the scope of the ETS will be extended in the future to several new industrial branches and sectors. In comparison with the current practice, in the period between 2013 and 2020, a much higher ratio of allowances will be auctioned, instead of a gratis distribution.

1.2 The Challenges of Accounting for Emission Rights

Accounting – as an area responsible for the external and internal data service of entities – is involved from several aspects in the emission rights and generally in the global climate change. One of the most important issues is: can the emission rights be presented as assets, and if so, which asset item should it be, and what is the value at which it is advisable to do so. From a theoretical side it is not clarified, and hence in practice it causes serious difficulties in identifying and classifying the emission rights properly. It is not clear for the entities whether this new item should be treated as rights falling into the category of intangibles or as securities or perhaps as inventories. In the current system, the organisations obtain most of the emission allowances free of charge when they are initially distributed, and only a small part is to be purchased in the EU ETS. Of the 26 largest polluters in the EU ETS, based on the 2008 statements, as many as 11 present the

allowances received through government distribution – i.e. the grants – as intangibles, 2 as inventories, and 6 as other assets, while the other enterprises do not disclose these figures. Similar proportions are found in the initial disclosure of emission rights [18]. The picture is varied regarding the measurement of liabilities and provisions arising due to the repayment of emission rights, in both theory and practice. The evaluation of emission allowances received as grants is a disputed area, but at the same time – because of the magnitude of distributed emission allowances – it may have a substantial impact on every entity's financial statements.

Already in the first trading period, two trends emerged in the disclosure and assessment of emission rights [1]:

- One of them recommended the showing of net position in the case of emission rights. In this event, only the purchased emission allowances may be presented in the balance sheet. In the first trading period, lacking any regulations, as many as 60% of the examined entities applied this net approach [14].
- The other trend was the gross method, basically in accordance with the experience gathered regarding the sulphur dioxide emission trading system launched by the US EPA in 1990. Accordingly, the emission allowances obtained as grants should be shown in the balance sheet just like the purchased allowances, and they are to be taken into consideration in the expenses when they are used as a compensation for the emitted pollution [22]. Therefore, the emission rights obtained free of charge are to be treated as a government grant, and they are to be shown at the fair value at the time of receipt. This creates a basis for the uniform handling of emission rights regardless of whether having been obtained by government distribution or by purchase.

According to the IETA² review of 2007, the gross method was used by only 5% of the companies, and this approach was reflected also by the IFRIC 3 published in 2004 and then withdrawn after less than six months. A review of the 2008 statements of the 26 largest polluters covered by the EU ETS confirms the finding already outlined, namely that contrary to the IFRIC 3 recommendation, most of the involved companies use the net method (15 out of the examined 26 enterprises) [18].

² International Emissions Trading Association

1.3 IFRIC 3 Interpretation about the Emission Rights

The IASB³ Interpretations Committee⁴ issued the IFRIC 3 Emission Rights Interpretation on 2 December 2004. In spite of the fact that IFRIC 3 was withdrawn by IASB less than six months later in June 2005, this interpretation has an impact until this day on the practice of accounting for emission rights [19] [20].

The European Financial Reporting Advisory Group (EFRAG) did not recommend the endorsing of IFRIC 3 [10], and on this basis, the European Commission did not approve the interpretation either, and in June 2005 it was withdrawn by IASB [12]. EFRAG's argument was that IFRIC 3 did not meet the requirements identified in association with the application of international accounting standards, i.e. the requirements laid down in Regulation 1606/2002 of the European Parliament and Council⁵, because

- it is contrary to the true and fair view principle (Directive 83/349/EEC, Article 16, clause (3), and Directive 78/660/EEC, Article 2 (3)), and
- it fails to meet those requirements of clarity, relevance, reliability and comparability which are expected of the financial information necessary for economic and responsible management decisions [10].

EFRAG has expressed its concerns also about the cost model, the revaluation model and the accounting entries after the compliance period. In the course of applying the cost model – resulting from the different evaluation of assets and associated liabilities – mismatch may arise in the balance sheet and in the profit and loss statement. The mismatch observed in the case of fair value accounting can be traced back to the revaluation of emission rights against equity and the evaluation of resulting liabilities against profit and loss. This mismatch also prevails after the compliance period, until the debt is settled. EFRAG's further criticism was that the companies were not allowed – in spite of this being in harmony with the standards – to calculate the result of the process at the end of the compliance period, including the net effect in the profits [10].

Further accounting opportunities featuring in the standards referred to by IFRIC 3

According to IAS 20 dealing with the accounting of government grants, two solutions are available in the case of non-monetary government grants:

- the assets, and the grants associated with the assets, can be presented at a fair value in the balance sheet (government grant approach, GGA) or
- both the assets and the grants can be shown at the nominal amount (nominal amount approach, NAA) [17].

³ International Accounting Standards Board

⁴ International Financial Reporting Interpretations Committee (IFRIC)

⁵ Regulation (EC) No 1606/2002 of the European Parliament and of the Council

The IFRIC 3 interpretation had specified accounting based on a fair value. If the company applied the nominal amount method, the emission rights and the received grant would also be presented at a nominal amount, which would be zero in this case, because no emission value prevails. The method practically leads to the same result as the net approach, because the grants received and the emission rights obtained by the grants balance out each other, i.e., both the so obtained emission rights and the grants received appear at a zero value in the balance sheet. If the company buys the emission rights, they are booked at the purchase price.

According to IAS 37, provisions can be generated in two ways: by the gross and net liability approaches. The IFRIC 3 interpretation advocated the recognition of provisions by the gross method, i.e., presented the liabilities applying to the handover of emission rights.

In the case of recognition provisions by the net approach, the companies do not recognise provisions until they have as coverage a volume of emission rights necessary for handing over a quota corresponding to the emissions in the period. If they do not have a quota to cover the emissions in the subject year, then through the application of the principle of best estimate, provisions must be generated for the lacking volume.

2 The Established Practice for the Accounting of Emission Rights

Painting a brief picture above in relation to the problems of emission rights disclosure and assessment underlines the justified requirement of companies involved in emission trade for clear guidelines in the accounting for emission rights [9]. In the following discussion, we shall review and analyse different solutions, and then by means of an example, we shall attempt to shed light on the conclusions that can be drawn from these methods. We examine how the proposals issued by the governments and professional accounting bodies of four countries affect the financial statements of companies. These proposals were issued by the following institutions: Instituto de Contabilidad y Auditoría de Cuentas (ICAC) in Spain [21], the Institut Deutscher Wirtschaftsprüfer (IDW) in Germany [13], the Austrian Financial Reporting and Auditing Committee (AFRAC) in Austria [1] and the HM Treasury [11] and the Department of Health [1] of the Government in Great Britain.

2.1 The Place of Emission Rights in the Balance Sheet

The emission rights are presented in each of the examined accounting models, and they appear in the statements, but their balance sheet classification and valuation can be very different depending on the statutory provisions of each nation and the

related opinions issued by various accounting bodies. Therefore, emission rights are shown within the non-current assets as intangible assets and also among the current assets.

According to the IDW model, the AFRAC model, and the UK fair value model, the emission rights are intangible assets which are to be presented in the balance sheet among the current assets [1] [1] [13]. According to the guideline, in the IDW model, the emission rights associated with the production process must be shown as inventories, and the other emission rights as other current assets [13]. In the AFRAC model, the emission rights are other current assets [1], and in the UK fair value model current asset investments [1]. In the ICAC model, emission rights are shown in the balance sheet within the non-current assets as intangible assets [21]. In Switzerland, Leibfried and Eisele present the emission rights in the balance sheet as non-current assets, among the intangibles [16].

2.2 The Initial Recognition of Emission Rights

The first recognition of emission rights depends on whether the entity has acquired the rights against a fair consideration or free of charge (or at a favourable rate) by government distribution. In the case of assets obtained against a fair consideration, practically no deviation is seen among the various solutions: the rights are entered at purchase cost. The IFRIC 3 as well as the British fair value model and the Spanish ICAC model also require showing at the market value the assets obtained without transferring consideration [1] [11] [15]. In the IDW and the AFRAC models, when presenting the emission rights initially, a business organisation may choose from two methods.

- In the German model, in the case of assets received without consideration, i.e., government grants, the assets can be entered at zero value (nominal amount) and also at the market value which prevailed at the time of distribution [13].
- In Austria, the Austrian Commercial Code (UGB) does not provide instructions about the evaluation of assets obtained without consideration. AFRAC, in its publication about the accounting presentation of emission rights, recommends that the rights obtained by a government distribution be capitalised at the market value prevailing at the time of subscription. However, as an alternative solution, in case the expected emission is higher than the quantity of distributed quotas, the entity may disregard capitalising the rights received by a government distribution, but it must disclose information about the market values [1].

2.3 The Initial Recognition of Government Grants

In the examined accounting systems, the initial recognition of government grants is in harmony with the valuation applied for emission rights. In case the emission rights obtained without a consideration is featured in the balance sheet at fair

value, then the government grant is also shown at fair value. If the emission rights are featured at the nominal amount, the government grant also appears in the balance sheet at the nominal amount, rights received without consideration are featured at a zero value.

2.4 The Sale of Emission Rights

The profits of selling emission rights generally fall into the category of operating profits. The only exception is the Spanish regulation, where the profits stemming from the sale of intangible assets appear as an extraordinary profit [21]. The German IDW recommends the presentation of profits resulting from sale as other operating revenues [13]. AFRAC recommends the accounting of sales by the gross method: cancellation is booked in material expenses, and the consideration in the category of sales revenues or other operating incomes [1].

In all accounting systems, simultaneously with the sale when the emission rights are cancelled, the government grant featuring on the liabilities side must also be proportionately cancelled.

2.5 Subsequent Measurement of Emission Rights

Except for the British fair value model, the emission rights were evaluated at the historical cost.

- In the Spanish ICAC model, emission rights are presented within the non-current assets, as intangibles, but the accounting of amortisation is not permitted. Impairment must be accounted for the emission rights if the recoverable amount determined on the basis of IAS 36 is lower than the book value of the assets. Impairment is accounted for as other operating expenses [21].
- Concerning Swiss entities, Leibfried and Eisele found examples for amortisation of emission rights, on the grounds that they have a defined and useful business cycle [16].
- In Germany and Austria, the “strict lower of cost or market” principle is applied in the course of the subsequent measurement of emission rights appearing among the current assets. This means that if the fair value on the balance sheet date is lower than the book value, a write-down is to be made to the fair value at the balance sheet date. In the case of emission rights registered at a market value, obtained by government distribution or purchased by the enterprise, this method is also applied. No impairment may be accounted for assets which are booked at zero value [1] [13].

The British fair value model assesses the emission rights featuring among the current assets at the fair value of the balance sheet date. In this case, revaluation is done against the government grant and not against the revaluation surplus [1].

2.6 Provisions Recognised to Deliver Allowances

The value of liabilities and provisions recognised to deliver the allowances corresponding to the actual emission of the period may show deviations in the financial statements. Basically, provisions can be generated in two ways, by the gross and net methods (IAS 37). The withdrawn IFRIC 3 used the gross approach, and showed the balance sheet date obligation corresponding to the actual emission at a fair value to be determined by the best estimate [15]. The difference between the evaluations of assets and provisions caused the striking problem that the profit impact associated with the given period appeared in several periods and therefore the underlying assumption of accrual basis was violated. For overcoming this problem, several solutions were developed in practice, as reflected also by the accounting recommendations of the examined nations.

- In Germany, Austria and Spain, an attempt was made to determine the recognition value of provisions (liabilities) in a way that the deviation between the book value of the rights to be transferred and the value of provisions is minimised. In determining the recognition value of provisions, the German and Spanish guidelines set out from the assumption that first the rights obtained through government grants are used up, and hence the historical cost of these rights is taken into consideration in the value of provisions, even if the historical cost of the rights is zero (see German nominal value model). In case the entity has obtained less emission rights through government grants than the actual emission, then as the next step, when determining the amount of provisions, it must take into consideration the historical cost of the emission rights purchased. If the entity has not bought in the reporting period additional emission rights, then according to the German guidelines, provisions for the missing quantity of rights are to be generated at the balance sheet date fair value of the emission rights, while the Spanish guidelines specify the application of the best estimate which can differ from the balance sheet date value. According to AFRAC 's guidelines, the determining of liabilities or provisions must follow the accepted cost formula (FIFO, weighted average, etc.) applied decreasing the emission rights, and the missing quantity of emission rights must be entered at the market value at the balance sheet date [1] [13] [21].
- Based on the guidelines of the British fair value model, provisions must be generated for the quantity of rights to be handed over, and the value of provisions must be determined at the fair value at the balance sheet date. Since the emission rights and the government grants are to be revalued to the balance sheet date fair value, at the time of handover – if the rights necessary for handover are available to the entity already before the balance sheet date – no difference is generated between the book value of the emission rights to be handed over and the value of recognised provisions [1]. A difference only emerges if the historical cost of the emission rights obtained (purchased or granted) after the balance sheet date deviates from the fair value of the balance sheet date.

2.7 The Subsequent Measurement of Government Grants

The German, Austrian and Spanish guidelines describe that the incomes resulting from the cancellation of government grants should be shown simultaneously with the provisions recognised to deliver allowances, the impairment accounted for the emission rights and the expenses arising due to the cancelling of emission rights [1] [13] [21].

According to the English fair value model, the value of government grants changes in the course of the subsequent measurement with the value of granted emission rights featuring in the balance sheet. A change in profit is only achieved if expenses in association with the emission rights were accounted for in the relevant period [1].

2.8 Deliver of Allowances

In general, the entities settle the accounts in accordance with their actual emissions with the responsible state authority in the business year following the reporting period. When the rights are handed over, simultaneously with the cancellation of rights, the provisions (liabilities) generated must be eliminated. A profit impact emerges if the book value of the assets to the cancelled deviates from the value of the provisions (liabilities). In the case of examined accounting recommendations and national regulations, this profit impact influences the reported operating/business profits of the entity [1] [1] [13] [21].

3 Case Study for the Accounting of Emission Rights on the Basis of the Presented Accounting Practice

In the following discussion, we shall show examples based on the German IDW and the British DH recommendations, as well as the Spanish ICAC resolution of 2006, regarding the accounting practices in relation to the emission rights.

Example⁶: In a government grant, a quota corresponding to 13,000 tonnes of CO₂ is credited to the account of an entity; the entity does not have a quota brought forward from previous years. At the time the quota is credited, the market rate of quotas is CU10. The entity's business year coincides with the calendar year. It draws up an interim report with the end date of 30 June, when the market value of quotas is CU12. Until the end date of the interim report, the entity emitted 5,500 tonnes of CO₂, and the expected annual emission is 12,000 tonnes. The entity sells in the first six months of the year a quota corresponding to 1,000 tonnes, at

⁶ Prepared on the basis of IFRIC 3

CU11.5. On and after the year-end date, when the emission rights are delivered, the market value of the quotas is CU11.

If the interim financial statement prepared in accordance with the various national solutions are compared with the original IFRIC 3 interpretation (Table 1), it is found that the total assets calculated according to the cost method crops up again in the German market value based model and according to the Spanish ICAC resolution, while the British fair value model results in the same balance-sheet total as the revaluation model in IFRIC 3. The entities keeping their books on the basis of the German nominal amount method significantly deviate from these methods. In their case, neither the emission rights nor the government grants appear in the balance sheet, and therefore this value is missing also from the balance sheet total of the entity.

The national guidelines recommend the gross method for the assessment of provisions [13] [21], but their values show deviations in the interim reports from the value recommended by IFRIC 3 which also used the gross method – except for the English model [1] [11]. The deviation is the consequence of the various measurement methods. In the German and Spanish models, the provisions – in harmony with the measurement procedure applied for the emission rights – are presented in the balance sheet at the historical cost of the emission rights. (Also in the German nominal amount method, but the value of provisions is zero, because the emission rights obtained as grants are also entered at this value). Again no mismatch emerges in the British model, because both the emission rights and the related provisions are evaluated at the balance sheet date fair value.

Deviations can be experienced also in the value of government grants. The balance sheet value of government grant is identical in the German market value based model and the Spanish model with the corresponding value based on IFRIC 3. The government grants are presented at the fair value of the emission rights at the grant date. In the case of the German nominal amount method, in accordance with the value of the emission rights, the balance sheet value of the government grant is zero. In the English fair value model, the balance sheet value of government grant is also in harmony with the value of the emission rights, and the government grants are shown at the balance sheet date fair value.

In all the three national models it can be seen that the balance sheet value of the emission rights is equal to the sum of balance sheet values in the liabilities side provisions and government grants category. Consequently, the national models – partly following a different practice – eliminated the mismatch resulting from the deviating measurement of liabilities and assets in the IFRIC 3 interpretation. The impact made on the profits is also unambiguous: the interim financial statement presents the actually realised profits stemming from the sale of emission rights. The difference is spectacular in comparison with the IFRIC 3 interpretation. While the business events of the first six months demonstrated in the example generated CU500 profits according to IFRIC 3, on the basis of the accountings of national

models this profit is uniformly CU11,500. Except for the Spanish method, this profit is manifest in the profits of the operating and business activities. In the Spanish model, the sale of intangible assets is qualified as an extraordinary event, i.e., it appears as an extraordinary profit.

The following explanations can be attached to the balance sheet values at the balance sheet date according to the national guidances. In the German market value method [13] and also in the Spanish method, the emission rights are featured at the historical cost [21]. According to the German nominal amount method, the balance sheet value of the emission rights obtained as a grant is zero. The British model shows the rights consistently at the balance sheet date fair value. The balance sheet date value of a government grant is zero in each method, because the grant has been used in the business year [1] [13] [21]. The balance sheet value of provisions is in line with the measurement method of emission rights. It can be noted in each method that the balance sheet value of provisions is CU5,500 higher than the balance sheet value of emission rights. And this amount is nothing else but the estimated value of the quota applying to the 500 tonne emission missing on the balance sheet date. This expense practically erodes the first six-month profits of the entity shown in the example. Already in association with our example related to the IFRIC 3 interpretation we have stated that the accumulated profit impact was CU6,000 (Table 1). This accumulated profit is shown generally in the national reports within the operating profit. The only exception is the Spanish statement, where the profit impact resulting from the sale of intangible assets is shown in the extraordinary profits [21].

In the case of IFRIC 3, in the statements of the business year following the balance sheet date, a significant profit impact appears in association with the previous year's accounting period of the quotas. In the financial statements based on the national guidances, however – in the case of an appropriate estimate – the profit impact indeed appears in the period with which it is associated and it does not influence the profits of the subsequent business years. This means that the examined national guidances eliminate the deficiency which IFRIC 3 has been accused of, because in this case the underlying assumption of accrual basis is manifest.

3.1 Models Based on Recognising Provisions by the Net Method

In this section, we extend our case study through two different accounting methods based on international accounting standards (Table 1); the government grants are presented at a fair value in the first one (GGA method) and at nominal amount in the second one (NAA method). In both cases the provisions are measured by the net method (on the basis of Leibfried *et al.* [16] and Lorson *et al.* [17])

Table 1
Accounting models for emission rights

Interim balance sheet	IFRIC 3		Germany (IDW)		AFRAC	Spain	UK model	Provisions - net method	
	Cost model	Revaluation model	Market value model	Nominal amount model	Market value model	ICAC model	Fair value model	GGA model	NAA model
Intangible assets (non-current)	120000	144000				120000		120000	
Inventories			120000						
Cash	11500	11500	11500	11500	11500	11500	11500	11500	11500
Other current assets					120000		144000		
Total assets	131500	155500	131500	11500	131500	131500	155500	131500	11500
Profit for the year	500	500	11500	1500	11500	11500	11500	11500	11500
Revaluation surplus		24000							
Provisions	66000	66000	55000		55000	55000	66000		
Government grants	65000	65000	65000	10000	65000	65000	78000	120000	
Equity and liabilities	131500	155500	131500	11500	131500	131500	155500	131500	11500
Balance sheet at the year end									
Intangible assets (non-current)	120000	132000				120000		120000	
Inventories			120000						
Cash	11500	11500	11500	11500	11500	11500	11500	11500	11500
Other current assets					120000		132000		
Total assets	131500	143500	131500	11500	131500	131500	143500	131500	11500
Profit for the year	-6000	-6000	6000	6000	6000	6000	6000	6000	6000
Revaluation surplus		12000							
Provisions	137500	137500	125500	5500	125500	125500	137500	5500	5500
Government grants	0	0	0	0	0	0	0	120000	
Equity and liabilities	131500	143500	131500	11500	131500	131500	143500	131500	11500
Balance sheet at surrender date									
Cash	6000	6000	6000	6000	6000	6000	6000	6000	6000
Total assets	6000	6000	6000	6000	6000	6000	6000	6000	6000
Profit for the year		0	0	0	0	0	0	0	0
Retained earnings	-6000	6000	6000	6000	6000	6000	6000	6000	6000
Equity and liabilities	6000	6000	6000	6000	6000	6000	6000	6000	6000

Applying the net method, no provisions are presented in the interim report, because the emission rights available will cover the actual emission [16] [17]. This measurement method of provisions – in case of the GGA method – makes an impact on the valuation of the government grant also: the deferred income is not cancelled because no expenses arise. The NAA method leads to a result identical with that of the German nominal amount method, because in that case the generated provisions – which will be zero at the time of applying the nominal amount – are determined based on the historical cost of emission rights.

In the case of the GGA method, it can be seen that the government grant which should appear as deferred income is also featured in the balance sheet on the balance sheet date at the market value at the grant date. This raises doubts, because pollution emission exceeding the government grant took place in the period, i.e., it would be justified to eliminate the government grant as a deferred income. This problem does not prevail in the case of the NAA method, because both government grants and emission rights are shown at zero value.

To summarize, in these models the full accumulated profit impact appears in the business year when the distributed quotas are actually used. In the subsequent year, when the rights are actually delivered, no profit impact is booked, when the emission rights, the generated provisions and the amount of government grant are cancelled against one another. In the NAA method, due to the zero value of the emission rights and the received grant, the purchased emission rights and the provisions have to be cancelled.

4 The Main Questions and Answers Relating to the Accounting for Emission Rights

On the basis of the presented models, the following main questions are outlined in association with the accounting for emission rights.

4.1 Emission Rights: Non-Current Assets or Current Assets?

Of the emission rights purchased or obtained through a government grant, those rights must always be classified as current assets which are realised within 12 months after the reporting period, in accordance with the definition of standard IAS 1 Presentation of financial statements. Can the rights reserved for a longer period be considered as non-current assets? IAS 1 (68) emphasises that the inventories “that are sold, consumed or realised as part of the normal operating cycle” must be shown among the current assets even if their realisation is not expected within 12 months after the reporting period. Could this cover the emission rights?

From the definition of inventories in IAS 2, it is unambiguous that emission rights held for a sales purpose are qualified as inventories, but the question is, can rights held for own use be treated as inventories? The emission rights relating to the production process behave like “materials and supplies” that are consumed in the production process. In case the emission rights held for use can also interpreted as inventories, then – according to IAS 1 – they must be presented as current assets, regardless of the intended period of use. The most common argument against recognising the emission rights for use as inventories is that these rights do not have a physical substance [19]. Presenting goods without physical substance among the inventories is commonly used, but if an asset without physical substance behaves as a material, this approach is indeed unusual. In this case, users may refer to the substance over form principle.

4.2 Government Grant and the so Obtained Emission Rights: at Fair Value or Nominal Amount?

The countries that permit accounting on the basis of the nominal amount method generally specify a disclosure obligation. Therefore, the necessary information about market values are available in the notes. [1] [13] In our view, it would globally better enhance the comparability of financial statements if these data appeared in the balance sheet.

4.3 How should the Subsequent Measurement of Emission Rights Take Place?

With the emission rights treated as inventories, the subsequent measurement can be brought in accordance with the IAS 2 regulations about the subsequent measurement of inventories: the inventories must be evaluated at the lower of the historical cost and the net realisable value. If the realisable value is below the historical cost, the value of the emission rights must be reduced to the lower market value. The realistic assessment of the emission rights and hence their revaluation to a higher market value is not possible on the basis of IAS 2.

4.4 Recognising Provisions by the Gross or Net Method?

We have demonstrated with the GGA method that as a result of the net approach of generating provisions, government grants are not fully eliminated at the end of the period because of the lack of relating expense, although it could be necessary based on the actual emission of greenhouse gases. In this case, the government grant is practically not a deferred income.

4.5 How should Provisions and Government Grants be Evaluated?

Government grants imply that the emission rights are based on subsidies, and therefore it is obvious that the grants should be presented in the balance sheet at the same value as the rights. The provisions – if they are shown by the gross method – embody liabilities applying to the handover of rights in association with a periodical emission. This obligation may apply also to emission rights obtained by a government grant, and in this case the debt part related to the handover of these rights must be featured at the same value as that of the assets serving as a coverage. If there is no harmony between the evaluations of the assets available and the liabilities associated therewith, this leads to mismatch; the most striking appearance of this is that the profit impacts do not appear in the period to which they actually relate, violating by this the underlying assumption of accrual basis. The national models examined during the research found various solutions for this problem.

We attempted to find a consistent solution for our proposals above also in the subsequent measurement of provisions and grants. The government grant and the so obtained emission rights must be presented at the same value in order to avoid any mismatch. In the course of a subsequent measurement, it may happen that the value of emission rights is reduced to the net realisable value, the government grant is cancelled simultaneously, and therefore the emission rights obtained by a grant and the government grant are featured at the same value in the balance sheet. A government grant remains in the balance sheet if the emission rights obtained through the grant and associated with the reporting period have not been fully utilised by the entity, i.e., its total emission in the period was below its permitted emission level. The actual emission is reflected by the value of generated provisions. The value of provisions depends on how the level of emission develops vis-à-vis the available emission rights, and how the entity obtained these covering rights.

Let us assume that an entity has emission allowances exclusively stemming from government grants, and they cover the actual emission of the entity. In this case, the balance sheet value of provisions must be equal to the value of emission rights obtained by a government grant and also handed over as a result of the emission. How to proceed if the entity has emission rights stemming from a government grant exclusively, but they do not cover the actual emission? The value of the provisions must be determined in a way that it approaches as closely as possible the value of the rights to be delivered. In case the entity has purchased the missing rights before preparing the balance sheet, harmony in the valuation of assets and liabilities can be created if the provisions are determined jointly at the book value of the rights available on the balance sheet date and at the historical cost of the missing rights obtained after the balance sheet date. If the entity obtains the missing rights after the reporting period, the value of provisions regarding the

missing rights must be determined with the best possible estimate based on the most precise information available at the time of preparing the balance sheet. The best estimate does not necessarily equal the balance sheet date market value of the emission rights.

In case the entity obtained the emission rights not only through a government grant, the value of provisions must be determined on the basis of the book value of available rights obtained or purchased. In case the available rights do not cover the actual emission, provisions must be generated for the missing rights through the application of the principle of best estimate. However, the situation raises many questions when the entity has more emission rights than necessary for the actual emission: How are the provisions determined and which value of the rights is to be considered as the basis for measurement?

It is only a seemingly appropriate solution to determine the provisions in such a case by the cost formulas of IAS 2, moving average price or the FIFO method, because these methods could again lead to mismatch. The government grant is to be shown as income of the reporting period, to such an extent by which the received grant is actually realised. An equilibrium situation prevails if the incomes realised due to the emission obtained as a grant is counterbalanced by the expenses arising through the provisions generated according to the emission level. This is only possible if the provisions, and hence also the expenses, are determined primarily at the book value of the emission rights obtained as a grant, and the value of the purchased emission rights is only taken into consideration in the value of the provisions if the rights obtained by the grant do not provide a coverage for the actual emission. The value of liabilities applying to the handover of purchased emission rights can then be determined by the moving average price or the FIFO method.

4.6 Where should the Profit Impact Related to the Emission Rights be Shown?

Since the emission of pollutants is part of the production process, regarding the quotas held for sale or usage, it is justified in all cases to present the impact on profits within the category of operating profits. In certain national regulations, the impact made on the profit by certain items is entered as an extraordinary profit. An example could be in Spain the impact made on the profit of selling intangible assets or in Hungary the showing of received grants as extraordinary revenues. Presentation within extraordinary profit distorts the impact made on the operating profit.

Conclusions

The differing accounting treatment of emission rights causes problems in the field of group accounting and comparability, and also places a very high administrative burden on companies. In our opinion, taking into consideration the role of

emission rights in the production process and the relevant specifications of the International Financial Reporting Standards:

- it is justified to show emission rights as inventories,
- it is justified that emission rights obtained through a government grant, the received government grant and the provision should be presented by the gross method,
- the subsequent measurement of emission rights is to be brought in accordance with the standard IAS 2 (Inventories),
- violating the underlying assumption of accrual basis can be avoided if the emission rights, the government grants and the provisions are evaluated in line with each other. To this end, the available emission rights must be reflected in the value of government grants and provisions. If the emission rights do not provide coverage for the actual emission, the liabilities applying to the missing rights must be determined by the principle of the best estimate (Table 2).

Acknowledgement

This work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the New Széchenyi Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

References

- [1] Accounting for the European Union Greenhouse Gas Emissions Trading Scheme, Department of Health UK, 2006, NHS Finance Manual, <http://www.info.doh.gov.uk/doh/finman.nsf/Admin%20Views%20%5C%20Stubs/Whatsnew>
- [2] AFRAC Stellungnahme „Bilanzierung von CO₂-Emissionszertifikaten gemäß österreichischem HGB“ (2006 Februar) der Arbeitsgruppe „CO₂-Emissionszertifikate“
- [3] Baricz, R.: Mérlegtan, Aula, Budapest, 1994, pp. 45-61
- [4] Bebbington, J., Larrinaga-González, C.: Carbon Trading Accounting and Reporting Issues, *European Accounting Review*, 17:4, 2008, pp. 697-717
- [5] Beck'scher Bilanzkommentar §248 note 70-81, §249 note 100, §255 note 325 Verlag C. H. Beck München, 2010
- [6] Coenenberg, A. G., Haller, A., Schultze, W.: Jahresabschluss und Jahresabschlussanalyse Betriebswirtschaftliche, handelsrechtliche, steuerrechtliche und internationale Grundsätze – HGB, IFRS, US-GAAP Schäffer-Poeschel Verlag Stuttgart, 2009
- [7] Cook, A.: Emission Rights: From Costless Activity to Market Operations, *Accounting, Organizations and Society*, 2009, 34(3-4)

Table 2
Main differences in the national guidelines about the accounting for emission rights

	German IDW models		Austrian AFRAC model	Spanish ICAC model	DH of UK government fair value model	Our conclusion
	Market value model	Nominal amount model	Market value model			
Balance sheet position of emission rights	Inventory/Other current assets		Other current assets	Intangibles (non-current assets)	Current asset investments (intangible assets)	Inventory
Initial recognition of emission rights granted	market value at the grant date	at zero value	market value at the grant date		fair value at the grant date	market value at the grant date
Initial recognition of government grants	market value at the grant date	at zero value	market value at the grant date		fair value at the grant date	market value at the grant date
Recognition of provisions	Gross method - First: carrying value of rights granted, then of rights purchased, and market value at BSD for missing rights		Gross method - The used measurement method (FIFO, weighted average) for existing rights and market value at BSD for missing rights	Gross method - First: carrying value of rights granted, then of rights purchased, and best estimate for missing rights	Gross method - fair value at the BSD	Gross method - First: carrying value of rights granted, then of rights purchased, and best estimate for missing rights
Subsequent measurement of emission rights	Lower of cost or market value at the BSD			based on IAS 36	fair value of emission rights at the BSD, revaluation of granted rights against government grants	based on IAS 2
Subsequent measurement of government grants	Decrease of government grants, when expenses are recognised relating to emission rights granted	-	Decrease of government grants, when expenses are recognised relating to emission rights granted		fair value at the BSD - profit impact only when expenses are recognised relating to emission rights granted	Decrease of government grants, when expenses are recognised relating to emission rights granted
P/L impact	Operating profit			Operating profit and extraordinary profit	Operating profit	Operating profit

- [8] Directive 2003/87/EC of the European Parliament and of the Council of 13 October 2003 establishing a scheme for greenhouse gas emission allowance trading within the Community and amending Council Directive 96/61/EC
- [9] Duh, M; Strukelj, T.: The Integration and Requisite Holism of the Enterprise' Governance and Management as Preconditions for Coping with Global Environmental Changes, *Acta Polytechnica Hungarica*, 2011, 8(1) pp. 41-60
- [10] EFRAG, Chairman Stig Enevoldsen: Adoption of IFRIC 3 Emission Rights, comment letter to Dr. Alexander Schaub, Director General, European Commission, <http://www.iasplus.com/interps/ifric003.htm>
- [11] Emission rights – worked examples, www.hm-treasury.gov.uk/d/emissions_worked_example.pdf
- [12] IASB: Discussion at the June 2005 IASB Meeting – Withdrawal of IFRIC 3, <http://www.iasplus.com/interps/ifric003.htm#withdraw>
- [13] IDW Stellungnahme zur Rechnungslegung: Bilanzierung von Schadstoffemissionsrechten nach HGB. IDW RS HFA 15
- [14] IETA: Trouble-Entry Accounting – Revisited. Uncertainty in Accounting for the Emission Trading Scheme and Certified Emission Reductions. International Emissions Trading Association, Geneva, 2007
- [15] IFRIC (International Financial Reporting Interpretations Committee): IFRIC interpretation 3: Emission rights. London: International Accounting Standards Board, 2004
- [16] Leibfried, P./Eisele, A.: Bilanzierung von Emissionsrechten nach IFRS und Swiss GAAP FER Der Schweizer Treuhänder, 2009/1-2
- [17] Lorson, P., Toebe, M.: Bilanzierungsfeld Emissionsrecht handel. *Zeitschrift für internationale und kapitalmarktorientierte Rechnungslegung KoR* 7-8/2008, pp. 498-510
- [18] Lovell, H., Sales de Aguiar, T., Bebbington, J., Larrinaga-Gonzalez, C.: Accounting for Carbon. Research report, Certified Accountants Educational Trust, London, 2010, 38 p
- [19] Reizinger-Ducsai, A.: Accounting for Emission Rights, *Periodica Polytechnica Social and Management Sciences*, 15/2 (2007)
- [20] Reizinger-Ducsai, A.: Managing Emission Rights in Financial Reports (in Hungarian) Budapest University of Technology and Economics, PhD Theses, 2011, pp. 66-71
- [21] Resolución de 8 de febrero de 2006, del Instituto de Contabilidad y Auditoría de Cuentas, www.boe.es/boe/dias/2006/02/22/pdfs/A07131-07135.pdf
- [22] Wambsganss, J. R., Stanford, B.: The Problem with Reporting Pollution Allowances. *Critical Perspectives on Accounting*, 1996, 7(6), pp. 643-652

Optimal Policy for the Replacement of Industrial Systems Subject to Technological Obsolescence – Using Genetic Algorithm

Mohamed Arezki Mellal¹, Smail Adjerid¹, Djamel Benazzouz¹, Sofiane Berrazouane², Edward J. Williams³

¹ LMSS, Faculty of Engineering Sciences (FSI), M'Hamed Bougara University, Avenue de l'Indépendance, 35000 Boumerdes, Algeria
mellal-mohamed@umbb.dz, adjerid_s@umbb.dz, dbenazzouz@umbb.dz

² Faculty of Engineering Sciences (FSI), M'Hamed Bougara University Avenue de l'Indépendance, 35000 Boumerdes, Algeria
berrazouane_so@umbb.dz

³ College of Business, Decision Sciences, University of Michigan
4901 Evergreen Road, Dearborn, 48126 Michigan, USA
williams@umd.umich.edu

Abstract: The technological obsolescence of industrial systems is characterized by the existence of challenger units possessing identical functionalities but with improved performance. This paper aims to define a new approach that makes it possible to obtain the optimal number of obsolete industrial systems which should be replaced by new-type units. This approach presents a new point of view compared with previous works available in the literature. The main idea and the originality of our approach is that we apply a genetic algorithm (GA) by considering the failure frequency, the influence of the environment/safety factors of the old-type systems and the purchase/implementation cost of the new-type units. These parameters are introduced in order to optimize this type of replacement in the context of engineering.

Keywords: technological obsolescence; industrial systems; replacement policy; failure frequency; safety/environment factors; genetic algorithm (GA)

1 Introduction

Often the behavior analysis of industrial systems in engineering is based on the study of monitoring and diagnostics, but technological obsolescence is neglected in the models. Nowadays, technological change is abrupt and the great majority of industrial systems are subject to obsolescence. An item becomes obsolete when a

new-type unit is available and performs the same functionalities but with improved performance. Hence, the necessity of an approach in order to deal with technological obsolescence is important for industrial firms. In most previous works devoted to the study of the dependability of industrial systems, the authors have not taken into consideration technological change.

In [1-4], several industrial plants were studied in order to achieve models of dependable installations, but the influence of technological obsolescence has not been considered in the approaches.

Technological improvement has an impact on the life-cycle of the industrial plants due to the unavailability of spare parts, and thus, if this problem is not considered, it will generate random and various consequences, such as accidents, stoppage of production, environmental disaster, etc.

The improvement in performance of the new technology items can be understood as smaller failure rates, lower pollution, more security, lower energy consumption, etc. At the same time, it is difficult to determine the optimal policy for the replacement of old technology units by new ones in the context of engineering (dependability study) and not only in the context of economics. On the other hand, it is economically more interesting for industrial firms to replace the old type units gradually to benefit from their residual lifetime. Most often, the authors studied the technological obsolescence under two contexts: the first one is strictly based on economic assumptions and other parameters are neglected (e.g., [5-7]). The second one aims to define an approach that takes into account various parameters of engineering and economic elements, such as failure frequencies, costs and strategy of maintenance, reliability, etc. We can cite [8-11].

The aim of our work is therefore to define a replacement policy of these obsolete industrial systems in the context of engineering and to help the decision maker find the optimal systems which should be replaced among them.

This paper summarizes and extends the works presented in previous papers. It is organized as follows: Section 2 describes a brief literature review of previous works, the assumptions on which they were based, and an overview. Section 3 illustrates our approach and the assumptions on which it is based, and we conclude this section with a case study and numerical results. Finally, we conclude this paper by suggesting some possible perspectives and extensions of our approach.

2 Thematic Review of the Literature

Several researchers have been studying the problem of technological obsolescence in industry from many points of view. The first paper was published by Elton and Gruber [5] in 1976. Their work considered one single component characterized by an annual income, purchase cost, resale value, which decreases with the age of the

component, and an aging factor, which reduces the income. Technological change generates efficiency which was given in the model by an increase of an annual income of the new-technology unit with a factor denoted g . The failures were not considered, but the age of the component was modeled by a linear decrease of the income factor, h , generated over time. The model was given as follows:

$$rT + e^{-rT} = 1 + \frac{r^2(I - S)}{(g + h) - rs} \quad (1)$$

where r is the discount rate for a period t , I is the purchase cost of the old-generation component, S is the purchase cost of the new-type unit, s is the resale value per time unit of the operating unit and T is the time interval for a component replacement. The authors of this work considered that the strategy which maximizes the income consists of replacing the component at regular interval T , where T is the solution of (1).

In [6], the authors considered one component subject to technological obsolescence. The model was proposed for a discretized time of replacement at the appearance of the new-type unit without taking into account the failure rates.

In [7], the authors considered a geometric technological change of several industrial systems and they solved a continuous-time optimization problem to define an optimal replacement policy by searching for the optimum of (3).

$$L(t) = \int_{d(t)}^t m(\tau) d\tau \quad (2)$$

$$\min I = \int_0^T e^{-rt} \left(\int_{d(t)}^t q(\tau, t) m(\tau) d\tau + p(t) m(t) \right) dt \quad (3)$$

where $L(t)$ is the number of systems in service, $a(t)$ is the installation time of the obsolete systems replaced at time t , $m(t)$ is the current investment (the number of new installed systems), $q(\tau, t)$ is the specific maintenance cost of the vintage τ at time t , $p(t)$ is the purchase price and installation cost of a new system, and r is the discounting factor, $r > 0$. Then $t - a(t)$ is the lifetime of the industrial plant (the age of the oldest system still in use). The constraint of (3) was given as follows:

$$0 \leq m(t) \leq M(t), \quad d(t) \leq t \quad (4)$$

where $M(t)$ is the number of the old-type units. The authors of this article neglected the failure rates and other paramount parameters.

In [8], the case of one single industrial item subject to aging and technological obsolescence was proposed in the model. The authors assumed that the time of the first failure of the component follows a Weibull distribution with two parameters. Several maintenances are undertaken at regular intervals and the repairs are considered. They assumed that the maintenance resets the component to the same

status at the beginning of the maintenance interval. The authors modeled the probability of failure of the component using a constant failure rate per part. The constant failure rate was given as follows:

$$\lambda^* = \left(\frac{1}{\alpha}\right)^\beta T^{\beta-1} \tag{5}$$

where α is the scale parameter, β is the shape parameter and $T^{\beta-1}$ is the maintenance interval. The authors assumed that the increase in this failure rate is due to the aging of the component and they formalized the issue of obsolescence in a quantifiable manner. To account for the various issues at stake, they postulated that as calendar time goes by new components are available on the market, and they are characterized by a failure rate which decreases exponentially. The authors concluded that this component in service can be either periodically maintained or preventively replaced by a new-type unit. They solved the problem by assessing the costs using Monte Carlo simulations.

In [11], a repairable system that operates continuously to the degraded state was studied. The model was presented by the following decisions about the interventions: do nothing, maintain or to replace by a new-type unit.

In [9, 10, 12, 13], the authors studied the following case: A set of N identical and independent industrial items.

The authors of [9] proposed that these components can be either preventively or correctively replaced by new-type units and the replacements take negligible time.

The works of Elmakis *et al.* [13] are characterized by this assumption: the failure rate λ_0 of each component is constant. The proposed approach in their paper is called the “ K strategy” (Fig. 1) and it is based as follows: first, new-type components are used only to replace failed old-type units; then, after K corrective actions of this kind, the $N - K$ old-type remaining components are preventively replaced by new-type ones at the time of the K^{th} corrective intervention.

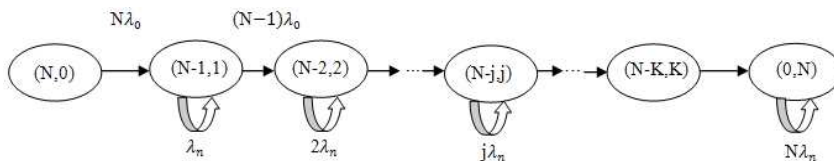


Figure 1
Diagram of K strategy

The “0” strategy represents the preventive replacement of all old-type components at the initial moment. To determine the value of K , the authors proposed a Monte Carlo simulation to assess the costs generated by each value of K .

In [10, 12, 14], the authors proposed extensions of the “ K strategy” by taking into consideration the failure frequencies as a Weibull law of this form.

In the contribution of the works presented in [10], the authors proposed a model for N identical components, but with several challengers. A probability of incompatibility was accounted to deal with the fact that the on-site implementation of new-type units could turn out to be problematic, and some replacements could not be immediately successful, as operators have no experience to rapidly implement the new-technology unit. In [15], a test case was performed using Petri nets to model the different replacement strategies proposed in [10].

All the works cited in this section proposed models without taking into account the influence of the old-system units on the environment and the safety criterion. These factors are recommended to be considered in the model, especially nowadays with the environmental problems and the industrial accidents.

3 Model Description

In this paper, we introduce a more realistic approach compared with the works illustrated in Section 2.

The case studied in our work is the following: A set of N different and independent obsolete industrial systems, one challenger per old-type unit is available, the industrial firm devotes a special budget to deal with technological obsolescence at the end of the year and the transition between the generations of the units will be done. To study the transition problem, we consider the following important data:

- Failure frequency per hour during the year of each old-type unit.
- An annual budget is intended to overcome the technical-economic impact of technological obsolescence.
- The purchase and implementation cost of each new-type unit (challenger) is fixed.
- We select only the compatible challenger for the replacement to avoid production delays.
- Each old-type unit is characterized by its environment and safety factors which vary in the scale $[0,1]$ (where a 100% non-polluting and secure system is assigned the value 1).

The data of each system are summarized as follows:

$$System_n \begin{cases} \lambda_n \\ C_n \\ E_n \\ S_n \end{cases} \quad (6)$$

where $System_n$ is the index of the system (for $n=1, \dots, N$), λ_n is the failure frequency per hour during the year, C_n is the purchase and implementation cost of the new-type type, E_n is the environment factor and S_n is the safety factor.

The aim of our work is to define a replacement policy for these obsolete systems in the context of dependability and to help the decision maker find an optimal strategy among them. We identify the optimal systems to be replaced by the new-type (challenger) but with these considerations: budget, optimal benefit from the residual lifetime of the old-type, and the environment/safety factors.

3.1 Genetic Algorithms Approach and Problem Formulation

To solve our optimization problem, we propose to develop an approach using a genetic algorithm (GA).

Genetic algorithms (GAs) are powerful bio-inspired algorithms that have been successfully used in several research problems: permutation flow shop [16], correcting the fine structure of surfaces [17], the synthesis of production-control systems [18], etc. The GA belongs to the soft computing technologies, who owe their name to their operational similarities with the biological and behavioral phenomena of living beings. Their primary target is the optimization of an assigned objective function (fitness).

GA was originally developed by Holland [19]. In general, genetic algorithms are based on the following steps [20]:

- (1) Creation of a random initial population of potential solutions to the problem and evaluation of these individuals in terms of their fitness, i.e. of their corresponding objective function values;
- (2) Selection of a pair of individuals as parents;
- (3) Crossover of the parents, with the generation of two children;
- (4) Replacement in the population, so as to maintain a constant population number;
- (5) Genetic mutation;
- (6) Repeat steps until satisfying solution is obtained.

The maximizing objective function (fitness) of our problem could be written as:

$$fitness = Max \sum_{n=1}^N \left[\lambda_n + \frac{1}{C_n} + \frac{1}{E_n} + \frac{1}{S_n} \right] \quad (7)$$

The fitness (7) is under the following constraint:

$$\sum_{n=1}^N C_n \leq Budget \quad (8)$$

The objective function (7) allows for identifying the optimal systems (high failure frequency, more polluting, less secure and lowest purchase cost of the challenger) recommended for the replacement by the new-type units. The solution is given under the constraint of the budget (8).

A binary coding is the most suitable in this case. The number of genes per individual (chromosome) is equal to the number of systems in order of appearance. Therefore, the size of the chromosome must be equal to N . Figure 2 shows an example of an eventual individual. A random initial population of potential solutions is given. If the system is marked in the individual, then it is assigned the value 1, otherwise 0. All systems which are assigned the value 1, their parameters will be implemented and evaluated in the fitness. A fixed number of iterations is stated after different times of run. During generations, our genetic algorithm seeks the optimal solution until convergence and stopping criterion.

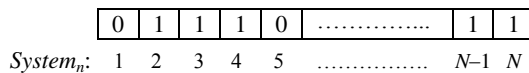


Figure 2

Example of an individual

In the individual shown in Fig. 2, the eventual solution is given by the replacement of $System_2$, $System_3$, $System_4, \dots, System_{N-1}$ and $System_N$. The parameters of these systems will be implemented and evaluated by the fitness.

Table 1

Comparison between main features of previous models and our approach

References	Context	Number of industrial systems	Environment and safety factors of the old-type units	Transition strategy
[5]	economics	1	–	replacement at fixed intervals
[7]	economics	1	–	replacement at fixed intervals
[6]	economics	1	–	replacement at fixed intervals
[8]	Engineering	1	–	replacement at the first failure
[13]	Engineering	N identical	–	K strategy
[9]	Engineering	N identical	–	K strategy
[14]	Engineering	N identical	–	K strategy
[10, 15]	Engineering	N identical	–	K strategy
[our model in this paper]	Engineering	N different	E_n, S_n	at the end of the year, according to the budget and the selection is made using genetic algorithms

After each generation, a new solution is given by the algorithm. The fitness function evaluates these solutions and they are ranked. This ranking is used in the selection procedure (standard roulette), which is performed in such a way that in the long run the best individuals will have a greater probability to be selected as parents, in resemblance to the natural principles of the “survival of the fittest”. Similarly, the ranking is used in the replacement procedures to decide who among the parents and the daughters should survive in the next population. An algorithm based on these procedures is often referred to as a steady-state GA [20].

We assume that there are a few random solutions beyond the budget; they will be kept to maintain diversity and to avoid stalling at local optima.

Table 1 summarizes the main features of different previous works and our proposition in this paper.

3.2 Case Study

In this subsection, we present a numerical application of our model. The case considered here is a set of ($N=14$) different industrial systems subject to technological change (obsolescence). The data of these systems are reported in Table 2 and ($Budget = 50 \times 10^2$ \$).

Table 2
Data of the systems

$System_n$	Failure frequency λ_n (hour ⁻¹ , during the year) $\times 10^{-6}$	Purchase and implementation cost of the challenger C_n (10^2 \$)	Safety factor S_n	Environment factor E_n
1	3.54	6.40	0.92	0.50
2	2.26	8.20	0.95	0.77
3	5.37	6.00	0.84	0.80
4	4.88	3.80	0.87	0.65
5	4.66	5.21	0.91	0.54
6	2.28	3.01	0.86	0.66
7	8.01	7.80	0.80	0.71
8	6.01	6.50	0.85	0.59
9	7.87	8.40	0.93	0.82
10	6.07	6.05	0.91	0.85
11	5.90	4.33	0.81	0.79
12	4.26	3.41	0.79	0.85
13	6.87	5.00	0.83	0.90
14	3.40	7.00	0.96	0.87

3.2.1 Problem Formulation

The fitness function of this case study can then be written as follows:

$$fitness = Max \sum_{n=1}^{14} \left[10^4 \times \lambda_n + 10 \times \frac{1}{C_n} + 10^{-2} \times \frac{1}{E_n} + 10^{-2} \times \frac{1}{S_n} \right] \quad (9)$$

where the weighting 10^4 , 10 and 10^{-2} are introduced to balance the fitness. This objective function tries to maximize the number of industrial systems which should be replaced by new-type units. The unidentified systems their residual lifetime will be exploited.

The fitness function (9) is subject to:

$$\sum_{n=1}^{N=14} C_n \leq 50 \times 10^2 \quad (10)$$

Table 2 contains the parameters related to the systems, whereas Table 3 contains the rules and the parameters for the GA implemented in order to solve the objective optimization problem. The values of the parameters were chosen after times of run to achieve a convergence.

Table 3
Genetic algorithm rules and parameters

GA property	Value
Number of genes for individual	14
Number of individuals (population size)	60
Number of generations (termination criterion)	500
Mutation probability	0.001
Selection technique	Standard Roulette

3.2.2 Results and Discussion

The results of the GA optimization process and the convergence of the fitness are shown in Fig. 3.

We remark that the convergence of the proposed algorithm is at 100 generations (see Fig. 3). The number of systems and their parameters has been controlled by the algorithm. As mentioned in Subsection 3.1, we assume that few random solutions exist beyond the budget, and they will be introduced during iterations to maintain diversity; hence, we obtained fitness values greater than the convergence value.

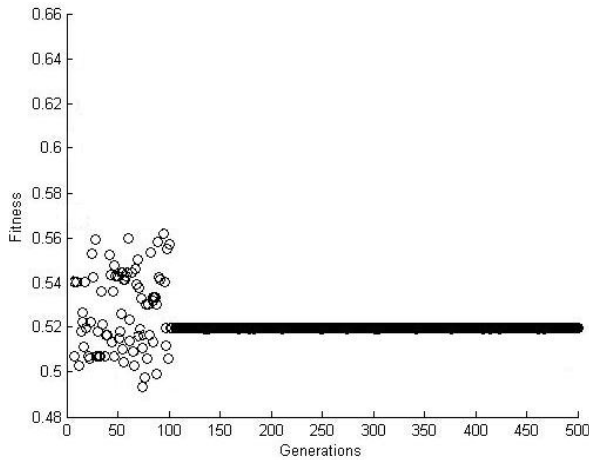


Figure 3

Result of the GA process for the optimal replacement strategy

The solution is identified in the individual represented in Fig. 4. All the systems assigned the value 1 are considered optimal for the replacement by the new-technology units.

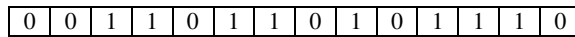


Figure 4

Individual of the solution

Table 4 illustrates the systems identified in the individual (see Fig. 4) of the optimal solution.

Table 4

Systems recommended for the replacement policy

<i>System_n</i>	
<i>System₃</i>	<i>System₉</i>
<i>System₄</i>	<i>System₁₁</i>
<i>System₆</i>	<i>System₁₂</i>
<i>System₇</i>	<i>System₁₃</i>

Conclusions

In this paper we proposed a model to deal with the obsolescence of industrial systems in the context of engineering. A genetic algorithm was elaborated for solving a case of several systems subject to technological obsolescence. The model was illustrated on a case study under the following considerations: environment/safety factors, failure frequencies of the old-types units and the purchase/implementation cost of the new-type units (challengers).

A major advantage of this model consists of the possibility to find an optimal replacement policy when we have a complex case study with many systems and several parameters. The difficulty persists in the choice of the parameter values of the algorithm and the program development.

Future work will need to define an optimal policy in the case of dependent systems, the impact of the new-challengers on the installation, and comparative results using other optimization methods.

References

- [1] M. A. Mellal, S. Adjerid, D. Benazzouz: Modeling and Simulation of Mechatronic System to Integrated Design of Supervision: Using a Bond Graph Approach, *Applied Mechanics and Materials*, Vol. 86, pp. 467-470, 2011
- [2] A. R. Conn, L. A. Deleris, J. R. M. Hosking, T. A. Thorstensen: A Simulation Model for Improving the Maintenance of High Cost Systems With Application to an Offshore Oil Installation, *Quality and Reliability Engineering International*, Vol. 26, No. 7, pp. 733-748, 2010
- [3] M. A. Mellal, S. Adjerid, D. Benazzouz: Modeling and Simulation of Mechatronic System to Integrated Design of Supervision: Using a Bond Graph Approach, in *Proceedings of ECMS'2011 European Council on Modelling and Simulation*, Krakow, Poland, pp. 370-373, 2011
- [4] E. P. Zafiropoulos, E. N. Dialynas: Reliability and Cost Optimization of Electronic Devices Considering the Component Failure Rate Uncertainty, *Reliability Engineering and System Safety*, Vol. 84, No. 3, pp. 271-284, 2004
- [5] E. J. Elton, M. J. Gruber: On the Optimality of an Equal Life Policy for Equipment Subject to Technological Improvement, *Operational Research*, Vol. 27, pp. 93-99, 1976
- [6] I. E. Schochetman, R. L. Smith: Infinite Horizon Optimality Criteria for Equipment Replacement Under Technological Change, *Operations Research Letters*, Vol. 35, No. 4, pp. 485-492, 2007
- [7] N. Hritonenko, Y. Yatsenko: Optimal Equipment Replacement Without Paradoxes: A Continuous Analysis, *Operations Research Letters*, Vol. 35, No. 2, pp. 245-250, 2007
- [8] E. Borgonovo, M. Marseguerra, E. Zio: A Monte Carlo Methodological Approach to Plant Availability Modeling with Maintenance-aging and Obsolescence, *Reliability Engineering and System Safety*, Vol. 67, No. 1, pp. 61-73, 2000
- [9] S. Mercier, P. E. Labeau: Optimal Replacement Policy for a Series System with Obsolescence, *Applied Stochastic Models in Business and Industry*, Vol. 20, No. 1, pp. 73-91, 2004

-
- [10] J. Clavareau, P. E. Labeau: Maintenance and Replacement Policies Under Technological Obsolescence, *Reliability Engineering and System Safety*, Vol. 94, No. 2, pp. 370-381, 2009
- [11] P. K. Nguyen Thi, T. G. Yeung, B. Castanier: Optimal Maintenance and Replacement Decisions Under Technological Change, in *Proceedings of ESREL'2010: European Safety and Reliability Conference*, Rhodes, Greece, 2010
- [12] O. Michel, P. E. Labeau, S. Mercier: Monte Carlo Optimization of the Replacement Strategy of Components Subject to Technological Obsolescence, in *Proceedings of the International Conference on Probabilistic Safety Assessment and Management*, Berlin, Germany, 2004
- [13] D. Elmakis, G. Leitin, A. Lisnianski: Optimal Scheduling for Replacement of Power System Equipment with New-type One, in *Proceedings of the 3rd International Conference on Mathematical Methods in Reliability*, Trondheim, Norway, 2002
- [14] S. Mercier: Optimal Replacement Policy for Obsolete Components with General Failure Rates, *Applied Stochastic Models in Business and Industry- Reliability*, Vol. 24, No. 3, pp. 221-235, 2008
- [15] J. Clavareau, P. E. Labeau: A Petri Net-based Modelling of Replacement Strategies Under Technological Obsolescence, *Reliability Engineering and System Safety*, Vol. 94, No. 2, pp. 357-369, 2009
- [16] K. Balazs, Z. Horvath, L. T. Koczy: Different Chromosome-based Evolutionary Approaches for the Permutation Flow Shop Problem, *Acta Polytechnica Hungarica*, Vol. 9, No. 2, pp. 115-138, 2012
- [17] G. Gyurecz, G. Renner: Correcting Fine Structure of Surfaces by Genetic Algorithm, *Acta Polytechnica Hungarica*, Vol. 8, No. 6, pp. 181-190, 2011
- [18] P. Y. Mok: Genetic Synthesis of Production-control Systems for Unreliable Manufacturing Systems with Variable Demands, *Computers and Industrial Engineering*, Vol. 61, No. 1, pp. 198-208, 2011
- [19] J. H. Holland: *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: University of Michigan Press, USA, 1975
- [20] S. Sumathi, P. Surekha: *Computational Intelligence Paradigms*, Taylor & Francis Group, London, United Kingdom, 12-15, 2010

Effect of Recycling on the Rheological, Mechanical and Optical Properties of Polycarbonate

Ferenc Ronkay

Department of Polymer Engineering
Budapest University of Technology and Economics
Műgyetem rkp. 3, H-1111 Budapest, Budapest, Hungary
e-mail: ronkay@pt.bme.hu

Abstract: The research was aimed at analyzing the polycarbonate scrap arising during production and its possible secondary utilization. The analysis of morphological, rheological and thermal data revealed significant differences between the original pellets and the reground material obtained from injection molded parts. Test specimens were injection molded from various mixtures of the virgin and the reground material, and their mechanical and physical properties were analyzed. Based on the results the reground material may be used in less than 20% proportion, as the mechanical properties of the products do not deviate significantly from those of the products made from virgin polycarbonate.

Keywords: polycarbonates; recycling; mechanical properties; optical properties; morphology

1 Introduction

In recent years, reprocessing of polymers has been widely used in plastics converting industries [1-3]. It is connected to the increasing awareness of environmental issues, to the desire to save resources, and to the high levels of scrap material generated during plastics conversion. There is a high demand for the recycling of scraps considering the relatively high cost of polymer production. To solve this problem the recycling of scrap material, and mixing it with virgin material, is the most common solution [4-5].

Polycarbonate (PC) is one of the important engineering plastics with a wide variety of applications due to the excellent mechanical properties, high impact strength, heat resistance and high modulus of elasticity, as well as due to its excellent balance of toughness, clarity, high thermal resistance and transparency

[6-7]. The recycling of this plastic material after the end of first life cycle has attracted attention recently [8-9].

Some physical and mechanical properties of PC can be severely reduced by recycling. Pérez et al. reported that after ten times of recycling the tensile strength reduced by 30% [10].

The rheological analysis of the dilute solution of macromolecular materials can give more information about the size and shape distribution of macromolecules. During the repetitive injection molding of polycarbonate, the molecular mass changes were studied by observing a rapid decrease in molecular weight, explained by two simultaneous degradation mechanisms [11].

In recent years the degradation of PC during accelerated aging tests has been studied by several researchers, among them the durability and a predictability of the properties to cover the whole lifecycle of the PC, as well as the degradation mechanism which occurs at the molecular level [12].

Long and Sokol studied the effect of moisture on the degradation of polycarbonate during injection molding [13]. It has been shown that even low moisture content during processing adversely affects the mechanical properties of the final product.

The effect of recycling on the properties of injection molded polycarbonate was studied by Shea and Nelson, who evaluated the extent of degradation by measuring melt flow rate, impact strength and molecular weight [14]. After ten times of recycling the value of melt flow rate increased five times.

Other works focused on the transparency of PC materials influenced by UV irradiation. The effect of UV irradiation was investigated on the structure and optical properties of polycarbonate material, and it was found that irradiation leads to a decrease of the optical energy gap of PC; they concluded that the decrease in optical energy gap could be due to the photo-degradation of PC and the formation of defects and clusters in the material [15].

The aim of the present study was to analyze the great amount of polycarbonate waste which arises during polymer processing and to find possibilities for secondary utilization. The aim was to determine the optimal rate of the recycled material as well, where the properties of the final product are still acceptable in respect of mechanical and optical properties, compared to the original polycarbonate product.

2 Experimental Work

2.1 Materials and Processing

Makrolon 1804 (Bayer) polycarbonate was used as virgin material in our study, and the reground obtained from the scrap of injection molded parts made of the same grade was used as recyclate. The ratio of the reground material to the original PC was changed from 0 to 100%.

Dumb-bell type test specimens were injection molded for the mechanical tests from the virgin material, from the reground recyclate and from various mixtures of the two. Injection molding was performed on an Arburg Allrounder 320C 600-250 injection molding machine. The zone temperatures were as follows from the feeding zone to the nozzle: 275/285/290/295/300°C. The mold temperature was 80°C, and the injection pressure was 1000 bar. In order to prevent hydrolytic degradation, the materials were dried before melting at 120°C for 4 hours.

2.2 Characterization Methods

Viscometric parameter determination was carried out at $25\pm 0.1^\circ\text{C}$, in chloroform solution, using an Ubbelohde 0B viscosimeter. For viscosity-average molecular weight determination, constants $K=0.012\text{ cm}^3/\text{g}$ and $a=0.82$ were employed [16-17].

Melt flow rate was measured on CEAST Melt Flow Modular Line equipment at 270°C with 2.16 kg load. The materials were dried before measurements at 120°C for different periods between 0 and 220 minutes.

In order to determine the glass transition temperature, DMA tests were made using Perkin Elmer DMA 7 type equipment in displacement-controlled mode, with 10 μm amplitude, at 1 Hz frequency. Three point bending mode was used for excitation.

Tensile tests were performed according to the EN ISO 527 standard using a ZWICK Z020 type universal tensile tester at a deformation rate of 20 mm/min, at room temperature.

Charpy impact tests were performed using CEAST Resil Impactor Junior type equipment with 15 J impact energy, a 20° starting angle and 0.589 m/s impact velocity.

Transmission optical tests were made with a JASCO V-530 UV/VIS type spectrometer. Prior to these tests, the surface of the test specimens was polished to an average of 1 μm roughness, using a Struers type polishing machine.

3 Results and Discussion

3.1 Changes of the Average Molecular Mass

Changes of the viscosity average molecular mass at various levels of processing are shown in Fig. 1. It can be observed that the virgin pellet exhibits the highest average molecular mass (18,000 g/mol); that of the reground is smaller (16,500 g/mol).

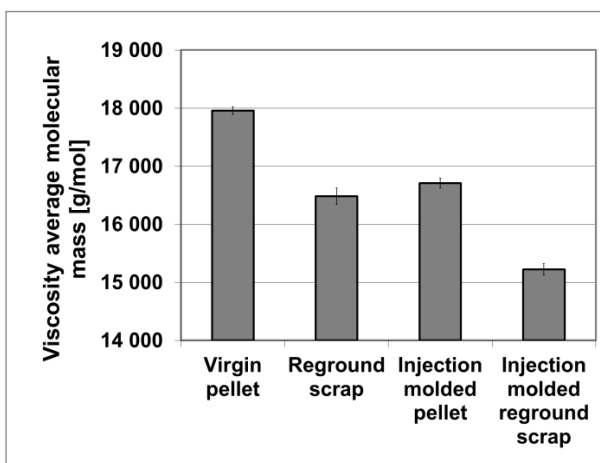


Figure 1

Viscosity average molecular mass of polycarbonates processed and reprocessed to various degrees

It is caused by the strong degrading effect of the shear forces and by the thermal impact encountered during injection molding. The average molecular mass of the test specimens injection molded from the virgin pellets was found to be similar (16,700 g/mol), the slight difference might be due the minor differences between the injection molding parameters used for preparing the test specimens and the parts. The lowest value (15,200 g/mol) was measured on test specimens injection molded from the regrind. By now the material has undergone two processing cycles, so the molecular chains are degraded to a higher degree. It can be established that the average molecular mass of the processed material decreases by about 8%, and that of reprocessed material by about 15%.

3.2 Changes in Melt Viscosity

The melt flow rates of the virgin pellet and of the recycled regrind were studied as a function of the drying time. The results are shown in Fig. 2. It can be observed that the flow rate of the recycled material without drying is higher (recyclate: 21.8 cm³/10 min; virgin: 17.5 cm³/10 min), and this difference persists after drying.

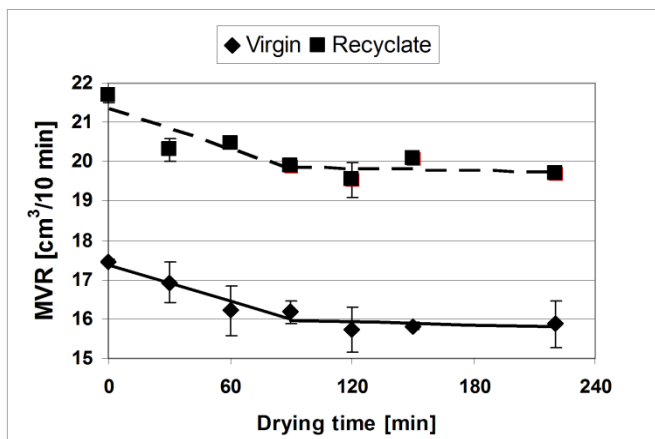


Figure 2

Volumetric melt flow rate of the virgin pellets and of the reground recyclate as a function of the drying time (drying temperature: 120°C)

The melt flow rate of both materials stabilized and became constant after about two hours of drying (recyclate: 19.8 cm³/10 min; virgin: 15.8 cm³/10 min), i.e. they exhibited similar behavior during drying. The difference can be explained by the degradation of the molecular chain: in the reground recyclate, shorter molecular chains can be found, which hinder the melt flow to a lesser degree. The melt flow rate of the reground recyclate is about 25% higher than that of the virgin material.

3.3 Changes in the Thermal Properties

The glass transition determined by DMA is ascribed to the maximum temperature of the mechanical loss (see Fig. 3). It can be observed that the glass transition temperature of the virgin material (135.8°C) is 1.1°C higher than that of the recycled material (134.7°C).

This shift can be explained by the shortening of the molecular chains: the movement of shorter chains' segments starts at lower temperature.

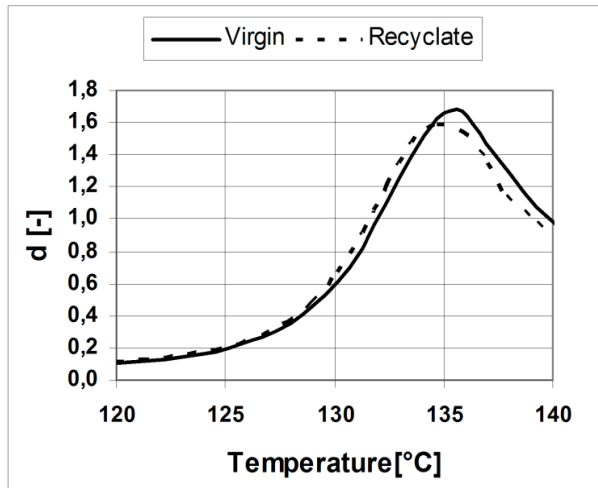


Figure 3
Change of the mechanical loss factor

3.4 Changes in the Mechanical Properties

3.4.1 Changes in the Tensile Strength

The dependence of the tensile strength on the amount of recycled material is shown in Fig. 4. Analyzing the plot it can be concluded that the value of the tensile strength increases slightly with the recycled material content. This increase is not significant: only 1-2% with respect to the virgin material.

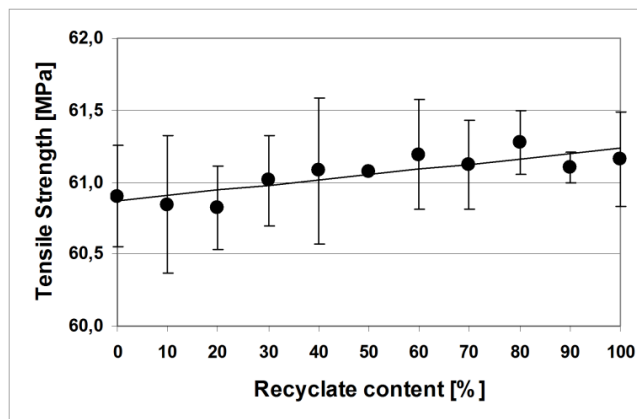


Figure 4
The tensile strength of the material as a function of the recyclate content

This trend seems to contradict somewhat earlier literature findings [18], according to which a decreasing molecular mass results in decreasing strength (according to equation 1).

$$\sigma_{recycled} = \sigma_{virgin} \left(1 - \frac{M_{difference}}{M_{virgin}}\right) \quad (1)$$

where $\sigma_{recycled}$ is the expected strength of the recycled material, σ_{virgin} is the strength of the original (virgin) material, $M_{difference}$ is the decrease of the average molecular mass, and M_{virgin} is the average molecular mass of the original (virgin) material.

Using equation (1) in our case, one would expect a 9% decrease between the strength of the recycled material and of the virgin material (equation 2).

$$\sigma_{recycled} = 60,9MPa \left(1 - \frac{16700 - 15200}{16700}\right) = 60,9MPa * 0,91 = 55,4MPa \quad (2)$$

The slight strength improvement observed in the tests may be due to the changes in the orientation of the amorphous macromolecular chains. Shorter chains may orient easier during injection molding along their long axis; thus they can bear more load during the tensile test. The prerequisite of this slight improvement is that the material be absolutely free of all contaminations, as even a small amount of contamination would serve as a defect site, which would decrease the strength.

3.4.2 Changes in the Tensile Modulus

The tendency of the change of the tensile modulus is similar that of the tensile strength (see Fig. 5).

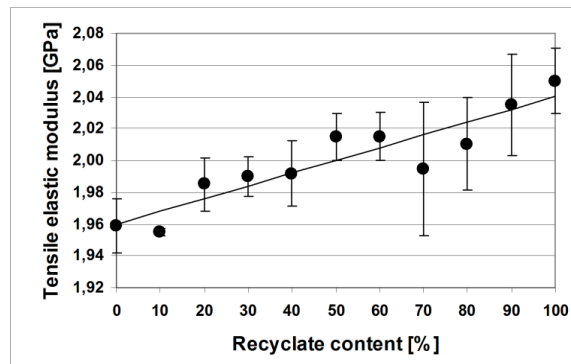


Figure 5

The tensile modulus of the material as a function of the recycle content

If compared to the modulus of the virgin material (1.96 GPa), the modulus of the recycle increased slightly, by 4.5%. It can be established that the rigidity of the material increases with the recycle content, but the improvement is insignificant.

3.4.3 Changes in the Elongation at Break Values

Elongation at break values measured during the tensile tests are shown in Fig. 6.

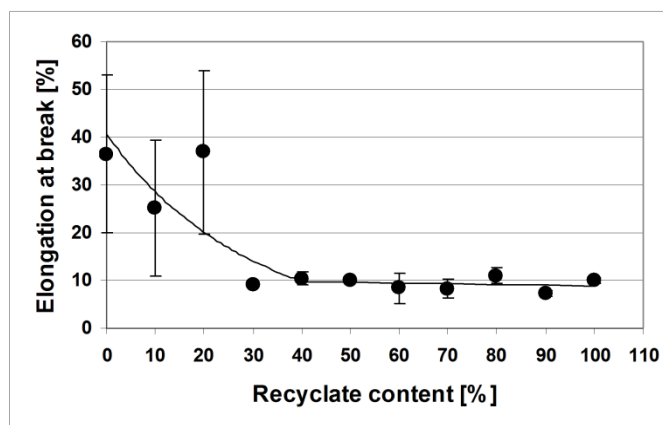


Figure 6

Change of the elongation at break as a function of the recycle content

The elongation at break decreased significantly with up to 30% recycle content. The decrease was about 75% with respect to the virgin material. The elongation at break is very sensitive to the change of the average molecular mass. The elongation at break of the virgin material and of the samples containing 10% or 20% recycle is 25-40% of the original length, although the results exhibit fairly large scatter. In the range of 30-100% recycle content the elongation at break is only 10% of the original length. In this range there is no significant change and the scatter around the average is also smaller.

The large drop in the elongation at break in the samples containing 30-100% recycle affects the quality and usefulness of the produced parts (e.g. snap-fit closures), and therefore it is not recommended to use more than 20% reground recycle.

3.4.4 Changes in the Impact Strength

Values of the impact strength calculated from the flexural impact test are shown in Fig. 7.

Decreasing impact strength values can be observed with increasing recycle content. The impact strength of the test specimens injection molded from pure recycle is 12% lower than that of the test specimen produced from the virgin material. Test results agree with the elastic modulus measured in the tensile test: with increasing recycle content the material becomes stiffer, and its ductility decreases.

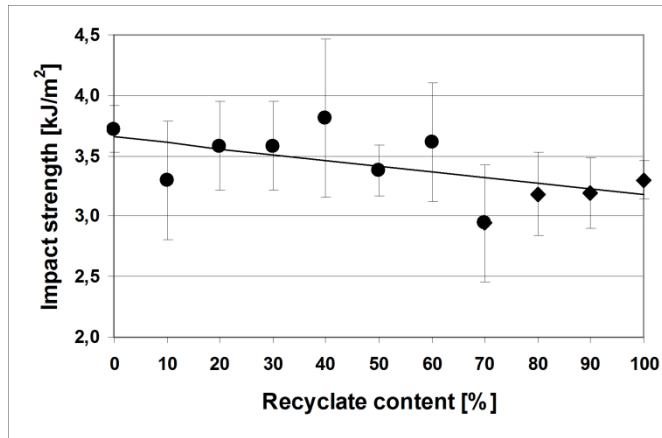


Figure 7

Impact strength as a function of the recycle content

3.5 Changes in the Optical Properties

Light transmittance of the virgin and recycled material is shown in Fig. 8. The transmittance curves exhibit a similar character in both cases: in the 380-408 nm range (belonging to the violet color) the materials transmit less light, but in the 408-760 nm range they transmit 80-90% of the light, so they can be regarded as transparent. It can be observed that in the 380-408 nm range the transmittance of the virgin and of the recycled material differs significantly: the original transmits better short wavelength (violet) rays.

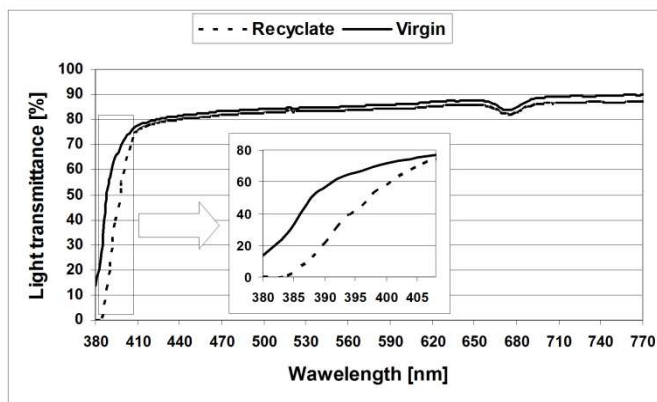


Figure 8

Light transmittance of the virgin and of the recycled test specimens as a function of the wavelength

If the full visible light spectrum (white light) is transmitted through a transparent material, the complementary color of the absorbed color will be amplified. The complementary color of violet is yellow; i.e., if the material absorbs more violet, it will appear more yellow to the human eye. The irradiation absorbance of polymers changes with the molecular weight reduction, not only in the UV range but also in the visible light range. In the case of polycarbonate, decreasing molecular mass causes higher UV absorption, which is presumably related to the increased number of end-groups. This phenomenon is similar to the way in which the photo-degradation of PC occurs, whereby the absorption varies in the same way in the range of wavelength between 250-400 nm [15]. Based on our measurements, 10-20% recyclate content does not deteriorate too strongly the UV transmission; above this concentration, however, the changing transmittance may cause distortion if the material is colored (the ratio of transmitted light is 57% at 290 nm in the case of original PET; 22% in the case of recycled PET; and 46% in the case of original PET with 20% recyclate content).

Conclusions

The degradation of polycarbonate during processing and its effects on the mechanical and optical properties of the material have been studied. It has been shown that the average molecular mass of polycarbonate decreases by about 8% during the first injection moulding and the subsequent grinding. Based on our test results, this 8% decrease in the average molecular mass causes about a 25% increase in the melt flow rate.

Changes in the mechanical properties were monitored by measuring the tensile and flexural impact properties. Test specimens were injection molded from various mixtures of the virgin pellets and reground material, using 10% steps. It has been established that the tensile strength and the tensile elastic modulus does not change too greatly, but the elongation at break and the impact strength values decrease significantly. Based on these findings, one can say that the use of more than 20% reground recyclate results in significant deterioration of the mechanical properties (especially of the impact strength) of the material.

In the optical studies, the transmittance of mixtures of various composition were studied in the full visible frequency range. Significant differences were found only in the first half of the violet range, where the absorbance of the recycled material is higher than that of the virgin material. The absorption of the violet light from the whole visible spectrum renders the material yellow for the human eye. Based on the study, a recyclate content above 10% causes a detectable difference in the violet absorption, although the color difference could not be detected by the naked eye between specimens made of various mixtures of the virgin and of the recycled material.

Based on these results, it can be concluded that the admixture of more than 20% reground recyclate may deteriorate the mechanical and optical properties of the product significantly.

Acknowledgement

This work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the New Széchenyi Plan (Project ID:TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

References

- [1] Al-Salem S. M., Lettieri P., Baeyens J. Recycling and Recovery Routes of Plastic Solid Waste (PSW): A Review. *Waste Manage* 2009; 29:2625-2643
- [2] Pegoretti A., Kolarik J., Slouf M. Phase Structure and Tensile Creep of Recycled Poly(Ethylene Terephthalate)/Short Glass Fibers/Impact Modifier Ternary Composites. *Express Polym Lett* 2009; 3:235-244
- [3] Huiting S., Pugh R. J., Forssberg E. A Review of Plastics Waste Recycling and the Flotation of Plastics. *Conserv Recy* 1999; 25:85-109
- [4] Saraiva Sanchez E. M. Ageing of PC/PBT Blend: Mechanical Properties and Recycling Possibility. *Polym Test* 2007; 26:378-387
- [5] Eguiazabal J. I., Nazabal J. Effect of Reprocessing on the Properties of Bisphenol-A Polycarbonate. *Eur Polym J* 1989; 25:891-893
- [6] Krawczak P. Plastics' Key Role in Energy-Efficient Building. *Express Polym Lett* 2009; 3:752
- [7] Cao K., Ma X., Zhang B., Wang Y., Wang Y. Tensile Behavior of Polycarbonate over a Wide Range of Strain Rates. *Mat Sci Eng A-Struc.* 2010; 527:4056-4060
- [8] Kahlen S., Wallner G. M., Lang R. W. Aging Behavior and Lifetime Modeling for Polycarbonate. *Sol. Energy* 2010; 84:755-759
- [9] Krivtsov V., Wagner P. A., Dacombe P., Gilgen P. W., Haven S., Hilty L. M. Analysis of Energy Footprints Associated with Recycling of Glass and Plastic – Case Studies for Industrial Ecology. *Ecol. Model.* 2004; 174:175-189
- [10] Pérez J. M., Vilas J. L., Lazaa J. M., Arnáizb S., Mijangosa F., Bilbaoc E., Rodrígueza M., León L. M. Effect of Reprocessing and Accelerated Ageing on Thermal and Mechanical Polycarbonate Properties. *J. Mater. Process. Tech.* 2010; 210:727-733
- [11] Glockner G. Polycarbonate Degradation under Processing Conditions. *Plaste und Kautschuk*, 1968; 15:632-635
- [12] Weibin G., Shimin H., Minjiao Y., Long J., Dan Y. The Effects of Hydrothermal Aging on Properties and Structure of Bisphenol A Polycarbonate. *Polym. Degrad. Stabil.* 2009; 94:13-17

- [13] Long T. S., Sokol R. J. Molding Polycarbonate: Moisture Degradation Effect on Physical and Chemical Properties. *Polym. Eng. Sci.* 1974; 14:817-822
- [14] Shea J. W., Nelson E. D., Cammons R. R. Effect of Recycling on the Properties of Injection Molded Polycarbonate. *Techn. Pap.-Soc. Plast. Eng.* 1975; 21:614-617
- [15] Migahed M. D., Zidan H. M. Influence of UV-Irradiation on the Structure and Optical Properties of Polycarbonate Films. *Curr. Appl. Phys.* 2006; 6:91-96
- [16] de Melo N. S., Weber R. P., Miguez Suarez J. S. Toughness Behavior of Gamma-irradiated Polycarbonate. *Polym. Test.* 2007; 26:315-322
- [17] Lia C., Zhang Y., Zhanga Y., Zhangb C. Blends of Polycarbonate and Ethylene-1-Octylene Copolymer. *Eur. Polym. J.* 2003; 39:305-311
- [18] Lawrence E. N., Robert F. L. "Mechanical Properties of Polymers and Composites" Marcel Dekker, New York 1994

The Shortest Path Planning for Manoeuvres of UAV

**Xian-Zhong Gao, Zhong-Xi Hou, Xiong-Feng Zhu,
Jun-Tao Zhang, Xiao-Qian Chen**

College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, 410073, P. R. China

E-mail: gaoxianzhong@nudt.edu.cn; hzx@nudt.edu.cn;
zhuxiongfeng@nudt.edu.cn; zhangjuntao@nudt.edu.cn;
chenxiaolian@nudt.edu.cn

Abstract: It is important to find the shortest path for manoeuvres of UAV, since the power consumed during manoeuvres is tightly coupled with the length of the flight path. In this paper, an algorithm that can find the shortest path during manoeuvres and improve the performance of UAV to follow waypoints is described. The shortest path for UAV during manoeuvres is derived firstly by the theory of Dubins curve. Secondly, in order to improve the ability of UAV to follow the derived optimal path, a real-time path planning algorithm is designed by transforming the constraints of Dubins curve into a dynamic equation. To demonstrate the applicability and performance of the proposed path planning algorithm, two numerical examples are presented. The results show that the proposed algorithm is promising to be applied in the path planning for manoeuvres of UAV.

Keywords: UAV; The shortest path; Path planning algorithm; Dubins curve set

1 Introduction

Nowadays, UAVs have been increasingly used in many applications, especially to replace the human presence in repetitive or dangerous missions [1], e.g., in environmental monitoring, security, military surveillance, crop and forest assessments, and so on [2].

A low-cost UAV in these missions must provide coverage of a certain region and investigate events of interested waypoints, so central for the development of UAV technology are the algorithms for the path planning and tracking [1]. It is important to find the shortest path for manoeuvres of a UAV, since the power consumed during manoeuvres is tightly coupled with the length of the flight path, which is determined by the planned path. Thus, it can be expected that the performance of a UAV may greatly benefit from the development of a path planning and tracking algorithm [3].

The problem of how to find the shortest path between two oriented points was first studied by Dubins [4]. Because it widely exists in applications, great attention was paid to this topic once it was proposed. Recently, variations of problems on this topic have been studied in literature. The problem is generally formulated as how to optimize the coverage costs, such as time [5, 6] or distance [3, 7] with the assumption that the location of targets is known [2]. In these cases, the manoeuvres of the aircraft lead by mission can be treated as a motion in a 2-D plane. The research results can be mainly categorized into two classes. The one is to classify Dubins curves, and the aim of real-time path planning is achieved by judging the initial and final states [8]. The other is to extend the problem proposed by Dubins to how to solve the shortest path when the robot can move forward and backward [9] and the UAV is impacted by the wind [10].

However, the problem studied by aforementioned papers is with the assumption that the orientation of the final point is fixed. In real applications, the circumstance that the orientation of the final point is unfixed is also general. In this paper, the method to solve the shortest path for the unfixed case is derived based on the conclusion of Dubins. In order to improve the ability of the UAV to follow the calculated optimal path, a real-time path planning algorithm is also designed.

The rest of this paper is organized as follows: In Section 2 the problem considered in this paper is formulated. A brief interpretation about the bounded curvature path (BCP) problem and the Dubins curves set is given in Section 3. The method to calculate the shortest path about the formulated problem is derived in Section 4. One new real-time path planning algorithm based on the results of Section 4 is developed in Section 5. The performance of the designed real-time path planning algorithm is analyzed and the numerical examples are carried out in different distributions of the waypoints in Section 6. Finally, the conclusions are given at the end of the paper.

2 Problem Formulation

The problem considered here can be stated as the following: given two oriented points (x_i, y_i, θ_i) and (x_f, y_f, θ_f) in the plane (x and y are the coordinates and the θ is the orientation), determine and compute the shortest piecewise paths joining them, along which the curvature is bounded everywhere by a given constant ρ_{\min} , which represents the manoeuvrability of aircraft.

If θ_f is fixed, this problem can be solved by the minimum principle of Pontryagin[9], and the results can be summarized in a Dubins curves set[8], which will be further explained in the next Section; If θ_f is unfixed, to our best knowledge the solution is still open. However, the later circumstance is always met in the path planning of UAVs, since the manoeuvres are constrained by admissible angles $[\theta_{\min}, \theta_{\max}]$ when flying along a path with multi-waypoints [2, 11], as shown in Fig. 1.

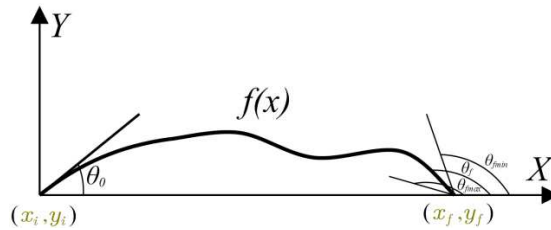


Figure 1

Schematic diagram of problem when θ_f is unfixed

The problem can be formulated as the following when the θ_f is unfixed:

$$\begin{aligned}
 \text{find } [f(x)] \quad \min : \mathbf{J}[f(x)] &= \int_{x_i}^{x_f} \sqrt{1 + f'^2(x)} dx \\
 \text{s.t. } \left\{ \begin{array}{l} f(x_i) = y_i, f(x_f) = y_f \\ f'(x_i) = \tan \theta_i, \tan \theta_{f \min} \leq f'(x_f) \leq \tan \theta_{f \max} \\ \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{3/2}} \leq \frac{1}{\rho_{\min}} \end{array} \right. & \quad (1)
 \end{aligned}$$

Because the initial orientation θ_i can be any angle in 2D plane and the UAV can be situated at any position, it is not convenient to discuss the method to solve the optimization problem formulated in Eq. (11). For the sake of clarity, all the possible cases are divided into sixteen categories [12], as listed in Table 1. Only the case that initial point is on the left side of final point is considered here, i.e. the case of I-LP. The results of the remaining cases can be obtained by a similar method.

Table 1
The classification of distributions of initial point and final point

	Long Path $x_f - x_i > 4\rho_{\min}$	Medium Path $2\rho_{\min} < x_f - x_i \leq 4\rho_{\min}$	Short Path $\rho_{\min} < x_f - x_i \leq 2\rho_{\min}$	Very Short Path $0 < x_f - x_i \leq \rho_{\min}$
Quadrant I $0 \leq \theta_0 < \pi/2$	I-LP	I-MP	I-SP	I-VSP
Quadrant II $\pi/2 \leq \theta_0 < \pi$	II-LP	II-MP	II-SP	II-VSP
Quadrant III $\pi \leq \theta_0 < 3\pi/2$	III-LP	III-MP	III-SP	III-VSP
Quadrant IV $3\pi/2 \leq \theta_0 < 2\pi$	IV-LP	IV-MP	IV-SP	IV-VSP

3 Bounded Curvature Path and Dubins Curves Set

3.1 Bounded Curvature Path

In order to solve the optimization problem formulated in Eq. (11), a preliminary problem should firstly be investigated. The preliminary problem can be formulated to find the shortest path from all the curves in the 2D plane, which pass initial point (x_i, y_i) and final point (x_f, y_f) with initial orientation θ_i and final orientation θ_f , and are subjected to minimal curvature radius ρ_{\min} , which is called as the problem of Bounded Curvature Path (BCP)[13]. A typical problem of BCP can be illustrated in Fig. 2:

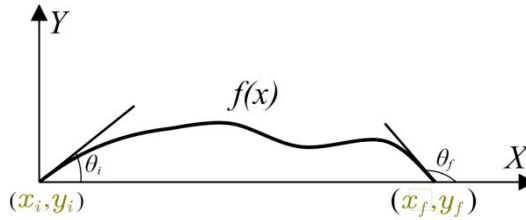


Figure 2

Schematic diagram of a typical bounded curvature path

For the problem of BCP, the mathematical formulation can be given as follows:

$$\begin{aligned}
 \text{find } [f(x)] \quad \min : \mathbf{J}[f(x)] &= \int_{x_i}^{x_f} \sqrt{1 + f'^2(x)} dx \\
 \text{s.t. } \begin{cases} f(x_i) = y_i, f(x_f) = y_f \\ f'(x_i) = \tan \theta_i, f'(x_f) = \tan \theta_f \\ \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{3/2}} \leq \frac{1}{\rho_{\min}} \end{cases} & \quad (2)
 \end{aligned}$$

3.2 Dubins Curves Set

The theoretical shortest path for BCP problems formulated above was firstly studied by L. E. Dubins in 1957 [4]. It is proved that for the problem presented in Section 3.1, the solution can be found among a finite set of curves. The set of curves consists of six elements, which are usually called Dubins curves. The Dubins curves set can be presented as [14]:

$$\mathbf{D} = \{LSL, RSR, RSL, LSR, RLR, LRL\} \quad (3)$$

where S represents a straight line segment, L denotes a circular arc to the left, and R is a circular arc to the right. The radius of of L and R arcs are exactly ρ_{\min} .

According to Dubins' result, the shortest path of BCP problem can be obtained by selecting the curve in the Dubins curves set with the shortest path length. Taking the BCP problem in Fig. 2, for example, by using the Dubins method, the shortest path can be obtained and is plotted in Fig. 3.

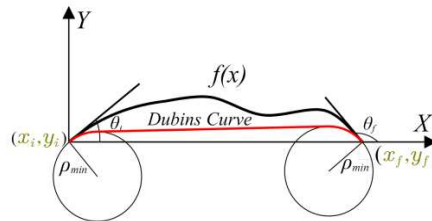


Figure 3
Schematic diagram of Dubins curves

4 Method to Find the Shortest Path

To solve the problem formulated in Eq. (11), the following theorem is given:

[THEOREM 1]:

For all the curves passing through initial point (x_i, y_i) and final point (x_f, y_f) with initial orientation θ_i and subjected to minimal curvature radius ρ_{\min} , if the final orientation θ_f is not fixed, as formulated in Eq. (11), $\mathbf{J}[f(x)]$ achieves the minimum when θ_f satisfies the following equation:

$$\theta_f^* = \arctan \frac{y_f - y_i^R}{x_f - x_i^R} - \arcsin \frac{\rho_{\min}}{\sqrt{(x_f - x_i^R)^2 + (y_f - y_i^R)^2}} \quad (4)$$

where (x_i^R, y_i^R) is the coordinate of the center of right circle, which crosses the initial point and is tangent with the vector of initial orientation.

[PROOF]

As shown in Fig. 4, the symbols of (x_i^R, y_i^R) and (x_f^R, y_f^R) are denoted as the coordinates of the centers of right circle, which cross the initial point and final point respectively, and are tangent with the vector of the initial and final orientation.

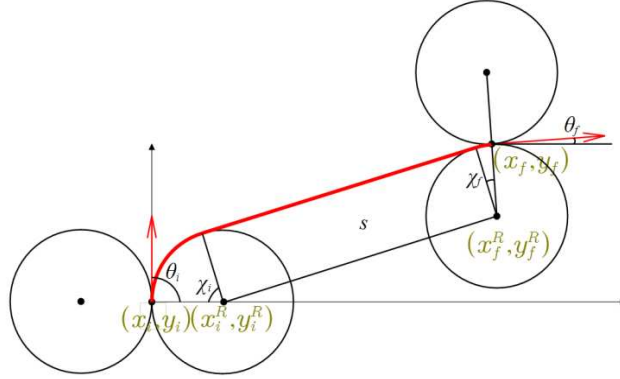


Figure 4

The centers of right circles which across initial and final point

From the conclusions of Dubins Curves Set, as described in Section 3.2, it can be derived that the shortest path in Fig. 4 is formed by the element of *RSR*. The geometry relationship shows that

$$\begin{cases} x_i^R = x_i + \rho_{\min} \sin \theta_i \\ y_i^R = y_i - \rho_{\min} \cos \theta_i \end{cases} \quad \begin{cases} x_f^R = x_f + \rho_{\min} \sin \theta_f \\ y_f^R = y_f - \rho_{\min} \cos \theta_f \end{cases} \quad (5)$$

The total length of path is

$$l = s + (\chi_i + \chi_f) \rho_{\min} \quad (6)$$

where the symbols of χ_i and χ_f represent the central angle of arc corresponding to initial point and final point respectively. s is the length of the straight line segment, which can be expressed as

$$s = \sqrt{(x_f^R - x_i^R)^2 + (y_f^R - y_i^R)^2} \quad (7)$$

The following equation can be derived from the geometry relationship

$$\chi_i + \chi_f = \theta_i - \theta_f \quad (8)$$

By substituting (55)(77) and (88) into (66), there is

$$l = \sqrt{(x_f + \rho_{\min} \sin \theta_f - x_i^R)^2 + (y_f - \rho_{\min} \cos \theta_f - y_i^R)^2} + (\theta_i - \theta_f) \rho_{\min} \quad (9)$$

The derivative of l with respect to θ_f can be expressed as follows

$$\frac{dl}{d\theta_f} = \frac{\rho_{\min}}{s} \left[(x_f + \rho_{\min} \sin \theta_f - x_i^R) \cos \theta_f + (y_f - \rho_{\min} \cos \theta_f - y_i^R) \sin \theta_f \right] - \rho_{\min} \quad (10)$$

Setting $\frac{dl}{d\theta_f} = 0$, and squaring both sides

$$\left[(x_f + \rho_{\min} \sin \theta_f - x_i^R) \cos \theta_f + (y_f - \rho_{\min} \cos \theta_f - y_i^R) \sin \theta_f \right]^2 = s^2 \quad (11)$$

Substituting (99) into (1111)

$$\begin{aligned} & \left[(x_f + \rho_{\min} \sin \theta_f - x_i^R) \cos \theta_f + (y_f - \rho_{\min} \cos \theta_f - y_i^R) \sin \theta_f \right]^2 \\ & = (x_f + \rho_{\min} \sin \theta_f - x_i^R)^2 + (y_f - \rho_{\min} \cos \theta_f - y_i^R)^2 \end{aligned} \quad (12)$$

Rearranging and simplifying (1212), the following expression can be obtained

$$(x_f - x_i^R) \sin \theta_f - (y_f - y_i^R) \cos \theta_f = -\rho_{\min} \quad (13)$$

Then the θ_f^* can be given as:

$$\theta_f^* = \arctan \frac{y_f - y_i^R}{x_f - x_i^R} - \arcsin \frac{\rho_{\min}}{\sqrt{(x_f - x_i^R)^2 + (y_f - y_i^R)^2}} \quad (14)$$

Thus the proof is complete.

By investigating the theorem, three remarks can be concluded:

[REMARK1]:

It can be seen from Eq. (1414) that the optimal final angle θ_f^* is only determined by the coordinate of right circle center of initial point (x_i^R, y_i^R) , the coordinate of final point (x_f, y_f) and the minimal curvature radius ρ_{\min} . Denoting:

$$\theta_f^a = \arctan \frac{y_f - y_i^R}{x_f - x_i^R} \quad \theta_f^b = \arcsin \frac{\rho_{\min}}{\sqrt{(x_f - x_i^R)^2 + (y_f - y_i^R)^2}} \quad (15)$$

As can be seen from Fig. 5, the geometric meaning of θ_f^a and θ_f^b are obvious. θ_f^a is the angle between the line d and the horizontal axis of coordinates x , where d is the line connected with center of right circle of initial point (x_i^R, y_i^R) and final point (x_f, y_f) . θ_f^b is the angle between the line s and the line d , where s is the straight line of path. It thus can be concluded that the optimal solution in Fig. 4 is a degenerated Dubins curve of RS, which is composed of an arc in the right circle of the initial point and a straight line segment. Substituting Eq. (44) into Eq. (99), the length of the shortest path can be calculated.

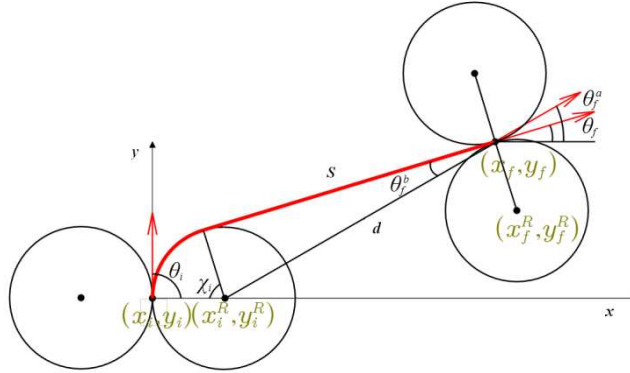


Figure 5

The angles composed of the optimal final angle

[REMARK2]:

It can be found from the proof of Theorem1 that the supposed final orientation θ_f is smaller than the optimal final angle θ_f^* , which indicates that the supposed shortest path is formed by the element of *RSR*; on the contrary, if the supposed final orientation θ_f is greater than the optimal final angle θ_f^* , as shown in Fig. 6, the supposed shortest path will be formed by the element of *RSL*, in which the same result can be obtained by the same method discussed above. Therefore, θ_f^* can be computed by Eq. (44), whatever the supposed θ_f is smaller or greater than θ_f^* .

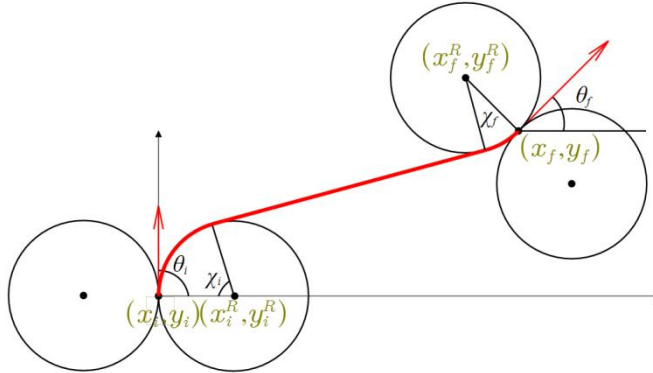


Figure 6

The case that θ_f is greater than θ_f^*

[REMARK3]:

In the discussion in Remark 1, the initial point is on the left side of final point, and thus the shortest path is composed by *RSC*; if the initial point is on the right side of

final point, the shortest path will be composed by *LSC*, in which case Eq. (44) must be changed into Eq. (16):

$$\theta_f^* = \arctan \frac{y_f - y_i^L}{x_f - x_i^L} - \arcsin \frac{\rho_{\min}}{\sqrt{(x_f - x_i^L)^2 + (y_f - y_i^L)^2}} \quad (16)$$

The proof method of Eq. (1616) is similar to that of Eq. (44), and thus is omitted here for the sake of clarity.

In the discussion in REMARK1 and REMARK3, θ_f is not subjected to any other constraints in 2D plane. In the following, a more general case is taken into account, in which θ_f lies in the interval $(\theta_{f\min}, \theta_{f\max}]$, where $-\pi < \theta_{f\min} < \theta_{f\max} < \pi$. Combining the result of Dubins and above discussion, the method to solve Eq. (11) can be concluded as follows:

[Method to solve the problem in Eq. (11)]

For all the curves passing initial point (x_i, y_i) and final point (x_f, y_f) with initial orientation θ_i and subjected to minimal curvature radius ρ_{\min} , if the final orientation θ_f is not fixed, as formulated in Eq. (11), the optimal final orientation θ_f^* can be calculated as in the following steps:

Step1:

Supposing the θ_f is a constant which can be any value in $(-\pi, \pi]$, according to results from Dubins, the element composed of the shortest path with the supposed θ_f can be determined.

Step2:

If the shortest path is composed by *RSR*, the optimal final orientation θ_f^* can be computed by Eq. (44); otherwise, if the shortest path is composed by *RSL*, the optimal final orientation θ_f^* is computed by Eq. (1616).

Step3:

If $\theta_{f\min} \leq \theta_f^* \leq \theta_{f\max}$, which means that the optimal final orientation θ_f^* is located in the arc *AB* as shown in Fig. 7, then $\theta_f = \theta_f^*$, and the shortest length of path can be computed by substituting θ_f into Eq. (99); if $-\pi \leq \theta_f^* \leq \theta_{f\min}$ or $\pi - (\theta_{f\min} + \theta_{f\max})/2 \leq \theta_f^* \leq \pi$, which means that the optimal final orientation is located in the arc *BC* of Fig. 7, then $\theta_f = \theta_{f\min}$ since $\theta_{f\min}$ is closer to θ_f^* than $\theta_{f\max}$, and the shortest length of path can be calculated by the result of Dubins; if $\theta_{f\max} \leq \theta_f^* \leq \pi + (\theta_{f\min} + \theta_{f\max})/2$, which means the optimal final azimuth is located in the arc *AC* of Fig. 7, then $\theta_f = \theta_{f\max}$ since $\theta_{f\max}$ is closer to θ_f^* than $\theta_{f\min}$, and the shortest length of path can be computed by the result of Dubins too.

To this end, the problem formulated in Eq. (11) can be solved.

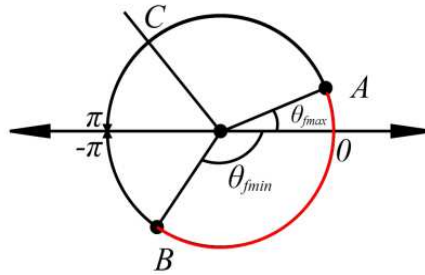


Figure 7
Relative distribution of θ_f^* , θ_{fmin} and θ_{fmax}

5 Path Planning Algorithm

5.1 The Structure of Algorithm

To apply the result in Section 4 in the path plan of a high altitude UAV, it is necessary to store the planned path into the UAV's onboard computer before take-off, then to track this planned path during flight, since the method in Section 4 to solve the problem would need to explicitly calculate the lengths of all arcs and straight line segment in the Dubins curve set, and then choose the shortest of the computed paths; furthermore, many judgments need to be considered. The time necessary for this calculation may become a bottleneck in real-time applications [8].

Taking an investigation on current path planning algorithm in non-holonomic and car-like robot [13, 15-17], multiple UAVs [18-20] and Dubins vehicles [21, 22], it can be seen that all of them are designed to plan the path by the current states and waypoints information, rather than by storing all the planned path on on-board computer. The main advantages are that, on one hand, it reduces the storage requirement of the on-board computer; on the other hand, it can adjust route in real-time when the waypoints are changed. This kind of path planning algorithm enhances the systems' intelligence, so it has been widely applied in actual systems.

In this Section, the real-time path planning algorithm base on the results of Section 4 will be designed. The structure of the algorithm is in Fig. 8.

It can be seen from Fig. 8 that this structure is analogical from real-time control system, and the path planning algorithm is equivalent with control law.

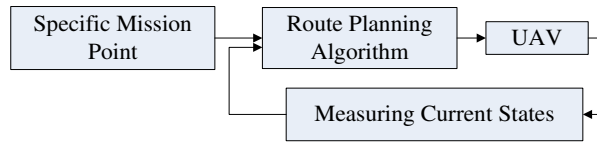


Figure 8

The structure of real-time path planning algorithm

5.2 The Real-time Path Planning Algorithm

In order to design the real-time path planning algorithm based on the results of Section 4, the so called Control Lyapunov Function (CLF) is adopted [23].

The state variables are selected as D_{LL}/D_{RR} , $\min(\alpha_{LL}, \alpha_{RR})$ and current orientation θ_i , as shown in Fig. 9. For clarity, $\min(\alpha_{LL}, \alpha_{RR})$ is denoted as α_L when D_{LL} is shorter than D_{RR} , and is denoted as α_R when D_{LL} is greater than D_{RR} .

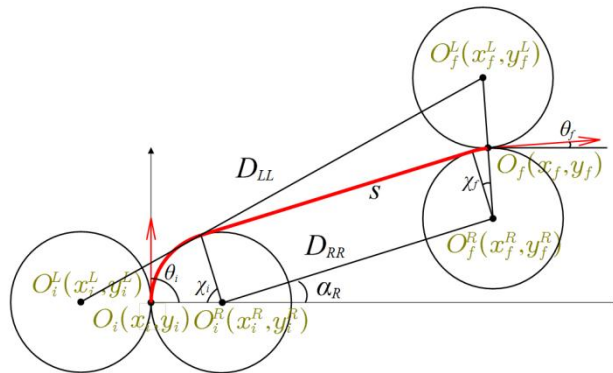


Figure 9

Schematic diagram of state variables

The physical meaning of Eq. (11) can be interpreted as the path planning problem for a UAV moving in the plane subject to the constraints of velocity and turning radius [24]. The state space formula can be presented as follows:

$$\begin{cases} \dot{x} = \cos \theta \\ \dot{y} = \sin \theta \\ \dot{\theta} = \frac{1}{\rho_{\min}} u \end{cases} \quad (17)$$

where, $-1 \leq u \leq 1$, representing the maneuverability constraints of UAVs, and the velocity of UAVs is supposed to be 1.

From the result of Dubins, it can be determined that Eq. (18) is satisfied when the path of UAV is the shortest one.

$$u \in \{-1, 0, 1\} \quad (18)$$

It has been proven in Ref. [23] that u^* is the optimal control law only if u^* can make $\min(\alpha_{LL}, \alpha_{RR})$ decrease and $\min(\alpha_{LL}, \alpha_{RR}) \rightarrow 0$ for the case of I-LP.

Without the loss of generality, the way to design a real-time path planning algorithm is demonstrated with the aid of Fig. 9. It also needs to be noted that:

$$\begin{aligned} -\pi < \theta_i, \theta_f \leq \pi \\ -\pi < \alpha_R \leq \pi \end{aligned} \quad (19)$$

According to Eq. (55):

$$D_{RR} = \sqrt{(x_f^R - x_i - \rho_{\min} \sin \theta_i)^2 + (y_f^R - y_i + \rho_{\min} \cos \theta_i)^2} \quad (20)$$

The difference of D_{RR} with respect to time can be expressed as follows

$$\begin{aligned} \dot{D}_{RR} &= \frac{1}{s} [(x_f^R - x_i - \rho_{\min} \sin \theta_i)(-\dot{x}_i - \rho_{\min} \cos \theta_i \dot{\theta}_i) \\ &\quad + (y_f^R - y_i + \rho_{\min} \cos \theta_i)(-\dot{y}_i - \rho_{\min} \sin \theta_i \dot{\theta}_i)] \\ &= -\frac{\rho_{\min}}{s} (1 + \dot{\theta}_i) [(x_f^R - x_i - \rho_{\min} \sin \theta_i) \cos \theta_i \\ &\quad + (y_f^R - y_i + \rho_{\min} \cos \theta_i) \sin \theta_i] \end{aligned} \quad (21)$$

From the geometry relationship, the following equation can be derived

$$\cos \alpha_R = \frac{x_f^R - x_i}{s} \quad \sin \alpha_R = \frac{y_f^R - y_i}{s} \quad (22)$$

Substituting Eq. (22) into Eq. (21)

$$\begin{aligned} \dot{D}_{RR} &= -\rho_{\min} (1 + \dot{\theta}_i) [\cos \alpha_R \cos \theta_i + \sin \alpha_R \sin \theta_i] \\ &= -\rho_{\min} (1 + \dot{\theta}_i) \cos(\alpha_R - \theta_i) \end{aligned} \quad (23)$$

Here, $-2\pi < \alpha_R - \theta_i \leq 2\pi$ since $-\pi < \alpha_R, \theta_i \leq \pi$. The range of $\alpha_R - \theta_i$ is shown in Fig. 10. For the reason that

$$1 + \dot{\theta}_i \in \{0, 1, 2\} \quad (24)$$

So, only if Eq. (25) or Eq. (26) satisfied, the left hand of Eq. (23) is smaller than zero, and $D_{RR} \rightarrow 0$.

$$\cos(\alpha_R - \theta_i) \geq 0 \quad \text{and} \quad 1 + \dot{\theta}_i \in \{1, 2\} \quad (25)$$

$$\cos(\alpha_R - \theta_i) < 0 \quad \text{and} \quad 1 + \dot{\theta}_i = 0 \quad (26)$$

Fig. 10 also shows that, UAV will fly along a straight line and make $D_{RR} \rightarrow 0$ when $\alpha_R = \theta_i$ or $\alpha_R = \theta_i \pm 2\pi$.

According to this result, Table 2 about the control function can be designed.

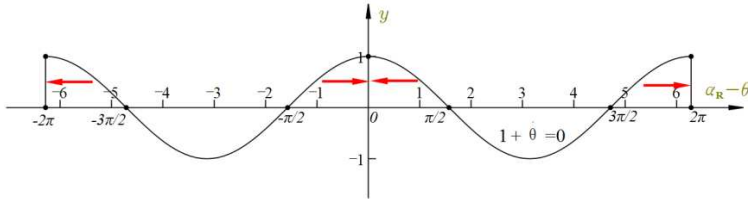


Figure 10
The range of $\alpha_R - \theta_i$

Table 2
Value table I of control function

Distribution range $\alpha_R - \theta_i$	Approaching value $\alpha_R - \theta_i$	Monotonicity $\alpha_R - \theta_i$	Monotonicity θ_i	Value $\dot{\theta}_i$
$(-2\pi, -3\pi/2]$	-2π	Decreasing	Increasing	1
$(-\pi/2, 0]$	0	Increasing	Decreasing	-1
0	--	--	--	0
$(0, \pi/2]$	0	Decreasing	Increasing	1
$(3\pi/2, 2\pi]$	2π	Increasing	Decreasing	-1
else	--	--	--	-1

However, the control function in Table 2 can only guarantee $D_{RR} \rightarrow 0$. Once $D_{RR} = 0$, $\min(\alpha_{LL}, \alpha_{RR})$ will be meaningless, but obviously, the aim has not been achieved yet, because θ_i is not equal to θ_f at this moment. Here, $(\theta_f - \theta_i)$ can also be picked up as a state variable when $D_{RR} = 0$, the goal is $(\theta_f - \theta_i) \rightarrow 0$ the value table of control function about $(\theta_f - \theta_i)$ can be designed as shown in Table 3.

Table 3
Value table II of control function

Distribution range $(\theta_f - \theta_i)$	Approaching value $(\theta_f - \theta_i)$	Monotonicity $(\theta_f - \theta_i)$	Monotonicity θ_i	Value $\dot{\theta}_i$
$(-2\pi, -\pi/2]$	-2π	Decrease	Increase	1
$(-\pi, 0]$	0	Increase	Decrease	-1
0	--	--	--	0
$(0, \pi]$	0	Decrease	Increase	1
$(\pi, 2\pi]$	2π	Increase	Decrease	-1

So far, Table 2 and Table 3 show a complete control law for the real-time path planning algorithm for UAVs.

5.3 Implementation Steps

The real-time path planning algorithm for the manoeuvres of UAVs can be summarized as follows with the steps to implement:

BEGIN

Step1:

The current position of UAV is denoted as P_i , and the next two waypoints are denoted as M_i and M_{i+1} . If $d(M_i, M_{i+1}) \geq 2\rho_{\min}$, the UAV flies along the path planned by Dubins curves set; if $d(M_i, M_{i+1}) < 2\rho_{\min}$, switch to **Step2**.

Step2:

The current position P_i is denoted as (x_i, y_i, θ_i) and the next waypoint position M_i is denoted as (x_f, y_f, θ_f) . The θ_f is calculated by the method in Section 4, then check whether $\min(\alpha_{LL}, \alpha_{RR})$ is zero; if yes, switch to **Step4**; if no, switch to **Step3**.

Step3:

Computing $\min(\alpha_{LL}, \alpha_{RR}) - \theta_i$, and obtaining the value of control function according to Table2. Switch to **Step2**.

Step4:

Checking whether $(\theta_f - \theta_i)$ is zero, if yes, switching to **Step5**; if no, obtaining the value of control function according to Table3, switching to **Step2**.

Step5:

Checking whether the task is complete, if yes, switching to **Step6**; if no, switching to **Step1**.

Step6:

END.

6 Simulation Examples

In this Section, the performance of the designed real-time path planning algorithm is analyzed. For the reason that the distribution of the waypoints has a great influence on the performance of the path planning algorithm, the simulation examples are carried out with different distributions of waypoints. Two types of quadrilateral routes are investigated here; for the other cases, similar discussions can be followed.

6.1 Case 1

In this subsection, the case of a quadrilateral route in which the distances of every two points are greater than $2\rho_{\min}$ is considered. The velocity of UAV is $V = 20\text{m/s}$, the initial orientation is $\theta_i = 90^\circ$, the time step of path planning algorithm is 1 second, and the constraint of manoeuvrability is $\Delta\theta = 10^\circ/\text{s}$. The equivalent minimal radius is

$$\rho_{\min} = \frac{V}{\Delta\theta} = \frac{20}{10\pi/180} = 114.6\text{m} \quad (27)$$

The coordinates of each waypoint are list in Table 4:

Table 4
Distribution of the waypoints in case 1

waypoints	x coordinate(m)	y coordinate(m)
A	0	0
B	100	500
C	500	500
D	200	0

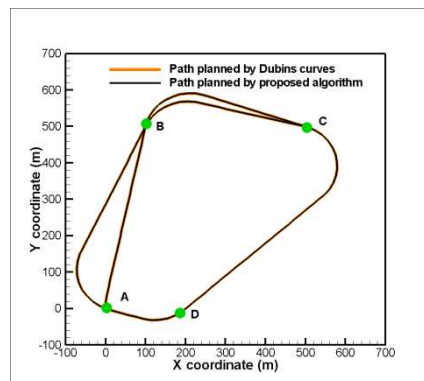


Figure 11

Compare between two algorithms for case 1

The comparison between the paths planned by Dubins curves and proposed real-time algorithm is shown in Fig. 11. Because the distances of every two points are greater than $2\rho_{\min}$ in this case, both of methods can find the shortest path to pass all of waypoints.

This result shows that the performance of the proposed real-time algorithm is equivalent to the Dubins curves in the case that the distances of every two points are greater than $2\rho_{\min}$.

6.2 Case 2

In this subsection, the case of a quadrilateral route in which some of the distances of two points are shorter than $2\rho_{\min}$ is considered. The simulation parameters are the same as those in subsection 6.1.

In this case, the coordinates of each waypoints are listed in Table 5; obviously, the distance of waypoint *C* and *D* is shorter than $2\rho_{\min}$, so the manoeuvres of the aircraft will be constrained by admissible angles $[\theta_{\min}, \theta_{\max}]$ when flying along a path passing the waypoints of *C* and *D*.

Table 5
Distribution of the waypoints in case 2

waypoints	<i>x</i> coordinate(m)	<i>y</i> coordinate(m)
<i>A</i>	0	0
<i>B</i>	100	500
<i>C</i>	500	500
<i>D</i>	500	350

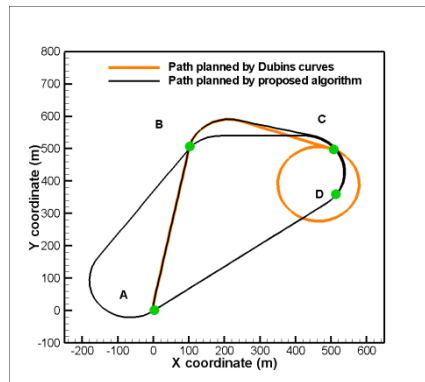


Figure12

Compare between two algorithms for case 2

It can be seen from Fig. 12 that the UAV cannot fly across waypoint *D* even by the maximal manoeuvrability when the path is planned by Dubins curves, since $d(C, D) < 2\rho_{\min}$ and Dubins curves cannot deal with the circumstance that the θ_f is not fixed and constrained by admissible angles $[\theta_{\min}, \theta_{\max}]$.

On the contrary, by the proposed algorithm, the UAV takes off from point *A*, and flies across point *B*, but the UAV flies along the way of *RSR* type of Dubins curves instead of flying toward point *C* directly, for the reason that $d(C, D) < 2\rho_{\min}$.

This result shows that the performance of the proposed real-time algorithm is better than the Dubins curves in this case.

Conclusions

The discussion about how to find the shortest path for manoeuvres of a UAV is present in this paper, and an algorithm that can find the shortest path during manoeuvres and improve the ability of the UAV to follow waypoints is described.

The method to calculate the shortest path for the UAV during manoeuvres is firstly derived by the theory of the Dubins curve set. Secondly, in order to improve the ability of the UAV to follow the calculated optimum path, a real-time path planning algorithm is designed by transforming the constraints of the Dubins curve into a dynamic equation.

To demonstrate the applicability and performance of the proposed path planning algorithm, some typical numerical examples are presented. The results show that the proposed algorithm is promising for application in the path planning for manoeuvres of UAVs.

References

- [1] Ambrosino G, Ariola M, Ciniglio U, Corraro F, Lellis ED, Pironti A. Path Generation and Tracking in 3D for UAVs. *IEEE Transactions On Control Systems Technology*. 2009; 17(4):980-8
- [2] Savla K, Bullo F, Frazzoli E. The Coverage Problem for Loitering Dubins Vehicles. *Proceedings of the 46th IEEE Conference on Decision and Control*; 1; New Orleans, LA, USA2007, pp. 1398-403
- [3] Said Z, Sundaraj K. Simulation of Nonholonomic Trajectory for a Car-Like Mobile Platform using Dubins Shortest Path Model. *IEEE Conference on Sustainable Utilization and Development in Engineering and Technology*; Selangor, Malaysia 2011, pp. 127-32
- [4] Dubins LE. On Curves of Minimal Length with a Constraint on Average Curvature, and with Prescribed Initial and Terminal Positions and Tangents. *American Journal of Mathematics*. 1957 1;79:497~516
- [5] Furtuna AA, Balkcom DJ. Generalizing Dubins Curves: Minimum-time Squences of Body-fixed Rotations and Translations in the Plane. *The International Journal of Robotics Research*. 2010;29
- [6] Chitsaz H, Lavalle SM. Time-optimal Paths for a Dubins airplane. *Proceedings of the 46th IEEE conference on Decision and Control*; New Orleans, LA, USA 2007, pp. 2379-84
- [7] Giordano PR, Vendittelli M. Shortest Paths to Obstacles for a Polygonal Dubins Car. *IEEE Transactions on Robotics*. 2009;25(5):1184-91
- [8] Shkel AM, Lumelsky V. Classification of the Dubins set. *Robotics and Autonomous Systems*. 2001 1;34:179-202
- [9] Boissonnat J-D, Cerezo A, Lenlond J. Shortest Paths of Bounded Curvature in the Plane. *Robotique, Image et Vision*. 1991 1;4:1~20

-
- [10] Bakolas E, Tsiotras P. Time-Optimal Synthesis for the Zermelo-Markov-Dubins Problem: the Constant Wind Case. American Control Conference; Baltimore, MD, USA 2010. p. 6163-8
- [11] Macharet DG, Neto AA, Campos MFM, Campos MFM. Nonholonomic Path Planning Optimization for Dubins' Vehicles. IEEE International Conference on Robotics and Automation; 1; Shanghai China 2011, pp. 4208-13
- [12] Hota S, Ghose D. A Modified Dubins Method for Optimal Path Planning of a Miniature Air Vehicle Converging to a Straight Line Path. American Control Conference; 1; St. Louis, MO, USA 2009, pp. 2397-402
- [13] Liang TC, Liu JS, Hung GT, Chang YZ. Practical and Flexible Path Planning for Car-Like Mobile Robot Using Maximal-Curvature Cubic Spiral. *Robotics and Autonomous Systems*. 2005 1;52:312-35
- [14] Minas AC, Urrutia S. Discrete Optimization Methods to Determine Trajectories for Dubins' Vehicles. *Electronic Notes in Discrete Mathematics*. 2010 1;36:17-24
- [15] Yong C, Barth EJ. Real-time Dynamic Path Planning for Dubins' Nonholonomic Robot. *Proceedings of the 45th IEEE Conference on Decision & Control* 1; San Diego, CA, USA 2006, pp. 2418-23
- [16] Scheuer A, Fraichard T. Planning Continuous-Curvature Paths for Car-Like Robots. *IEEE/RST Int Conf on Intelligent Robots and Systems*; 1; Osaka, Japan 1996. p. 1304~11
- [17] Tang G, Wang Z, Williams AL. On the Construction of an Optimal Feedback Control Law for the Shortest Path Problem for the Dubins Car-like Robot. *Electrical Engineering*. 1998 1:280~4
- [18] Shanmugavel M, Ã AT, White B, Z R. Control Engineering Practice Co-Operative Path Planning of Multiple UAVs Using Dubins Paths with Clothoid Arcs. *Control Engineering Practice*. 2010 1;18:1084-92
- [19] Jeyaraman S, Tsourdos A, Zbikowski R, White B. Formal Techniques for the Modelling and Validation of a Co-operating UAV Team that uses Dubins Set for Path Planning. American Control Conference; 1; Portland, OR, USA 2005, pp. 4690-5
- [20] Jeyaraman S, Tsourdos A, Rabbath CA, Gagnon E. Formalised Hybrid Control Scheme for a UAV Group using Dubins Set and Model Checking. *Conference on Decision and Control*; 1; Atlantis, Paradis Islan, Bahamas 2004, pp. 4299-304
- [21] Hanson C, Richardson J, Girard A. Path Planning of a Dubins Vehicle for Sequential Target Observation with Ranged Sensors. American Control Conference; 1; San Francisco, CA, USA 2011, pp. 1698-703

- [22] Balluchi A, Bicchi A, Piccoli B. Stability and Robustness of Optimal Synthesis for Route Tracking by Dubins's Vehicles. Proceedings of the 39th IEEE Conference on Decision and Control 1; Sydney, Australia 2000, pp. 581-6
- [23] Chao Y, Barth EJ, editors. Real-time Dynamic Path Planning for Dubins' Nonholonomic Robot. Proceedings of the 45th IEEE conference on Decision&Control; 2006; San Diego, CA, USA, December 13-15
- [24] Bui X-N, Soueres P, Boissonnat J-D, Laumond J-p. The Shortest Path Synthesis for Non-holonomic Robots Moving Forwards. INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE. 1993 1:1~33