Acta Universitatis Sapientiae

Electrical and Mechanical Engineering

Volume 1, 2009

Sapientia Hungarian University of Transylvania
Scientia Publishing House

Contents

Broadcast Technology

Signal Processing

Mechatronics & Industry Applications

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 5-28

A Survey of Speed and Flux Control Structures of Squirrel-Cage Induction Motor Drives

Maria IMECS

Department of Electrical Drives and Robots, Faculty of Electrical Engineering,
Technical University of Cluj-Napoca, Cluj-Napoca, Romania,
e-mail: imecs@edr.utcluj.ro

Manuscript received April 25, 2009; revised June 30, 2009.

Abstract: The paper presents an overview of the adjustable speed induction motors with short-circuited rotor from the classical V-Hz open-loop- to the field-oriented closed-loop methods. Scalar-control structures are presented based on direct and indirect flux regulation versus vector-control strategies with direct and indirect field-orientation, for voltage- and current-source frequency converter fed drives. Synthesis about DC-link frequency converters, pulse modulation procedures and mechanical characteristics of the flux-controlled machine are included. Details regarding generation, computation and identification of feedforward and feedback control variables are treated. A new vector control structure is proposed for voltage-controlled drives, which combines the advantages of rotor- and stator-field-orientation procedures.

Keywords: Vector and scalar control, direct and indirect flux control, stator- and rotor-orientation field, direct and indirect field-orientation.

Dedication:
In memoriam Professor Arpad Kelemen, my mentor in power electronics and electrical drives.

1. Introduction

It is estimated that more than 75% of all electrical drive applications require adjustable-speed. Nowadays the most wide-spread electric drive is based on the squirrel-cage short-circuited-rotor induction machine (SqC-IM) due to its low cost, robustness, reduced size and simple maintenance (it is in fact a brushless machine). An AC drive with rotating magnetic field, due to its nonlinear and highly interacting multivariable mathematical model, as an actuator, behaves considerably with more difficulty in a control structure compared to a compensated DC machine, where the torque is controlled directly with the

armature current, while the motor resultant flux is kept constant by means of the field winding. In case of the induction motor (IM), such a decoupling inherently can not exist. Electro-magnetically the induction motor was the electric machine (EM) most difficult to analyze. Nevertheless, the SqC-IM drives are employed in various industrial fields and dominate the power range from hundred Watts to tens of MWs. This is due to the innovation in power electronic converters (PEC), namely such as new topologies realized with high frequency power electronic devices, and the evolution of the drive-dedicated signal processing equipments, which permit the implementation of advanced control theories leading to a highly improved performance of economical industrial AC drives.

2. Generalities about control strategies of the cage induction motors

The area of adjustable AC drives is complex and multi-disciplinary, involving both, power- and signal-electronics, PEC circuits, microprocessors, EMs, various control procedures, system theory, measurement technique, etc. In the control of a technological process, the EM is the actuator of the mechanical load, but in the control of the EM the PEC will be the actuator. There are natural intrinsic feedback effects, as the effect of the mechanical load upon the motor (e.g. the load-torque dependence on the speed value and sense and/or on position), and the motor effect upon the PEC. Such a system in AC drives usually can not operate in feedforward control, due to its multivariable, parameter-varying and nonlinear structure with coupling effects of the state variables. Both the steady state characteristics and the dynamic behavior of the system may be analyzed only by means of a mathematical model based on the space-phasor theory [1].

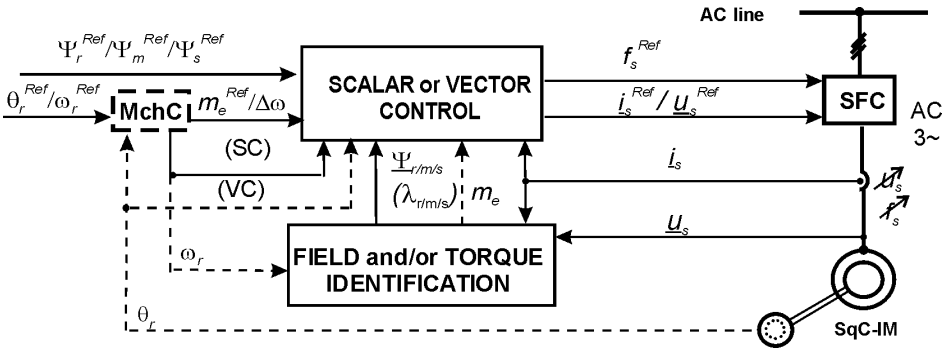


Figure 1: The general block diagram of a controlled cage-rotor induction motor drive.

In advanced AC drives a static frequency converter (block SFC in *Fig. 1*) controls the motor by means of two input variables; these are the amplitude and

frequency of the supply voltage. Consequently, mathematically it is possible to impose also *two reference values* in the control system of the SqC-IM. In a motor control system usually a loop is dedicated to the mechanical quantities, such as position (θ_m), speed (ω_m) and torque (m_e), and another to the magnetic quantities, which may be one of the resultant fields, i.e. belonging to the stator Ψ_s, air-gap Ψ_m or rotor Ψ_r [2], [24].

The motor control system needs information about the technological process and about the drive. That means it requires sensing and computing of the mechanical, electrical and magnetic quantities. In general, the rotor position or speed, the stator currents and voltages are measured, while the torque and magnetic field may be only identified, using estimators or observers.

The control structure depends on the procedures of flux control, field-orientation and identification of the feedback quantities, the PEC type, including its pulse modulation method, and the character of the mechanical load.

The generation of the control quantities for the static frequency converter may be based on scalar control (SC) or vector control (VC) principle.

3. Comparison of the scalar and vector control procedures

A scalar AC-drive system controls only the magnitude of the prescribed quantities, without taking into account the relative position (phase shift) of the current-, voltage- and flux space phasors, which correspond to the three-phase variables of each quantity. Consequently, only the module of the controlled flux vector should be identified. In SC schemes the two control-loops work independently as is shown in *Table 1*:

Table 1: Scalar control (SC) strategy of induction motor drives

Imposed reference quantities	Decoupled control loops	Intermediate control variables	Decoupled control loops	Motor control variables
Position / Speed / Torque	→	**Absolute Slip**	→	**Frequency – f_s**
Stator / Air-gap/Rotor Flux	→	**Stator Current**	→	**Voltage – U_s**

The mechanical loop generates the actual working frequency usually by means of slip compensation, while the reactive loop provides the amplitude or r.m.s. value of the control variable for the stator current (neglecting its torque-producing active character), resulting directly or inherently the stator voltage.

The VC procedure is based on the field-orientation principle, and it needs the identification of a resultant flux as a vector. That means not only its magnitude (amplitude), but also the position (phase) angle (λ), because the stator-current control variable (with its active-/torque producing and reactive-/field producing components) will be generated in field-oriented (FO) axis frame. In VC

structures, as it is presented in *Table 2*, the two intermediate control variables are different from those of a SC scheme, because they are special ones, i.e. the decoupled FO components of the stator-current space-phasor (SPh).

Usually the stator-current FO components are directly generated from the controllers in the decoupled control loops. In order to generate the SFC control variables, the two control loops will be re-coupled due to the reverse transformation of the FO components into natural (i.e. stator-oriented) ones, which needs the feedback variable λ (the position angle of the rotating orientation flux). In fact it realizes the self-commutation of the IM by means of the current- or voltage- space phasor with respect to the orientation flux one. As a consequence, a VC system achieves high performance considering its static stability and dynamic behavior.

Table 2: Vector control (VC) strategy of induction motor drives

Imposed reference quantities	Decoupled control loops	Intermediate field-oriented control variables	Re-coupled control loops	Motor control variables
Position / Speed / Torque	→	**Stator Current Active Component**	**X**	**Voltage vector** $\underline{u}_S$ **(U_S & γ_S)**
Stator / Air-gap/Rotor flux	→	**Stator Current Reactive Component**		

The control variable of the supplying stator-voltage ($\underline{u}_s=U_s\ e^{j\gamma_s}$) contains information inherently about the actual working frequency f_s, because the synchronous speed – at last in steady state – will be $\omega_{sy} = 2\pi f_s = d\gamma_s/dt = d\lambda/dt$ and it is determined by the re-orientation angle λ.

In the VC structures the natural behavior of the IM is taken into account by means of the FO state-space (dynamic) model of the SqC-IM, based on the space-phasor theory, in contrast with the SC ones, where this aspect is ignored.

4. Comparison of direct and indirect torque control

The direct control of the variables is made with proper controllers. Due to their natural correlation, the position, speed and torque are controlled in the same *active control loop*, as it is shown in *Fig. 2* [2], [24]. It has the output reference variables according to *Table 1* and *2*: $\Delta\omega$ the absolute slip in SC and i_A the active component of the current in VC, respectively, which are proportional to the torque, as it is shown in the next sections.

If the position control is not required, the proportional (P) position controller may be disconnected and the speed reference value will be prescribed. The torque (PI) controller may be also missing in case of regulation by the speed

loop error or if the torque is controlled indirectly (see *Figure 3*) by dividing the reference torque value with the flux amplitude (in VC) or its square value (in SC), according to the expressions given in the next sections.

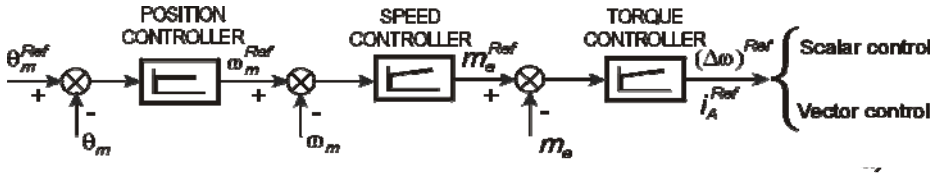


Figure 2: The complete active control loop of the electromechanical variables with direct control of the induction motor torque.

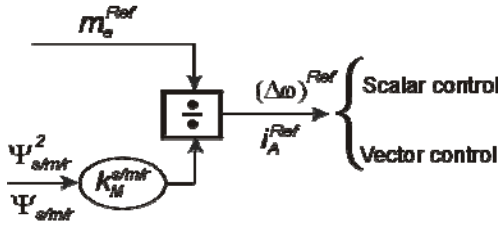


Figure 3: Indirect torque control of the cage-rotor induction motor.

In the middle of the ‘80s – fifteen years after Germany had developed the classical field-oriented VC – the so-called *direct torque (& flux) control* (DTC) was introduced, which needs both stator-flux and motor torque identification. It is a very simple and robust VC method for AC motors supplied from PWM-VSI with space-vector modulation (SVM) method, based on SPh theory.

5. Comparison of direct and indirect flux control procedures

The indirect flux control (IFC) is made without any controller and mainly it is characteristic to SC structures. The r.m.s. values of the control variables – like stator current or stator voltage – are computed based on the steady-state mathematical model of the SqC-IM (i.e. the classical time-phasor equations), the desired flux value resulting inherently. In *Fig. 4* are shown two basic procedures of indirect flux control, which are achieved by means of function generators. One of them has the input the absolute slip and the other the synchronous speed, before and after the slip compensation, respectively.

The stator flux may be kept at a quasi-constant value by the so-called control at U/f= ct. (denoted as V-Hz procedure as well). It results from the stator-voltage equation. Near the rated working point (i.e. U_{sN} – f_{sN}) the stator resistance may be neglected, and this leads to a linear expression:

$$U_s^{Ref} = \frac{U_{sN}}{f_{sN}} f_s^{Ref} \quad . \tag{1}$$

For lower speed range of the drive, the stator resistance R_s has to be taken into account. A possible solution is as follows:

$$U_s^{Ref} = R_s I_s + \frac{U_{sN} - R_s I_s}{f_{sN}} f_s^{Ref} \quad , \tag{2}$$

where I_s is the measured feedback stator current [2], [24]. It is an R_s-compensation procedure with variable slope characteristic.

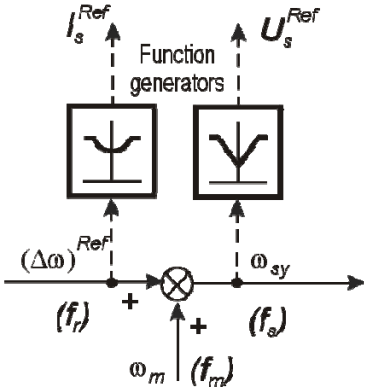


Figure 4: Indirect rotor- and stator-flux control, before and after slip-compensation in scalar control structures.

In V-Hz control, according to (1) and (2), the reference of the stator-voltage r.m.s. value (or amplitude) is generated depending on the frequency reference. This procedure may be applied to all types of AC machines, including synchronous ones, too.

The rotor flux of the SqC-IM may be also controlled indirectly, based on the rotor-voltage equation, by computing the stator-current, depending on $\Delta\omega$ the absolute slip. Considering the rated value of the rotor-flux-based magnetizing current I_{mrN}, the reference value of the stator current may be generated according to the following expression:

$$I_s^{Ref} = I_{mrN}^{Ref} \sqrt{(\tau_r * \Delta\omega)^2 + 1} \text{ , where } I_{mrN}^{Ref} = \frac{\Psi_{rN}^{Ref}}{L_m} \tag{3}$$

and $\tau_r = L_r / R_r$ is the rotor time constant. The angular speeds of the IM may be expressed with the corresponding frequencies as follows:

– the synchronous angular speed

$$\omega_{sy} = 2\pi f_s , \tag{4}$$

where f_s is the frequency of the stator quantities (voltages, currents, etc.);

– the absolute angular slip

$$\Delta\omega = 2\pi f_r , \tag{5}$$

where f_r is the frequency of the rotor quantities (EMFs, currents, etc.);

– the rotor electrical angular speed (usually it is measured):

$$\omega_m = 2\pi f_m , , \tag{6}$$

where f_m is the mechanically rotating rotor frequency for z_p = 1 pole-pairs. Consequently, the slip compensation may be written as follows:

$$\omega_{sy} = \Delta\omega + \omega_m \quad \text{or} \quad f_s = f_r + f_m , \tag{7}$$

The direct flux control (DFC) may be achieved by using a controller, which is basically of PI-type, and it needs an imposed reference value i.e. one of the resultant field values: stator-, air-gap or rotor-flux, as it is shown in *Table 1, Table 2* and also in *Fig. 5* ($\Psi_{s/m/r}$).

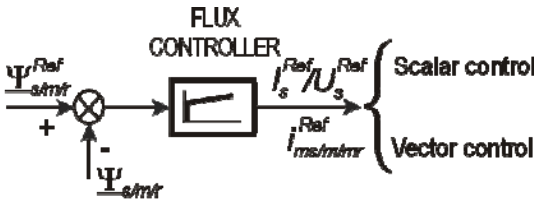


Figure 5: Direct flux control with PI controller in the reactive loop.

The flux controller in SC schemes generates the reference value of the stator voltage or current in amplitude or r.m.s. value. In VC structures, the flux controller will provide the corresponding magnetizing current, ($i_{ms/m/mr}$), which may be equal or not to the reactive component of the field-oriented stator-current space phasor.

6. Comparison of the stator- and rotor-flux control

The well-known *Kloss*'s equation gives the analytical expression of the IM static mechanical characteristics (SMC) at constant stator voltage U_s and frequency f_s. If the pull-out critical torque is kept at constant value by adjusting the value of U_s according to the actual value of f_s, the SMCs have different

shapes. In *Fig. 6* two U_s = ct characteristics are represented – torque versus the absolute slip $\Delta\Omega$ (measured in electrical rad/s) – for two supplying frequencies, i.e. f_{sN} at U_{sN} (both rated values) and $f_s = 0$ at U_{so}, which provide the same break-down torque. The zero value of the rotor speed corresponds to $\Delta\Omega = 314$ rad/s. For different frequencies at U_s = ct, the speed-torque SMCs – due to the different feature of the slip curves – are not parallel, but at constant resultant flux they will become parallel [3].

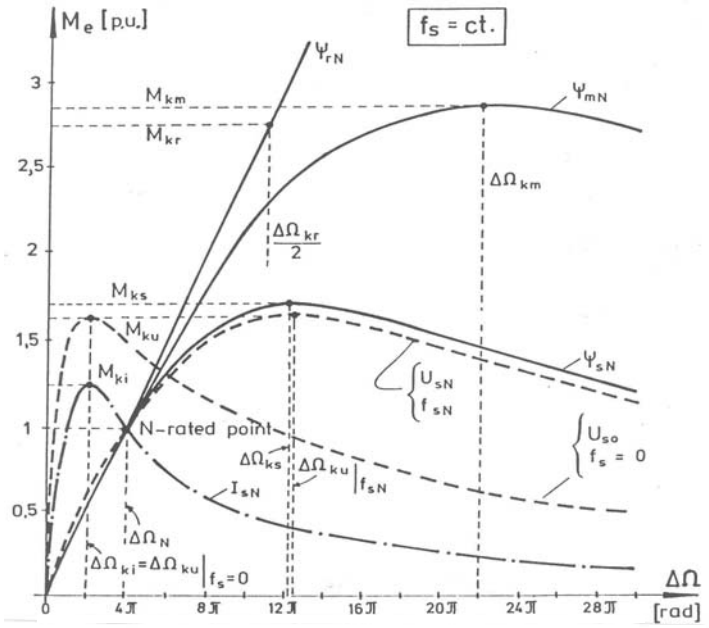


Figure 6: Stator-voltage, stator- and rotor-flux-controlled mechanical characteristics: the electromagnetic torque versus the absolute slip.

Fig. 6 also presents the mechanical characteristics for $I_s = I_{sN}$, $\Psi_s = \Psi_{sN}$, $\Psi_m = \Psi_{mN}$ and $\Psi_r = \Psi_{rN}$, which are useful in FO structures. In a VC scheme the magnitude (module) of the orientation flux vector is usually controlled directly.

A. Stator-flux controlled (SFC) characteristics

For constant stator flux it results a simplified *Kloss*'s type expression that is no more depending on the frequency, as follows [3]:

$$M_e = 2M_{k_S}\left(\frac{\Delta\Omega_{k_s}}{\Delta\Omega} + \frac{\Delta\Omega}{\Delta\Omega_{k_s}}\right)^{-1}, \text{ where } M_{k_S} = k_M \frac{\Psi_s^2}{2L_m}\cdot\frac{1-\sigma}{\sigma(1+\sigma_S)}; \quad \Delta\Omega_{k_s} = \frac{1}{\sigma\tau_r} \tag{8}$$

are the critical (pull-out) torque and slip, $\sigma_s = L_{\sigma s} / L_m$ is the stator leakage coefficient and σ is the resultant one. The self-cyclic inductance L_m corresponds also to the three-phase mutual magnetic effect between the stator and rotor, and it gives the useful resultant field in the air-gap:

$$\underline{\Psi}_m = L_m \underline{i}_m \text{, where } \underline{i}_m = \underline{i}_s + \underline{i}_r \tag{9}$$

is the conventional magnetizing current.

The torque-slip SMCs at Ψ_s = ct from *Fig. 6* are valid for any stator frequency. In spite of being a combination of a linear- and a hyperbolic shape, it leads to parallel speed-torque characteristics for different stator frequencies, excepting flux-weakening region [3], as is shown in *Fig. 7*.

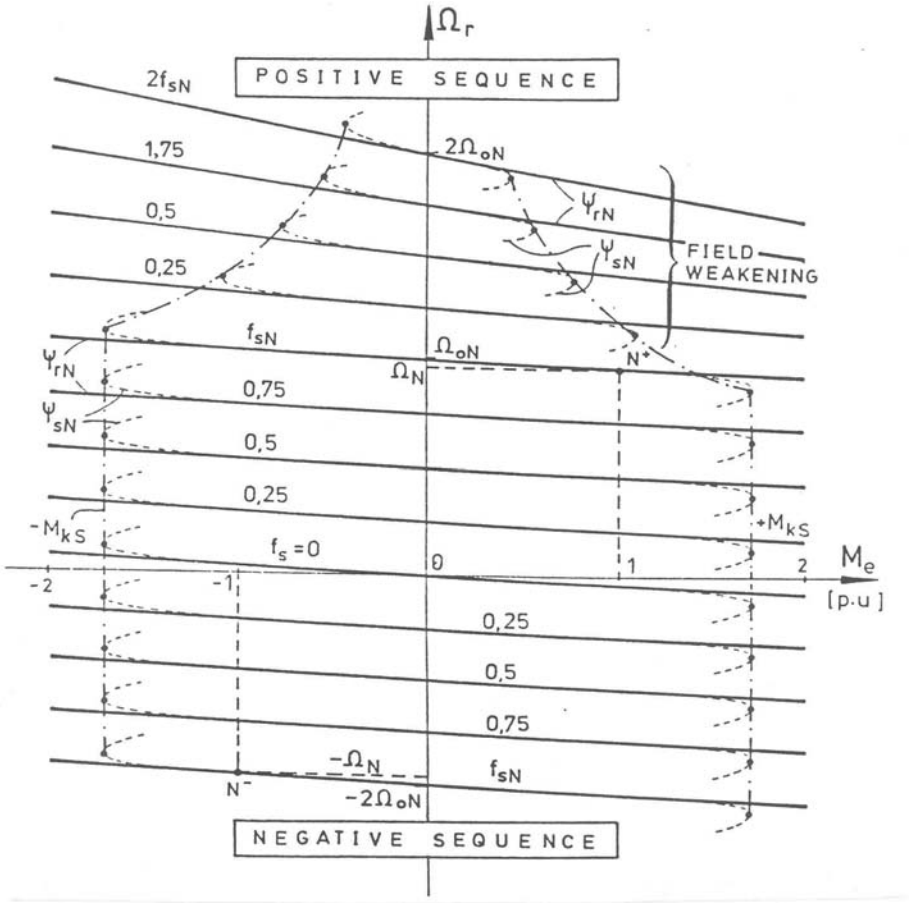


Figure 7: Stator- and rotor-flux-controlled mechanical characteristics: the angular speed versus the electromagnetic torque.

B. Rotor-flux controlled (RFC) characteristics

At constant rotor-flux, the SMC becomes linear without any hyperbolic effect, as in case of a compensated separately excited DC machine. The torque-slip characteristic is given by the following expression:

$$M_e = 2M_{kr}\frac{\Delta\Omega}{\Delta\Omega_{kr}} = \frac{k_M \Psi_r^2}{R_r}\Delta\Omega \ . \tag{10}$$

and they are represented in *Fig. 7*, too.

Due to the linearity of the SMCs, the RFC ensures more stability in behavior of the IM with respect to SFC-ed drives, where the SMCs present pull-out critical torque, due to the so called "*Kloss*" feature given by a hyperbolic shape.

The torque coefficient in (8) and (10) is $k_M = z_p\ 3/2$, if the flux is corresponding to tits peak value, or $k_M = 3\ z_p$, if the flux is expressed by the r.m.s. value of Ψ_s or Ψ_r., respectively.

7. Comparison of the rotor- and stator-field orientation

The FO principle was initially proposed by *Blaschke* in 1971 [4], and it referred to the decoupled control of the mechanical and magnetic phenomena of the short-circuited rotor IM by means of the stator-current rotor-field-oriented components. In fact, field-orientation means change of variables corresponding to phase- (3/2) and coordinate- (complex plane) transformations of the control and feedback variables in a VC structure [6].

A. Rotor-field orientation (RFO)

The classical RFO is usually applied for SqC-IM drives. That means that the direct axis of the complex plane, denoted with $d\lambda_r$, is oriented in the direction of the resultant rotor-flux $\underline{\Psi}_r$, as it is shown in *Fig. 8.*

As a consequence, the flux components result according to (11.1) and (12.1) from *Table 3*.

In case of the SqC-IM ($u_r = 0$), if Ψ_r may be considered at a constant value (that means steady-state or Ψ_r is a controlled variable), the rotor-current $\underline{i}_r$ and rotor-flux $\underline{\Psi}_r$ space phasors are perpendicular one to other. This property led to the idea of the original FO principle based on the rotor-flux-oriented axis frame, in which the stator-current space phasor may be split into two components, as in (13.1), where the RFO components of the stator-current SPh result according to (14.1), (15.1) and (16.1) in *Table 3*. Consequently, the rotor-flux controller may generate directly the field-producing (reactive) component ($i_{sd\lambda r}$) in a control

structure, because this component is equal to the rotor-flux-based magnetizing current (i_{mr}); the speed or torque controller will generate the torque producing (active) quadrature component ($i_{sq\lambda r}$) of the stator current.

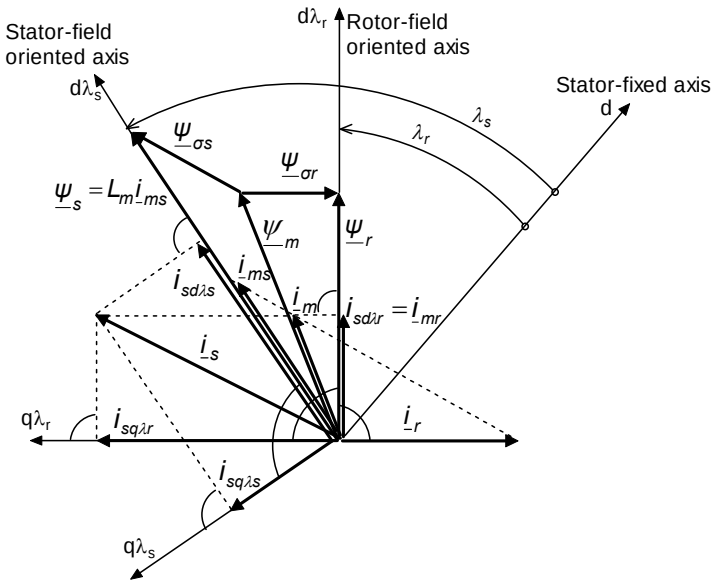


Figure 8: Phasor diagram of the magnetizing currents, fluxes and stator-current field-orientated components.

If the frequency converter is controlled in current, there is no need for model-based computation of the control variables, because they are generated directly by the controllers. If the IM is controlled in voltage, the computation of the stator-voltage components based on the RFO model is highly complex and motor parameter dependent [1], [7], [8], [9], [10].

B. Stator-field orientation (SFO)

SFO means that the direct axis (denoted $d\lambda_s$) of the coordinate frame is oriented in the direction of the resultant stator flux vector $\underline{\Psi}_s$ (see *Fig. 8*), therefore its components are according to (11.2) and (12.2) in *Table 3*.

In the stator-field-oriented (SFO-ed) axes frame the stator-current SPh components from (13.2) can be expressed using (14.2) and (15.2).

Comparing the SFO-ed components with RFO-ed ones, it must be remarked that although the active component in both cases is proportional to the electromagnetic torque, according to (17.1) and (17.2), the reactive one in SFO

is no more equal to the stator-flux-based magnetizing current (i_{ms}), according to (16.2), as *Fig. 8* also shows.

Table 3: Comparison of rotor- and stator-field orientation

Rotor-Field-Orientation (RFO)	**Stator-Field-Orientation (SFO)**
$\Psi_{rd\lambda r} = \underline{\Psi}_r = \lvert\underline{\Psi}_r\rvert = \Psi_r$ (11.1)	$\Psi_{sd\lambda s} = \underline{\Psi}_s = \lvert\underline{\Psi}_s\rvert = \Psi_s$ (11.2)
$\Psi_{rq\lambda r} = 0$ (12.1)	$\Psi_{sq\lambda s} = 0$ (12.2)
Rotor-field-oriented stator-current components:	Stator-field-oriented stator-current components:
$\underline{\boldsymbol{i}}_{s\lambda r} = i_{sd\lambda r} + \underline{\mathrm{j}}\, i_{sq\lambda r}$ (13.1)	$\underline{\boldsymbol{i}}_{s\lambda s} = i_{sd\lambda s} + \underline{\mathrm{j}}\, i_{sq\lambda s}$ (13.2)
where	where
$i_{sd\lambda r} = i_{mr} = \Psi_r / L_m$ (14.1)	$i_{sd\lambda s} \neq i_{ms} = \Psi_s / L_m$ (14.2)
$i_{sq\lambda r} = m_e / K_{Mr}\, \Psi_r = -(1+\sigma_r)\, i_r$ (15.1)	$i_{sq\lambda s} = m_e / K_M\, \Psi_s$ (15.2)
Rotor-flux-based magnetizing current:	Stator-flux-based magnetizing current:
$\underline{i}_{mr} = \underline{\Psi}_r / L_m = i_{sd\lambda r}$ (16.1)	$\underline{i}_{ms} = \underline{\Psi}_s / L_m \neq i_{sd\lambda s}$ (16.2)
Electromagnetic torque:	Electromagnetic torque:
$m_e = K_{Mr}\, \Psi_r\, i_{sq\lambda r} = -K_M\, \Psi_r\, i_r$ (17.1)	$m_e = K_M\, \Psi_s\, i_{sq\lambda s}$ (17.2)
where the torque coefficient is	where the torque coefficient is
$K_{Mr} = K_M / (1+\sigma_r)$ (18.1)	$K_M = (3/2)\, z_p$ (18.2)
Synchronous angular speed:	Synchronous angular speed:
$\omega_{\lambda r} = \mathrm{d}\lambda_r/\mathrm{d}t$ (19.1)	$\omega_{\lambda s} = \mathrm{d}\lambda_s/\mathrm{d}t$ (19.2)
Orientation-field angle:	Orientation-field angle:
$\lambda_r = \int \omega_{\lambda r}\, \mathrm{d}t$ (20.1)	$\lambda_s = \int \omega_{\lambda s}\, \mathrm{d}t$ (20.2)

On the other hand, in SFO schemes the stator-voltage equation provides a more simple computation of the voltage control variables for a voltage-source inverter (VSI) compared to RFO, because in SFO axis frame, the stator-flux SPh has only one component (the direct one), which is equal to its module [1], [7], [11]:

$$u_{sd\lambda s} = R_s\, i_{sd\lambda s} + e_{sd\lambda s} \text{ and } u_{sq\lambda s} = R_s\, i_{sq\lambda s} + e_{sq\lambda s}\,, \tag{21}$$

where EMFs are

$$e_{sd\lambda s} = \mathrm{d}\Psi_s/\mathrm{d}t \text{ and } e_{sq\lambda s} = \omega_{\lambda s}\,\Psi_s. \tag{22}$$

The direct component ($e_{sd\lambda s}$) is the self-induced EMF, which becomes zero in steady state. This is due to the variation in magnitude of the Ψ_s. The quadrature component ($e_{sq\lambda s}$) is generated by the rotation of the stator field with the synchronous speed $\omega_{\lambda s}$ – given by (19.1) –, and it can be computed from the SPh components of the identified orientation field.

For voltage-PWM-VSI-fed drives – due to a simpler voltage model –, SFO is recommended [1], [7], [12]. The computation of the control variables can be made based on expressions (21) and (22). These expressions are affected only by the stator resistance R_s, which may be identified online as well.

SFO was extended also to the synchronous motor drives [5].

8. Comparison of stator and rotor field identification

Because the initially proposed direct flux sensing (see [13]) is no more recommended, nowadays the indirect flux sensing is applied almost exclusively, which is based on the computation of the orientation field from other measured variables. There are two basic field-identification procedures of the field: the so called I-Ω (stator-current & rotor-speed) method for rotor-flux identification and the integration of the stator-voltage equation for stator-flux computation.

A. Stator-flux identification (SFI)

In the ‘70s and ‘80s this flux identification method could be applied only for AC drives supplied from a current-source inverter (CSI), which operates with full-wave currents and quasi-sine-wave terminal voltages, determined by the freely induced rotating EMFs [1], [14], [15], [23], [25]. However, in the last two decades it became a possible method for PWM-inverter-fed drives as well, which are operating with relatively high sampling frequency. Nowadays this procedure seems to be the simplest one for the calculation of the resultant stator flux.

This flux identification procedure is based on the stator-voltage model, written with natural two-phase components in the stator-fixed axis frame. First, the stator EMFs are computed according to equations:

$$\mathrm{d}\Psi_{sd}/\mathrm{d}t = e_{sd} = u_{sd} - R_s\, i_{sd} \text{ and } \mathrm{d}\Psi_{sq}/\mathrm{d}t = e_{sq} = u_{sq} - R_s\, i_{sq}\,, \tag{23}$$

and then it is followed by the direct integration of them, obtaining the flux components:

$$\Psi_{sd} = \int e_{sd}\,\mathrm{d}t \quad \text{and} \quad \Psi_{sq} = \int e_{sq}\,\mathrm{d}t\,. \tag{24}$$

The inputs of the EMF computation block are the two-phase feedback variables of the measured stator-currents and the identified stator voltages computed from the measured DC-link voltage at the input of the inverter and the PWM logic signals generated by the inverter control block.

Today it seems that this method is the most preferable for field identification, due to the fact that it is not affected by the motor parameters, excepting R_s. If it is necessary, the stator resistance may be measured on-line. The applicability of this flux identification depends first of all on the quality of the integration procedure [16].

B. Rotor-flux identification (RFI)

Still in the '80s, the rotor-model-based I-Ω flux identification procedure was preferable for IM drives supplied from PWM-inverters. It was introduced by *Hasse* in 1969 [7]. According to this procedure, there are two possibilities to perform RFI: either with natural (stator-fixed) stator-current components or with RFO ones. The latter procedure needs slip compensation [1], [8], [9]. Both I-Ω methods are strongly affected by the rotor parameters.

Nowadays it is preferable the rotor-flux computation by compensation of the identified stator-flux, using the expressions of the leakage fluxes depending on the measured stator currents. The unknown rotor current from the expression of the rotor leakage flux is eliminated based on the magnetizing current equation (9). The compensation is made without any cross effect between the d-q components, which in synthesized form yields to:

$$\Psi_{rd/q} = (1+\sigma_r)\,\Psi_{sd/q} - (\sigma_s + \sigma_s\sigma_r + \sigma_r)\,L_m\,i_{sd/q}, \tag{25}$$

where $\sigma_r = L_{\sigma r}/L_m$ is the rotor leakage coefficient. The coefficient of L_m is equal to $(1-\sigma)^{-1}$, where σ is the resultant leakage coefficient. The stator-flux components are obtained based on the direct integration of the stator-voltage equation according to the procedure, which was presented in the previous *A* subheading.

9. Comparison of direct (DFO) and indirect (IFO) field orientation

In VC structures the recoupling of the active and reactive control loops is made by means of a reverse coordinate transformation (CooT), which calculates the natural two-phase components of the stator current from the input field-oriented ones. This needs as input variable also the angle of the orientation-field (λ_r or λ_s, respectively, according to *Fig. 8* and *Table 3*). The identification procedure of the orientation angle determines the field-orientation character, which may be direct or indirect.

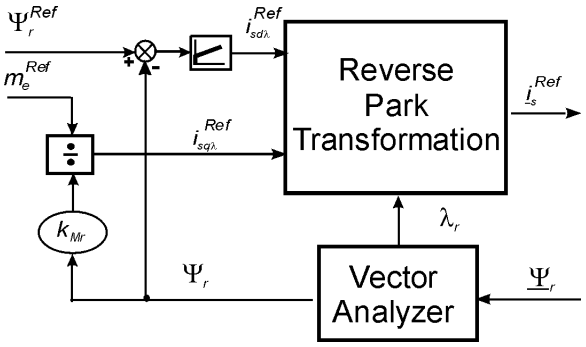


Figure 9: Direct field-orientation (DFO) realized with field-orientation angle computed in a vector analyzer.

A. Direct field-orientation (DFO) procedure

Fig. 9 shows the simplest recoupling of the rotor-field-oriented IM supplied from a current-controlled PEC. It needs the three-phase stator-currents as references (represented symbolically with the space-phasor $\underline{i}_s^{Ref}$), which are computed in the reverse Park transformation (combined from a coordinate- and phase transformation) block.

The orientation-field angle is identified in a vector analyzer (VA), which has as inputs the stator-fixed / stator-oriented two-phase coordinates of the rotor flux (in *Fig. 9* it is represented symbolically by the space-phasor $\underline{\Psi}_r$). This is the direct field-orientation (DFO) procedure, where the orientation-field angle is computed based on the stator-fixed axis frame.

Flux identification procedures based on the stator-oriented axis frame lead to DFO, i.e. the direct integration of the stator-voltage equation and the I-Ω procedure calculated with the rotor-voltage equation written with stator-oriented coordinates.

B. Indirect field-orientation (IFO) procedure

Slip compensation is used not only in SC (see *Fig.4*), but also in VC structures. The indirect field-orientation (IFO) procedure means, that the field-orientation angle is computed by integration of the synchronous speed, which is usually obtained by slip compensation, as in (7) [1], [7], [8], [9]:

$$\omega_{\lambda r} = \mathrm{d}\lambda_r/\mathrm{d}t = \Delta\omega + \omega_m\,, \tag{26}$$

The absolute slip of the IM with short-circuited rotor windings ($\underline{u}_r = 0$) is computed from the rotor-field-oriented voltage equation for $\mathrm{d}\Psi_r / \mathrm{d}t = 0$ (i.e. steady state or controlled rotor flux), as follows [1], [8]:

$$\Delta\omega = \tau_r^{-1}\, i_{sq\lambda r}/i_{sd\lambda r} \quad . \tag{27}$$

The I-Ω flux identification procedure based on the rotor-voltage equation written with RFO components leads to IFO, where the current components ($i_{sq\lambda r}$ and $i_{sd\lambda r}$) are computed from the measured stator currents on the feedback side.

In rotor- or stator-flux-oriented VC structures for IFO the orientation angle (λ_r or λ_s) may be also calculated after the integration of the absolute slip, as is presented in *Fig. 10*.

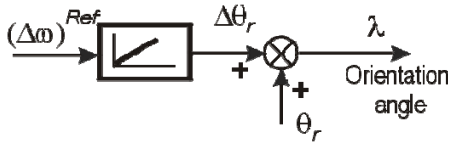


Figure 10: Absolute slip computation in the control loop for indirect field-orientation (IFO).

First it is obtained the slip angle $\Delta\theta_r$, to that is added the measured or estimated rotor position θ_r, then achieving the orientation angle, as follows:

$$\lambda = \Delta\,\theta + \theta_r\,. \tag{28}$$

In VC systems, the field-oriented current components generated in the active and reactive control loops may also serve for the computation of the absolute slip according to expression (27), as it is shown in *Fig. 11*.

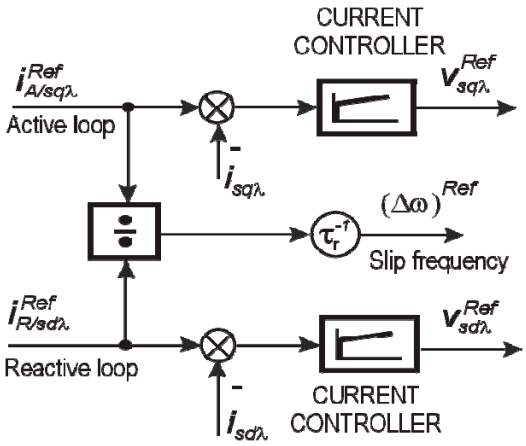


Figure 11: Absolute slip computation in the control loop for indirect field-orientation (IFO).

In case of a voltage controlled PEC (like VSI with feed-forward voltage-PWM), current controllers are recommended, which control the field-oriented stator-current components before the computation of the stator-voltage control variables. These will provide the re-coupling of the two control loops by means of a coordinate transformation block.

The synchronous speed $\omega_{\lambda r}$ is usually needed as input variable in the computation block of the stator-voltage references. It may be also generated from the absolute slip (computed in *Fig. 11*) by slip compensation, according to (26).

10. Computation of the PEC actuator control variables

The computation of the actuator control variables depends on the PEC type and, above all, its pulse modulation procedure, which can be one of two fundamental ones: pulse amplitude modulation (PAM) and pulse-width modulation (PWM). In *Table 4* the control variables of the DC-link frequency converters are given, which are depending on the inverter (voltage- or current-source) type and its pulse modulation method.

Table 4: Inverter types, pulse modulation procedures and control variables of the DC-link frequency converters

Converter control type	VECTOR CONTROL				SCALAR CONTROL
Inverter type	MOS-FET/BT/IGBT-VSI			GTO-CSI	Thy-CSI
Converter output	Voltage-source character		Current-source character		
Pulse modulation method	Open-loop feedforward voltage-PWM		Closed-loop current-PWM	PAM	PAM
Inverter control procedure	Carrier-wave modulation	Space-vector modulation SVM	Bang-bang current control	DC-link current control	
Converter control variables	Instantaneous three-phase voltages: $u_{a,b,c}$	Voltage-amplitude U and γ phase angle	Instantaneous three-phase currents: $i_{a,b,c}$	Current-amplitude I and ε phase-angle	Current: amplitude I and f frequency

In VC structures they can be computed in four ways, according to the current- or voltage space-phasor expression with polar- or with three-phase coordinates (which inherently keep the vector character) as follows:

$$\underline{i} = i e^{j\varepsilon} = k_{Ph}\left(i_a + \underline{a}\, i_b + \underline{a}^2 i_c\right) \text{ and } \underline{u} = u e^{j\gamma} = k_{Ph}\left(u_a + \underline{a}\, u_b + \underline{a}^2 u_c\right), \quad (29)$$

where $i_{a,b,c}$ and $u_{a,b,c}$ are the instantaneous values of the three-phase currents and voltages, respectively. Angles ε and γ are the electrical phase positions of the respective SPhs, which inherently contain "information" about the imposed motor supply frequency, because the inverter output frequency is equal to the derivative of these position angles. The space-phasor coefficient is usually k_{Ph}=2/3. In this case the module of the space phasor will be equal to the amplitude (peak value) of the sine-wave phase variables [1], [8].

The three-phase natural coordinates of the stator control variables result from the field-oriented two-phase ones using a reverse Park-transformation, as in *Fig. 9.* The polar coordinates may be calculated by means of a reverse coordinate transformation block, followed by a vector analyzer (VA).

11. Scalar control structures

The synthesis of computation of the control variables in SC structures with IFC and DFC is represented in *Fig. 12*. The mechanical control loop is represented in its complete form based on the slip compensation procedure. The integrator generates the phase angles γ_s or ε_s of the voltage and current sine-wave three-phase references, respectively.

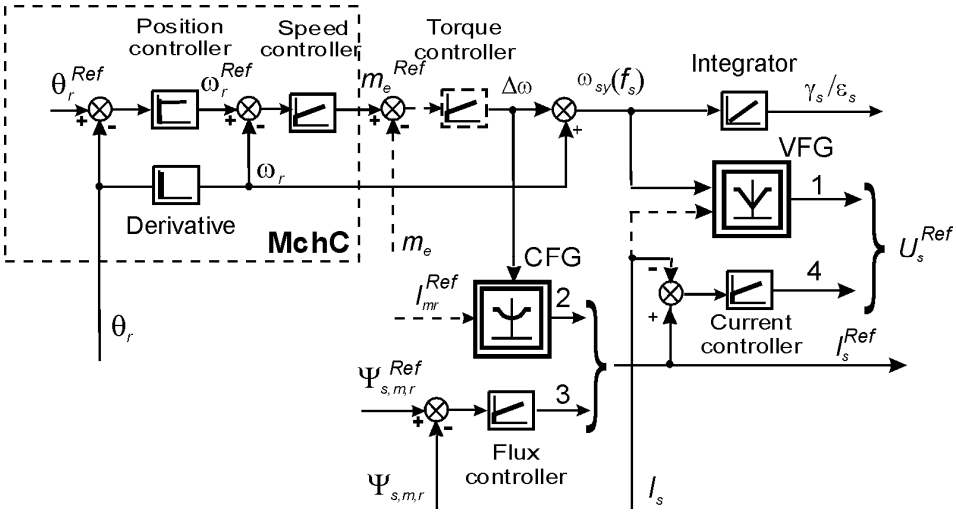


Figure 12: Synthesis of scalar control structures with DFC and IFC of the SqC-IM controlled in current or voltage.

Variant 1 at the output of the voltage function generator (VFG) corresponds to the well-known V-Hz control (IFC of the stator flux). It may have voltage drop compensation by means of the feedback stator current (dashed line).

Variant 2 at the output of the current function generator (CFG) corresponds to IFC of the rotor flux, as it was presented in Section 5.

The DFC procedures correspond to Variant 3. Both current control variables (outputs 2 and 3) may be transformed into a voltage control one by using a current controller with output 4.

The simplest drive control structures are those, which use current mode-controlled inverter [21].

A simple SC structure is presented in *Fig. 13.* It is provided with DFC of the stator flux and speed control (without torque control). The sine-wave generator (SWM) has a scalar character due to the fact that it contains information only about the motor supply frequency. On the feedback path there are two phase transformation blocks (3/2 – PhT), which operate with matrix [A].

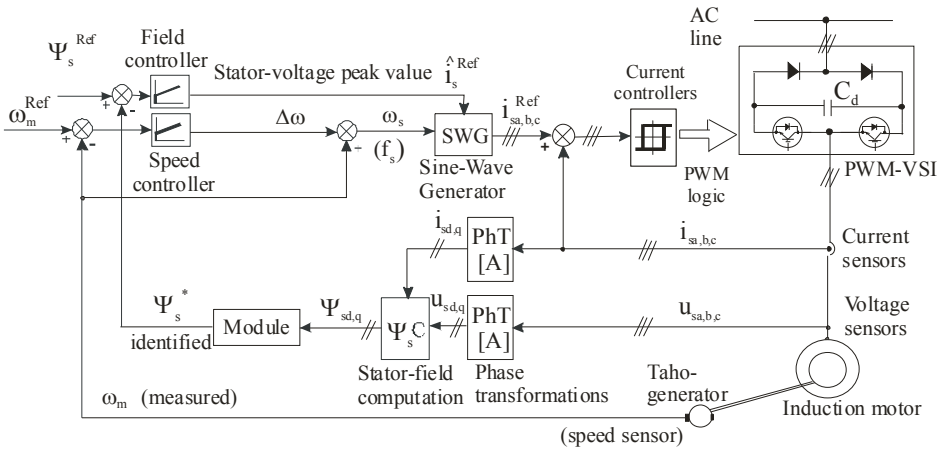


Figure 13: The simplest scalar control structure with DFC of the IM drive supplied from a current-feedback PWM controlled VSI.

The structure in *Fig. 13* may be adapted for voltage control (by means of carrier-wave or SVM) of the PWM-VSI, according to scheme shown in *Fig. 12*.

12. Vector control structures

The simplest VC structure of the SqC-IM presented in *Fig. 14* is achieved by current controlled static frequency converter, rotor-field orientation (RFO) and rotor flux control (RFC). In comparison with structure from *Fig. 13*, it has in addition flux compensation of the identified stator flux, according to (25).

The reverse Park transformation is in fact a sine-wave generator with vectorial character. This structure is not affected by the motor parameters (excepting field identification and controller tuning). Furthermore, such a control system presents the best performance compared with the schemes with stator-field orientation (SFO), stator-flux control (SFC) drives and/or IM supplied from voltage-mode controlled inverter [1], [2], [11], [17].

Some motor-control-oriented digital signal processing (DSP) equipments present on the market do not dispose of implementation possibility of the current-feedback PWM, suitable for current-controlled VSIs, only of possibility of the voltage-feedforward ones, like carrier-wave and SVM. That means the IM can be supplied only by a voltage-source inverter (VSI) with voltage-control [26], [27].

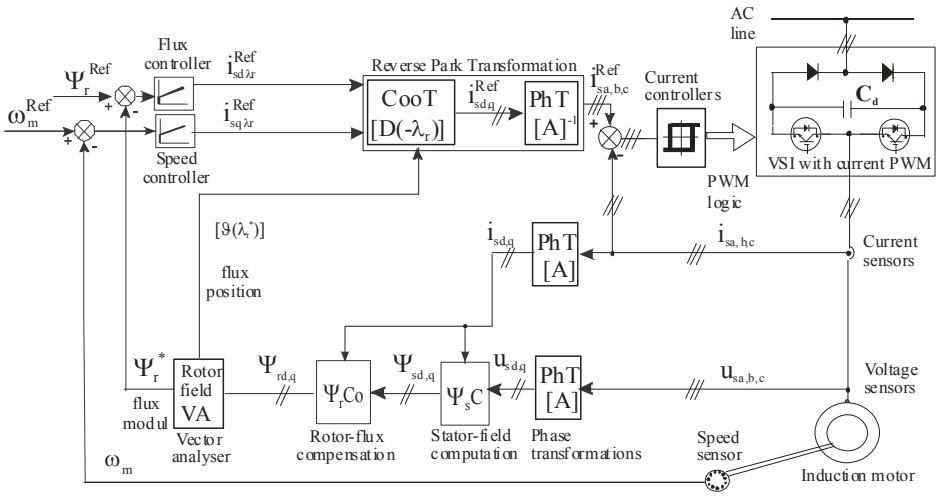


Figure 14: The simplest vector control structure with rotor-field-orientation of the short-circuited IM drive supplied from a current-feedback PWM controlled VSI.

In RFO schemes the computation of the voltage control variables is sophisticated and affected by the motor parameters such as rotor resistance (R_r), rotor time constant τ_r, leakage coefficients and others. Consequently, the drive control performance may be lightly damaged. This problem is usually solved by renouncing the RFC and applying SFO, which leads to a much simpler stator-voltage computation, dependent only on the stator resistance (R_s).

Fig. 15 presents the simplest VC structure for voltage-controlled IM with stator-field-orientation, which is less affected by motor parameter than the RFO one. The stator-flux-based magnetizing current i_{ms} may be also generated at the output of the flux controller, as *Table 3* shows in Section 7. This structure has a somewhat sluggish response to speed reversal, torque command and

perturbation. The structure in *Fig. 15* has no current-control, but it may be completed by including the scheme detail from *Fig. 11*.

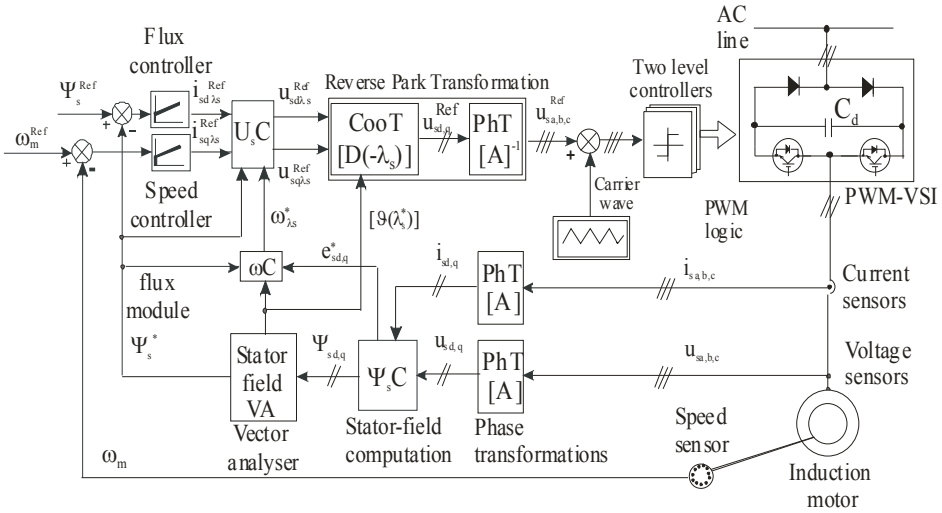


Figure 15: Vector control structure with stator-field-orientation of the short-circuited IM drive supplied from a voltage-feedforward PWM controlled VSI.

Based on the above-mentioned reasons, a new vector control structure was proposed for the induction motor drive fed by a voltage-controlled static frequency converter. It is carried out with double field orientation (DFO) as follows [20]: RFC and RFO of the stator-current components are generated by the speed and flux controllers at the decoupled control side, and then they are transformed into SFO variables for stator-current control and stator-voltage computation at the re-coupling side of the control scheme (see *Fig.16*):

a) The direct RFC ensures a good static stability due to the linearity of the mechanical characteristics at $\Psi_r = \text{ct.}$;

b) The decoupling control of the mechanical and magnetic phenomena realized by means of the RFO-ed components of the stator-current provides a good dynamic to the IM drive;

c) Based on SFO-ed two-phase model, the computation of the stator-voltage control variables is made in the simplest manner realized by the separation of the two kinds of EMFs according to equations (22).

The two kinds of control variables are coupled with the field-oriented stator-current components, by means of a CooT block, indicated with matrix operator $[D(\lambda_s\text{-}\lambda_r)]$. Its inputs contain the RFO-ed components $i_{sd\lambda r} - i_{sq\lambda r}$ and its outputs are the SFO-ed ones $i_{sd\lambda s} - i_{sq\lambda s}$. The deviation angle $\lambda_s\text{-}\lambda_r$ between the two orientation fluxes (see *Fig. 8*) is computed in another CooT block with inputs

[o(λ_s)] and [o(λ_r)], resulting from the VAs of the stator- and rotor-fluxes. The trigonometric functions required for the CooT blocks are symbolized with an "oscillatory" matrix containing two elements: [o(λ)] = [cos(λ), sin(λ)]^t [1].

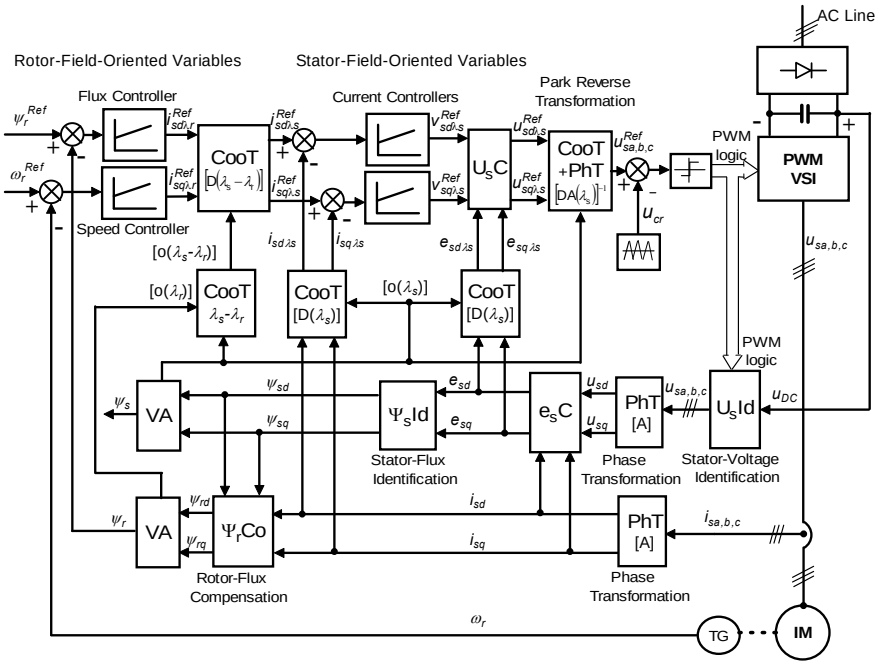


Figure 16: Vector control structure with dual-field-orientation of the short-circuited IM drive supplied from a voltage-feedforward PWM controlled VSI.

The stator-voltage control variables are computed in the U_sC block based on equations (9), where the input EMFs result from the feedback side and the *Ohm*'s law voltage drops are generated by the controllers of the SFO-ed current components. This structure eliminates the influence of the rotor parameters.

13. Conclusion

The RFC-ed IM with RFO-ed structure behaves similarly to a DC machine, both from point of view of dynamics and stability, due to the linear mechanical characteristics. The best control scheme seems to be a RFO with RFC and current-controlled converter as actuator. Compared to other structures, its dynamic response is superior, the computation requirements are reduced, and it is less dependent on the motor parameters. But the implementation of the current-feedback PWM presents difficulties.

Voltage-controlled VSI-fed drives (usually with SFC, either SC or VC structure), generally can not ensure the same performance, neither regarding stability and torque ripple, nor dynamics compared to RFC achieved by current-controlled VSI. This is due to the natural behavior of the IM, considering the magnetizing and torque producing phenomena.

The RFO with RFC for voltage-controlled converter-fed drives requires the highest computational capacity of the DSP, and – in addition –, the quality of the operation may suffer from the sensitivity to motor parameters, especially the coefficients of leakage and rotor time constant.

The SFO with SFC, especially used for voltage controlled converter-fed drives, is less computationally demanding and more robust, but the reaction to torque commands in low-inertia drives can lead to stability problems.

The DFO combines the advantages of the two types of field-orientation procedures for voltage-controlled IM drives, on the one hand of the RFC and RFO and on the other hand of the SFO. This combination ensures reduced computational demand, increased stability, a good dynamic and robustness, avoiding the influence of the rotor parameters.

Acknowledgements

Thanks to my colleagues, Ioan Iov INCZE, PhD and Csaba SZABÓ, PhD, co-authors of numerous published scientific papers, for the experimental background achieved by simulation and implementation of the control structures.

References

[1] Kelemen, Á., Imecs, M., "Vector control of AC drives", Vol. 1: "Vector control of induction machine drives", *OMIKK-Publisher*, Budapest, 1991.

[2] Imecs, M., Szabó, Cs., "Control structures of induction motor drives - state of the art", *WESIC 2003* Lillafüred, Ed. Miskolc University, pp. 495-510, 2003.

[3] Imecs, M., "How to correlate the mechanical load characteristics, PWM and field-orientation methods in vector control systems of AC drives", *Bulletin of Polytechnic Institute of Iassy*, Tomul XLVI (L), Fasc. 5, pp. 21-30, 2000.

[4] Blaschke, F., "Das Prinzip der Feldorientierung, die Grundlage für die Transvector-Regelung von Drehfeldmaschinen" (in German), *Siemens-Zeitschrift* 45, Heft 10, 1971.

[5] Bayer, K. H., Waldmann, H., Weibelzahl, H. D., "Die Transvektor-Regelung für den feldorientierten Betrieb einer Synchronmaschine", *Siemens-Zeitschrift* 45, Heft 10, 1971.

[6] Flöter, W. & Ripperger, H., "Das Transvector-Regelung für den feldorientierten Betrieb einer Asynchronmaschine" (in German), *Siemens-Zeitschrift* 45, Heft 10, 1971.

[7] Hasse, K., "Zur Dynamik drehzahlgeregelter Antriebe mit stromrichtergespeiste Asynchron-Kurzschlussläufer-maschinen" (in German), *PhD Dissertation, T. H. Darmstadt*, 1969.

[8] Leonhard, W., "Control of Electrical Drives", Springer Verlag, Berlin, 1985.

[9] Späth, H., "Steuerverfahren für Drehstrommaschinen" (in German), Springer Verlag, Berlin, 1977.

[10] Beierke S., "Rapid implementation of a field-oriented control method for fixed-point DSP controlled asynchronous servo drives", *EPE Chapter Symposium on Electric Drive Design and Applications*, Edited by EPFL Lausanne, pp. 361-365, 1994.
[11] Imecs, M., Incze, I. I., Szabó, Cs., "Control strategies of induction motor fed by a tandem DC link frequency converter", *Proceedings of the 9th EPE '01*, Graz, Austria, 2001, pp. L1b-7 (Abstract) & CD-ROM (Full paper).
[12] Kazmierkowski, M. P., Malinowsky, M., Sobczuk, D. L., Blaabjerg, F., Pedersen, J. K., "Simplified stator flux oriented control", *ISIE 1999*, Bled, Slovenia, pp. 474-479, 1999.
[13] Langweiler, F., Richter, M., "Flusserfassung in Asynchronmaschinen", *Siemens-Zeitschrift* 45, Heft 10, 1971.
[14] Böhm, K., Wesselak, F., "Drehzahlregelbare Drehstromantriebe mit Umrichterspeisung", *Siemens-Zeitschrift* 45, Heft 10, 1971.
[15] Járdán, R. K., Horváth, M., "Flux control in current source inverter drives", *International Conference on Electrical Machines ICEM`80*, Athens, Greece, 1980.
[16] Incze, I. I., Imecs, M., Szabó, Cs., Vásárhelyi, J., "Orientation-field identification in asynchronous motor drive systems", *6th IEEE-ICCC International Carpathian Control Conference*, Lillafüred, Ed. Uni of Miskolc, Vol I, pp. 131-136, 2005.
[17] Imecs, M., Trzynadlowski, A. M., Incze I. I., Szabó, Cs., "Vector control schemes for tandem-converter fed induction motor drives", *IEEE Transactions on Power Electronics*, Vol. 20, No. 2, pp. 493-501, 2005.
[18] Imecs, M., "Synthesis about pulse modulation methods in electrical drives, Part 1 and 2", *Proceedings of CNAE '98*, Craiova, pp. 19-33, 1998.
[19] Imecs, M., "Open-loop voltage-controlled PWM procedures", *Proceedings of ELECTROMOTION '99*, Patras, Greece, Vol. I, pp. 285-290, 1999.
[20] Imecs, M., Szabó, Cs., Incze, I. I., "Vector control of the cage induction motor with dual field orientation", *CINTI 2008*, Budapest, pp. 47-58, 2008.
[21] Imecs, M., Patriciu, N., Benk, E., "Synthesis about modelling and simulation of the scalar and vector control systems for induction motors", *Proceedings of International Conference ELECTROMOTION'97*, Cluj-Napoca, pp. 121-126, 1997.
[22] Imecs, M., "Villamos hajtások szabályozása mai szemmel" (in Hungarian), *Proceedings of the First International Conference on Energetics and Electrotechnics ENELKO'2000*, Ed. by EMT, Cluj-Napoca, pp. 7-16, 2000.
[23] Imecs, M., Szabó, Cs., Incze, I. I., "Four quadrant drives for AC machines fed by frequency converters" (in Hungarian) *Proceedings of 3th International Conference of Energetics and Electrical Engineering ENELKO*, Ed. by EMT, Cluj-Napoca, pp. 53-59, 2002.
[24] Imecs, M., Incze, I. I., Szabó, Cs., Ádám, T., "Scalar and vector control structures of AC drives" (in Hungarian), *Proceedings of 4th Conference on Energetic and Electrical Engineering ENELKO*, Ed. by EMT, Cluj-Napoca, pp. 82-98, 2003.
[25] Imecs, M., Incze, I. I., Szabó, Cs., Ádám, T., Szőke Benk E., "Line-friendly DC-link frequency converters for low and high power AC drives", Plenary paper (in Hungarian), *Proceedings of 5th International Conference of Energetics and Electrical Engineering ENELKO'2004*, Ed. by EMT, Cluj-Napoca, pp. 86-96, 2004.
[26] Incze, I. I., Imecs, M., Mátis, St., Szabó, Cs., "Up-to-date experimental rig for development of controlled AC electrical drives" (in Hungarian), *Proceedings of 6th International Conference of Energetics and Electrical Engineering ENELKO*, Ed. by EMT, Cluj-Napoca, pp. 62-68, 2005.
[27] Incze, I. I., Szabó, Cs. Mátis I., Imecs, M., Zoltán, E., "Implementation on experimental rig of control for AC electrical drives" (in Hungarian), *Proceedings of 6th International Conference of Energetics and Electrical Engineering ENELKO*, Ed. by EMT, Cluj-Napoca, pp. 69-75, 2005.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 29-40

Experimental Investigation on Robust Control of Induction Motor Using H_∞ Output Feedback Controller

Dénes FODOR

Department of Electrical Engineering and Information Systems,
Faculty of Information Technology
University of Pannonia, Veszprém, Hungary,
e-mail: fodor@almos.vein.hu

Manuscript received March 15, 2009; revised June 15, 2009.

Abstract: This paper deals with H_∞ controller design and real-time experimentation of it for three-phase induction motor control. The model of the motor is given in its state space representation in d-q reference frame. An H_∞ feedback controller is used in the speed loop, which was synthesized by combining a full information controller with an estimator to achieve the desired results. Simplification on the model during the design process has been considered as modelling noise of the system. There are exogenous inputs considered as disturbances, which are not correlated with the measurement noise. The controller was synthesized to minimize the effects of the disturbances entering the plant and the influence of the measurement noise and modelling errors. The desired controller is given by a state space model. The simulation of the system was done in MATLAB/Simulink. The controller is realized as a Simulink embedded S-function. The real-time implementation of the proposed structure has been done on the dSPACE DS1102 DSP development board, with very promising results, showing good reference tracking and good dynamical behaviour.

Keywords: Robust control, induction motor, field-oriented control, estimation technique, implementation issues.

1. Introduction

In the past few years the performance of computers has increased dramatically. With this increased performance it is possible to develop more complex control systems for real life applications. These new control methods are more robust and more reliable than the others, because they can handle complex plant models. The H_∞ theory is a new method and only a few scientific papers with the application to induction motors can be found in the scientific literature [1]. The goal is to demonstrate the practical applicability of the H_∞ controller in an industrial environment. The electrical drive with three phase asynchronous motor is widely used in industry, its main drawback is the

difficult control possibility, so it is ideal for a benchmark application to prove the usefulness of H_∞ control theory.

2. Induction motor model

Table 1 lists the symbols used in this article. Symbols representing vectors are underlined.

Table 1: List of symbols

Meaning	*Notation*	*Meaning*	*Notation*
Control input	*u, u(t)*	Stator current (d,q)	i_{Sd}, i_{Sq}
Disturbances, noise	*w, w(t)*	Rotor speed	ω
Output	*y, y(t)*	Flux speed, angle	ω_{mR}, ε
Measurement	*m, m(t)*	Noise, reference	*n, r*
Stator voltage (d,q)	u_{Sd}, u_{Sq}	Leakage factor	σ
Magnetizing current	i_{mR}	Rotor, stator time constant	T_R, T_S
Number of pole pairs	*p*	Inertia	*J*
Load torque	m_L	Rotor flux vector	$\underline{\Psi}_R$

The dynamic behaviour of the induction motor is described by a set of nonlinear so called general equations. Differential equations (1)-(11) describe the behaviour of the AC motor in the rotor-field-oriented synchronously rotating d-q reference frame [2].

$$\frac{d}{dt} i_{Sd}(t) = \eta_1 i_{Sd}(t) - \eta_2 i_{mR}(t) + \omega_{mR} i_{Sq}(t) + \eta_3 u_{Sd}(t); \tag{1}$$

$$\frac{d}{dt} i_{mR}(t) = \frac{1}{T_R} i_{Sd}(t) - \frac{1}{T_R} i_{mR}(t); \tag{2}$$

$$\frac{d}{dt} i_{Sq}(t) = \eta_4 i_{Sq}(t) - \eta_5 \omega_{mR}(t) i_{mR}(t) - \omega_{mR}(t) i_{Sd}(t) + \eta_3 u_{Sq}(t); \tag{3}$$

$$\frac{d}{dt} \omega(t) = \frac{2}{3} \frac{p^2}{J} (1-\sigma) L_S i_{mR}(t) i_{Sq}(t) - \frac{z_p}{J} m_L; \tag{4}$$

$$\underline{i}_{mR}(t) = \frac{1}{L_H} \underline{\Psi}_R(t) e^{j\varepsilon(t)}; \tag{5}$$

$$\omega = \omega_{mR} - \frac{i_{Sq}}{T_R i_{mR}}; \tag{6}$$

$$\eta_1 = -\frac{1}{\sigma T_S} - \frac{(1-\sigma)}{\sigma}\frac{1}{T_R}; \tag{7}$$

$$\eta_2 = -\frac{(1-\sigma)}{\sigma}\frac{1}{T_R}; \tag{8}$$

$$\eta_3 = \frac{1}{\sigma L_S}; \tag{9}$$

$$\eta_4 = -\frac{1}{\sigma T_S}; \tag{10}$$

$$\eta_5 = -\frac{(1-\sigma)}{\sigma}; \tag{11}$$

The above equations result in a nonlinear, time-variant state-space representation. For the controller synthesis let us assume that during normal operation of the drive the modulus of the flux is constant. From (5) it results that $i_{mR}=i_{mR}^{ref}$ is constant, and so $\frac{d}{dt} i_{mR}(t) = 0$ in (2). Then i_{Sd} is equal to i_{mR}, and so $\frac{d}{dt} i_{Sd}(t) = 0$. (1) is then not a differential equation, but an algebraic one. Substituting (6) into (3) and (4), and using $i_{mR}=i_{mR}^{ref}$, $i_{Sd}= i_{mR}$, it results the following simplified state space representation of the AC drive:

$$\frac{d}{dt}\begin{bmatrix} i_{Sq} \\ \omega \end{bmatrix} = A\begin{bmatrix} i_{Sq} \\ \omega \end{bmatrix} + B_u\left[u_{Sq}\right] + B_w\begin{bmatrix} m_L \\ n \end{bmatrix}; \tag{12}$$

$$A = \begin{bmatrix} -\frac{1}{\sigma T_S} - \frac{1-\sigma}{\sigma}\frac{1}{T_R} - \frac{1}{T_R} & -\frac{1-\sigma}{\sigma} i_{mR}^{ref} - i_{mR}^{ref} \\ \frac{2}{3}\frac{p^2}{J}(1-\sigma)L_S i_{mR}^{ref} & 0 \end{bmatrix}; \tag{13}$$

$$B_u = \begin{bmatrix} \frac{1}{\sigma L_S} \\ 0 \end{bmatrix}; \tag{14}$$

$$B_w = \begin{bmatrix} 0 & N_1 \\ -\frac{p}{J} & N_2 \end{bmatrix}; \tag{15}$$

The load torque is assumed to be a disturbance, and n represents the disturbances resulting from the imprecise definition of the reference value i_{mR}^{ref}, cross effects between the state variables, the general system noise and the unmodelled dynamics of the system. Coefficients are bounded and can be determined by taking the upper bound of the disturbances modulus. The model presented in (12) is the nominal plant, which will be used for controller design. The controller will be used for the original model (1)-(11), that is the perturbed plant presented in *Fig. 1*.

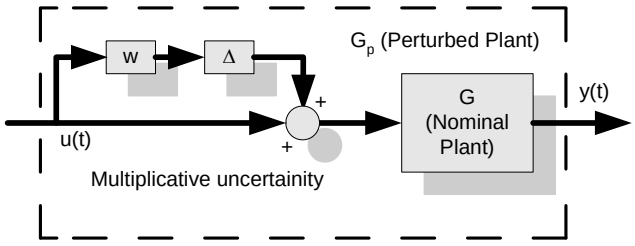


Figure 1: Model used in controller design.

The only input of the model (12) is the q component of the stator voltage.

3. The H_∞ controller design

The goal is to design an H_∞ controller for the three-phase asynchronous motor in the speed loop. The assumption is that there is noise entering the plant, so the estimator part of the controller was used, too. The plant is described in general case by the following equations [3], [4]:

$$\dot{x}(t) = Ax(t) + \begin{bmatrix} B_u & B_w \end{bmatrix} \begin{bmatrix} u(t) \\ w(t) \end{bmatrix} \tag{16}$$

$$\begin{bmatrix} m(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} C_m \\ C_y \end{bmatrix} x(t) + \begin{bmatrix} 0 & D_{mw} \\ D_{yu} & 0 \end{bmatrix} \begin{bmatrix} u(t) \\ w(t) \end{bmatrix} \tag{17}$$

The following conditions need to be satisfied [5]:

- $D_{mw} B_w^T = 0$;
- $D_{mw} D_{mw}^T = I$;
- $D_{yu}^T C_y = 0$;
- $D_{yu}^T D_{yu}^T = I$;

- The plant is controllable from the control input and from the disturbance input;
- The plant is observable from the measured output and from the reference output.

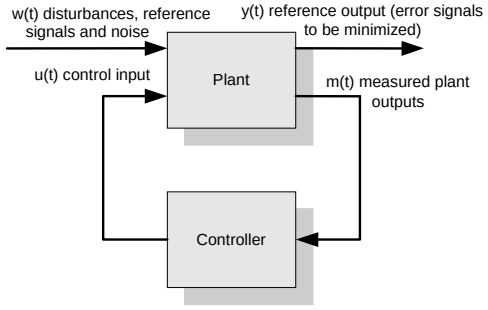


Figure 2: General control configuration.

If these requirements are satisfied, a controller can be designed according to the general control configuration in *Fig. 2*. The internal structure of the controller is shown in *Fig. 3*. The estimator part estimates the state and the feedback part generates the control input. The solution of the finite-time steady state suboptimal H_∞ output control problem can be given by solving two Riccati equations. The suboptimal controller is defined according to (18), the matrices can be calculated using the solutions of the algebraic Riccati equations [3].

$$\begin{aligned} \dot{x}_c(t) &= A_c(t)x_c(t) + B_c(t)m(t) \\ u(t) &= C_c(t)x_c(t) \end{aligned} \quad . \tag{18}$$

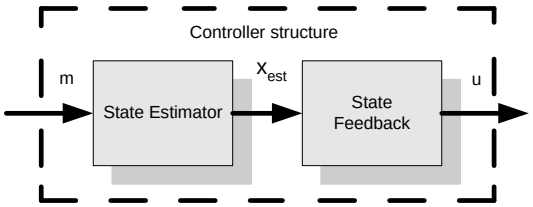


Figure 3: Structure of the controller.

4. Adaptation of the motor model to the H_∞ control method

The model presented in (12) is appropriate for the control synthesis. To guarantee good reference tracking, the speed-reference input is considered as a disturbance input and all coefficients of this input in the equation (12) are set to zero, since the reference input has no direct influence on the equations of the motor. The output equations need to be formulated to have a representation as in

(16) and (17). Note that, since all parameters are fixed, the description is time-invariant. The measured value is the difference between the reference speed and the actual speed together with the measurement noise. The reference output consists of the difference between reference and actual speed (without noise), i_{Sq}, and u_{Sq}. The controller should keep the control input finite and should not try to set it to zero, so the weights for the stator current and voltage in the reference output are kept small. Note that the disturbance input is included in the reference output, which is not consistent with the H_∞ problem statement as in (16) and (17). This drawback can be solved by applying more general suboptimal control formulas [6]. In order to make calculations easier, a steady state gain was used to design the controller. Some performance limitations have to be made [5] to guarantee good reference tracking (after the design ad hoc integral action was added [3]) and the output signals were bounded. The final model of the system for controller design results:

$$\frac{d}{dt}\begin{bmatrix} i_{Sq} \\ \omega \end{bmatrix} = A\begin{bmatrix} i_{Sq} \\ \omega \end{bmatrix} + B_u\begin{bmatrix} u_{Sq} \end{bmatrix} + B_{wf}\begin{bmatrix} m_L \\ n \\ r \end{bmatrix}; \tag{19}$$

$$\begin{bmatrix} m(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0.01 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} i_{Sq} \\ \omega \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} u_{Sq} \\ m_L \\ n \\ r \end{bmatrix}; \tag{20}$$

$$B_{wf} = \begin{bmatrix} 0 & N_1 & 0 \\ -\frac{p}{J} & N_2 & 0 \end{bmatrix}. \tag{21}$$

The controller has the form (18). After solving the Riccati equations, the matrices of the controller A_c, B_c, C_c can be calculated using the solutions of the equations and the state matrices of the model. A performance bound required for the solution was chosen approximately 10% over theoretical optimum.

5. Simulation results

Parameters of the induction drive used in the simulation:

$L_s = 0.13$ H, $L_h = 0.12$ H, $L_r = 0.13$ H, $R_r = 3.0\ \Omega$, $R_s = 1.86\ \Omega$, $p = 2$.

The simulation was made with MATLAB/Simulink, according to the structure presented in *Fig. 4*. The model representing the AC drive is a time-

variant model described by the equations (1)-(11), and not the simplistic model used for controller design. The inputs of the motor are three phase stator voltages and the load torque. The controller provides the value of u_{sq}, while u_{sd} is kept constant. Flux computation, presented in *Fig. 5*, was used in order to get the angle of the rotor flux, which would allow the transformation of the voltages from d-q reference with reverse Park- and reverse Clarke-transformation into the a, b, c three-phase components.

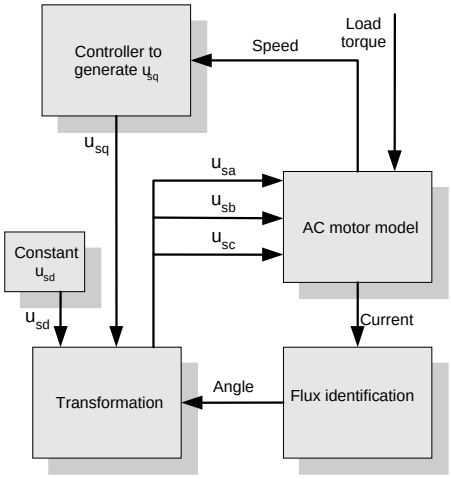


Figure 4: The structure of the drive model.

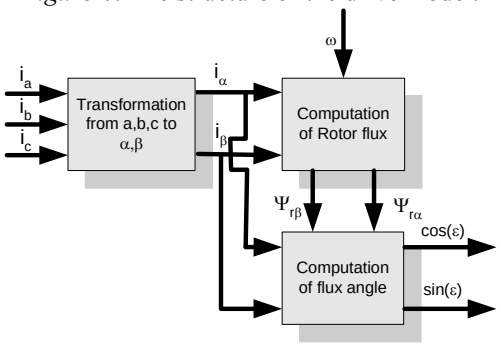


Figure 5: Flux identification used in the simulation.

The simulation results are presented in *Fig. 6-8*. *Fig. 6* shows the reference tracking capability, even with large measurement noise. The load torque is presented in *Fig. 7*. The simulation results show that the H_∞ controller works well, and the system with rapidly changing reference shows good dynamic behaviour.

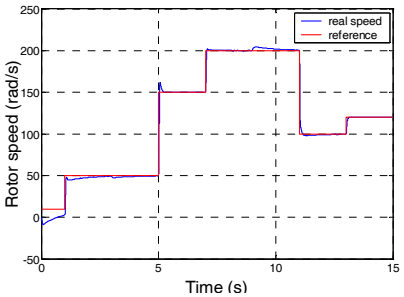


Figure 6: Rotor speed.

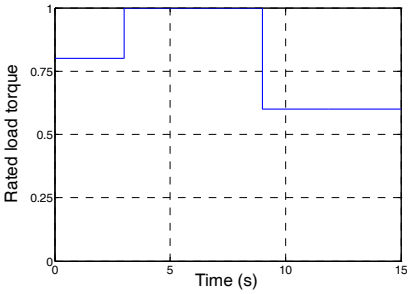


Figure 7: Load torque.

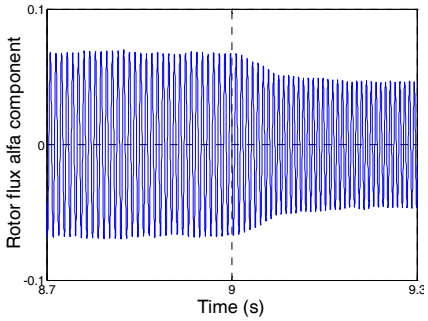


Figure 8: Rotor flux *d* component.

From the simulation results presented in *Fig. 9-10*, it is clear that the controller shows good results even when load torque changes rapidly. Note, that during controller design, the load was considered as disturbance; according to (4), it influences the rotor speed more than the measurement noise. Thanks to the robustness of the H_∞ controller, the measurement errors and modeling errors are compensated.

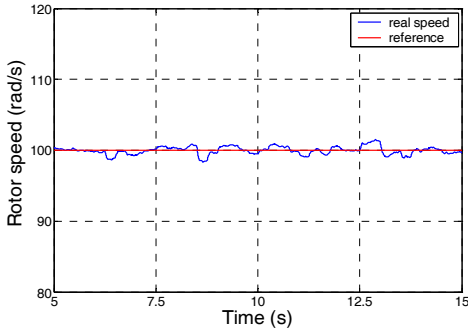


Figure 9: Reference tracking (100 rad/s) during rapid variation of the load.

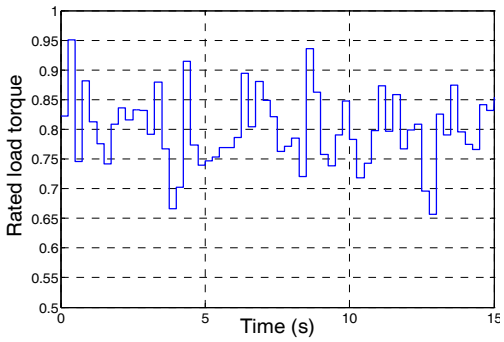


Figure 10: Rapidly changing load.

6. Testing the dynamic behaviour of the induction motor

With the experimental setup based on the dSPACE DS1102 development board, it is possible to study the dynamic behaviour of the motor in real-time. Note, that the motor model is tested without controller; supplied by three sine-wave generators, with amplitude and frequency values programmable via a graphical user interface. The value of the load torque can be also arbitrarily changed (see *Fig. 11)*. The values of the rotor flux, stator current and speed are displayed on plotters. Robustness tests can be made by changing the position of the slider bar indicating the value of the rotor resistance. By changing this value, the parameters of the transfer functions are also changed. These are displayed numerically on the right.

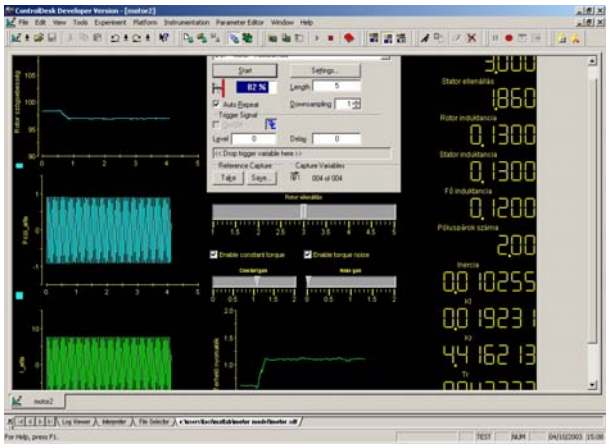


Figure 11: Test of the motor model.

7. Experimental results on the DS1102

In the second experiment, the motor model is tested with the controller whose design was presented in section 3 of the paper. *Fig. 13* shows the measured values when executing the real-time code. There are four plotters on the screen. The top-left shows the reference and actual speed values (it can be said that the controller shows good tracking capabilities). The other displays show the value of the rotor flux, the tracking error and the load torque. There are two slide bars at the bottom. The left slide bar enables the user to set the value of the load torque during code execution. The right slide bar can be used to set the magnetic operating point (the desired flux value) and thus test how it affects the control process. With these and other similar GUI-s functionalities, all aspects and effects of parameter changes can be studied in real-time. The functionality of the system will be shortly discussed as follows.

The ds1102 board (with TMS320C31 processor) is a PC card designed for development of high-speed multivariable digital controllers equipped with analogue to digital and digital to analogue converters. ControlDesk software [7] provides a framework to manage the DSP board [8]. It has a real-time interface (RTI) to MATLAB/Simulink and additional blocksets as well. With this interface, real-time code can be easily built, downloaded and executed. Instrument panels (also called layouts, presented later) can be easily made using ControlDesk, allowing the change of parameters and real-time display of all state-variables and signals in the system. The main advantage of the ControlDesk is that the code -executable on the DSP- can be directly generated from the Simulink model of the system. This can be done in the following way. The model of the system is built in Simulink using the original Simulink blocks (and using only the blocks which are fully supported by ControlDesk), and blocks handling the board's hardware such as in- and outputs, interrupts. As it was mentioned before, there are blocks, which can not be used with RTI. These blocks must be replaced by combinations of other blocks before the code is generated. Using RTI, several different settings are possible. The generated code can be single- or multitasking, block reduction can be switched on or off, signals can be reused, and so on. The best results can be achieved when all parameters are *tuneable*.

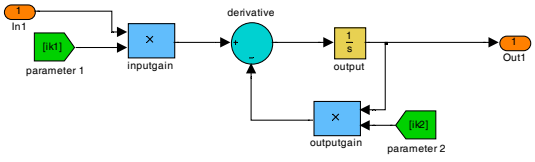


Figure 12: Block diagram of the transfer function.

In order to do this, it is necessary to replace transfer functions with parameter coefficients (such as 1/(*Ts*+1) for example) with a combination of blocks from integrators and gain blocks (see *Fig. 12*). To be able to change the parameters on-line, the parameters have to be masked (or else they are not accessible from ControlDesk during code execution). By changing parameters, it becomes possible to make robustness tests (by changing the induction motor parameters and thus simulating parametric uncertainty), as well as to tune the controller and to give different input signals to the systems input. With the user interface, parameters can be easily tuned, values displayed or saved. First, the controller is simulated with the mathematical model of the system built in Simulink. After having good results, the blocks, representing the theoretical model of the induction motor, are replaced by blocks representing the real induction motor (such blocks as inputs and outputs, interrupt handler blocks and so on). After the blocks are replaced, the system can be tested with the real induction motor.

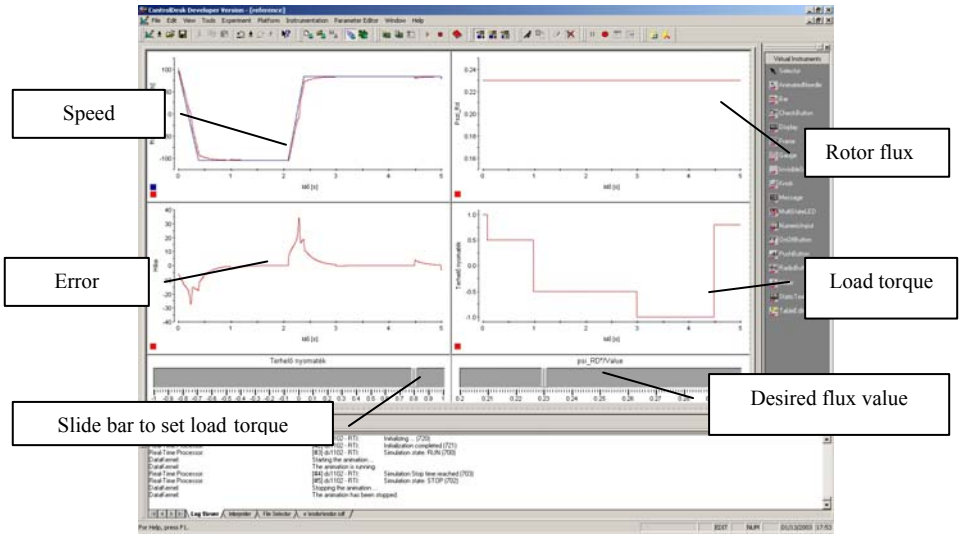


Figure 13: Experimental results displayed during real-time code execution.

8. Conclusion

According to the theoretical background presented in section 2 and 3, an H_∞ controller was designed by making some simplifications on the motor model and then the controller's behaviour was tested with it. It was shown that all assumptions, simplifications made during the design process are valid. The system showed good dynamic behaviour and it was shown how the performance (noise rejection, robustness, and tracking) depends on the design parameters. In

consequence, the controller can be implemented for real-time usage as the last sections show. The controller can be further developed by taking into account more disturbances. Remark that actuator errors (resulting from the switching behaviour of the inverter) and modelling noise were not used in the controller design, and current measurements were assumed to be precise (it was assumed to measure them without noise). The performance can be further improved by using a flux model, which is further developed so that it gives more precise values during transient processes, or by using a speed estimator [9]. Since the motor model is nonlinear, a parameter-variant or nonlinear controller can be taken into account as well [10].

References

[1] L. Rambault, C. Chaigne, G. Champenois, S. Cauet, "Linearization and H_∞ controller applied to an induction motor", EPE, 2001-Graz.

[2] "Field Oriented Control of 3-Phase AC-motor", *Texas Instruments Europe*, 1998.

[3] Burl, J. B., "Linear Optimal Control H_2 and H_∞ methods", *Addison Wesley Longman Inc.*, 1999.

[4] J. C. Doyle, K. Glover, P.P. Khargonekar, B.A. Francis ,"State-Space Solutions to Standard H_2 and H_∞ Control Problems", *IEEE Trans. Automat. Contr.* , vol. 34, no. 8, August 1989.

[5] Skogestad, S., Postlethwaite, I., "Multivariable Feedback Control Analysis and Design", *John Wiley & Sons Ltd.*, 1996

[6] K. Zhou, J. C. Doyle, K. Glover, "Robust and Optimal Control", *Prentice-Hall*, 1996.

[7] Controldesk Experiment Guide, *dSPACE GmbH* (2001).

[8] DS1102 User Guide, *dSPACE GmbH* (2001).

[9] Brunsbach, B.-J., Henneberger, G. (1990). Einsatz eines Kalman-Filters zum feldorientierten Betrieb einer Asynchronmaschine ohne mechanische Sensoren, *Archiv für Elektrotechnik.*

[10] Prempain, E., Postlethwaite, I., Benchaib, A. (2002). A linear parameter variant H_∞ control design for an induction motor, *Control Engineering Practice.*

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 41-52

Lyapunov-Based Frequency-Shift Power Control of Induction-Heating Converters with Hybrid Resonant Load

András KELEMEN[1], Nimród KUTASI[2]

[1, 2] Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tîrgu Mureş, Romania,
e-mail: kandras@ms.sapientia.ro, kutasi@ms.sapientia.ro

Manuscript received March 15, 2009; revised June 25, 2009.

Abstract: The frequency-shift method is an attractive choice for power control of induction-heating inverters due to the simplicity of the power circuit. However, frequency-shift control proves to be a challenging task in case of practical resonant loads with high quality factor and uncertain circuit parameters. The paper presents a bilinear large-signal model of the induction-heating inverter with hybrid LLC resonant load. A control law is proposed, which is based on the Lyapunov stability theory. Moreover, an adaptive control method is presented to handle the uncertainty concerning the nominal values of the state variables. The theoretical results are illustrated by numerical simulation.

Keywords: Induction heating, resonant inverter, energy in the increment, Lyapunov stability, d-q model.

1. Introduction

The technological task is heating metals by means of eddy currents induced directly in the work-piece by a strong and variable magnetic field produced by a high-intensity alternative current flowing through an inductor. The inductor is a component of a resonant circuit fed by a power electronic load-resonant inverter. The inductor is designed to fulfill the main technological requirement of heating the work-piece. For this purpose a magnetic field is necessary with a certain distribution in space and evolution in time, able to produce by eddy current losses the required heat pattern.

The power electronic converter has to deliver power at frequencies that are close to the resonance frequency of the load. The task is finding power control methods able to create low loss switching conditions of the power semiconductor devices, while keeping reduced complexity of the heating equipment [1],[2].

This paper proposes a power control method of load-resonant inverters by means of the operation frequency, named "frequency-shift control", based on the Lyapunov stability theory.

The paper is organized as follows. Section 2 presents challenges of frequency-shift control of induction-heating voltage inverters. The bilinear large-signal low-frequency d-q model of the high-frequency inverter with hybrid LLC resonant load is introduced in Section 3. Construction of Lyapunov-based control laws for the above system is presented in Section 4 in case of known circuit parameters and known steady-state (nominal) values of the state variables. Section 5 introduces an estimation method for handling uncertain nominal values of the state variables. The proposed methods are verified by MatLab Simulink simulation. Finally, concluding remarks are presented in Section 7.

2. Considerations on frequency-shift power control

Schematic of the induction-heating converter with voltage-fed load-resonant inverter is shown in *Fig. 1.a*. The frequency-shift power control method is based on the frequency characteristics of the resonant load (*Fig. 1.b* presents the frequency characteristics of a particular load). Consequently the control characteristics are very much influenced by the load parameters and the power control range is relatively reduced. Some of these parameters, like L_s and C_p are known accurately and are subject only to small variations during the heating process. The L_{is} inductance of the heating inductor is generally known with less accuracy, and is subject to relatively small changes during the heating process. These changes are reflected in relatively small variations of the resonance frequency. However, large variation of power may result especially in case of high quality factor. The most uncertain load parameter, which is also most influenced during the heating process, is the R_{is} equivalent resistance of the inductor with work-piece. This parameter may change by an order of magnitude due to large variations with temperature of the resistivity and permeability, which are anyway known with low accuracy. In the same time, permeability is much dependent on the intensity of the magnetic field.

A classical frequency-shift control structure is shown in *Fig. 2*.

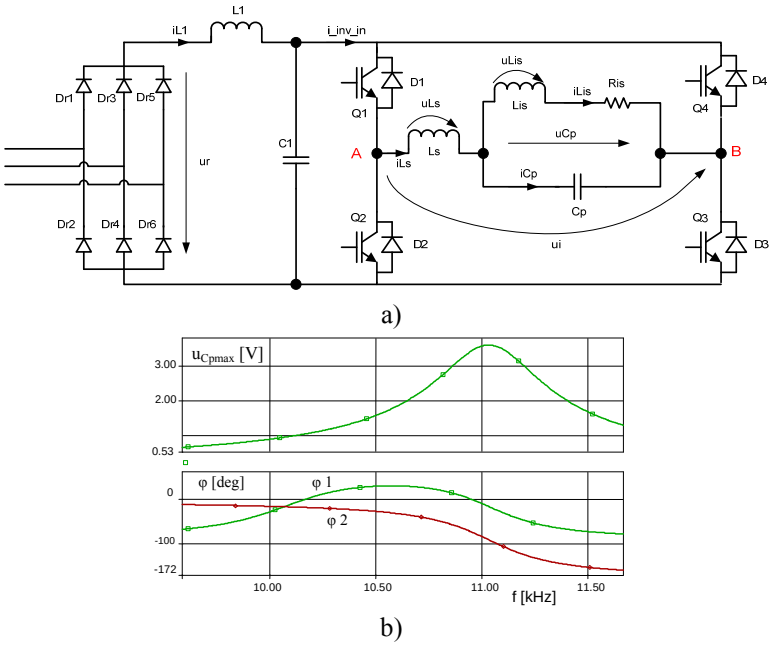


Figure 1: Schematic of the induction-heating converter (a) and frequency characteristics of the LLC hybrid resonant load (b). $u_{Cp\max}$ is the amplitude of the capacitor tank voltage, φ_1 denotes the phase shift between the inverter output current and voltage, while φ_2 denotes the phase shift between the capacitor tank voltage and inverter output voltage.

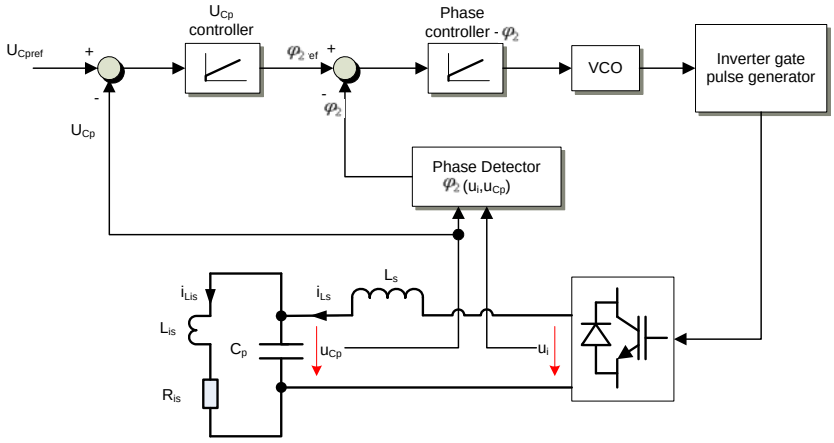


Figure 2: Traditional cascade-type frequency-shift control structure.

3. The model of the load-resonant inverter

For investigation of power control by means of frequency, we need converter models that are simple enough to be included in dynamic models of frequency and power control structures. Modeling of power electronic converters generally needs a hierarchic approach and development of modeling techniques tailored to the specific problem to be solved. In our approach rules are established that assure certain desired (soft) switching conditions by proper choice of the switching instant. Only those power control methods are accepted, that follow these rules. From this moment, control design, i.e. analysis of long-term closed-loop operation is made caring neither about the switching process, nor about the instantaneous values of the circuit variables.

In case of the frequency-shift control, the resonant load is fed by the inverter with full square-wave output voltage, thus the switching instant is strongly coupled with the switching frequency. The aim is the derivation of a model for the resonant inverter, which allows the analysis of the envelope of circuit variables (only variation of magnitudes is of concern). The model is built for continuous current mode operation, valid in case of frequency-shift control. The inverter output voltage is represented by its fundamental component.

The low-frequency model (*Fig. 4b* and *Fig. 5*) is obtained by eliminating the high-frequency terms from the equations written for the complex high-frequency circuit (*Fig. 4a*), which is derived from the original and the orthogonal circuit (*Fig. 3*) [3].

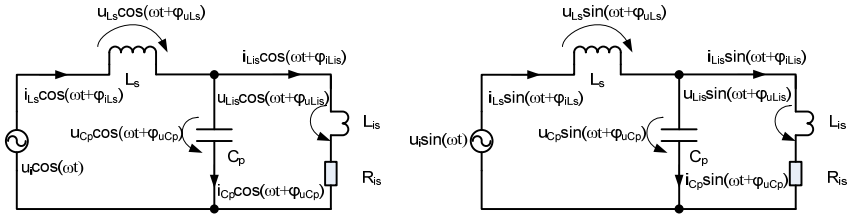
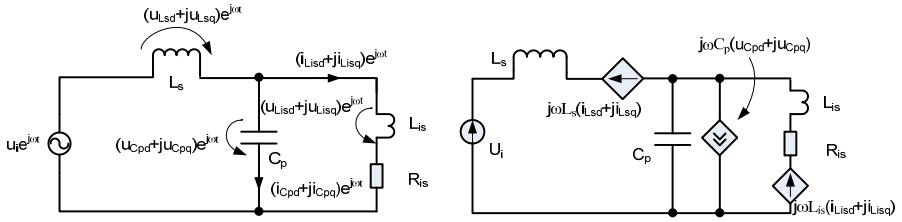


Figure 3: The inverter with resonant load and its orthogonal circuit.



a.) b.)

Figure 4: The complex representation (a) and the low-frequency complex d-q model of the resonant inverter (b).

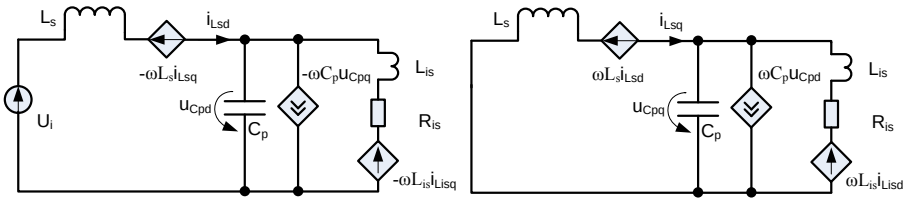


Figure 5: Low-frequency d and q circuits.

4. The Lyapunov-based frequency-shift power control

Using the non-linear model introduced in the previous section, we present an approach for development of the frequency-shift control law based on the Lyapunov stability theory.

The inverter has the structure from *Fig. 1.a*, with circuit parameters:

L_s= 20μH, L_{is}= 3.95μH, R_{is}= 0.03Ω, C_p= 63μF.

The amplitude of the rectangular inverter output voltage is assumed to be constant, with the peak value of its fundamental: u_{i1max}= 210V (lower-case notation has been used on purpose, to indicate the time-variable character of the amplitude).

The frequency characteristics of the inverter output current and load capacitor voltage are shown in *Fig. 1.b*.

Denoting by $\tilde{x}$ the -not necessarily small- deviation (increment) of a state variable x_T from its steady-state value x_n, it results:

$$[x_T]=[x_n]+[\tilde{x}] \tag{1}$$

In a similar way, in case of the angular frequency it results:

$$\omega=\omega_n+\tilde{\omega} \tag{2}$$

The state space equation system of the increments has the bilinear matrix form

$$[\dot{\tilde{x}}]=[A][\tilde{x}]+([B][\tilde{x}]+[b])\tilde{\omega}\,, \tag{3}$$

with matrices detailed in (6).

A possible and advantageous choice of Lyapunov function candidate is the "energy in the increment" [4], [5]:

$$V(\tilde{x})=\frac{1}{2}[\tilde{x}]^T[Q][\tilde{x}]\,, \tag{4}$$

with the positive defined diagonal matrix:

$$[Q] = diag\left([L_s\ L_s\ C_p\ C_p\ L_{is}\ L_{is}]\right). \tag{5}$$

$$[A] = \begin{bmatrix} 0 & \omega_n & -\frac{1}{L_s} & 0 & 0 & 0 \\ -\omega_n & 0 & 0 & -\frac{1}{L_s} & 0 & 0 \\ \frac{1}{C_p} & 0 & 0 & \omega_n & -\frac{1}{C_p} & 0 \\ 0 & \frac{1}{C_p} & -\omega_n & 0 & 0 & -\frac{1}{C_p} \\ 0 & 0 & \frac{1}{L_{is}} & 0 & -\frac{R_{is}}{L_{is}} & \omega_n \\ 0 & 0 & 0 & \frac{1}{L_{is}} & -\omega_n & -\frac{R_{is}}{L_{is}} \end{bmatrix}, \quad [B] = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$$[\tilde{x}] = [\tilde{i}_{Lsd}\ \tilde{i}_{Lsq}\ \tilde{u}_{Cpd}\ \tilde{u}_{Cpq}\ \tilde{i}_{Lisd}\ \tilde{i}_{Lisq}]^T, \tag{6}$$

$$[b] = [I_{Lsqn}\ -I_{Lsdn}\ U_{Cpqn}\ -U_{Cpdn}\ I_{Lisqn}\ -I_{Lisdn}]^T$$

Generally, for a positive scalar α, global stability of the bilinear system (3) is guaranteed by the control law:

$$\tilde{\omega} = -\alpha([B][\tilde{x}] + [b])^T [Q_1][\tilde{x}], \tag{7}$$

which assures a negative derivative of the function (4) along the system trajectories [5]. A short proof of this control law is given in [6], based on

$$\frac{d}{dt}(V([\tilde{x}])) = \frac{1}{2}[\tilde{x}]^T([A]^T[Q_1] + [Q_1][A])[\tilde{x}] + \frac{1}{2}\tilde{\omega}[\tilde{x}]^T([B]^T[Q_1] + [Q_1][B])[\tilde{x}] + \frac{1}{2}\tilde{\omega}([b]^T[Q_1][\tilde{x}] + [\tilde{x}]^T[Q_1][b]). \tag{8}$$

The existence of a symmetric, positive defined Q_1 that makes negative the first term of the derivative is guaranteed according to the Lyapunov equation:

$$[A]^T[Q_1] + [Q_1][A] = -[P][P]^T, \tag{9}$$

where $\{[P]^T, [A]\}$ is an observable pair.

The sum of the other two terms, equal to

$$N = \tilde{\omega}([\tilde{x}]^T[B]^T[Q_1][\tilde{x}] + [b]^T[Q_1][\tilde{x}]) \tag{10}$$

can be made negative by choosing the control law:

$$\tilde{\omega} = -\alpha([\tilde{x}]^T[B]^T[Q_1][\tilde{x}] + [b]^T[Q_1][\tilde{x}]) = -\alpha([B][\tilde{x}] + [b])^T[Q_1][\tilde{x}]. \tag{11}$$

The equation (9) is satisfied choosing $[Q_1] = [Q]$ (12).

$$[A]^T[Q] + [Q][A] = -R_{is}\,diag[0\ 0\ 0\ 0\ 1\ 1] \le 0, \tag{12}$$

a natural result taking into account that the resonant load is dissipative.

By direct calculation,

$$[B]^T[Q]+[Q][B]=[0] \tag{13}$$

and the control law takes the form

$$\tilde{\omega}=-\alpha[b]^T[Q][\tilde{x}]. \tag{14}$$

In order to assure monotonic control characteristics, the operation frequency is lower-limited by $\omega_{\inf}$, larger than the resonance frequency, and the control law becomes

$$\tilde{\omega}=\begin{cases}-\alpha[b]^T[Q][\tilde{x}] & for \quad -\alpha[b]^T[Q][\tilde{x}]>\omega_{\inf}-\omega_n \\ \omega_{\inf}-\omega_n & for \quad -\alpha[b]^T[Q][\tilde{x}]\le\omega_{\inf}-\omega_n\end{cases}. \tag{15}$$

Indeed, $\omega=\omega_n+\tilde{\omega}\ge\omega_{\inf}$ implies that $\tilde{\omega}\ge\omega_{\inf}-\omega_n$.

V(x) remains a Lyapunov function even in the saturation range of (15) because for positive values of α it results

$$-\alpha[b]^T[Q][\tilde{x}]\le\omega_{\inf}-\omega_n \quad\Rightarrow\quad [b]^T[Q][\tilde{x}]\ge\frac{\omega_n-\omega_{\inf}}{\alpha}>0, \tag{16}$$

and the third term of (8) is

$$\tilde{\omega}\left([b]^T[Q][\tilde{x}]\right)\le(\omega_{\inf}-\omega_n)\frac{\omega_n-\omega_{\inf}}{\alpha}\le 0, \tag{17}$$

while $[b]^T[Q][\tilde{x}]=0$ implies $\tilde{\omega}=0$, and $\tilde{x}=0$.

Substituting the terms from (6) into (14), it results:

$$\tilde{\omega}=-\alpha\left(L_sI_{Lsqn}\tilde{i}_{Lsd}-L_sI_{Lsdn}\tilde{i}_{Lsq}+C_pU_{Cpqn}\tilde{u}_{Cpd}-C_pU_{Cpdn}\tilde{u}_{Cpq}+L_{is}I_{Lisqn}\tilde{i}_{Lisd}-L_{is}I_{Lisdn}\tilde{i}_{Lisq}\right) \tag{18}$$

The control law has been implemented in Matlab Simulink. *Fig. 6* shows the open-loop response of the LLC load to a step variation of the inverter output voltage, along with closed-loop evolution. *Fig. 6* shows the evolution of the u_{Cpmax} voltage towards a frequency which is lower (*Fig. 6.a*), respectively larger (*Fig. 6.*b) than the resonance frequency. It should be mentioned that knowledge of the steady-state values of the variables has been assumed. In case of the inverter shown in *Fig. 1* these are known from steady-state simulation results:
I_{Lsdn} = 182.2 A, I_{Lsqn} = -26.4 A, U_{Cpdn} = 176.3 V, U_{Cpqn} = -242.7 V,
I_{Lisdn} = -835.6 A, I_{Lisqn} = -765.9 A, ω_n = 66571 rad/sec, u_{i1maxn}=211.5V.

However, the main drawback of the above control method is the poor knowledge of the steady-state values, which explicitly appear in the control law (18). One reason of this uncertainty is the variation of the inductor parameters during the heating process.

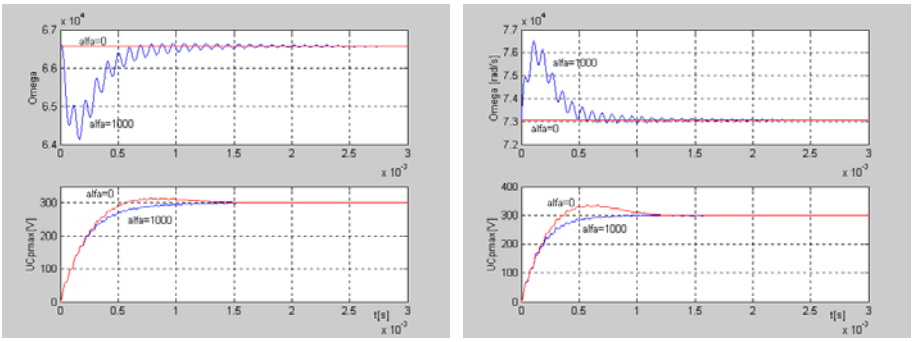


Figure 6: Comparison of the inverter's step responses for different gains α of the control law (18). α=0 means open-loop operation, i.e. response of the load to an "a.c. voltage step". In case (a) this means u_{i1max}=211V amplitude, ω_n=66750 rad/s angular frequency of the inverter output voltage fundamental, while the resonance occures at ω_0=69138 rad/s. The time diagrams show the evolution of the angular frequency (upper) and u_{Cpmax} load capacitor voltage amplitude (lower). In case (b), u_{i1max}=266 V, ω_n=73062 rad/s.

The next section presents a modified control method of the resonant capacitor tank voltage, able to estimate these values.

5. Estimation of the steady-state values

In order to handle the uncertainties of the steady-state values, the control law is modified in a manner that it is based on the estimated steady-state values and assures both global stability of the system and convergence of the estimation error towards zero. The Lyapunov candidate is modified (19)

$$V(\tilde{x}) = \frac{1}{2}[\tilde{x}]^T[Q][\tilde{x}] + \frac{1}{2}[\delta x_n]^T[K][\delta x_n], \tag{19}$$

with $[K]$ symmetric and positive defined, in order to include the estimation error vector

$$[\delta x_n] = [\tilde{x}_n] - [x_n]. \tag{20}$$

The derivative of the Lyapunov candidate becomes

$$\begin{aligned}\frac{dV(\tilde{x})}{dt} &= \frac{1}{2}[\tilde{x}]^T\left([A]^T[Q] + [Q][A]\right)[\tilde{x}] + \frac{1}{2}[\tilde{x}]^T\left([B]^T[Q] + [Q][B]\right)[\tilde{x}]\tilde{\omega} \\ &+ \frac{1}{2}\left([b]^T[Q][\tilde{x}] + [\tilde{x}]^T[Q][b]\right)\tilde{\omega} + \frac{1}{2}[\delta\dot{x}_n]^T[K][\delta x_n] + \frac{1}{2}[\delta x_n]^T[K][\delta\dot{x}_n]\end{aligned} \tag{21}$$

The sum of the last two terms from (21) can be brought to the form:

$$L=\frac{1}{2}[\delta\dot{x}_n]^T[K][\delta x_n]+\frac{1}{2}[\delta\dot{x}_n]^T[K][\delta\dot{x}_n]=L_1\tilde{\omega} \text{ if} \tag{22}$$

$$[\delta\dot{x}_n]=\tilde{\omega}[F]. \tag{23}$$

It results [6]:

$$L=\frac{1}{2}\tilde{\omega}\left([F]^T[K][\delta x_n]+[\delta x_n]^T[K][F]\right). \tag{24}$$

The sum of the last five terms from (21) is made negative by the choice:

$$\tilde{\omega}=-\alpha\left([\tilde{x}]^T[B]^T[Q][\tilde{x}]+[b]^T[Q][\tilde{x}]+L_1\right)=-\alpha\left(([B][\tilde{x}]+[b])^T[Q][\tilde{x}]+L_1\right) \tag{25}$$

Generally the quantity $\left(([B][\tilde{x}]+[b])^T[Q]\right)^T=[Q]([B][\tilde{x}]+[b])$ can be measured easily [5] and $([\tilde{x}]-[\delta x_n])=[x_T]-[x_n]-[\delta x_n]=[x_T]-([x_n]+[\delta x_n])=[x_T]-[\tilde{x}_n]$ is known, because $[x_T]$ is the measured state vector and $[\tilde{x}_n]$ is the estimate of the steady state. Consequently, it would be useful if (25) was brought to the form

$$\tilde{\omega}=-\alpha([B][\tilde{x}]+[b])^T[Q]([\tilde{x}]-[\delta x_n]), \text{ choosing} \tag{26}$$

$$L_1=-([B][\tilde{x}]+[b])^T[Q][\delta x_n] \tag{27}$$

Thus, we need

$$L=\frac{1}{2}\tilde{\omega}\left([F]^T[K][\delta\dot{x}_n]+[\delta\dot{x}_n]^T[K][F]\right)=\tilde{\omega}L_1=-\tilde{\omega}([B][\tilde{x}]+[b])^T[Q][\delta\dot{x}_n], \tag{28}$$

which is satisfied if

$$[F]=-[K]^{-1}[Q]([B][\tilde{x}]+[b]). \tag{29}$$

According to (23) the update law of the estimate becomes

$$[\delta\dot{x}_n]=\tilde{\omega}[F]=-[K]^{-1}[Q]([B][\tilde{x}]+[b])\tilde{\omega} \tag{30}$$

This update law, together with the control law (26) assures the stability of the extended system

$$\begin{bmatrix}[\dot{\tilde{x}}]\\ [\delta\dot{x}_n]\end{bmatrix}=\begin{bmatrix}[A] & [0]\\ [0] & [0]\end{bmatrix}\begin{bmatrix}[\tilde{x}]\\ [\delta x_n]\end{bmatrix}+\begin{bmatrix}[B][\tilde{x}]+[b]\\ -[K]^{-1}[Q]([B][\tilde{x}]+[b])\end{bmatrix}\tilde{\omega} \tag{31}$$

Substituting the matrices of the system (6) into (26) and using the notation (1) and (2) it results:

$$([B][\tilde{x}]+[b])=\begin{bmatrix}0&1&0&0&0&0\\-1&0&0&0&0&0\\0&0&0&1&0&0\\0&0&-1&0&0&0\\0&0&0&0&0&1\\0&0&0&0&-1&0\end{bmatrix}\begin{bmatrix}\tilde{i}_{Lsd}\\ \tilde{i}_{Lsq}\\ \tilde{u}_{Cpd}\\ \tilde{u}_{Cpq}\\ \tilde{i}_{Lisd}\\ \tilde{i}_{Lisq}\end{bmatrix}+\begin{bmatrix}I_{Lsqn}\\ -I_{Lsdn}\\ U_{Cpqn}\\ -U_{Cpdn}\\ I_{Lisqn}\\ -I_{Lisdn}\end{bmatrix}=\begin{bmatrix}\tilde{i}_{Lsq}+I_{Lsqn}\\ -\tilde{i}_{Lsd}-I_{Lsdn}\\ \tilde{u}_{Cpq}+U_{Cpqn}\\ -\tilde{u}_{Cpd}-U_{Cpdn}\\ \tilde{i}_{Lisq}+I_{Lisqn}\\ -\tilde{i}_{Lisd}-I_{Lisdn}\end{bmatrix}=\begin{bmatrix}i_{LsqT}\\ -i_{LsdT}\\ u_{CpqT}\\ -u_{CpdT}\\ i_{LisqT}\\ -i_{LisdT}\end{bmatrix} \tag{32}$$

Thus, the control law (26) becomes:

$$\tilde{\omega} = -\alpha \begin{pmatrix} i_{LsqT} L_s \left(i_{LsdT} - \tilde{I}_{Lsdn}\right) - i_{LsdT} L_s \left(i_{LsqT} - \tilde{I}_{Lsqn}\right) + u_{CpqT} C_p \left(u_{CpdT} - \tilde{U}_{Cpdn}\right) - \\ - u_{CpdT} C_p \left(u_{CpqT} - \tilde{U}_{Cpqn}\right) + i_{LisqT} L_{is} \left(i_{LisdT} - \tilde{I}_{Lisdn}\right) - i_{LisdT} L_{is} \left(i_{LisqT} - \tilde{I}_{Lisqn}\right) \end{pmatrix} \tag{33}$$

Assuming that $[K]$ is diagonal, with the diagonal elements equal to *k*, the update law of the estimate (30) becomes:

$$\begin{cases} \delta\dot{\tilde{I}}_{Lsdn} = -k^{-1} L_s i_{LsqT} \tilde{\omega} \\ \delta\dot{\tilde{I}}_{Lsqn} = k^{-1} L_s i_{LsdT} \tilde{\omega} \\ \delta\dot{\tilde{U}}_{Cpdn} = -k^{-1} C_p u_{CpqT} \tilde{\omega} \\ \delta\dot{\tilde{U}}_{Cpqn} = k^{-1} C_p u_{CpdT} \tilde{\omega} \\ \delta\dot{\tilde{I}}_{Lisdn} = -k^{-1} L_{is} i_{LisqT} \tilde{\omega} \\ \delta\dot{\tilde{I}}_{Lisqn} = k^{-1} L_{is} i_{LisdT} \tilde{\omega} \end{cases} \tag{34}$$

The estimated steady-state value of the variable x results:

$$\tilde{X}_n = \tilde{X}_{n0} + \int_0^t \delta\dot{\tilde{X}}_n dt, \cdots with\ \tilde{X}_{n0} = \tilde{X}_n\Big|_{t=0} \tag{35}$$

Fig. 7 shows the control structure of U_{Cp} load capacitor voltage, based on the control law (33). In this control structure, the circuit variables are estimated according to formulae (34) and (35), while the steady-state angular velocity estimate is updated by an integrator according to:

$$\tilde{\omega}_n = \tilde{\omega}_n\Big|_{t=0} + K_{iUCp} \int_0^t \left(U_{Cp} - U_{Cpref}\right) dt`, \tag{36}$$

valid for the upper frequency domain from *Fig. 1.b*.

Fig. 8 shows comparison of the inverter start-up (set value $U_{Cpref} = 300V$), according to the above adaptive control law and start-up with PI controller tuned for the nominal inductor parameters in the case when the inductor is almost "empty", i.e. its equivalent resistance is one third of the nominal value.

The nominal parameters of the resonant load are:
L_s= 20μH, L_{is}= 3.95μH, R_{is}= 0.03Ω, C_p= 63μF.

The amplitude of the inverter output voltage's fundamental is u_{i1max}=266V.
Fig. 9 shows the update process of the estimated steady-state values starting from the initial values:

$$\tilde{I}_{Lsdn0} = 121\ A, \tilde{I}_{Lsqn0} = -348\ A, \tilde{U}_{Cpdn0} = -243\ V, \tilde{U}_{Cpqn0} = -177\ V,$$
$$\tilde{I}_{Lisdn0} = -692\ A, \tilde{I}_{Lisqn0} = 770\ A, \omega_{n0} = 80{,}000\ rad/s$$

These values belong to the $\omega_{n1} = 73062\ rad/s$ operating angular frequency for which $u_{Cp\max ref} = U_{Cpref} = 300\,V$ in case of nominal circuit parameters.

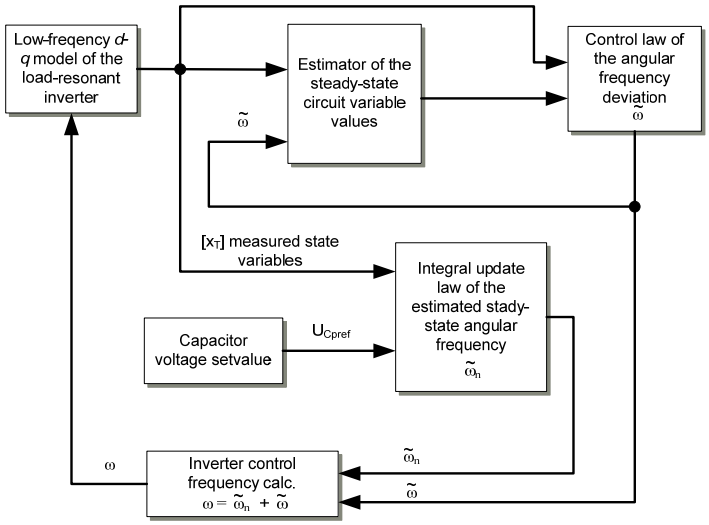
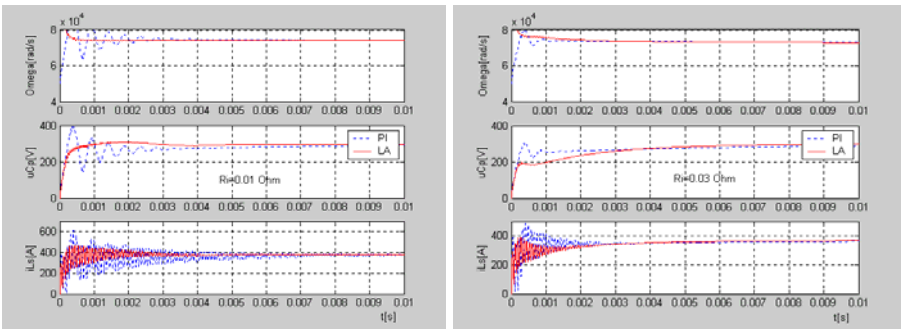


Figure 7: Control structure of the load capacitor tank voltage based on Lyapunov stability theory, with steady-state estimation.

Figures 8 and *9* show operation according to the proposed control law with controller parameters: $K = 0.02$, $\alpha = 1000$, $K_{iUCp} = 20000\,1/Vs^2$.

The main practical difficulty is the measurement of the fundamental amplitudes and phases of three electrical quantities: u_{Cp}, i_{Ls}, i_{Lis}.



a) b)

Figure 8: Inverter start-up (R_{is}=0.01 Ω in case (a), R_{is}=0.03 Ω in case (b)) with Lyapunov-based control with steady-state estimation (LA), and PI control. In both cases the initial angular frequency is 80,000 rad/s.

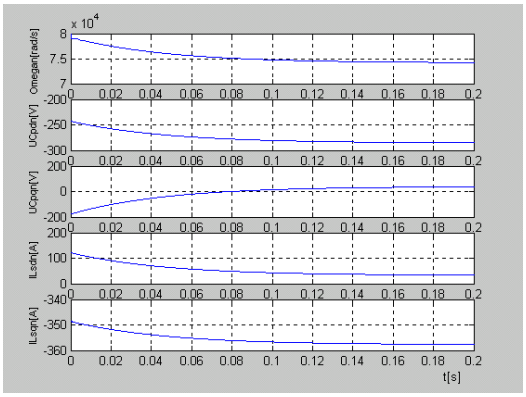


Figure 9 : Update process of the estimated steady-state values in case of R_{is}=0.01 Ω, much different from its "nominal" value.

6. Conclusions

A load-resonant inverter is a nonlinear system, which is difficult to control in case of high quality factor of the inductor. Lyapunov-based control methods have been proposed and demonstrated by simulation both for known and uncertain load parameters.

A possible further development consists in reduction of the number of measured state variables based on the circuit equations with "nominal" parameters.

References

[1] Dieckerhoff, S., Ryan, M.J., De Doncker, R.W., "Design of an IGBT-based LCL-resonant inverter for high-frequency induction heating", in *Proc. Thirty-Fourth IEEE Industry Applic. Conf. IAS*, Vol. 3 pp. 2039-2045, 1999.

[2] Kelemen, A., Kutasi, N., Mátyási, Sz., "Control strategies for a voltage source induction heating inverter with hybrid LLC resonant load", in *Proc. ICCC2005 – Miskolc, Hungary*, pp. 63-70, 2005.

[3] Kelemen, A., Kutasi, N., "Induction heating voltage inverter with hybrid LLC resonant load, the D-Q model", *Pollack Periodica*, Vol.2, No.1, pp. 27-37, 2007.

[4] Kawasaki, N., Nomura, H., Masuhiro, M., "A new control law for bilinear DC-DC converters developed by direct application of Lyapunov", *IEEE Trans. Pow. Electr.*, Vol. 10, No. 3, pp. 318-325, May 1995.

[5] Sanders, S.R., Verghese, G.C., "Lyapunov-based control for switched power converters", *IEEE Trans. Pow. Electr.*, Vol. 7, No. 1, pp. 17-24, Jan. 1992.

[6] Kelemen, A., "Reglarea puterii convertoarelor electronice din instalațiile de încălzire prin inducție", *PhD Thesis*, Transylvania University of Braşov, 2007.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 53-63

Design and Simulation of a Shunt Active Filter in Application for Control of Harmonic Levels

Adrian GLIGOR

Department of Electrical Engineering, Faculty of Engineering,
"Petru Maior" University of Tîrgu Mureş, Tîrgu Mureş, Romania,
e-mail: agligor@upm.ro

Manuscript received March 15, 2009; revised June 28, 2009.

Abstract: Nowadays, the active filters represent a viable alternative for controlling harmonic levels in industrial consumers' electrical installations. It must be noted the availability of many different types of filter configurations that can be used but there is no standard method for rating the active filters. This paper focuses on describing the shunt active filter structure and design. The theoretical concepts underlying the design of shunt active filters are presented. To validate and highlight the performance of shunt active filters a Matlab-Simulink model was developed. Simulation results are also presented.

Keywords: Shunt active filters, harmonic analysis, nonlinear control, instantaneous power theory.

1. Introduction

After a brief analysis performed on evolution of electric power consumption during the last two decades, it can be observed a change mainly on nature of electric power consumption and profile of consumers. The main causes are represented by introduction of new equipment and facilities to increase comfort in civil construction, new appliances and equipment in order to raise efficiency and diversification of production for industrial consumers, or coexistence in the same building of both households and some industrial consumers. We must also note the impact of the new sources of energy that can easily transform the consumer into power supplier. However, all these changes have led to the emergence of undesirable phenomena in all power system, accounting for the

new challenges to be addressed by engineers and scientists involved in the power system design and management.

Among the measures required there must be mentioned the need to adapt the existing electrical network to the new requirements and the introduction of new advanced methods of control, management and monitoring, in order to ensure the efficiency of electricity use.

The aims of this paper are to present a solution to improve the operation of consumers' electrical installations, to reduce the electric power consumption and default costs allocated for the purchase of electricity and removing unwanted effects caused by the presence of harmonics. In order to achieve this, the main goal is to increase the power quality available for consumers. In the case of power consumers affected by the presence of harmonic pollution, power quality improvement can be achieved by implementing systems based on active filtering of the unwanted components. This type of automated system based on shunt active filter is presented in the following sections.

2. Operation principle of system based on shunt active filter

Fig. 1 shows the schematic implementation of active power filter with static power converter.

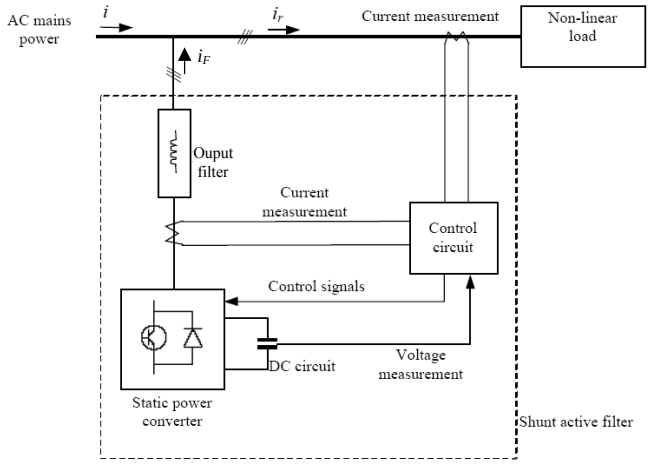


Figure 1: The configuration of a system with shunt active filter.

In order to compensate harmonic pollution caused by a nonlinear receiver, the parallel active filter consists of a DC-link static power converter and an energy storage element.

The control circuit performs synthesis of the reference currents of the filter in a manner to compensate the undesired mains current components.

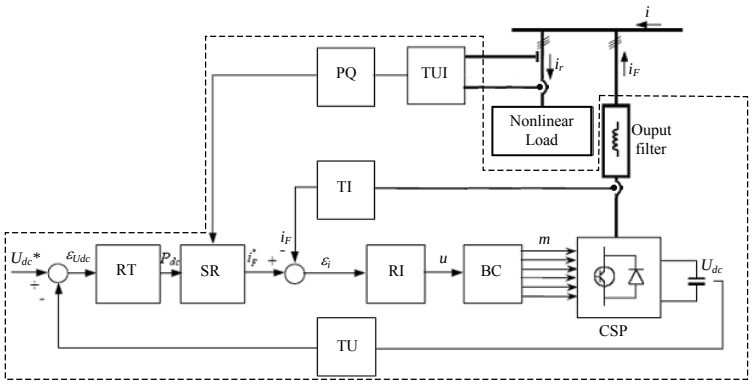
Since currents synthesized by an active filter depend on the average voltage of the storage element, this one should be kept constant. This voltage control has to be provided by the filter control algorithm.

3. Structure of the active filter configured for control of harmonics levels

The controller has both the task of controlling the DC-link voltage and the task of controlling the three-phase current system of the active filter.

This requires a complex control structure with two control loops, one for the i_F current, which has to be synthesized and another one for the DC-link voltage.

The structure with two control loops is shown in *Fig. 2*. In this scheme the two controllers can be highlighted: RI – current controller – from the current control loop, and RT voltage controller from the DC-link voltage control loop.



RT – voltage controller; SR – reference synthesis current block; RI – current controller; BC – gate-control block; CSP – power static converter; TU – voltage transducer; TI – current sensor; PQ – instantaneous power calculus block; TUI – voltage and current voltage transducer

Figure 2: Block diagram of the system for control of harmonic current level based on active shunt filter.

The voltage controller generates the signal P_{dc} based on the reference signal U_{dc}* and on the feedback signal U_{dc} provided by the voltage sensor TU. SR block, which generates the reference currents, based on signal P_{dc} and on instantaneous powers determined by PQ block, provides the reference for the current control loop. The PQ block has at its input the current and voltage

signals provided by the TUI sensor from the circuit which supply the non-linear load.

The RI controller from the current control loop synthesizes the control signal u, which is generated from the error signal ε_i. This error signal (ε_i) is obtained by comparing the signal provided by the SR block with the current measured at the input of static power converter (CSP). The control signal u is applied to the BC block, which generates the logic signal m needed to control the CSP block.

A. Synthesis of references from current compensation control loop based on the instantaneous power theory

Instantaneous power theory introduced by Akagi offers the methodology for determining the harmonic distortion [1], [2], [3], [4], [5].

According to the notation from *Fig. 1*:

$$\begin{aligned} i_{ra} &= I_{r1a}\sqrt{2}\sin\omega t + \tilde{i}_{ra} \\ i_{rb} &= I_{r1b}\sqrt{2}\sin\left(\omega t - \frac{2\pi}{3}\right) + \tilde{i}_{rb}\,, \\ i_{rc} &= I_{r1c}\sqrt{2}\sin\left(\omega t - \frac{4\pi}{3}\right) + \tilde{i}_{rc} \end{aligned} \tag{1}$$

where I_{r1a} represents the r.m.s value of the load fundamental currents i_{ra}, i_{rb}, i_{rc}, and $\tilde{i}_{ra}$, $\tilde{i}_{rb}$, $\tilde{i}_{rc}$ are the polluting load current components.

In the following there will be noted:

$$\mathbf{u} = \begin{bmatrix} u_a \\ u_b \\ u_c \end{bmatrix},\ i = \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix},\ \mathbf{i}_r = \begin{bmatrix} i_{ra} \\ i_{rb} \\ i_{rc} \end{bmatrix},\ \mathbf{u}_r = \begin{bmatrix} u_{ra} \\ u_{rb} \\ u_{rc} \end{bmatrix},\ \mathbf{i}_F = \begin{bmatrix} i_{Fa} \\ i_{Fb} \\ i_{Fc} \end{bmatrix}. \tag{2}$$

Converting them into *(α-β)* coordinates, it results:

$$\begin{bmatrix} u_{r\alpha} \\ u_{r\beta} \end{bmatrix} = \mathbf{C}\begin{bmatrix} u_{ra} \\ u_{rb} \\ u_{rc} \end{bmatrix},\ \begin{bmatrix} i_{r\alpha} \\ i_{r\beta} \end{bmatrix} = \mathbf{C}\begin{bmatrix} i_{ra} \\ i_{rb} \\ i_{rc} \end{bmatrix},\ \text{where: } \mathbf{C} = \sqrt{\frac{2}{3}}\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \tag{3}$$

Assuming that the zero-sequence components of the three-phase systems are missing, the C matrix is given by:

$$\mathbf{C} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix}. \tag{4}$$

In the case of the new two-phase coordinates, the instantaneous power is given by:

$$p = u_{r\alpha}\, i_{r\alpha} + u_{r\beta}\, i_{r\beta} \tag{5}$$

If a new quantity is introduced, the so-called instantaneous imaginary power, denoted with q, this is defined as (*see Fig. 3*):

$$q\vec{k} = u_{r\alpha}\vec{i} \times i_{r\beta}\vec{j} + u_{r\beta}\vec{j} \times i_{r\alpha}\vec{i} \tag{6}$$

Equation (6) may be rewritten as module as well:

$$q = u_{r\alpha} i_{r\beta} - u_{r\beta}\, i_{r\alpha}\,. \tag{7}$$

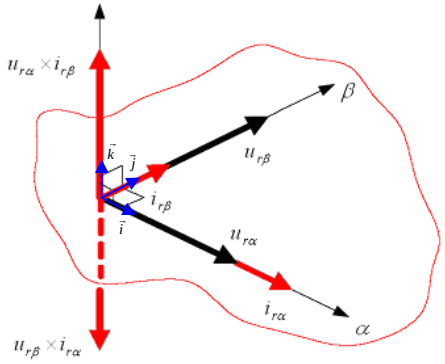


Figure 3: Determination of the instantaneous imaginary power.

Equations (5) and (7) may also be rewritten in the form of a matrix as follows:

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} u_{r\alpha} & u_{r\beta} \\ -u_{r\beta} & u_{r\alpha} \end{bmatrix} \begin{bmatrix} i_{r\alpha} \\ i_{r\beta} \end{bmatrix}, \tag{8}$$

which results in the expression of current i_r in (α,β) system:

$$\begin{bmatrix} i_{r\alpha} \\ i_{r\beta} \end{bmatrix} = \begin{bmatrix} u_{r\alpha} & u_{r\beta} \\ -u_{r\beta} & u_{r\alpha} \end{bmatrix}^{-1} \begin{bmatrix} p \\ q \end{bmatrix} = \frac{1}{u_{r\alpha}^2 + u_{r\beta}^2} \begin{bmatrix} u_{r\alpha} & -u_{r\beta} \\ u_{r\beta} & u_{r\alpha} \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \\ = \frac{1}{u_{r\alpha}^2 + u_{r\beta}^2} \left\{ \begin{bmatrix} u_{r\alpha} & -u_{r\beta} \\ u_{r\beta} & u_{r\alpha} \end{bmatrix} \begin{bmatrix} p \\ 0 \end{bmatrix} + \begin{bmatrix} u_{r\alpha} & -u_{r\beta} \\ u_{r\beta} & u_{r\alpha} \end{bmatrix} \begin{bmatrix} 0 \\ q \end{bmatrix} \right\} \tag{9}$$

Considering that:

$$p = \bar{p} + \tilde{p} \text{ and } q = \bar{q} + \tilde{q} \tag{10}$$

where:

– $\bar{p}$ represents the component of the instantaneous active power absorbed by the nonlinear load associated to the fundamental of current $\boldsymbol{i_r}$, and voltage $\boldsymbol{u_r}$;

– $\tilde{p}$ represents the component of the instantaneous power absorbed by the nonlinear load associated to the harmonics of current $\boldsymbol{i_r}$ and voltage $\boldsymbol{u_r}$;

– $\bar{q}$ represents the component of the instantaneous imaginary power corresponding to the reactive power associated to the fundamentals of current $\boldsymbol{i_r}$ and voltage $\boldsymbol{u_r}$;

– $\tilde{q}$ represents the component of the instantaneous imaginary power corresponding to the reactive power associated to the harmonics of the current $\boldsymbol{i_r}$ and voltage $\boldsymbol{u_r}$.

If i_{rA} represents the fundamental active component of the absorbed current:

$$\begin{bmatrix} i_{rA\alpha} \\ i_{rA\beta} \end{bmatrix} = \begin{bmatrix} u_{r\alpha} & u_{r\beta} \\ -u_{r\beta} & u_{r\alpha} \end{bmatrix}^{-1} \begin{bmatrix} \bar{p} \\ 0 \end{bmatrix} = \frac{1}{u_{r\alpha}^2 + u_{r\beta}^2} \begin{bmatrix} u_{r\alpha} & -u_{r\beta} \\ u_{r\beta} & u_{r\alpha} \end{bmatrix} \begin{bmatrix} \bar{p} \\ 0 \end{bmatrix} \tag{11}$$

then the reference system of currents in coordinates (α-β) can be obtained in the following form:

$$\begin{bmatrix} i_{F\alpha}^* \\ i_{F\beta}^* \end{bmatrix} = \begin{bmatrix} i_{r\alpha} \\ i_{r\beta} \end{bmatrix} - \begin{bmatrix} i_{rA\alpha} \\ i_{rA\beta} \end{bmatrix} = \begin{bmatrix} u_{r\alpha} & u_{r\beta} \\ -u_{r\beta} & u_{r\alpha} \end{bmatrix}^{-1} \begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} u_{r\alpha} & u_{r\beta} \\ -u_{r\beta} & u_{r\alpha} \end{bmatrix}^{-1} \begin{bmatrix} \bar{p} \\ 0 \end{bmatrix} = \\ = \frac{1}{u_{r\alpha}^2 + u_{r\beta}^2} \begin{bmatrix} u_{r\alpha} & -u_{r\beta} \\ u_{r\beta} & u_{r\alpha} \end{bmatrix} \begin{bmatrix} \tilde{p} \\ \bar{q} + \tilde{q} \end{bmatrix} \tag{12}$$

and the equivalent three-phase components will be, respectively:

$$\begin{bmatrix} i_{Fa}^* \\ i_{Fb}^* \\ i_{Fc}^* \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} i_{F\alpha}^* \\ i_{F\beta}^* \end{bmatrix} \tag{13}$$

B. Harmonic current compensation by delta modulation

In case of a hysteresis current control the switching frequency is not well defined. Therefore, it was introduced the concept of the average switching frequency. In principle, the increase of the frequency of power converter leads to a better current compensation. However, the increase of the switching frequency is limited by the switching losses of the power devices.

The operation of the hysteresis current controller is shown in Fig. 4.

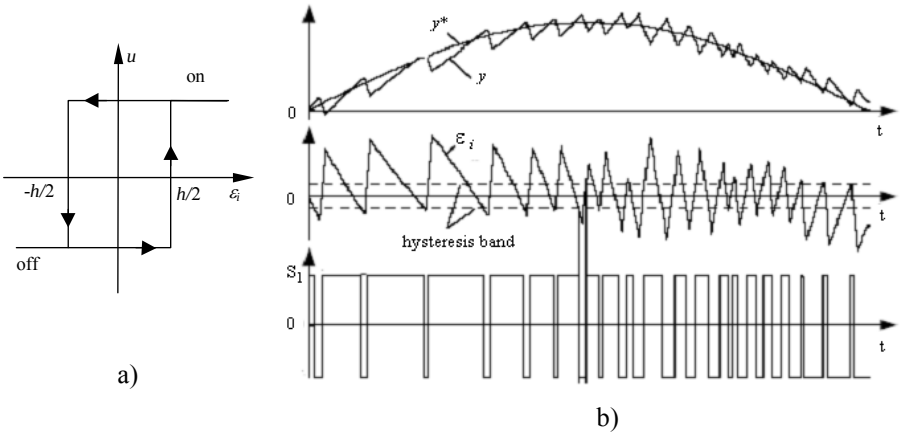
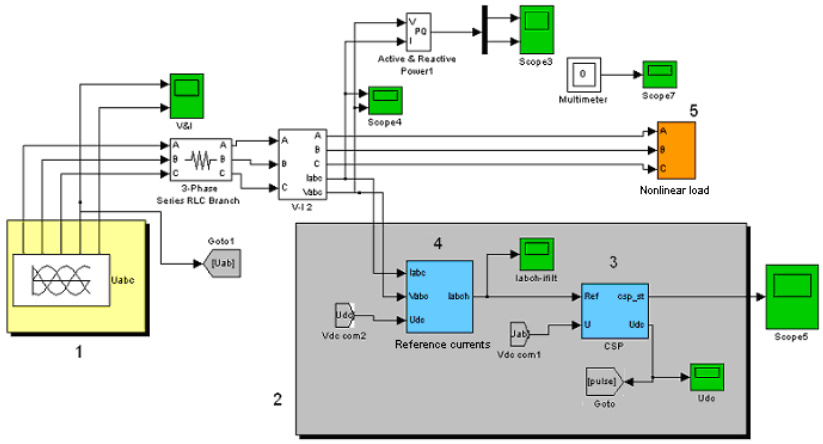


Figure 4: Input-output characteristic (a) and the operating principle of the hysteresis controller.

4. Results obtained by numerical simulations

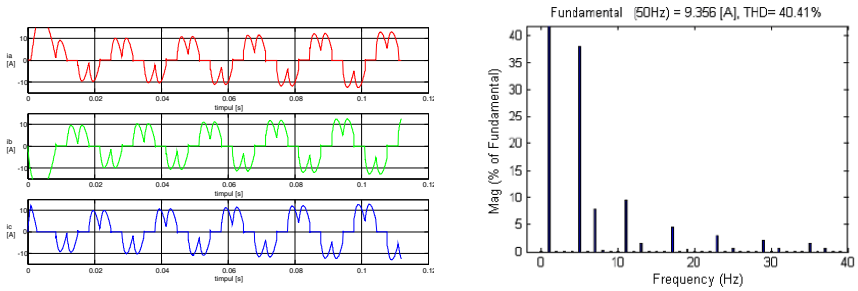
The performance analysis of the system with active filtering was realized based on the data obtained by simulation in Matlab-Simulink environment. *Fig. 5* presents the Simulink model used for study.



1 – power system, 2 – shunt active filter, 3 – power static converter, 4 – reference current signal generator, 5 – nonlinear load

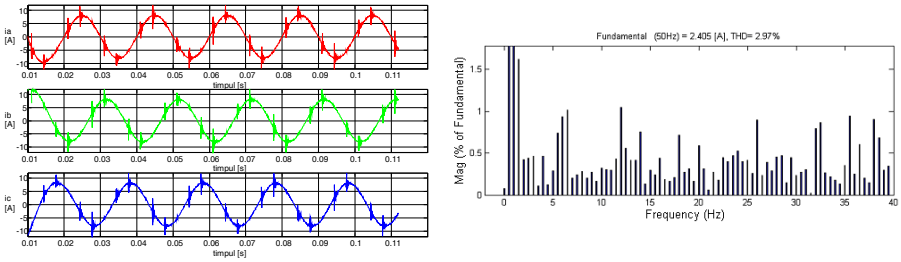
Figure 5: Simulink model.

Fig. 6 and *8* show the current waveforms and their harmonic spectrum in case of a non-linear load taken for study, whereas *Fig. 7* and *9* present the compensated current waveforms and harmonic spectrum resulted after compensation.



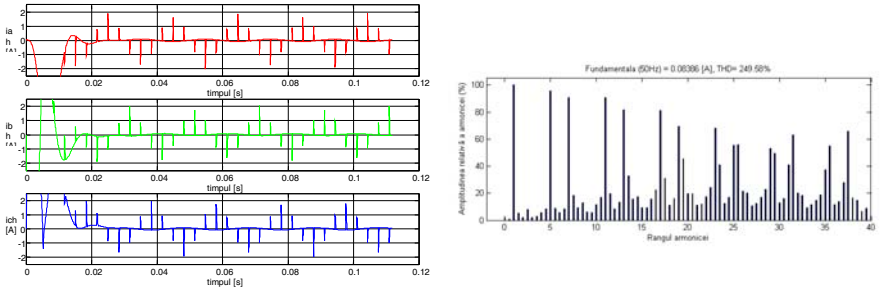
a) b)

Figure 6: Current waveforms of a nonlinear load represented by a controlled rectifier in case of a control angle equal with 30°, (a) and the harmonic spectrum of these currents (b).



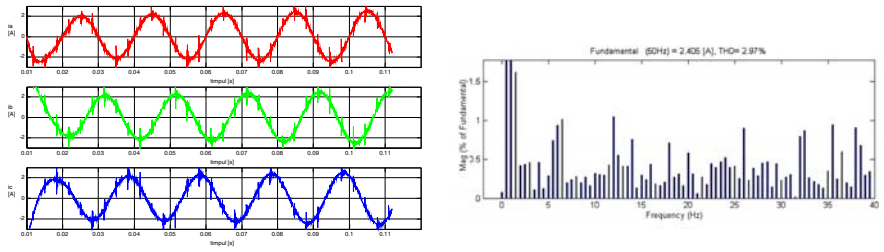
a) b)

Figure 7: Compensated mains currents in case of load currents from Fig.6: waveforms (a) and harmonic spectrum (b)



a) b)

Figure 8: Current waveforms of a rectifier in case of control angle equal with 10° (a) and the harmonic spectrum of these currents (b).



a) b)

Figure 9: Compensated mains currents in case of load from Fig. 8: waveforms (a) and harmonic spectrum (b).

In *Table 1, 2* and *3* there are synthesized the main results considering a nonlinear load represented by a controlled rectifier in case of a control angle equal with 30°, and P = 1.54 kW. The automatic system with active filtering achieves a reduction in the level of odd harmonic of 50 ÷97% and a reduction of the THD_i factor by over 94%.

In case of a control angle of 10°, one may observe an increased level of current harmonic distortion (THD_i=249.58%). In this case, the automatic system with active filtering achieves a reduction to 2.97% of THD_i.

Table 1: THD_i of the uncompensated system.

Phase	THD_i	harmonic ratio [%]							
		5^{th}	7^{th}	11^{th}	13^{th}	17^{th}	19^{th}	23^{th}	25^{th}
A	53,47	48,55	18,38	9,82	5,91	3,94	2,69	2,06	1,44
B	53,55	48,58	18,56	9,79	5,96	3,90	2,74	2,04	1,48
C	53,44	48,55	18,31	9,86	5,85	3,93	2,67	2,06	1,41

Table 2: THD_i of the compensated system.

Phase	THD_i	harmonic ratio [%]							
		5^{th}	7^{th}	11^{th}	13^{th}	17^{th}	19^{th}	23^{th}	25^{th}
A	2,74	1,55	1,33	0,18	0,17	0,27	0,11	0,16	0,25
B	2,82	1,34	1,33	0,13	0,41	0,11	0,33	0,20	0,11
C	2,50	1,08	1,43	0,29	0,56	0,21	0,44	0,17	0,35

Table 3: Reduction of THD_i in case of using compensation [%].

Phase	Reduction of THD_i	reduction of the harmonic components [%]							
		5^{th}	7^{th}	11^{th}	13^{th}	17^{th}	19^{th}	23^{th}	25^{th}
A	94,88	96,81	92,76	98,17	97,12	93,15	95,91	92,23	82,64
B	94,73	97,24	92,83	98,67	93,12	97,18	87,96	90,20	92,57
C	95,32	97,78	92,19	97,06	90,43	94,66	83,52	91,75	75,18

5. Conclusion

Proliferation of the power electronic equipments leads to an increasing harmonic contamination in power transmission or distribution systems. Many researchers from the field of the power systems and automation have searched for different approaches to solve the problem. One way was open by introducing the harmonic compensation by using active filters.

This paper presents an automatic system based on active filtering for harmonic current reduction with direct applicability in the civil and industrial electrical installations affected by harmonics.

The proposed system is based on the new theory of instantaneous power introduced by Akagi and on delta modulation control technique of the static power converter.

Simulation results were obtained before and after the use of the automatic system based on active filtering. From the analysis of the experimental data, in case of a nonlinear load of rectifier type, one may observe that there are different levels of current distortion produced depending on the load and its control mode, with high values of the total current harmonic distortion and low power factor.

Using the active filter, the experimental data show that the total harmonic distortion of current (THD_i) decreases to 1%-4%, and the power factor rises up to 0.98-1. A 30% decrease of the r.m.s. value of the current was also recorded.

Analyzing the harmonic spectra of the compensated currents, it results that the weight of the 5th and 7th harmonics are close to 1% and the rest of upper harmonics are below 1%.

References

[1] Akagi, H., "Modern active filters and traditional passive filters"; *Bulletin of The Polish Academy of Sciences Technical Sciences*; Vol. 54, No. 3, pp. 255-269; 2006.

[2] Akagi, H., "New trends in active filters for improving power quality", in *Proc. of the International Conference on Power Electronics, Drives and Energy Systems for Industrial Growth*, *1996*,Vol. 1, Issue 8-11, Jan. 1996, pp. 417 - 425.

[3] Gligor, A., "Contribuţii privind sistemele avansate de conducere şi optimizare a proceselor energetice în instalaţiile electrice la consumatori", *PhD Thesis,* Universitatea Tehnică Cluj-Napoca, 2007.

[4] Codoiu, R., "Selection of the representative non-active power theories for power conditioning", *Proceedings of the International Scientific Conference Inter-Ing. 2007*, Tg. Mureş, 2007, pp.V-6-1 - V-6-14.

[5] Codoiu, R., Gligor, A., "Current and power components simulation using the most recent power theories", *Proceedings of the International Scientific Conference Inter-Ing. 2007*, Tg. Mureş, 2007, pp.V-7-1 - V-7-8.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 65-75

Measurement Considerations on Some Parameters of Supercapacitors

Ana Maria PUŞCAŞ[1], Marius CARP[1],
Paul BORZA[1], Iuliu SZEKELY[1]

[1] Department of Electronics and Computers, Faculty of Electrical Engineering and Computer Science, "Transilvania" University of Braşov, Braşov,
e-mail: ana_maria.puscas@yahoo.com, marius.carp@yahoo.com, borzapn@unitbv.ro, szekelyi@vega.unitbv.ro

Manuscript received May 30, 2009; revised June 30, 2009.

Abstract: The paper is focused on identification and measurement of the main parameters of the stacked supercapacitors (SC). The volt-ampere method was used to measure the charge/discharge characteristic, impedance spectroscopy method for measurement of the equivalent serial resistance and analysis of the dynamic parameters. The measured and calculated parameters are: equivalent serial resistance (ESR), capacitance, self discharge resistance, maximum power and energy, energy efficiency.

Keywords: Supercapacitor, self discharge, parameters, impedance spectroscopy.

1. Introduction

Energy storage represents one of the key problems in present day research to increase the energy efficiency of processes. The development of nanotechnology creates premises for the use of new types of devices (Li-ion batteries, supercapacitors, different kinds of fuel cells), which are more adequate for modern energy efficient solutions.

The supercapacitors have a bidirectional power flow exchange efficiency up to 98%, available only for superconductivity applications. There is a lack of knowledge in the characterization of these new devices, like how the device is integrated into a system. Some parameters can be defined to characterize the ability of integration of devices in a system: life span, aging and the finite character of the provided power / energy. The development of nanotechnology led to high performances of symmetrical supercapacitors, such as: equivalent serial resistance (ESR), self discharge resistance, resistance of losses, power and energy density.

In the transportation area supercapacitors proved their important role in combined solutions (supercapacitors, batteries and intelligent controllers) to increase the energy efficiency. Solutions consisting in a simple parallel connection of batteries and supercapacitors proved to be inefficient and new intelligent control systems must be developed to improve the energy efficiency. In the development of a test system for the analysis of the parameters of supercapacitors the benefits of new acquisition systems have to be taken into account which can assure portability, reliability, availability and autonomy. Supercapacitors are able to fill the gap between batteries and classical capacitors (see also Ragone diagram) [2]. A system containing battery and supercapacitor provides not only high energy density but also high power density. They also allow to fulfil a carbon-free desiderate by "eco-footprint" solutions [3].

2. Supercapacitors (SC)

Supercapacitors are rapid release storage devices with capacitance up to thousands of Farad. The main applications of the supercapacitors are related to high power density, large peak power requirements, load smoothing, backup power for mobile applications and different kinds of recovering systems. Supercapacitors can be easily and rapidly charged from various electrical supply sources and also, their voltage - current characteristics and charging / discharging behaviours are not influenced by the over-current, deep discharge and temperature variations (-40^{o}C to +70^{o}C [4]). Supercapacitors are able to support up to millions of lifecycles [5].

In mobile applications the batteries are often under high peak power stresses, which can dramatically reduce their life span. The combination of batteries with supercapacitors can improve dramatically the life span of batteries.

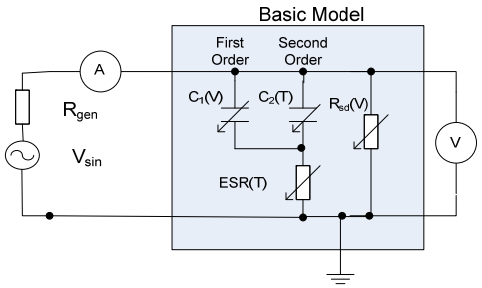


Figure 1: Basic equivalent model of a supercapacitor.

The amount of the stored energy is different, depending on the type of the supercapacitor, the size of the ions, the electrode surface and the level of the electrolyte decomposition voltage [6]. In supercapacitors the charge transfer is performed by electrons and ions: at the collector (metal electrodes) the charge transfer is assured by electrons and in the electrolyte (inside of the supercapacitor) the charge transfer is ionic.

Taking into account the above mentioned considerations we suggested a basic model presented in *Figure 1*, model which was used for definition of parameters in our experiments.

The experiments started with the following assumptions: the equivalent serial resistance (ESR) depends on temperature, voltage and frequency; the resistance of losses (self discharge resistance - R_{sd}) is voltage dependent; the parallel connected equivalent capacitances are voltage and temperature dependent; C_1, C_2 are the equivalent serial capacitances of the double layer supercapacitor. Generally ESR has values of tens to hundreds of mΩ and the resistance of losses has values of tens to hundreds of kΩ.

3. Parameter evaluation for supercapacitors

3.1 Data acquisition system, impedance spectroscopy

A symmetric (stacked) supercapacitor contains multiple cells connected in series and in parallel. As an improper overvoltage is applied, a particular cell could be damaged, therefore, electronic control methods have to be developed. The test setup developed in our laboratory permanently controls the voltage of the supercapacitor and assures the voltage limitation below the preset value.

The data acquisition system, which also fulfils the impedance spectroscopy requirements, is presented in *Figure 2*.

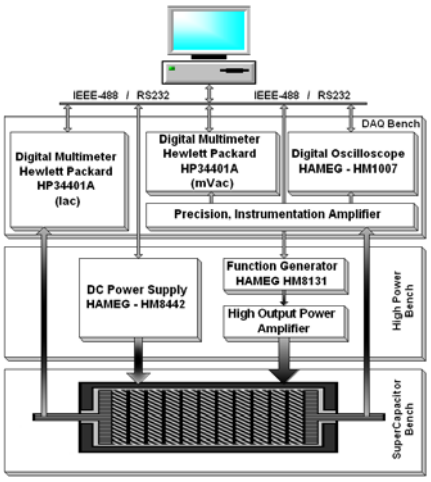


Figure 2: The configuration of the data acquisition system.

The data acquisition system has the following components: DC Power supply driven by an ATMega128 microcontroller-based system, HP 34401A multimeter, digital oscilloscope, function generator, amplifiers. All the acquired data were transmitted to a computer and the stored data were processed with Matlab software.

3.2 Capacitance measurement and its temperature dependence

A volt-ampere method was used to determine the capacitance of two types of aqueous electrolyte stacked supercapacitors of 40F/14V and of 350F/14V, manufactured in Russia by ECOND Ltd.

If the requirements related to the voltage are respected, the performances of the supercapacitors are practically unchanged in a wide range of temperature, as experimental diagrams show in *Figure 3* (supercapacitor of 40F/14V).

The capacitance can be determined using the basic equation:

$$i = \frac{dq}{dt} = C \cdot \frac{du}{dt} \approx C \cdot \frac{\Delta u}{\Delta t}, \tag{1}$$

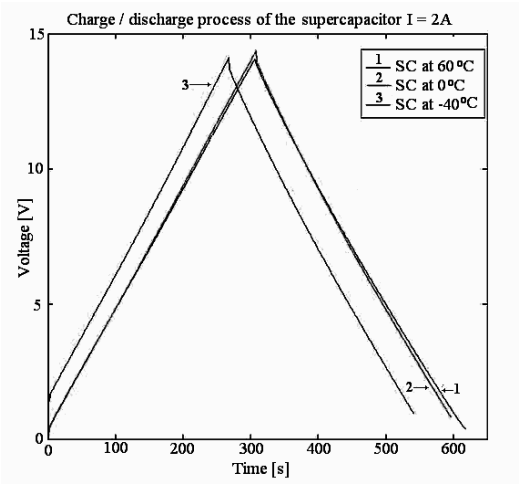


Figure 3: Charge / discharge process at constant current and different temperatures.

which gives the relation:

$$C = I \cdot \frac{t_2 - t_1}{U_2 - U_1}, \tag{2}$$

valid for I = constant. The indices 1 and 2 indicate two arbitrary moments of time (and the corresponding voltage values) either on the charging or on the discharging segment of the supercapacitor.

The capacitance of the 40F /14V supercapacitor (a charge / discharge process at constant slope) is practically not influenced by the temperature (in the range of -40°C to +60°C). The measurements for the charge / discharge process at a constant current of 2A were performed. In the model, ESR and C_2 are the parameters of the supercapacitors which are temperature dependent. Their influence on the equivalent total capacitance is practically negligible [7].

3.3 Determining the ESR parameter

ESR is an important parameter of the supercapacitors and depends on many factors such as the resistance of the electrode materials, the resistance of the electrolyte, the resistance of wires, the voltage and the frequency.

The influence of ESR occurs in the first milliseconds of the self discharge process, following the disconnection of the supercapacitor from the charging system. A plot of the self discharge process of supercapacitors is presented in *Figure 4*.

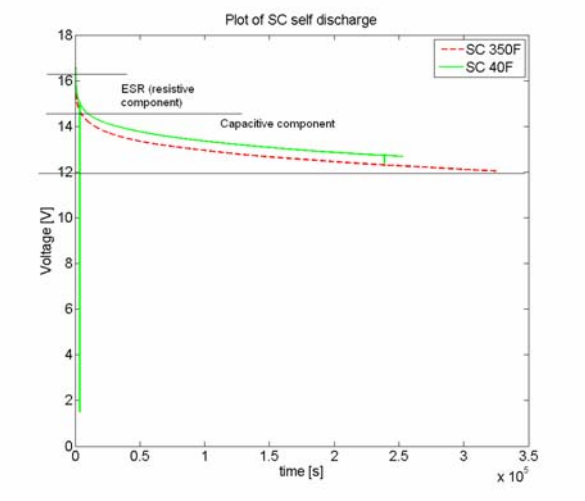


Figure 4: The self discharge process of the 350F and 40F supercapacitors.

Thus, after the ESR influence, the supercapacitor discharges on the internal resistance of losses. In order to observe that, both supercapacitors were charged to 16V with a 2A active load and the self discharge rate was monitored with the data acquisition system described above (the process was monitored during more than four days).

Impedance spectroscopy is one of the most important methods used to determine the equivalent impedance value of a storage device at various frequencies.

In our tests, in order to determine the ESR variation in function of frequency, we used the impedance spectroscopy method. We approximated the equivalent impedance measured by the impedance spectroscopy method with the value of the ESR, as at these very high values of the capacitance (40F and 350F) the capacitive reactance is negligible (at AC sinewave signal of 10Hz to 10 kHz). The test setup for impedance spectroscopy was already presented in *Figure 2.* Manufacturers usually specify the ESR value at a fixed frequency of 1 kHz [9].

Figure 5 presents voltage dependence of ESR of the 40F/14V supercapacitor at different frequency values.

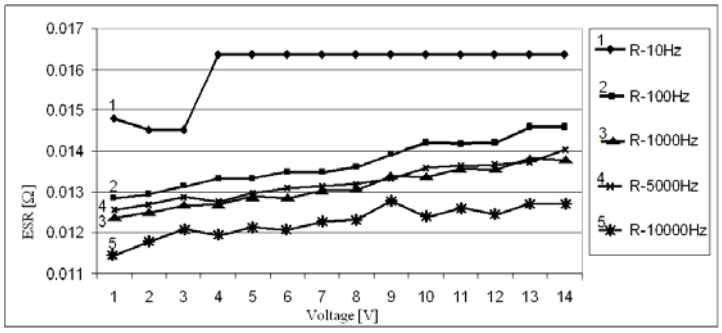


Figure 5: Voltage and frequency dependence of ESR (SC of 40F).

As *Figure 5* shows, the ESR value decreases with the frequency. The values of the recorded data are presented in *Table 1*.

Table 1: ESR at different frequencies for the 40F/14V supercapacitor.

Frequency [Hz]	Average ESR [mΩ]
10Hz	15.99
100Hz	13.69
1kHz	13.24
5kHz	13.23
10kHz	12.26

Table 2: ESR at different frequencies for the 350F/14V supercapacitor.

Frequency [Hz]	Average ESR [mΩ]
10Hz	1.8
100Hz	2.6
1kHz	4.1
5kHz	7.3
10kHz	8.7

In *Figure 6:* The ESR – voltage dependence of the 350F/14V supercapacitor is presented at different frequency values.

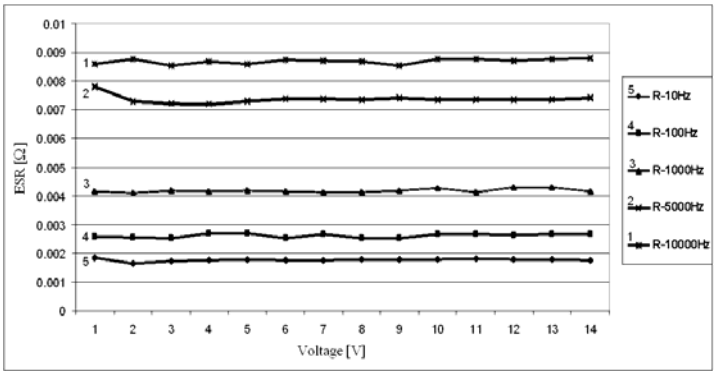


Figure 6: Voltage and frequency dependence of ESR (SC of 350F).

As *Figure 6* shows, the values of the ESR increase with the frequency and the values of the recorded data are presented in *Table 2*.

Depending on the final application, supercapacitors can be designed for high nominal voltage or high capacitance by connecting multiple packages in series and/or in parallel. Therefore internal structure of the supercapacitors is different depending on their voltage and capacitance, so the equivalent impedance can have different behaviours with the frequency (as it can be seen in *Table 1* and *Table 2*), originated in the internal parasitic impedances of SC.

3.4 Resistance of losses (self discharge resistance R_{sd})

The resistance of losses depends on the dielectric resistance and can be determined using the equation:

$$U_{SC}(t) = U_{SC_{\max}} \cdot e^{-\frac{\Delta t}{R_{sd} \cdot C}} \tag{3}$$

For the 350F/14V supercapacitor the average resistance of losses was determined as $7.9 \cdot 10^3$ Ω, with the average standard deviation of 0.32%. For the 40F/14V supercapacitor the value obtained was $7.5\ 10^4$ Ω, with the average standard deviation of 0.3%.

Figure 7 illustrates the variation of the losses resistance in different time intervals. In order to obtain this variation, five windows of 18,000 samples (3,000 considered, 15,000 ignored) from the self discharge process illustrated in *Figure 4* were considered. The average values of all samples from the five considered windows were calculated and are illustrated in *Figure 7*.

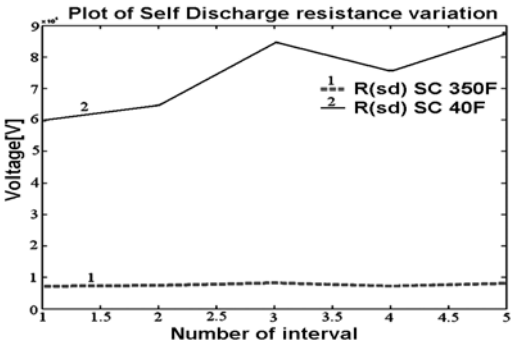


Figure 7: Variation of the self discharge resistance.

As *Figure 7* shows, the resistance of losses varies in time. Thus, in the case of the 350F/14V supercapacitor, for the first considered interval the resistance of losses was $7.1 \cdot 10^3$ Ω with a standard deviation of 0.39% and for the last considered interval the resistance of losses value was $8.3 \cdot 10^3$ Ω with a standard deviation of 0.26%. In the case of the 40F/14V supercapacitor the resistance of losses for the first considered interval was $5.9 \cdot 10^4$ Ω with a standard deviation of 0.42% and for the last considered interval the resistance of losses was $8.8 \cdot 10^4$ Ω with a standard deviation of 0.24%.

4. Energy efficiency

The power density expresses the maximum amount of energy transferred in a time unit, divided by the weight/volume, as the energy density represents the total energy that can be extracted from a supercapacitor, divided by the weight/volume. In order to improve the energy efficiency, the minimum voltage on SC during the discharging process has to be limited by the extraction factor (*d*), as it follows:

$$d = \frac{U_{\min}}{U_{\max}} \cdot 100\,\% \tag{4}$$

Usually, the value of the extraction factor is d = 50%. Thus, the useful energy can be computed using the equation:

$$W_u = W_{SC} - W_{\min} = W_{SC}\left(1 - \left(\frac{d}{100}\right)^2\right) \tag{5}$$

The energy efficiency is determined by:

$$\eta = \frac{W_u}{W_{SC}} \tag{6}$$

Table 3: Energy of the two types of supercapacitors.

	40F/14V	350F/14V
W_{SC} [Ws]	3920	34300
W_u [Ws]	980	8575

The electrical energy stored in the SC is determined as $W_{SC} = \frac{C \cdot U^2}{2}$. The considered extraction factor was 50% and thus it was calculated the useful energy that can be provided by the supercapacitors [10].

As it can be seen in Table 3, both supercapacitors can provide increased energy. More than that, if a supercapacitor is used with a battery, the system can assure the necessary amount of power and energy for many applications with high peak power requirements.

5. Conclusions

In order to determine the behaviour of the parameters of supercapacitors two different measurement methods have been implemented: volt-ampere method with constant-current charge and discharge process and impedance spectroscopy method. The measurements demonstrate a different behaviour of the SC under investigation (40F and 350F, 14V SCs), in relation with the voltage value and the test frequency applied. A different behaviour of the supercapacitors is observed depending on the internal serial and parallel connections of the capacitor cells (packages).

The measurements also demonstrate an insignificant influence of the temperature on the capacitance of SCs, in the temperature range of -40^{o}C to +60^{o}C. Also energy efficiency was determined.

Supercapacitors are suitable for multiple applications for a large range of temperatures, assuring high power requirements, having a long life span, low maintenance costs, and light weight compared with their capacitance.

Acknowledgements

The paper is a part of the national research project "TRANS-SUPERCAP" no. 21-018/2007 PNII/P4 CNMP, currently under development at "Transilvania" University of Braşov.

References

[1] Puşcaş, A. M., Coquery, G., Borza, P. N., Szekely, I., Carp, M., "State of the Art in Mobile Systems' Energy Management and Embedded Solutions for Improving the Energy Efficiency", *Bulletin of the Transilvania University of Brasov*, Vol. I (50), pp. 383-390, 2008.

[2] Taberna, P. L., Simon, P., "The role of the interfaces on supercapacitor performances" *ESSCAP'06*, Lausanne, Switzerland, 2006.

[3] Wackernagel, M. and Rees, W., "Our Ecological Footprint: Reducing Human Impact on the Earth", Gabriola Island, 1996.

[4] Sorjef, D., Borza, P. N., "Comparison of high-voltage supercapacitor approaches and case study in diesel locomotive starting system", *ESSCAP'08, 3rd European Symposium on Supercapacitors and applications*, Roma, 2008.

[5] Chen, J. H., Li, W. Z., Huang, Z. P., Wang, D. Z., Yang, S. X., Wen, J. G., Ren, Z. F., "Electrochemistry of carbon nanotubes and their potential application in supercapacitors",

Proceedings of the 197th Meeting of Electrochemical Society, Toronto, Canada, May 14-18, 2000.

[6] Barrade, P., Rufer, A., "Energy Storage and applications with supercapacitors", *ESSCAP'07*, 2007.

[7] Puşcaş, A. M., Muşat, R., Carp, M. C., Borza, P., Helerea, E., "Energetical Monitoring of the storage devices by using sensors networks placed inside the mobile systems", "AFASES -2009", *The international Session of XI-th scientific papers, Scientific research and education in air force*, 20-22 May, 2009.

[8] ***, Project "TRANS - SUPERCAP", PNII-P4, 21-018/14.09.2007,2007.

[9] ***, Linear Technology, "Supercapacitors can replace a backup battery for power ride/through applications", Design Notes, design note 450.

[10] Sikora, A., "Idea of supercapacitor supporting modules for alternative power solutions", *ESSCAP'06*, Lausanne, Switzerland, 2006.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 77-92

Investigating a Novel Model of Human Blood Glucose System at Molecular Levels from Control Theory Point of View

András GYÖRGY[1], Levente KOVÁCS[1],
Tamás HAIDEGGER[1], Balázs BENYÓ[1]

[1]Department of Control Engineering and Information Technology,
Faculty of Electrical Engineering and Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: andrew.gyorgy@gmail.com, lkovacs@iit.bme.hu,
haidegger@gmail.com, bbenyo@iit.bme.hu

Manuscript received March 15, 2009; revised May 08, 2009.

Abstract: According to the data provided by the World Health Organization (WHO) diabetes has become an endemic of these days. There are several nonlinear models describing the dynamic of glucose-insulin of diabetes mellitus, like the simplest one with only three state variables, also known as the model of Bergman, and the most complex with 19 state variables, the model of Sorensen. Their common characteristic is that they describe type 1 diabetes physiologically. A recently published theoretical model [1] is capable of describing human blood glucose system at molecular levels. This paper is based on its analysis from a control theory point of view with multiple purposes: nonlinear analysis, rank reduction possibilities with physiological explanations, defining physiological working points for further polytopic modeling, analyzing control properties of the linear systems in the defined working points.

Keywords: Diabetes, nonlinear analysis, model reduction, physiologic working points.

1. Introduction

The normal blood glucose concentration level in the human body varies in a narrow range (70 - 110 mg/dL). If for some reason the human body is unable to control the normal glucose-insulin interaction (e.g. the glucose concentration level is constantly out of the above mentioned range), diabetes is diagnosed. The consequences of diabetes are mostly long-term: among others, diabetes increases the risk of cardiovascular diseases, neuropathy and retinopathy. Four

types of diabetes are known: type 1 (also known as insulin-dependent diabetes mellitus), type 2 (or insulin- independent diabetes mellitus), gestational diabetes and other special types, like genetic deflections. Consequently, diabetes mellitus is a serious metabolic disease, which should be artificially regulated.

The newest statistics of the World Health Organization (WHO) predate an increase of adult diabetes population from 4% (in 2000, meaning 171 million people) to 5.4% (366 million worldwide) by the year 2030 [2]. This is a warning that diabetes could be the "disease of the future", especially in developing countries (due to stress and unhealthy lifestyle).

To design an appropriate control, an adequate model is necessary. In the last decades several models appeared for type 1 diabetic patients [3]. The most widely used and also the simplest one proved to be the minimal model of Bergman [4], for type 1 diabetic patients under intensive care, and its extension, the three-state minimal model [5]. However, the simplicity of the model proved to be its disadvantage too, since in its formulation a lot of components of the glucose-insulin interaction were neglected.

Besides the Bergman-model other models appeared in the literature [6]-[8], which are more general, but more complicated. The most complex one proved to be the 19^{th} order Sorensen-model [6], which is based on the earlier model of [8]. Even if the Sorensen-model describes the human blood glucose dynamics in a very exact way, it is rarely used in research problems due to its complexity.

2. The molecular model

In contrast with the earlier phenomenological aspect, the model applies a more accurate approach [1] published in 2008; it describes the human blood glucose system at molecular levels. Consequently, the cause-effect relations are more plausible and different functions and processes can be separated. The considered model is approximately halfway from Bergman's model [4]-[5] to Sorensen's [6] with its 8 state variables and it can be naturally divided into three subsystems: the transition subsystem of glucagon and insulin, the receptor binding subsystem and the glucose subsystem. Parameters of the model can be found in [1].

A. Transition subsystem

We assume that plasma insulin does not act directly on glucose metabolism but through cellular insulin [9]. Let s_1^p and s_2^p denote concentrations of plasma glucagon and insulin, respectively. Complementing equations of [10] with transition delay the subsystem can be described with

$$\frac{ds_j^p}{dt} = -\left(k_{j,1}^p + k_{j,2}^p\right)s_j^p + w_j \qquad j = 1,2 \tag{1}$$

where w_1 and w_2 stand for glucagon and insulin produced by the pancreas. The equations show that the hormones of pancreas have a positive effect on their plasma concentrations, while the hormones in plasma can be interpreted as a negative feedback.

The positive constants $k_{j,1}^p$ denote transition rates and $k_{j,2}^p$ the degradation rates (j=1,2). Contrary to [10], we suppose that intracellular insulin cannot go back to plasma, which is in harmony with Bergman's minimal model [4]-[5].

B. Receptor binding subsystem

Let s_1 and s_2 denote intracellular concentrations of glucagon and insulin, whereas r_1 and r_2 stand for concentrations of glucagon- and insulin-bound receptors, respectively. Assuming that the receptor recycling system is closed intracellular concentrations can be described with

$$\frac{ds_j}{dt} = -k_{j,1}^s s_j\left(R_j^0 - r_j\right) - k_{j,2}^s s_j + k_{j,1}^p s_j^p V_p V^{-1} \qquad j = 1,2\,, \tag{2}$$

$$\frac{dr_j}{dt} = k_{j,1}^s s_j\left(R_j^0 - r_j\right) - k_j^r r_j \qquad j = 1,2 \tag{3}$$

where R_1^0 and R_2^0 denote total concentrations of receptors, $k_{j,1}^s$ stand for the hormone-receptor association rates, $k_{j,2}^s$ the degradation rates, k_j^r the inactivation rates (j=1,2). V_p is plasma volume, whereas V is intracellular volume.

C. Glucose subsystem

Blood glucose has two sources: endogenous hepatic production with glycogen transformation and exogenous meal intake. Glucose utilization can be divided into two groups: insulin-independent (brain and nerve cells) and insulin-dependent (muscle and adipose tissues).

Insulin-independent part [12] can be modelled by

$$f_1(g_2) = U_b\left(1 - e^{-\frac{q_2}{C_2}}\right) \tag{4}$$

that saturates at 500 mg/l (q_2 denotes glucose concentration).

Insulin-dependent part can be calculated by the product

$$f_2(q_2)f_3(s_2)=\frac{q_2}{C_3}\left\{U_0+(U_m-U_0)\left(\frac{s_2}{C_4}\right)^{\beta}\left[1+\left(\frac{s_2}{C_4}\right)^{\beta}\right]^{-1}\right\} \tag{5}$$

which was originally used in [13]. $f_3(s_2)$ saturates at insulin concentration 500 mU/l.

Concluding the assumptions the glucose subsystem can be described with

$$\frac{dq_1}{dt}=\frac{k_1r_2}{1+k_2r_1}\frac{V_{\max}^{gs}q_2}{K_m^{gs}+q_2}-k_3r_1\frac{V_{\max}^{gp}q_1}{K_m^{gp}+q_1}, \tag{6}$$

$$\frac{dq_2}{dt}=-\frac{k_1r_2}{1+k_2r_1}\frac{V_{\max}^{gs}q_2}{K_m^{gs}+q_2}+k_3r_1\frac{V_{\max}^{gp}q_1}{K_m^{gp}+q_1}-f_1(q_2)-f_2(q_2)f_3(s_2)+G_{in}, \tag{7}$$

where q_1 and q_2 denote glycogen and glucose concentration, v^{gp} and v^{gs} stand for reaction rate of glycogen phosphorylase and glycogen synthase, respectively. $V_{\max}^{gp}$ and $V_{\max}^{gs}$ are maximal reaction rates of the enzymes whereas K_m^{gp} and K_m^{gs} are their Michaelis-Menten constants. Exogenous glucose intake is denoted by G_{in}.

D. Pancreatic control

Hormones of the pancreas have a cardinal role in blood glucose regulation and homeostatic stability, since negative feedback of glucagon and insulin assures controllability. Control mechanism of the pancreas [9] is described with

$$w_1(q_2)=\frac{G_m}{1+b_1e^{a_1(q_2-C_5)}}, \tag{8}$$

$$w_2(q_2)=\frac{R_m}{1+b_2e^{a_2(C_1-q_2)}}, \tag{9}$$

where w_1 and w_2 denote glucagon (GIR) and insulin infusion rates (IIR), respectively.

E. Aspect of control theory

As for inputs, exogenous insulin (u_1) is completely disposable, since it is in daily use in the form of injection (type 1 diabetes is treated this way). Glucose

taken as meal (G_{in}) represents disturbance for the model, but as a result of more profound consideration it can be regarded as control input. Healthy people use it more or less to regulate their blood glucose level, but in case of diabetic patients the situation is crystal clear: it is strictly prescribed for them what to eat and when to eat, so exogenous glucose is treated as a second control input henceforward (u_2). In order to analyze the model in a quantitative manner, a physiologically correct exogenous glucose input has to be defined. According to the literature a widely used absorption curve is applied which was recorded under extremely strict and precise conditions [14].

As for outputs, blood glucose level (q_2) is essential to characterize the system (in addition it can be measured easily). Concentration of plasma insulin (s_2^p) is only measurable under laboratory conditions, but any controller designed to regulate pathologic blood glucose system has to be qualified by the amount of injected insulin. Summarizing the considerations, the outputs of the model are plasma insulin (y_1) and blood glucose level (y_2).

3. Nonlinear analysis

In this section global characteristics of the molecular model are observed from a differential geometric point of view. Differential geometry deals with differential equations defined over differentiable manifolds, hence dynamic systems and their trajectories can be analyzed. Main definitions and ideas of differential geometry can be found in [15].

A. Nonlinear model

The molecular model presented in Section 2 has to be formulated exactly. Let f, g_1 and g_2 denote vector fields over an eight dimensional manifold, h_1 and h_2 stand for real-valued functions. In this case, the input-affine nonlinear system Σ can be described with

$$\begin{aligned} \dot{x} &= f(x) + \sum_{i=1}^{2} u_i g_i(x) \\ y_i &= h_i(x) \qquad i = 1,2 \end{aligned}, \tag{10}$$

where

$$f = \begin{bmatrix} -\left(k_{11}^{p} + k_{12}^{p}\right)x_1 + w_1 \\ -\left(k_{21}^{p} + k_{22}^{p}\right)x_2 + w_2 \\ -k_{11}^{s}x_3\left(R_1^0 - r_1\right) - k_{12}^{s}x_3 + k_{11}^{p}x_1\dfrac{V_p}{V} \\ -k_{21}^{s}x_4\left(R_2^0 - r_2\right) - k_{22}^{s}x_4 + k_{21}^{p}x_2\dfrac{V_p}{V} \\ k_{11}^{s}x_3\left(R_1^0 - x_5\right) - k_1^{r}x_5 \\ k_{21}^{s}x_4\left(R_2^0 - x_6\right) - k_2^{r}x_6 \\ \dfrac{k_1x_6}{1+k_2x_5}\dfrac{V_{\max}^{gs}x_8}{K_m^{gs}+x_8} - k_3x_5\dfrac{V_{\max}^{gp}x_7}{K_m^{g1}+x_7} \\ -\dfrac{k_1x_6}{1+k_2x_5}\dfrac{V_{\max}^{gs}x_8}{K_m^{gs}+x_8} + k_3r_1\dfrac{V_{\max}^{gp}x_7}{K_m^{g1}+x_7} - f_1\left(x_8\right) - f_2\left(x_8\right)f_3\left(x_4\right) \end{bmatrix}, \tag{11}$$

$$g = \begin{bmatrix} g_1 & g_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T, \tag{12}$$

$$h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} s_2^p \\ g_2 \end{bmatrix}. \tag{13}$$

B. Reachability

Let Δ^C be a nonsingular involutive distribution of dimension *d* and assume that Δ^C is invariant under the vector fields $f, g_1, g_2, \ldots, g_m$. Moreover, suppose that the distribution $span\{g_1, \ldots, g_m\}$ is contained in Δ^C. Then, for each point x_0 it is possible to find a neighborhood U_0 of x_0 and a local coordinate transformation $z = \Psi(x)$ defined on U_0 in such a way that in the new coordinates, the control system Σ is represented by equations of the form

$$\begin{aligned} \dot{\varsigma}_1 &= f_1\left(\varsigma_1, \varsigma_2\right) + \sum_{i=1}^{m} g_{1i}\left(\varsigma_1, \varsigma_2\right)u_i \\ \dot{\varsigma}_2 &= f_2\left(\varsigma_2\right) \\ y_i &= h_i\left(\varsigma_1, \varsigma_2\right) \end{aligned}, \tag{14}$$

where $\varsigma_1 = \left(z_1, z_2, \ldots, z_d\right)$ and $\varsigma_2 = \left(z_{d+1}, z_{d+2}, \ldots, z_n\right)$ [15].

In this case ς_1 is locally reachable, since it can be manipulated by the inputs of Σ while ς_2 cannot be controlled. Consequently, the number of reachable states is equal to the rank of distribution Δ^C.

Construction of distribution Δ^C:

- initialization $\Delta_0^C = span\{f, g_1, \ldots, g_m\}$,
- expansion of distribution Δ^C

$$\Delta_{k+1}^C = \Delta_k^C + \sum_{i=1}^{q} [\tau_i, \Delta_k^C], \tag{15}$$

until rank Δ_{k+1}^C >rank Δ_k^C, where $\tau_i \in \Delta_k^C$, $i = 1,2,\ldots,q$ ($\dim \Delta_k^C = q$).

C. Observability

Let $d\Delta^O(x) \subset (R^n)^*$ denote the subspace containing the row vectors $d\alpha(x)$ for $\forall x \in X$, where $\alpha \in O$ (observation space). Moreover, suppose that for each point x_0 it is possible to find a neighborhood U_0 of x_0 so that $d\Delta^O(x) = d < n$, $\forall x \in U_0$. In this case a local coordinate transformation $z = \Psi(x)$ defined on U_0 transforms the control system Σ to the form

$$\begin{aligned} \dot{\varsigma}_1 &= f_1(\varsigma_1) + \sum_{i=1}^{m} g_{1i}(\varsigma_1) u_i \\ \dot{\varsigma}_2 &= f_2(\varsigma_1, \varsigma_2) + \sum_{i=1}^{m} g_{2i}(\varsigma_1, \varsigma_2) u_i \\ y_i &= h_i(\varsigma_1) \end{aligned} \tag{16}$$

where $\varsigma_1 = (z_1, z_2, \ldots, z_d)$ and $\varsigma_2 = (z_{d+1}, z_{d+2}, \ldots, z_n)$ [15].

In this case ς_1 is locally observable since it appears in the outputs of Σ, while ς_2 cannot be observed because it does not show up either in the outputs of Σ or in ς_1. Consequently, the number of observable states is equal to the rank of codistribution $d\Delta^O$.

Construction of codistribution $d\Delta^O$: expansion of observation space O with Lie-derivatives until the rank of $d\Delta^O$ increases.

C. Input-output linearization

If there is a relative degree vector $r = (r_1, r_2, \ldots, r_m)$, open set $U(x_0)$, assignment $v = q(x) + S(x)u$, smooth function $q: U \to R^m$ and $S: U \to R^{m \times m}$ where $\det S(x_0) \neq 0$ for the nonlinear system Σ so that $y_i^{(r_i)} = v_i$, $i = 1,2,\ldots,m$,

then the system can be decomposed into *m* subsystems with r_i integrators in the i^{th} subsystem [15].

If there is no zero dynamics, the system can be input-output linearized with the static feedback $u(x) = -S^{-1}(x)q(x) + S^{-1}(x)v$ where v is the new input vector, q is feedback, so the linearized system can be transformed into Brunovsky-form.

D. Global control characteristics

To be able investigating the global control characteristics of the molecular model we have implemented under MATLAB the algorithms presented above (sections 3.A and 3.B). The following results were obtained:

- completely reachable system, since $rank\,\Delta^C = 8$,
- number of the observable states of the model is 4, since $rank\,d\Delta^O = 4$,
- static feedback results in such complex vector fields that MATLAB is unable to handle them (manually it is also too complex), so this question cannot be answered this way. Linearization with dynamic feedback (dynamic extension, Cartan fields) has the same problem.

4. Linear analysis

Global characteristics of the molecular model are examined in Section 3 and led to the conclusion that despite the great importance of the achieved results practical application is very difficult because of the extreme complexity of the generated vector fields. Furthermore, ulterior aim of the research is polytopic modeling of the system, hence linearization and model reduction possibilities are observed in this section. In this manner local characteristics of the molecular model can be determined.

A. Steady state linearization

Linearization is carried out by applying Jacobian matrices in a steady state. The linearized form of system

$$\begin{aligned} \dot{x} &= f(x,u) \\ y &= h(x,u) \end{aligned} \tag{17}$$

in steady state (x_0, u_0) with output y_0 is

$$\begin{aligned} \dot{\tilde{x}} &= A\tilde{x} + B\tilde{u} \\ \tilde{y} &= C\tilde{x} + D\tilde{u} \end{aligned}, \tag{18}$$

where

$$\begin{aligned} \tilde{x} &= x - x_0 \\ \tilde{u} &= u - u_0 \\ \tilde{y} &= y - y_0 \end{aligned} \tag{19}$$

and

$$A = \left.\frac{\partial f}{\partial x}\right|_{(x_0,u_0)}, \tag{20}$$

$$B = \left.\frac{\partial f}{\partial u}\right|_{(x_0,u_0)}, \tag{21}$$

$$C = \left.\frac{\partial h}{\partial x}\right|_{(x_0,u_0)}, \tag{22}$$

$$D = \left.\frac{\partial h}{\partial u}\right|_{(x_0,u_0)}. \tag{23}$$

B. Corner points

Observing simulation results it can be seen that blood glucose varies between 700-1800 mg/l. An obvious choice could be searching for steady points within this range and approximating the nonlinear model with these steady states. With a resolution of 100 mg/l the twelve determined steady states (in our terms corner points) are stable, completely controllable and completely observable.

In order to approximate the nonlinear system third degree polynomial functions can be fit to the corner points interpolating the non-determined steady states. Applying the glucose and insulin input presented in Section 2 to the system that is determined by linearizing the nonlinear model along the approximation function in each variable, the system becomes instable and its responses are meaningless. This is because linearization is only precise in the neighborhood of the actual steady state, but the presented method tries to describe the system in distant regions of the state space with only one variable (blood glucose) which is far too imprecise. Consequently, another method has to be chosen in order to reduce complexity of the nonlinear model.

C. Physiologic working points and further LPV modeling

Polytopic approach of LPV modeling can only be applied if the linear models are stable and cover the operating area in a more or less uniform way. In order to fulfill these conditions physiologic working points (PWPs) are defined: these are state vectors that are not exact solutions of the differential equations describing the molecular model but derived from a steady state.

Applying only five different values for each state variable (which is a rather rough quantization) almost 400000 PWPs should be considered. Exponential explosion is down-to-earth, hence complexity reduction is crucial.

Normalization of the trajectories of the molecular model (see *Fig. 1 and Fig. 2*) to [0,1] results in a valuable experiment: variables can be divided into two groups. One of them is the glucagon-type variables (see *Fig. 1*) whereas the other is the group of insulin-type variables (see *Fig. 2*). Glycogen is the only variable that does not fit perfectly into either group (but it can be categorized as an insulin-type variable), which is not surprising, since glycogen is the stored form of glucose that can be interpreted as the integral of the glucose excess (saturation can be remarked after linear phase).

As a result of biochemical and physiological considerations variables are divided into two groups: glucagon- and insulin-type. PWPs are generated by multiplying the normoglycaemic values [1] by [0.25 0.5 0.75 1 1.25 1.5 2 4] in case of both group, value of glycogen is not modified. The created 64 PWPs are stable, completely controllable and observable, hence further polytopic modeling can be fulfilled.

D. Model reduction

Considering the results it is probable that the complexity of the model can be reduced since variables are not independent in a physiologic sense. Many methods have been published in the subject of linear model reduction, one of the most widely used is based on state space transformation and projection to a subspace [16].

The aim is to determine a minimal set of state variables producing almost the same input-output behavior as the original system. Input-output behavior of a linear, autonomous system remains unchanged after a linear, nonsingular state space transformation.

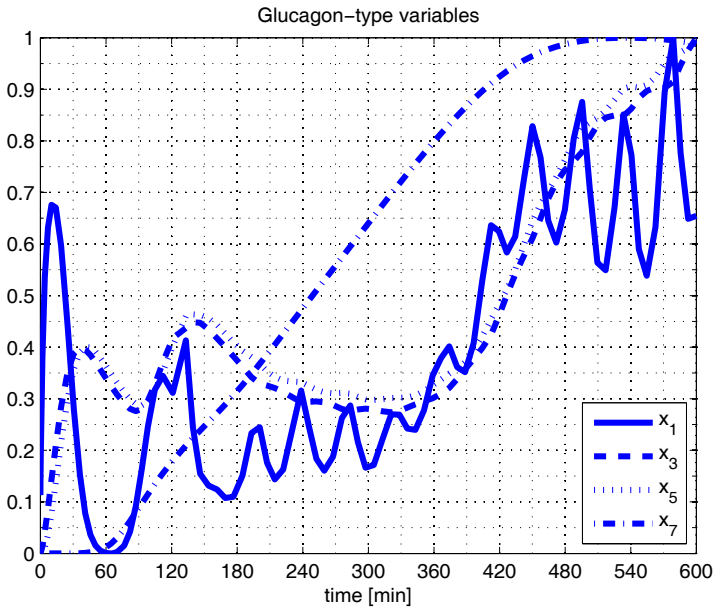


Figure 1: Glucagon-type variables.

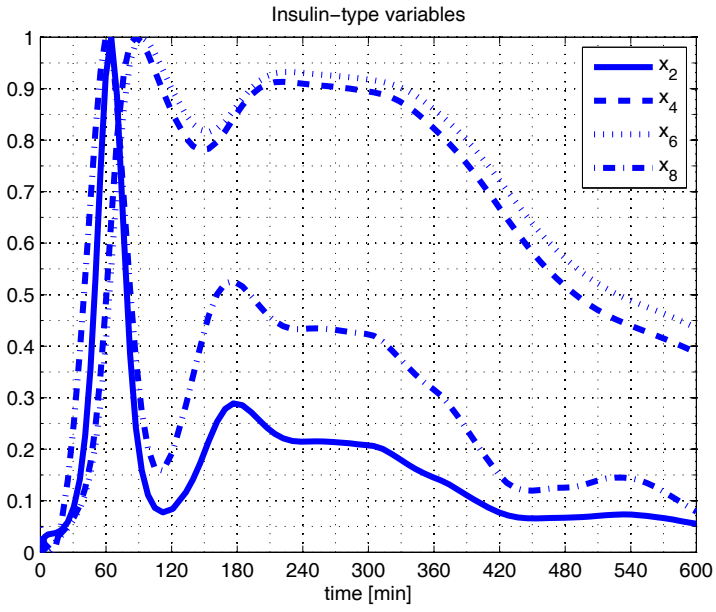


Figure 2: Insulin-type variables.

The method is supported by MATLAB Control System and Robust Toolbox. In case of linearized systems model reduction can be realized as follows: numeric conditioning (ssbal), input-output balancing (obalreal), model reduction based on frequency domain (modred) and determination of state space transformation.

Function balmr() of the MATLAB Robust Control toolbox executes model reduction by minimizing the difference between the H_∞ norms of the original and the reduced systems in frequency domain, but information on state space transformation is lost. This information can be gathered by applying the above-mentioned algorithm.

As a result of reduction Hankel Singular Values of the transformed system Ξ (obtained by MATLAB) are: 9.64, 5.89, 1.37, 1.10, $1.38 \cdot 10^{-2}$, $4.32 \cdot 10^{-3}$, $7.03 \cdot 10^{-5}$ and $1.31 \cdot 10^{-6}$. Hankel Singular Values represent the relative importance of the state variables independently from realization. Consequently, model reduction can be fulfilled by omitting the state variables with small Hankel Singular Values. It can be seen that the structure of the state variables is the same as in the previous subsection: the first two Hankel Singular Values are much greater than the others validating the considerations applied in case of determination of PWPs.

E. Analyzing the results of reduction

Linearized model in the normoglycaemic state is reduced with different ranks. Responses of the reduced models for 5% perturbation in the initial conditions can be seen in *Fig. 3* and *Fig. 4*. Observing the trajectories it can be seen that the structure described above is plausible since the behavior of the original model can be more or less imitated with only two state variables and with the model of rank four no significant development is achieved.

Results of the model reduction are examined in time and frequency domain. Since the models have two inputs and two outputs four transfer functions can be defined. In time domain impulse responses of the original, linearized model and the reduced models are compared (see *Fig. 5* for one of the four possible transfer functions), whereas in frequency domain Bode diagrams are collated (see *Fig. 6* for one of the four possible transfer functions).

The investigation is realized in the normoglycaemic steady state, but it should be done in case of any steady state. The important frequency range is 0.0002-0.2 rad/min [17]: noise dominates in higher frequencies and the dynamics of the sensor and the actuator (insulin pump) takes place in this range.

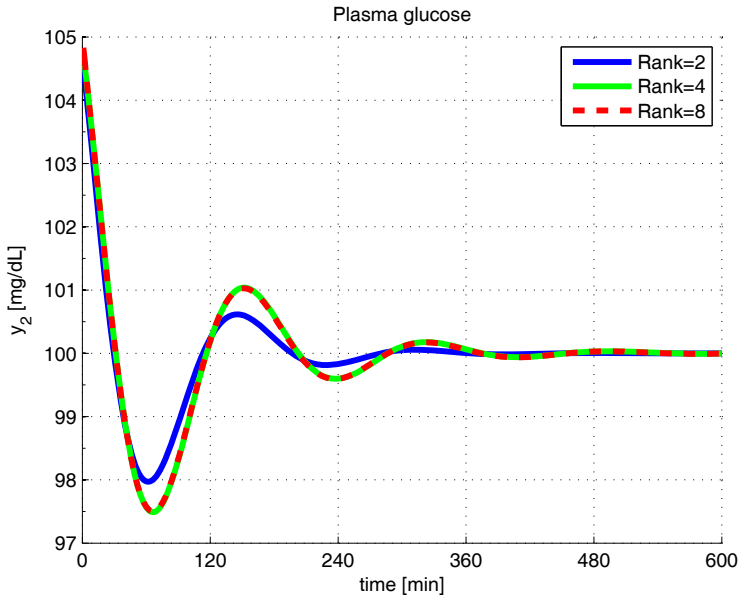


Figure 3: Plasma glucose concentration responses for 5% deviation at the normoglycaemic steady state of the reduced models of different ranks.

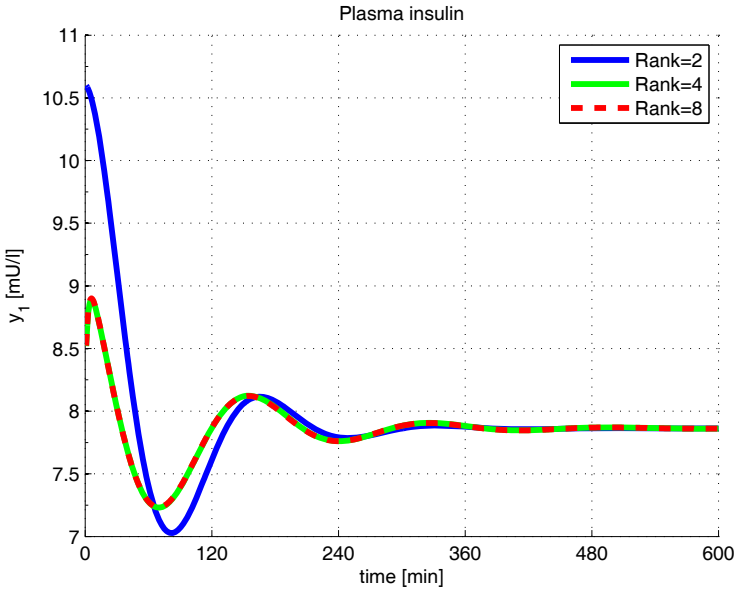


Figure 4: Plasma insulin concentration responses for 5% deviation at the normoglycaemic steady state of the reduced models of different ranks.

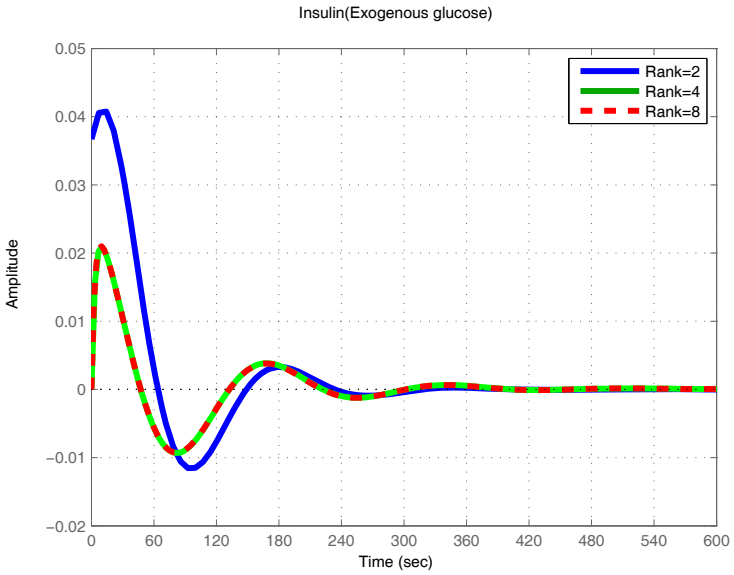


Figure 5: Impulse responses of Insulin(Exogenous glucose) transfer functions of the linearized models of different ranks.

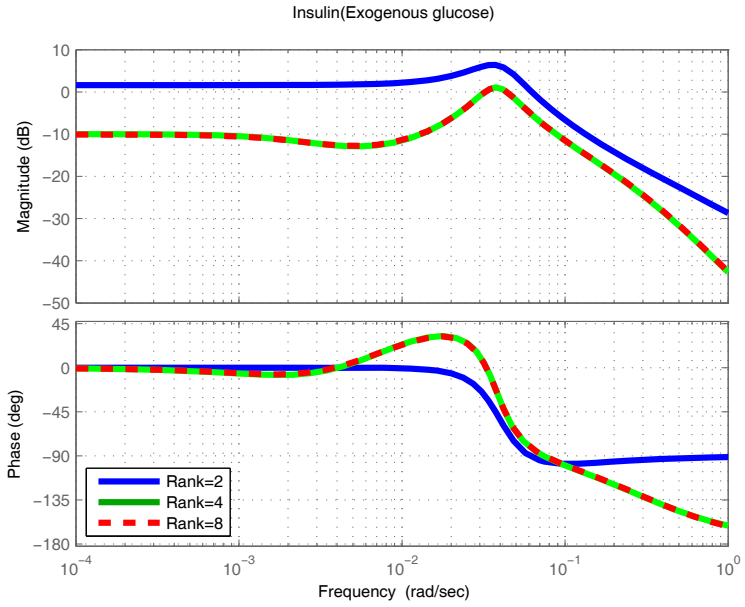


Figure 6: Bode-diagrams of Insulin(Exogenous glucose) transfer functions of the linearized models of different ranks.

Examination results in reassuring observations: the second order model is more or less similar to the original behavior, but the precision is not enough. In contrast with this, the fourth order approximation produces almost identical behavior that the original, linearized model in frequency and time domain as well.

Summarizing the achieved results, we can state that second order approximation is not enough, but fourth order is almost perfect which is not surprising considering biochemical and physiological principles.

5. Conclusion

This paper focused on a novel model offering a radical change in approach. As a result of the applied molecular point of view the cause-effect relations are more plausible and the processes can be described in a more exact and precise way.

After a brief review of the earlier results, the molecular model was presented and described in detail from the aspect of control theory. Global control properties were determined by nonlinear analysis. Nonlinear analysis was followed by steady state linearization. First corner points were defined, but this approach could not ensure proper approximation of the model, hence physiological working points (PWPs) were defined for further LPV modeling. In order to reduce complexity model reduction possibilities were observed with physiological concerns as well as with mathematical ones and the results agreed.

Physiological, biochemical and mathematical approaches were applied and conclusions were made by synchronizing the principles of the different fields of study.

Acknowledgements

This research is supported in part by Hungarian National Scientific Research Foundation, Grants No. OTKA T69055.

References

[1] Liu, W., and Fusheng, T., “Modeling a simplified regulatory system of blood glucose at molecular levels”, *Journal of Theoretical Biology*, Vol. 252, pp. 608-620, 2008.

[2] Wild, S., Roglic, G., Green, A., Sicree, R., and King, H., “Global prevalence of diabetes - Estimates for the year 2000 and projections for 2030,” *Diabetes Care*, Vol. 27, No. 5, pp. 1047–1053, May 2004.

[3] Chee, F., and Fernando, T., “Closed-loop control of blood glucose”, Springer, Berlin, 2007.

[4] Bergman, B. N., Ider, Y. Z., Bowden, C. R., and Cobelli, C., "Quantitive estimation of insulin sensitivity," *American Journal of Physiology*, Vol. 236, pp. 667–677, Jun. 1979.
[5] Bergman, R. N., Philips, L. S., and Cobelli, C., "Physiologic evaluation of factors controlling glucose tolerance in man," *Journal of Clinical Investigation*, Vol. 68, pp. 1456–1467, Dec. 1981.
[6] Sorensen, J. T., "A physiologic model of glucose metabolism in man and its use to design and assess improved insulin therapies for diabetes," *PhD Thesis, Dept. of Chemical Eng. Massachusetts Institute of Technology*, Cambridge, 1985.
[7] Hovorka, R., Shojaee-Moradie, F., Carroll, P. V., Chassin, L. J., Gowrie, I. J., Jackson, N. C., Tudor, R. S., Umpleby, A. M., and Jones, R. H., "Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT," *American Journal Physiology Endocrinology Metabolism*, Vol. 282, pp. 992–1007, Jan. 2002.
[8] Guyton, J. R., Foster, R. O., Soeldner, J. S., Tan, M. H., Kahn, C. B., Koncz, L., and Gleason, R. E., "A model of glucose-insulin homeostasis in man that incorporates the heterogeneous fast pool theory of pancreatic insulin release," *Diabetes*, Vol. 27, 1027, Oct. 1978.
[9] Bergman, R. N., Finegood, D. T., and Ader, M., "Assessment of insulin sensitivity in vivo", *Endocrine Rev*., Vol. 6, pp. 45–86, 1985.
[10] Sturis, J., Polonsky, K. S., Mosekilde, E., and Cauter, E. V., "Computer model for mechanisms underlying ultradian oscillations of insulin and glucose", *American Journal of Physiology, Endocrinology and Metabolism*, 260, pp. 801–809, 1991.
[11] Sedaghat, A. R., Sherman, A., Quon, M. J., "A mathematical model of metabolic insulin signaling pathways" *American Journal of Phisiology, Endocrinology and Metabolism*, Vol. 283, pp. 1084–1101, 2002.
[12] Turner, R. C., Holman, R. R. Matthews, D., Hockaday, T. D., and Peto, J., "Insulin deficiency and insulin resistance interaction in diabetes: estimation of their relative contribution by feedback analysis from basal plasma insulin and glucose concentrations", *Metabolism*, Vol. 28 (11), pp. 1086–1096, 1979.
[13] Rizza, R. A., Mandarino, L. J., and Gerich, J. E., "Dose-response characteristics for effects of insulin on production and utilization of glucose in man", *American Journal of Physiology, Endocrinology and Metabolism*, Vol. 240, pp. 630–639, 1981.
[14] Korach-André, M., Roth, H., Barnoud, D., Péan, M., Péronnet, F., and Leverve, X., "Glucose appearance in the peripheral circulation and liver glucose output in men after a large 13C starch meal", *American Journal of Clinical Nutrition*, 80, pp. 881-886, 2004.
[15] Isidori, A., "Nonlinear control systems", Springer, Berlin, 1995.
[16] Willcox, K., and Peraire, J., "Balanced model reduction via the proper orthogonal decomposition", *AIAA Journal*, Vol. 40, No. 11, 2002.
[17] Parker, R. S., Doyle III, F. J., Ward, J. H., and Peppas, N. A. "Robust H∞ glucose control in diabetes using a physiological model", *AIChE Journal*, 46 (12), pp. 2537-2549, 2000.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 93-104

FPGA Implementation of Fuzzy Controllers and Simulation Based on a Fuzzy Controlled Mobile Robot

Sándor Tihamér BRASSAI

Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tîrgu Mureş, Romania,
e-mail: tiha@ms.sapientia.ro

Manuscript received March 15, 2009; revised June 30, 2009.

Abstract: Fuzzy controllers offer a simple solution to the problem of controlling a system. Combining the simplicity of fuzzy systems with the parallel implementation on FPGA circuits, a very fast controller can be obtained. This paper presents a fuzzy controller implemented on an FPGA circuit, with membership functions, rule table, inference system and defuzzification modules implemented on this hardware tool. In the final part of the paper the control of a mobile robot with a fuzzy controller is exemplified.

Keywords: FPGA, hardware implementation, fuzzy control, mobile robot.

1. General information

The main purpose of the paper is presentation and practical application of a fuzzy controller implemented on a digital reconfigurable hardware (FPGA). There are several methods for implementation of fuzzy controllers or other discrete controllers, such as implementation on PC, microcontrollers, DSP digital signal processors, DSC digital signal controllers, PLC, on ASIC circuits or FPGAs. In this paper, a fuzzy controller implemented on FPGA is discussed. Some of the FPGA implementation advantages are: parallelization possibility, the advantages resulting from reconfigurability (scalable size of the fuzzy processor) and very high speed operation.

For the description of the controller's structure, VHDL hardware [1], [2], [3] description language was used with XILINX ISE development software. As a target device, a NEXYS-2 development board with SPARTAN3E 1200 FPGA

chip is proposed. In the paper, a short overview of the components of a fuzzy controller (fuzzification, defuzzification, inference system, rule table [4], [5]) as well as the FPGA-based implementation of these components are described. The controller system is proposed to be used for a mobile robot track-following problem.

In the first step of the controller design, the controller applicability was tested for the mentioned track-following problem by using a model for the robot simulation, which will also be discussed. The robot model will also be presented in the work. The results of the simulation are described. The paper ends with the conclusions and future ideas.

2. The control structure and the robot model

For the control of a mobile robot, a fuzzy controller was chosen because it is an efficient method for system control. In case of the fuzzy controller, the control strategy can be described based on simple rules such as (the examples will be given for the mobile robot):

- *if* the target point is on the right hand side in front of the robot*, then* the robot has to move forward rotating to the right
- *if* the target point for the robot is behind, *then* the robot has to turn back (see *Fig. 1.*)

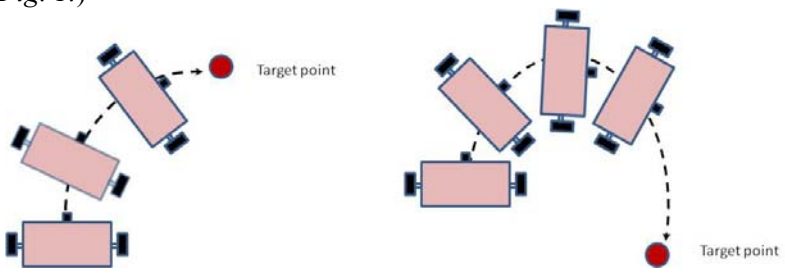


Figure 1: Rule exemplification.

The main objective is to control the robot to follow a predefined target trajectory. A robot with three wheels has been used. The left and right side wheels are driven by DC motors, the third wheel is a free wheel. The robot model used in simulations is presented in the following equations [6].

$$x[k+1] = x[k] + L\frac{v2+v1}{\xi + |v1-v2|}\sin(\varphi[k]) \tag{1}$$

$$y[k+1] = y[k] + L\frac{v2+v1}{\xi + |v1-v2|}\cos(\varphi[k]) \tag{2}$$

$$\varphi[k+1]=\varphi[k]+\frac{v2-v1}{2L}T \tag{3}$$

In *Fig. 2.* the XY coordinate system in which the robot's position is measured and the X_RY_R coordinate system attached to the robot are shown, where:

(x_r,y_r) - the robot's position in the main XY coordinate system;

φ -the robot's orientation in the XY coordinate system;

L - the distance between the robot's side wheels;

θ_r - the next target point orientation measured in the X_RY_R coordinate system attached to the robot.

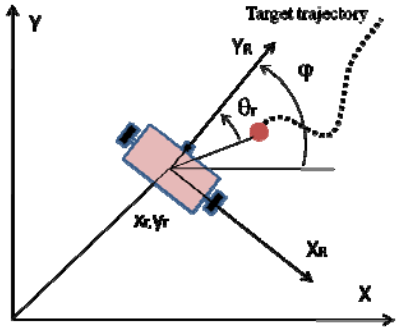


Figure 2: System variables specifications.

The structure of a fuzzy controller is presented in *Fig. 3*. One of the two inputs of the controller is the distance between the origin of the robot coordinate system (X_RY_R) and the target point. The second input is the robot orientation (θ_r). The two output signals of the controller represent the control signals for the two DC motors. The DC motors are controlled by pulse modulation with a pair of H bridges. A FUJITSU based DICE-KIT development board with 16 bit MB90F352 microcontroller is used for the PWM signal generation. [7], [8].

3. Fuzzy controller structure

The Fuzzy controller structure with fuzzification, defuzzification, inference system and rule base module is presented in *Fig. 3*.

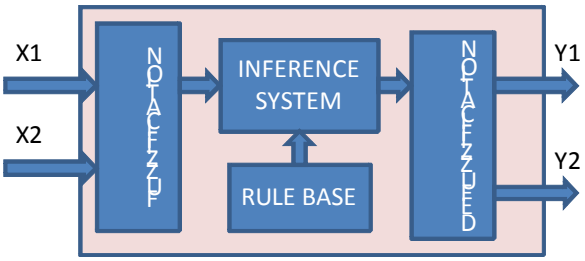


Figure 3: Fuzzy controller structure.

During fuzzification the inputs are converted into fuzzy variables (linguistic variables). The input spaces and the output universes are covered with membership functions (MF). Using the fuzzification procedure for an input within the input universe, based on fuzzy set theory, a membership value will result for each active MF. The membership values are connected to the outputs through the inference system and the rule-base table. The rule-base table defines which output MF will be selected for a pair of active input MFs. As it can be seen in *Fig. 4.*, two MFs are selected for both input spaces. The *a*, *b* and *c* vector variables represent the MF definition parameters.

The fuzzy rule base is composed by IF-THEN statements. The IF-part of the rule represents the antecedence and composes the input conditions. The THEN-part represents the consequence of the rule [4].

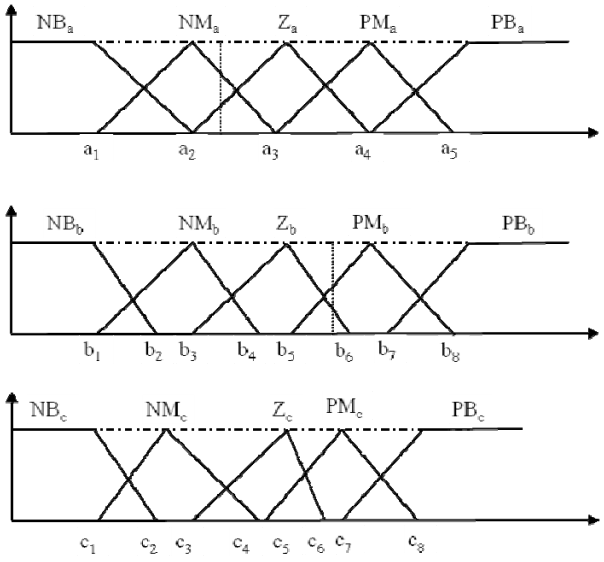


Figure 4: Input and output universes membership functions.

The fuzzy rule sets usually have as many antecedents as the number of input variables and as many consequences as the number of output variables, these are described using fuzzy IF-THEN statements:

$$\text{IF } x_1=A_i \text{ and } x_2=B_j \text{ THEN } y=C_k. \tag{4}$$

An input-output relation is described with a 3 variable fuzzy rule, where x_1 and x_2 represent the inputs and y the output.

For the antecedent part are used the:

$$A_i \quad \mu_i : X_i \to [0,1];\ i = 1 \cdots n_1,\ B_j \quad \mu_j : X_j \to [0,1];\ j = 1 \cdots n_2 \tag{5}$$

fuzzy sets, respectively for the consequence part the

$$C_k \quad \mu_k : X_k \to [0,1];\ k = 1 \cdots n_3 \text{ fuzzy set.} \tag{6}$$

The number of membership functions for the two inputs are n_1 and n_2, and for the output variable is n_3.

If for each of the two inputs two MFs can be activated, then looking to the rule base table (based on (4)), maximum four rules can be considered. For the 'and' logic operator the 'min' function is considered, and for the 'or' logic operator the 'max' function is used.

An example of the active rules of the mobile robot, is in *Fig 5*.

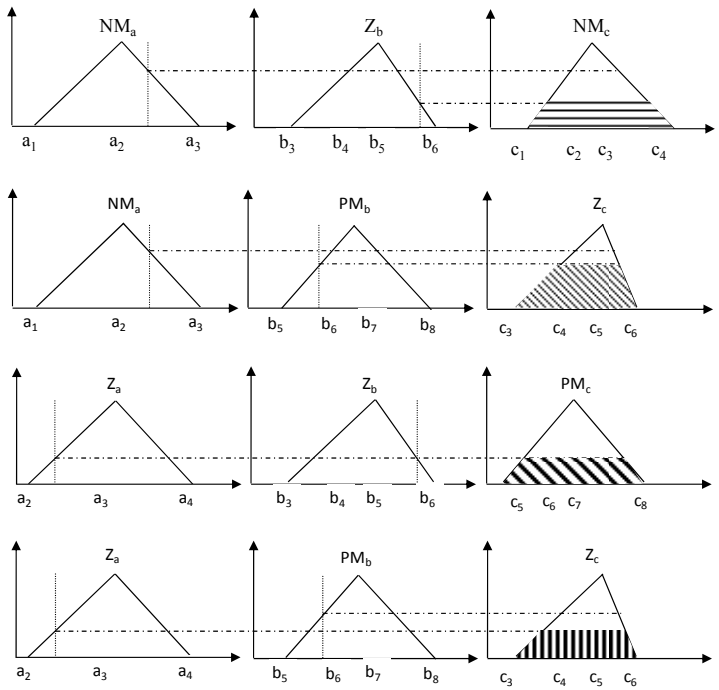


Figure 5: Active rule example.

The result for a fuzzy inference system is a fuzzy set (*Fig. 6.*). To compute the output value, it is necessary to evaluate, defuzzificate the resulted fuzzy set. For defuzzification, a large number of defuzzification algorithms can be used. In this paper the center of gravity (COG) defuzzification algorithm is used.

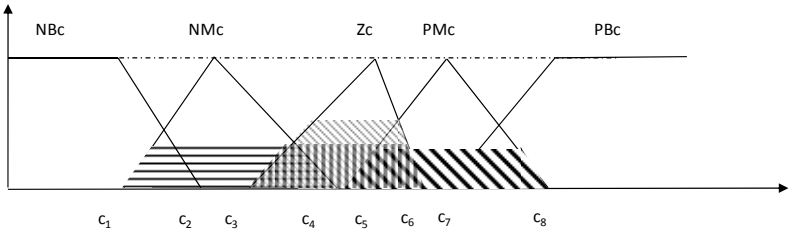


Figure 6: The resulting fuzzy sets.

4. Hardware implementation of a fuzzy controller

A. Membership functions implementation

In the project, triangle shaped MFs are used. The MFs are stored in BRAM memories. For each MF the function form and other parameters, such as the displacement distance of the MF measured to the input space origin, the center of MF, the length of the MF and the BRAM starting address for function form storage are stored. A memory composed of multiple BRAMs is used for function form storage and another BRAM memory is used for parameter storage. In *Table 1* and *Fig. 7.* the MF's stored parameters are presented.

Table 1: Membership function parameters stored in BRAM

offset	length	center	storing address

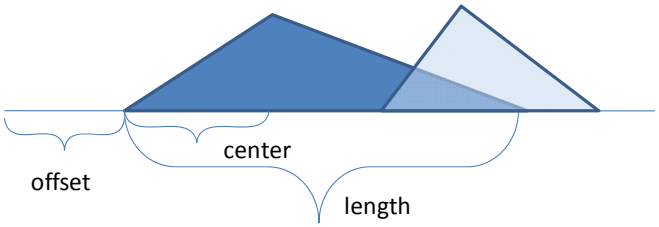


Figure 7: Membership function parameter specifications.

B. Rule-base table implementation

For a system with two inputs a number of maximum four rules are active for an input pair.

The rules are stored in a BRAM memory. The rule of BRAM addressing is composed based on a pair (A, B) of input MF indexes stored in index registers. In the rule table, the indexes of the MFs from the output universe are stored.

C. Simplified steps for the fuzzy controller

In *Fig. 8.*, the data path for controller output processing for an input data pair is presented. The controller output process is composed of the following steps:

a.) Membership value calculation and the active MF index computing. With the MFs offsets and length, the active MF index and member value are determined and stored in index and member value registers. The variables i, j, k, l represent the indexes and μ_i, μ_j, μ_k, μ_l represent the membership value registers.

b.) In step two the rule memory is indexed.

c.) The active output MFs index is extracted.

d.) Based on output MF indexes, these parameters are extracted from the output MF parameter storage BRAM and stored in output parameter registers with alpha cut results.

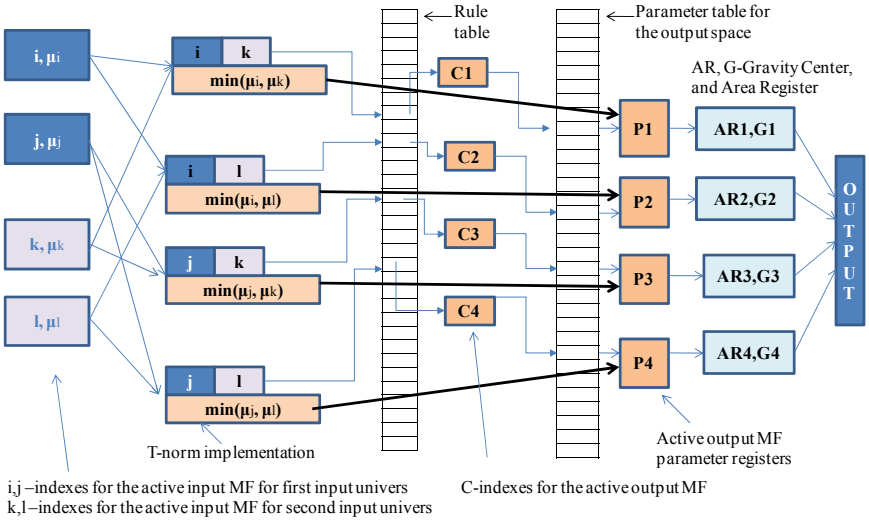


Figure 8: Simplified data path for the fuzzy controller.

All active output MFs are alpha cut and, as mentioned before, these are also stored in MF output parameter registers.

e.) In the next step the gravity centers and MF alpha cut area are processed.

f.) In the last step the controller output is processed using the center of gravity defuzzification algorithm.

E. Defuzzification

The center of gravity algorithm was used for defuzzification. For triangle type membership functions, two equations for the defuzzification algorithm are presented, one for MF gravity center estimation and another one for alpha cut area computing presented in the following equations:

$$A_i = \mu_{C_i} L_i \left(1 - \frac{\mu_{C_i}}{H}\right) \tag{7}$$

$$g_i = \frac{a_i - b_i}{6} \mu_{C_i} + \frac{a_i - b_i}{2}, \tag{8}$$

where A_i is the area, g_i is the center of gravity of i-th conclusion alpha cut membership function, μ_{C_i} -alpha cut value, H –max. value of membership function (resulted from bits used for MF value coding), $L_i = a_i - b_i$ -length of the membership function. For a general triangle based membership function with alpha cut value increase, the center of gravity (g_i) is disposed on a non-linear curve. In our algoritm, this curve is approximated with a line, whose parameters are very simply composed based on $a_i - b_i$ and μ_{C_i} (*Fig. 9.*).

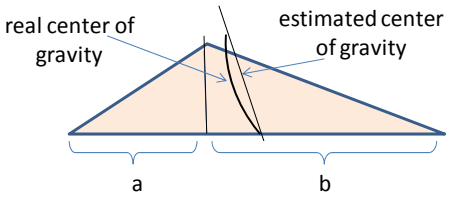


Figure 9: Estimated center of gravity for an asymmetric triangular type MF.

Using the calculated area and estimated gravity for output fuzzy sets, the controller output is processed using the following equation:

$$u = \frac{\sum_{i=1}^{N} A_i g_i}{\sum_{i=1}^{N} A_i} \tag{9}$$

The defuzzification algorithm can be simplified by using the singleton type MFs for the output space. In this case, the output MFs center and areas do not need to be computed. The MF center is known, and instead of the area, the alpha cut values are used.

F. Fuzzy controller parameters

Input spaces are coded on 12 bits. The MFs are stored in a memory with a 13-bit address space and 10 data bits, composed of 5 Block RAMs, 13 address bits and two data bits.

Out of the total number of Block RAMs, 3x5 are used for membership function storage and 1 for rule table storage. For data representation, integer based arithmetic was used. Floating point arithmetic can be used by including floating point IP CORES. The system will be tested on NEXYS-2 and VIRTEX 2PRO development boards. The FPGA implemented fuzzy controller is connected to the PC on the development board's USB port [2], [3], [9], [10]. The presented structure can also be implemented on microcontroller based systems. For data coding and MFs representation, the ideas used in neural networks for data coding and activation functions implementation can be applied [1].

5. Simulation-based measurement results

Simulations were made using the following membership functions for two input universes, represented by the robot's distance to the next target point and the robot's orientation (*Fig. 10.*) respectively the output space for the DC motors' control signals (*Fig. 11.*).

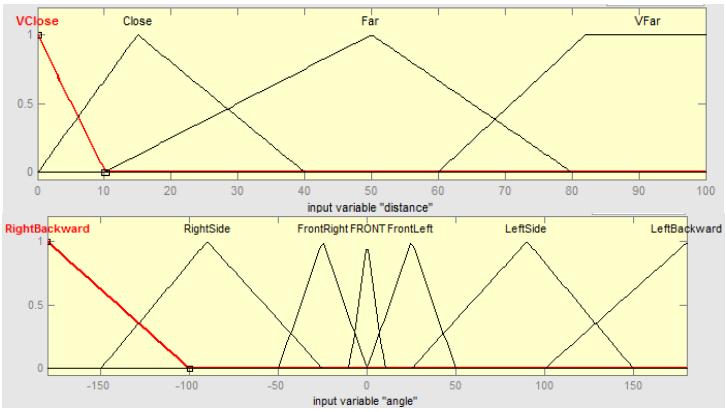


Figure 10: Membership functions for the input universe used in simulation.

The two output spaces are identical, and only the membership function of a single output space is presented (*Fig. 11.*). For the simulation experiment, the robot model described in equations (1), (2), (3) has been used.

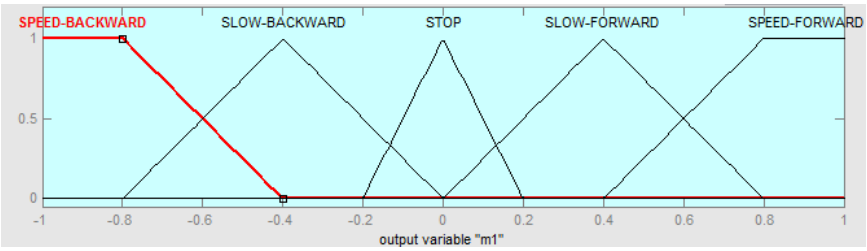


Figure 11: Membership functions for output universe used in simulation.

The membership functions for the distance input space are VClose, Close, Far and VFar, meaning that the robot is very close, close, far and very far to the target point. The second input space (robot orientation) was covered with the following MFs: RightBackward, RightSide, FrontRight, Front, Front Left, LeftSide and LeftBackward. RightBackward and LeftBackward mean that the target point is behind the robot on the right or the left part. RightSide and Left Side mean that the target point is on one side of the robot. In case of FrontRight and FrontLeft MFs, the target point is in front of the robot either on the right or on the left. The Front MFs means that the target point is directly in the front of the robot. For output MFs coding, a number of five membership functions were used (STOP, SLOW BACKWARD and SLOW FORWARD, SPEED BACKWARD and SPEED FORWARD) with a lower and a higher wheel speed control for backward and forward direction.

The robot's position for a predefined sinusoidal trajectory is presented in *Fig. 12.* for axis X and *Fig. 13.* for axis Y.

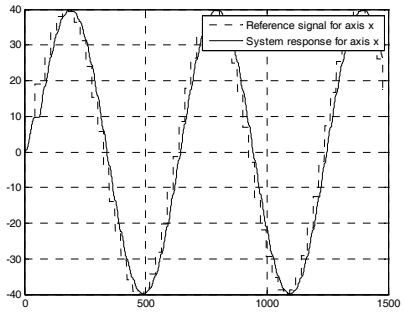


Figure 12: Reference and target trajectory for axis X.

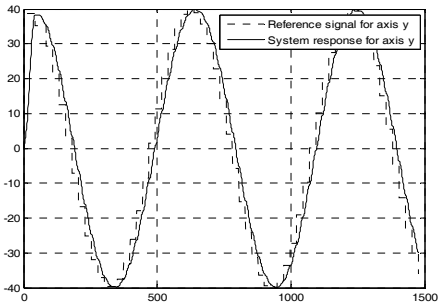


Figure 13: Reference and target trajectory for axis Y.

6. Conclusion

The hardware structure proposed for FPGA implementation presents some advantages, such as:

- Data representation on different bus sizes: in VHDL coding, the user can change/define the number of bits used for different data paths representation, such as input space coding, output space coding, membership function value coding.
- Flexible membership function programming: for all input and output spaces membership functions with different parameters can be defined.
- High speed controller output processing: the proposed mixed parallel pipeline structure permits a high order of data processing parallelism.
- Accelerated algorithm for defuzzification implementation: for triangular membership functions, the presented method reduces the defuzzification processing time.
- Easy rule base programmability: the rule base table and the membership functions can be reprogrammed from a host computer.

Acknowledgements

This paper presents a part of the achievements of a research project supported by the Institute for Research Programmes of the Sapientia University.

References

[1] Omondi, A. R., Rajapakse, J. C., FPGA Implementations of Neural Networks, *Springer*, 2006.
[2] Volnei A. Pedroni, Circuit Design with VHDL, *MIT Press, Cambridge, Massachusetts, London*, England, 2004
[3] Pong. P. Chu, FPGA Prototyping by VHDL Examples XILINX SPARTAN-3 Version, *John Whiley & Sons*, Hoboken, New Jersey, 2008
[4] Ross, J. T., Fuzzy logic with engineering applications / *Timothy J. Ross, John Wiley*, Chichester, 2004
[5] Lantos B., Fuzzy systems and genetic algorithms : lecture notes / Béla Lantos ; Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai Kar, *Műegyetemi Kiadó*, Budapest, 2002
[6] S. T. Brassai, Neuroadaptive Systems based on FPGA circuits with application in automatic control, *Phd thesis*, Transylvania University of Brasov, Brasov, 2008
[7] http://mcu.emea.fujitsu.com/mcu_product/detail/MB90F352SPMC.htm
[8] F2MC-16LX 16-BIT MICROCONTROLLER MB90350 Series, HARDWARE MANUAL, FUJITSU SEMICONDUCTOR, CM44-10132-1E
[9] Pong. P. Chu, RTL Hardware Design Using VHDL, *John Whiley & Sons*, Hoboken, New Jersey, 2006
[10] Spartan-3 FPGA Family Data Sheet, Product Specification, XILINX, June 25, 2008

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 105-112

Extending WS-Security to Implement Security Protocols for Web Services

Béla GENGE, Piroska HALLER

Department of Electrical Engineering, Faculty of Engineering,
"Petru Maior" University of Tîrgu Mureş, Tîrgu Mureş, Romania,
e-mail: {bgenge, phaller}@engineering.upm.ro

Manuscript received March 15, 2009; revised May 24, 2009.

Abstract: Web services use tokens provided by the WS-Security standard to implement security protocols. We propose several extensions to the WS-Security standard, including name types, key and random number extensions. The extensions are used to implement existing protocols such as ISO9798, Kerberos or BAN-Lowe. The advantages of using these implementations rather than the existing, binary ones, are inherited from the advantages of using Web service technologies, such as extensibility and end-to-end security across multiple environments that do not support a connection-based communication.

Keywords: Security protocols, web services, WS-Security.

1. Introduction

Security protocols are "communication protocols dedicated to achieving security goals" (C.J.F. Cremers and S. Mauw, 2005) [1] such as confidentiality, integrity or availability. Existing technologies such as the Security Assertions Markup Language [2] or WS-Security [3] provide a unifying solution for the authentication and authorization issues through the use of predefined protocols. By implementing these protocols, Web services authenticate users and provide authorized access to resources. However, in order to integrate new protocols, such as key-exchange or confidentiality protocols, we need to extend the WS-Security standard with new components.

In this paper we propose several extensions to the WS-Security standard including name types, key and random number extensions. The extensions were

used to implement existing protocols such as ISO9798 [4], that makes use of the Diffie-Hellman [5] key exchange protocol with digital signatures, or the Kerberos V5 [6] symmetric key-based security protocol. The advantages of using these implementations rather than the existing, binary ones, are inherited from the advantages of using Web service technologies. From these we mention extensibility and end-to-end security across multiple environments that do not support a connection-based communication. In addition, by adding new tokens to the existing ones, message components can be further categorized and specialized, providing an increased security of these protocols because of the additional information that accompanies each component [9], [10].

The implementations were made according to the specifications of the SOAP [7] standard, which embodies the WS-Security components in its header. The execution timings revealed the possible use of these protocols in a wide variety of systems, ranging from e-commerce to multimedia streaming.

The paper is structured in four parts. After the introduction, section 2 illustrates the proposed extensions through the form of XML schemas. In section 3 we present our experimental results, clearly showing that the proposed extensions can be used to implement applications that require authentication, key exchange or confidentiality protocols. We end our paper with a conclusion and future work in section 4.

2. WS-Security extensions

WS-Security provides a set of tokens for implementing security properties such as authentication, integrity and non-repudiation [9], [11]. These properties are used by Web services to construct security protocols providing inter-domain authentication. These are predefined, static protocols that must be implemented by all communicating parties. In order to implement other authentication protocols or other types of security protocols, the tokens provided by WS-Security must be extended with several new ones.

We consulted a large number of security protocols from the SPORE [12] library and the library of protocols presented by John Clark [8]. Based on these, we identified four *basic sets* containing terms used by protocol participants to construct messages: P, N, denoting the set of nonces (i.e. “number once used”); K, denoting the set of cryptographic keys and M denoting user-defined components.

The set of participant names P is further specialized with the following disjoint sets: $\mathsf{P}_{DN} \subseteq \mathsf{P}$, denoting the set of distinguished names; $\mathsf{P}_{UD} \subseteq \mathsf{P}$, denoting the set of user-domain names; $\mathsf{P}_{IP} \subseteq \mathsf{P}$, denoting the set of user-IP

names; $P_D \subseteq P$, denoting the set of domain names; $P_U \subseteq P$, denoting the set of remaining user name types.

The set of nonces is also further specialized with two subsets: $N_R \subseteq N$, the set of random numbers and $N_T \subseteq N$, denoting the set of timestamps.

Based on the above-defined sets and subsets, in the remaining of this section we provide the XML representation of the terms corresponding to the implementation of each element. The WS-Security standard provides a single XML element for defining user names, through the form of *wsse:UsernameToken*. For example, in order to define a user name, the following syntax is required:

```
<wsse:UsernameToken>Denumire utilizator</wsse:UsernameToken>
```

Distinguished names are usually found in user certificates and they provide information related to the organization, country, domain and several other categories characterizing a user. In order to define user names in this format, we define the following XML schema:

```
<complexType name="DistinguishedNameToken">
    <sequence>
        <element name="Organization" type="string"/>
        <element name="OrganizationalUnit" type="string"/>
        <element name="CommonName" type="string"/>
        <element name="Country" type="string"/>
    </sequence>
</complexType>
```

User-domain names have the following structure: user@host.domain or user.host.domain. The schema for such a user name must include the user name and one or more host or domain names separated by dots. The resulting schema is the following:

```
<complexType name="UserDomainNameToken">
  <sequence>
      <element name="UserName" type="string"></element>
      <element name="DomainName">
          <simpleType>
              <restriction base="string">
                  <pattern value="(\w+\.\w+)+"></pattern>
              </restriction>
          </simpleType>
      </element>
  </sequence>
</complexType>
```

IP addresses and identifying machine names must include support for both IPv4 and IPv6 address formats. The resulted XML schema makes use of regular expressions to describe the rules for constructing such names:

```
<complexType name="UserIPNameToken">
  <choice>
    <element name="IPV4">
     <simpleType>
      <restriction base="string">
            <pattern value="\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}"/>
      </restriction>
     </simpleType>
    </element>
    <element name="IPv6">
     <simpleType>
      <restriction base="string">
          <pattern value="([0-9a-fA-F]{1,4}:){7}[0-9a-fA-F]{1,4}"/>
      </restriction>
     </simpleType>
    </element>
  </choice>
</complexType>
```

For names containing exclusive domain names, we use the following schema:

```
<simpleType name="DomainNameToken">
      <restriction base="string">
            <pattern value="(\w+\.\w+)+"></pattern>
      </restriction>
</simpleType>
```

Random numbers are transmitted as binary tokens, for which a security token is already provided by the WS-Standard. Transmitting timestamps is also possible by using existing tokens provided by WS-Security. However, in order to send and receive encrypted binary keys we use an XML schema that defines the key value and the encoding type used:

```
<complexType name="KeyToken">
    <sequence>
          <element name="KeyValue" type="string"/>
    </sequence>
    <attribute name="type">
          <simpleType>
           <restriction base="string">
                <enumeration value="base64Binary"/>
                <enumeration value="hexBinary"/>
```

```
            </restriction>
          </simpleType>
    </attribute>
</complexType>
```

3. Experimental results

The proposed extensions were used to implement protocols with security properties ranging from authentication to key exchange and message confidentiality. The protocols were constructed from participants exchanging *terms*. Terms were constructed from the elements belonging to the basic sets provided in the previous section:

$$\mathsf{T} ::= . \mid \mathsf{P} \mid \mathsf{N} \mid \mathsf{K} \mid \mathsf{M} \mid (\mathsf{T}, \mathsf{T}) \mid \{\mathsf{T}\}_{FuncName(\mathsf{T})}, \tag{1}$$

where *FuncName* defines the set of function names used to encrypt terms:

$$\begin{aligned} NumeFunc ::= {} & sk && (symmetricencryption) \\ & \mid pk && (asymmetric\ encryption) \\ & \mid h && (hash\ encryption) \\ & \mid hmac && (keyed\ hash\ encryption) \end{aligned} \tag{2}$$

By using the above definitions, protocol messages can be constructed as in the following examples:

- $\{A, B, N_a, K\}_{sk(K_{ab})}$, where $A, B \in \mathsf{P}$, $N_a \in \mathsf{N}$, $K \in \mathsf{K}$;
- $\{\{A, N_a\}_h\}_{pk(PK_a)}, A, N_a$, where $A \in \mathsf{P}$ şi $N_a \in \mathsf{N}$, $PK_a \in \mathsf{K}$.

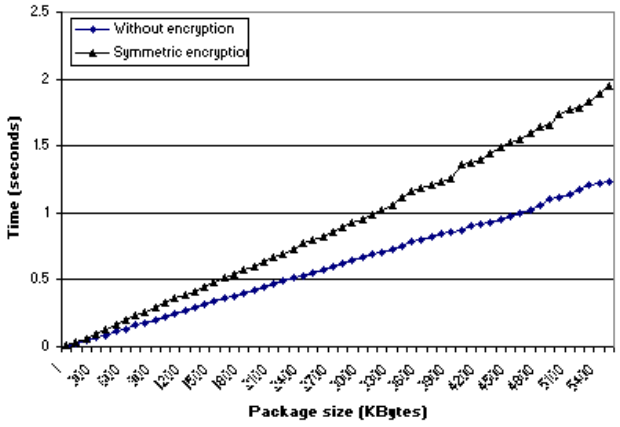


Figure 1: Symmetric encryption versus no encryption.

The implementation of these messages replaces each component with its corresponding security token provided by the proposed extensions. The performance of the implementations is strongly dependent on the type of encryption function used in the process. For example, there is an obvious difference between an implementation that uses symmetric encryption and one that does not use encryption at all. This is the case illustrated in *Fig. 1*, where the encrypted message is $\{M\}_{sk(K_{ab})}$, with $M \in \mathsf{M}$. The figure illustrates the time required to construct, encrypt and send a message using the proposed tokens and the already existing ones.

In our experiments, messages were encoded in the SOAP [7] header, according to the WS-Security standard. Because of their size, as seen in *Fig. 1*, the XML structures influence the performance of the implemented protocols. This is also influenced by the type of encryption used, as shown in *Fig. 2*.

The illustrated values correspond to the execution time for constructed messages using symmetric and asymmetric cryptography. The symmetric encryption-based protocol is clearly much more performant than the asymmetric encryption-based protocol. This is why, the first protocol is usually used for data transfer, while the second one for encrypting small sized messages, usually in key exchange and authentication protocols.

The experimental results given in *Fig. 1* and *2* show that the performance of the implemented protocols is not only influenced by the size of the encrypted messages, but also by the encryption algorithm type. We have implemented several other protocols, for which the execution timings are given in *Table 1*. We identified several participants for each protocol.

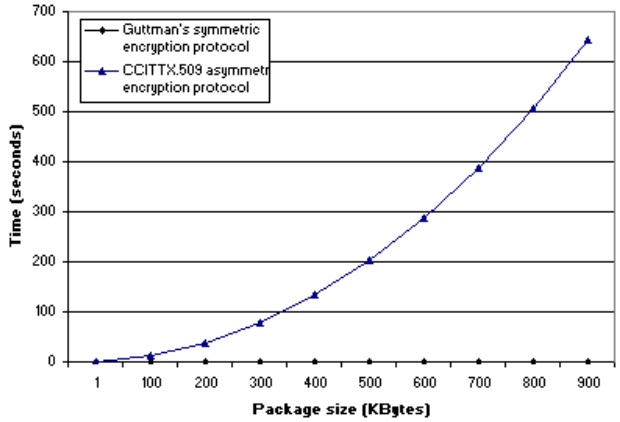


Figure 2: Symmetric versus asymmetric encryption.

We measured the construction and the processing time of messages for each participant; the measured values were added together, resulting the total time.

We can see a clear difference between protocols that use symmetric algorithms (e.g. Lowe-BAN, Kerberos, Andrew RPC) and protocols that use asymmetric algorithms (e.g. ISO9798, CCITT X.509). For some protocols, the processing or construction timings are 0 because the sub-protocols we identified do not require operations. Based on these measurements, we can clearly state that using the proposed WS-Security extensions, we can implement key exchange, authentication and user-defined data exchange protocols. Implementing such protocols with existing WS-Security tokens is possible only for authentication protocols, for which the WS-Trust standard (using WS-Security) defines several predefined protocols.

Table 1: Execution time of security protocols

Participant role	Message construction (ms)	Message processing (ms)	Total participant (ms)	Total (ms)
Lowe-BAN Initiator	11.81	3.68	15.49	19.97
Lowe-BAN Respondent	2.86	1.62	4.48	
ISO9798 Initiator	35.78	23.30	59.08	78.19
ISO9798 Respondent	6.87	12.24	19.11	
Kerberos 1 Initiator	0.83	0.00	0.83	27.69
Kerberos 2 Initiator	0.55	1.58	2.13	
Kerberos 3 Initiator	3.34	0.94	4.28	
Kerberos 1 Respondent	0.00	0.41	0.41	
Kerberos 2 Respondent	3.37	1.67	5.04	
Kerberos 3 Respondent	11.41	3.59	15	
CCITT X.509 Initiator	7.85	0.00	7.85	82.27
CCITT X.509 Respondent	0.00	74.42	74.42	
Andrew RPC Initiator	12.56	5.04	17.6	36.54
Andrew RPC Respondent	14.04	4.9	18.94	

4. Conclusions and future work

Existing tokens from the WS-Security standard provide the possibility for implementing a reduced set of security protocols. In order to enable the implementation of a wide range protocols, we proposed several token extensions for user name types and cryptographic keys.

The protocol implementations maintain their security properties by respecting the requirements given in the WS-Security standard. These requirements indicate the use of the SOAP header for transporting security

tokens and the use of the SOAP body for other message components. The implementations we have developed show that protocol performance is influenced by the XML constructions and by cryptographic functions used in the process. Based on our experimental results, we can clearly state that the proposed extensions offer security for protocols used in various applications, such as multimedia or eCommerce.

In the future we intend to use the proposed extensions to implement several security protocols for multimedia applications and to prove that our implementations can be used to transfer audio and video messages without loss of quality.

References

[1] Cremers, C. J. F., Mauw, S., "Checking secrecy by means of partial order reduction", *Revised selected papers LNCS*, Vol. 3466, pp. 171-188, 2005.

[2] Organization for the Advancement of Structured Information Standards, "SAML V2.0 OASIS Standard Specification", http://saml.xml.org/, 2007.

[3] Organization for the Advancement of Structured Information Standards, "OASIS Web Services Security (WSS)", http://saml.xml.org/, 2006.

[4] International Organization for Standardization, "Information technology – Security techniques – Entity authentication – Part 2: Mechanisms using symmetric encipherment algorithms", *ISO/IEC 9798-2*, Geneva, Switzerland, 1994.

[5] Diffie, W., and Hellman, M. E., "New directions in cryptography", *IEEE Transactions on Information Theory*, IT-22(6), pp. 644–654, 1976.

[6] Neuman, C., Yu, T., Hartman, S., Raeburn, K., "The Kerberos Network Authentication Service (V5)", http://www.ietf.org/rfc/rfc4120, 2005.

[7] World Wide Web Consortium, "Simple Object Access Protocol (SOAP) 1.2", http://www.w3.org/TR/soap/, April 2007.

[8] Clark, J., Jacob, J., "A Survey of Authentication Protocol Literature: Version 1.0", *York University*, 1997.

[9] Gavin Lowe, "Some new attacks upon security protocols", in *Proceedings of the 9th CSFW, IEEE Computer Society Press*, 1996, pp. 162-169.

[10] Cremers, C. J. F., "Compositionality of Security Protocols: A Research Agenda", *Electr. Notes Theor. Comput. Sci.*, 142, pp. 99-110, 2006.

[11] Cremers, C. J. F., Mauw, S., E. P. de Vink, "Injective Synchronization: an extension of the authentication hierarchy", *TCS 6186, Special issue on ARSPA'05, Editors: P. Degano and L. Vigano*, 2006.

[12] SPORE, Security Protocol Open Repository, http://www.lsv.ens-cachan.fr/spore.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 113-123

Certificate-Based Single Sign-on Mechanism for Multi-Platform Distributed Systems

Attila MAGYARI[1], Béla GENGE[2], Piroska HALLER[2]

"Petru Maior" University of Tîrgu Mureş, Tîrgu Mureş, Romania,
e-mail:[1] atti86@gmail.com, [2] {bgenge,phaller}@engineering.upm.ro

Manuscript received March 15, 2009; revised May 24, 2009.

Abstract: We propose a certificate-based single sign-on mechanism in distributed systems. The proposed security protocols and authentication mechanisms are integrated in a middleware. The novelty of our middleware lies on the use of XPCOM components. This way we provide different services that can be used on every platform where Mozilla is available. The component based architecture of the implemented services allows using the authentication components separately.

Keywords: Single sign-on, authentication, security protocols, cryptography, multi-platform.

1. Introduction

In this paper we propose a single sign-on mechanism based on certificates generated on request for client applications. Single sign-on mechanisms ensure the use of user credentials for accessing multiple resources where the user is requested to enter its credentials only once. This ensures a reduction of the number of passwords used which can significantly improve security of systems by minimizing the likelihood of a password being compromised [1]. Communication between client applications and servers is done using secure channels based on security protocols. In order to minimize the overhead needed for accessing multiple servers, instead of using protocols such as SSL [2] or its more recent version TLS [3], we designed a set of new protocols based on Guttman's authentication tests [4, 5]. The protocols have been implemented using the existing security library OpenSSL [6], which, together with the protocol descriptions, ensures the correct implementation of the designed protocols.

In order to provide a minimal effort for developing single sign-on mechanisms in distributed systems, we developed a middleware that

implements the proposed security protocols and single sign-on mechanism. Existing single sign-on mechanisms are either implemented to function on a single platform, such as Active Directory [7] for Microsoft Windows or eDirectory [8] for Unix systems, or they rely on a centralized directory structure such as LDAP [9], to which servers must be connected in order to authenticate users. The novelty of our middleware lies on the use of XPCOM [10] components provided by the Mozilla platform to encapsulate the communication layer. This way, we do not only provide a single sign-on mechanism for a single platform, but also a mechanism that can be used on every platform where Mozilla is available.

The rest of the paper is structured as follows: in the next section we describe the architecture of the middleware: the requirements, the software stack and the security protocols.

2. Middleware architecture

2.1 Requirements

Network users typically maintain a set of authentication credentials (usually a username/password pair) with every Service Provider (SP) they are registered with. In the context of this paper, a service provider is any entity that provides some kind of service or content to a user. Examples of SPs include web services, messenger services, FTP/web sites, and streaming media providers. The number of such SPs with which users usually interact has grown beyond the point at which most users can memorize the required credentials. The most common solution for users is to use the same password with every SP with which they register — a tradeoff between security and usability in favor of the latter. A solution for this security issue is Single Sign-On (SSO), a technique whereby users authenticate themselves once only and are automatically logged into SPs as necessary, without requiring further manual interaction [11].

There are several approaches to create a SSO network. The Kerberos based [12] systems initially prompt the user for credentials, emitting a Kerberos ticket-granting ticket (TGT). Drawbacks of the Kerberos based system include the centralized architecture: when the Kerberos server is down, no one can log in. Kerberos requires the clocks of the involved hosts to be synchronized, the tickets have a time availability period, which is 10 minutes by default configuration, and if the host clock is not synchronized with the Kerberos server clock, the authentication will fail. Furthermore, the secret keys for all users are stored on the central server, so a compromise of that server will compromise all users' secret keys. Another approach would be a smart card based

authentication: an integrated circuit, which can process data, is embedded in a plastic card, which will be used to identify its owner. The necessity of this hardware, which can be easily damaged, stolen or compromised, excluded the smart card method from our list. Some other possibilities include the use of one-time passwords (OTP) or the integrated windows authentication, but we have chosen a client certificate based configuration for our model. First of all, the X.509 certificates we have been using are ITU-T standardized, which widens the possibilities of the implementations or further developing. These certificates are based on the RSA encryption algorithm, providing the necessary security. The certificates are relatively easily generated and due to their small size, their storage and transport over the network is also easy. The X.509 certificates store several predefined information about their owner, but can also contain custom data. We use these fields to store each client's permissions in the network. An immediate disadvantage of such an approach is the support for a single encryption algorithm at a time. It was shown that the algorithm can be broken if there are enough resources used, but the use larger keys (1024 or 2048 bit) makes this very hard, if not impossible, with existing technologies. Another drawback of RSA encryption is its processing power and execution time, compared to other algorithms, such as AES, 3DES, Blowfish or RC6. This is why we try to minimize its usage, and – when possible –, replace it with a more resource-friendly encryption algorithm.

Single sign-on mechanisms already exist, and they are widely used, for instance the above-mentioned Active Directory for Microsoft Windows or eDirectory for UNIX systems. However, they are platform-specific. Our goal was to create a mechanism that runs on a wide variety of platforms, hence we have chosen XPCOM. It stands for Cross Platform Component Object Model, and it is a framework for writing multi-platform, modular software. The core of the components is written using the NSPR (Netscape Portable Runtime [13]) libraries, as shown in *Fig. 1* [14]. As an application, it uses a set of core XPCOM libraries to selectively load and manipulate XPCOM components. It is open source, and it supports just about any platform that hosts a C++ compiler, including Microsoft Windows, Linux, HP-UX, AIX, Solaris, OpenVMS, MacOS, and BSD.

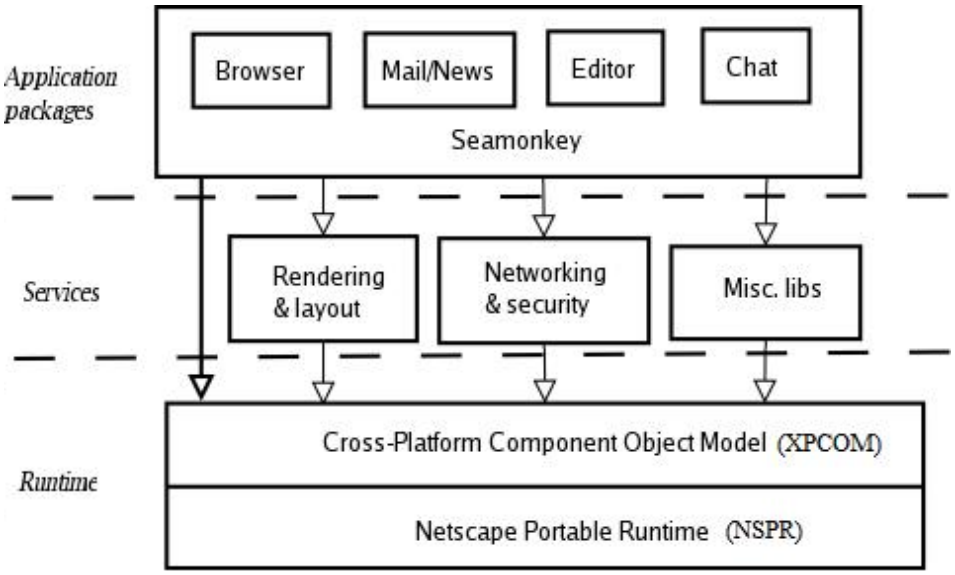


Figure 1: Top Level Conceptual Architecture of Mozilla Application Suite.

2.2 Software Stack

The middleware structure has four layers, as shown in *Fig. 2.*

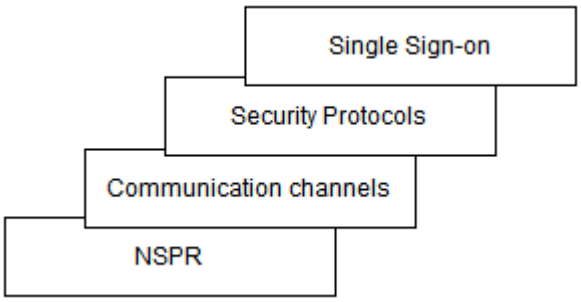


Figure 2: Middleware structure.

2.2.1 NSPR

The NSPR layer of the middleware is implemented using various classes and objects, such as threads, sockets, coders, parsers, timers, several data structures, and other implementations, which altogether constitute the

foundation of the whole platform. These components were written using the NSPR libraries. Netscape Portable Runtime (NSPR) provides platform independence for non-GUI operating system facilities. These facilities include threads, thread synchronization, normal file and network I/O, interval timing and calendar time, basic memory management and shared library linking. The current implementation supports Macintosh (PPC), WIN-32 (WinNT, Win9x) and 20 versions of UNIX and is still expanding.

2.2.2 Communication Channels

The communication channels are built on top of the NSPR layer to create more advanced data transportation mechanisms. The channels are created dynamically and managed by channel handlers. They support customary, predefined structured messages, but also raw data.

2.2.3 Single Sign-on

Single sign-on (SSO) is a mechanism whereby a single action of user authentication and authorization allows access to all computers and systems where authorization rights have been verified, without the need to enter multiple passwords. Single sign-on reduces human error, a major component of systems failure and is therefore highly desirable.

Our proposed system is composed of two types of participants: clients and servers. *Fig. 3* illustrates a simple network with 3 servers and two clients: one already connected and another who is in the authentication process. The communication lines between the nodes may be unstable and in most cases unsafe, which exposes our messages to different threats like spoofing, replicating or simple message loss. We designed the system to prevent any of these attacks, and to be easy to implement and use. Each server can host many and different services, but for our model we only need an authentication service and a resource service. The services are of request-response type, and all the data sent is confidential. The authentication service provides two types of authentication mechanisms: the first one requires the use of a username and password, while the second one requires the use of the generated certificates. In order to gain access to a Service Provider (SP), a client first has to register at one server called the *home server*. Each server can be a home server and *resource server* at the same time; it is only relative to the client. The registration can take any form; in our model we assume that there is a secure database, where every client is already registered. The requester contacts its home server, and sends over credentials (Step 1 in *Fig. 3*); this is the only time the user has to manually log in. The home server will generate a certificate, containing user

data (e.g. username, location, organization name, e-mail address, etc.), expiration date, but also information about the issuer, to verify its genuineness. The certificate also contains information about the user's permissions, following a role-based access control (RBAC) model. Since users are not assigned permissions directly, but only acquire them through their role (or roles), management of individual user rights becomes a matter of simply assigning appropriate roles to the user. This simplifies common operations, such as adding a user, or changing a user's department. In Step 2 (*Fig. 3)*, the client receives the certificate. The next two steps, 3 and 4 in *Fig. 3*, are to contact the desired SP, sending the certificate, and exchanging a session key, which will be used to encrypt data from that moment on. RSA encryption algorithms, which we have used so far, require more processing power, so we will use the triple DES algorithm, with a new key each session to maximize security and performance. If the client wants to access a different SP, the certificate has to be sent only once, and a new session key will be generated. As long as the certificate is not expired, it can connect to every SP in the network, otherwise it will have to repeat the first step and obtain a new certificate.

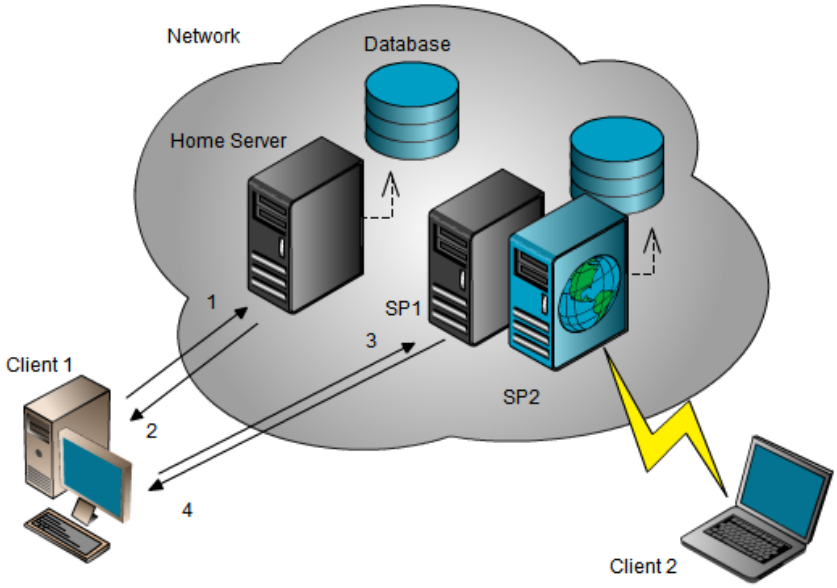


Figure 3: System setup.

2.3 Security Protocols

In the proposed middleware, there was a need for authentication protocols that satisfied security requirements, such as confidentiality in an insecure environment, supporting message loss, certificate and key generation. We developed several security protocols, based on Guttman's authentication tests. The implementation of these protocols was done using the OpenSSL security libraries. A combination of symmetric and asymmetric cryptographic algorithms was used to achieve a balance between security and performance. The authentication consists of two phases: acquiring the certificate from the home server, and authentication at the resource server with the newly generated credentials.

In order to achieve a valid certificate and key, the client (A) needs to contact its home server (B). This is where the first phase of the authentication protocol takes place (*Fig. 4*), initiated by the client who sends the username, requesting a connection. If the server finds the username in its database, and the system is capable of accepting a new connection, it generates a 1024 bit length nonce (N, random). A hash function (h) is applied on this nonce, and is sent to the client, together with a message informing the other participant that the next step is allowed. Then the client sends the username and password, and a single secret key is generated (K_{AB}), which is used to encrypt the next message from the server. The received hash of the nonce is hashed again, and together with the username, password and the generated symmetric key, they are encrypted using the server's public key (pk_B). Upon receiving the data from the client, the server hashes the nonce once again and compares it to the previously saved data. If they match, meaning the message is fresh, it verifies the username and password and a new certificate will be generated, along with the RSA inverse keys. The secret key (sk_A) will be encrypted with the key received from the client. The keys, the certificate and the nonce are digitally signed, and sent back to the client. This will verify the nonce and the signature, and if everything is valid, the certificate and the secret key are decoded and decrypted, finalizing the first phase of the authentication.

The second phase of the authentication (*Fig. 4*) starts after acquiring a certificate. The client contacts the desired resource server, communicating his intentions on getting access to the resources. If the server is willing to accept new connections, it will generate and send a 1024 bit nonce (N), informing the client about the connection being accepted. Receiving this message, the client hashes and signs the received nonce with his own private key (sk_A), and attaches the certificate to the message. The server can verify the signed nonce with the received certificate, but this certificate will also be verified to ensure it

was emitted by a trusted authority, in this case, the client's home server. If no problems occur, the server proceeds to generate a session key (K_{AB}), which will be used for further data encryption. This key and the nonce will be encrypted with the client's public key (pk_A), and also signed by the server, to protect its contents. The whole message is encrypted again with the server's public key, to prevent any modifications on the data.

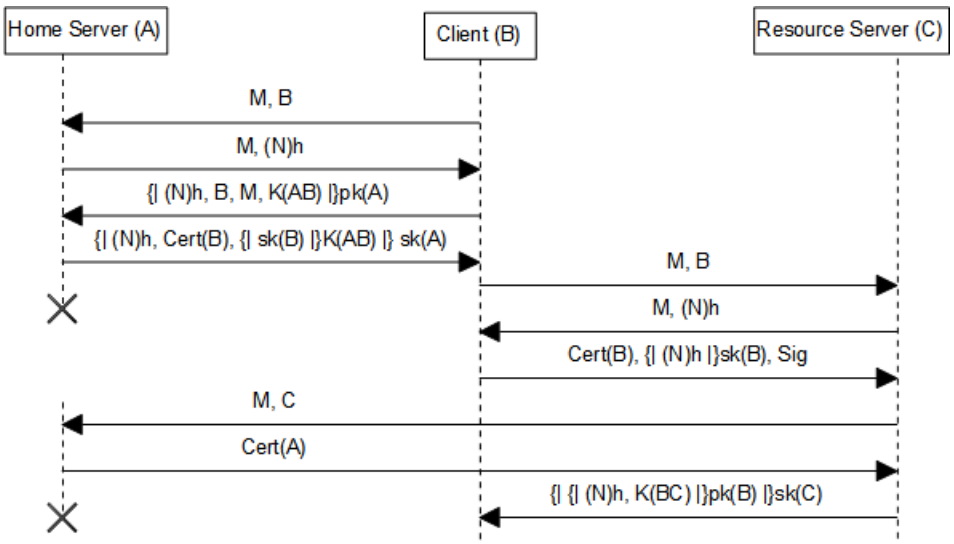


Figure 4: Authentication protocol.

3. Experimental Results

The tests were performed on a Microsoft Windows machine, 2800 MHz dual core CPU. As you can see in *Fig. 5*, the RSA key generations use the most resources. When the number of clients is lower than 10, the delay could vary between 150 to 1200 milliseconds, but if more than 10 clients try to request certificates simultaneously, the waiting time can go over 1-2 seconds, as you can see in *Fig. 7.* This wouldn't be a problem, but in a populated network, we cannot limit the number of clients to 10, there could be hundreds or even thousands of requests at the same time, and could create a bottleneck in the servers. To improve performance, and to avoid complications, we could add more servers, distributing the load across the system. The key generating time is directly proportional with the processing power of the CPU, so upgrading our

hardware can speed up the acquiring process. There are several other ways to improve the overall performance of the system:

- Using a dedicated processor for RSA key generation, optimized only for this algorithm;
- Developing either a new library algorithm, or improving the current one;
- Introducing a new type of server in our system, this could analyze each server's load and balance the system by sending clients to less busy servers.

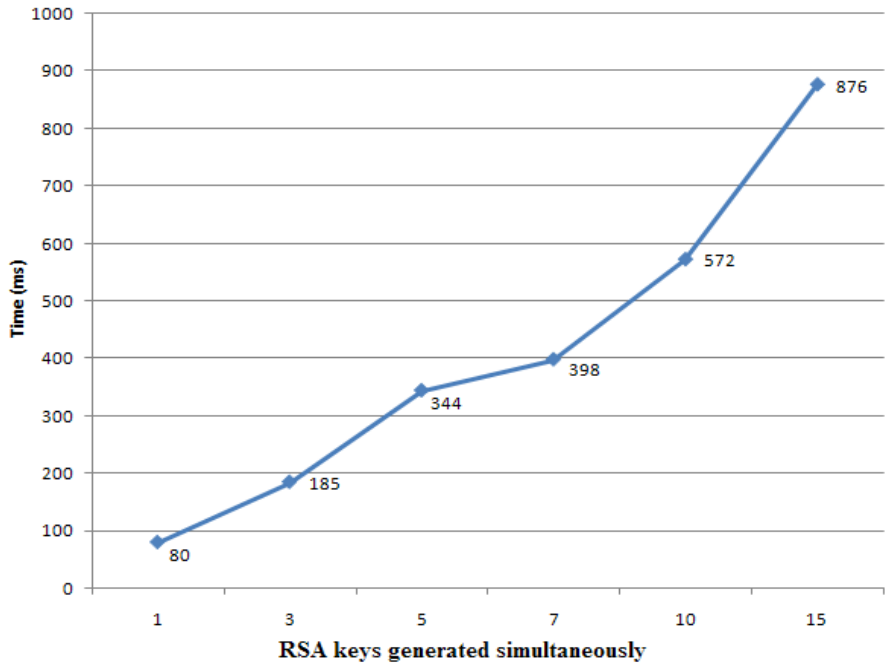


Figure 5: RSA key generation in time.

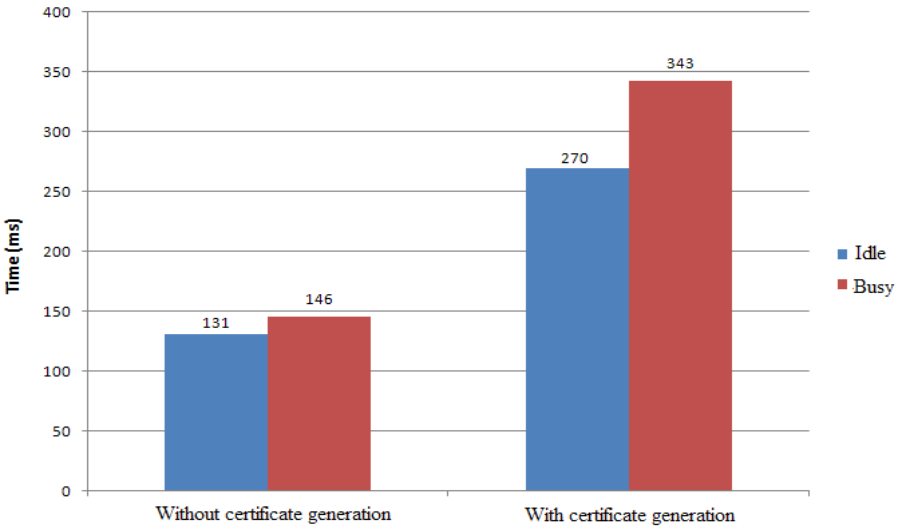


Figure 6: Authentication time, with busy and idle servers.

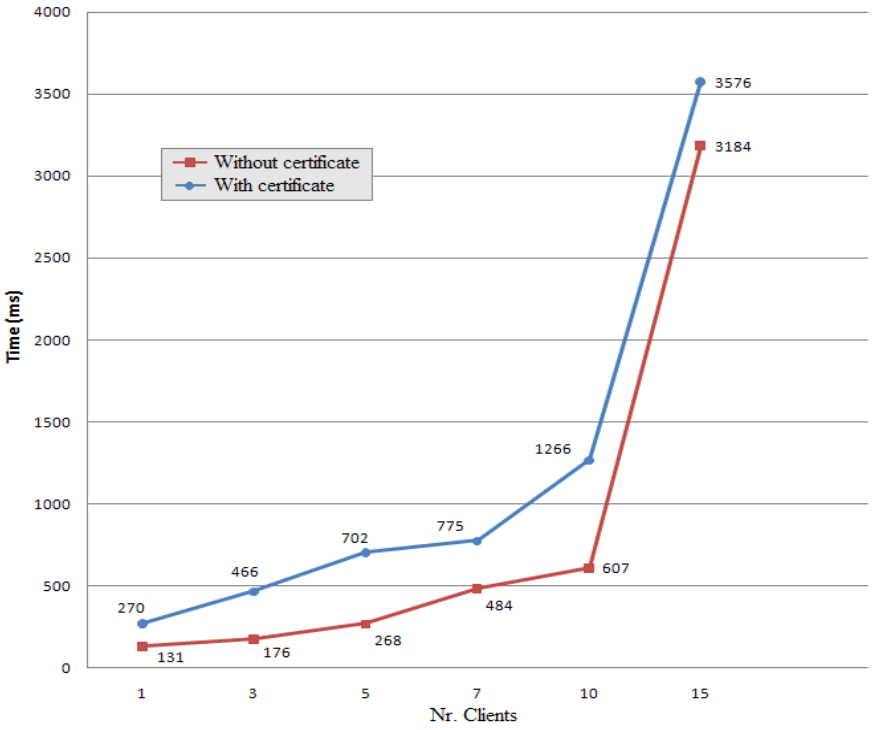


Figure 7: Certificate acquiring time by client.

4. Conclusions

We have implemented a middleware platform based on XPCOM components to assure different services for platform independent distributed application. The proposed authentication protocol as part of the middleware was designed to work in an insecure environment, supporting message loss, certificate and key generation. The implemented protocols have high computational requirements, but the proposed distributed architecture of the services can guarantee this.

References

[1] Lampson, B., Abadi, M., Burrows, M., Wobber, E., "Authentication in distributed systems: Theory and practice", *ACM Trans. Computer Systems 10, 4*, pp. 265-310, Nov. 1992.

[2] Freier, A., Karlton, P., Kocher, P., "The SSL protocol, Version 3.0, draft-ietf-tls-sslversion3-00.txt, Internet-draft", *Transport Layer Security Working Group*, Nov. 1996.

[3] Dierks, T., Allen, C., "The TLS protocol, Version 1.0, Request for comments: 2246", *Network Working Group,* Jan. 1999.

[4] Guttman, J. D., Javier, F., Fabrega, T., "Authentication tests and the structure of bundles", *Theoretical Computer Science*, Vol. 283, No. 2, pp. 333-380, June 2002.

[5] Guttman, J. D., "Security protocol design via authentication tests", in *Proc. of the 15th IEEE Computer Security Foundations Workshop, IEEE CS Press*, June 2002, pp. 92..

[6] "OpenSSL Project, version 0.9.8h", available at http://www.openssl.org/, 2008.

[7] Hunter, L., "Active directory user guide", *Springer-Verlag*, 2005.

[8] Killpack, R., "eDirectory field guide", *Springer-Verlag*, 2006.

[9] "OpenLDAP, version 2.4.15", http://www.openldap.org/, 2008.

[10] "Mozilla Corporation, XPCOM, Cross platform component model", http://www.mozilla.org/projects/xpcom/, 2008.

[11] Pashalidis, A., Mitchell, C. J., "A taxonomy of single sign-on systems", Vol. 2727/2003, *Springer Berlin / Heidelberg*, 2003.

[12] The Kerberos Network Authentication Service, http://www.kerberos.info/

[13] Mozilla Corporation, NSPR, Netscape Portable Runtime, http://www.mozilla.org/projects/nspr/, 2008.

[14] D'souza, A., Hildebrand, K., Israeli, G., "Conceptual architecture of Mozilla", Sept. 30, 2004.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 125-132

Rate Control in Open-Loop MPEG Video Transcoder

Bence FORMANEK[1], Tihamér ÁDÁM [2]

[1] R&D Department, CableWorld Ltd., Budapest, Hungary,
e-mail: formanek.bence@cableworld.hu
[2] Department of Automation, University of Miskolc, Miskolc, Hungary,
e-mail: adam@mazsola.iit.uni-miskolc.hu

Manuscript received March 15, 2009; revised June 06, 2009.

Abstract: Video transcoding is intended to provide transmission flexibility to pre-encoded bit streams by dynamically adjusting the bit rate of these bit streams according to new bandwidth constraints that were unknown at the time of encoding. What makes transcoding different from video encoding is that transcoding has access to many coding parameters, which can be obtained from the input compressed stream. They may be used not only to simplify the computation, but also to improve the video quality. In this paper, we propose a low complexity rate-control method for open loop MPEG-2 video transcoders, working entirely in the frequency domain.

Keywords: Video, MPEG, MPEG-2, bitrate, transcoder, DVB, codec.

1. Introduction

The process of converting between different compression formats and/or further reducing the bit rate of a previously compressed signal is known as transcoding, and it can introduce significant impairments in video quality if performed without due care. There are three basic requirements in transcoding. The information in the original bitstream should be exploited as much as possible. The resulting video quality of the new bitstream should be as high as possible, or as close as possible to the bitstream created by coding the original source video at the reduced rate [1], [2]. In real-time applications, the transcoding delay and memory requirement should be minimized to meet real-time constraints.

2. Video transcoding

The most straightforward transcoding architecture is to cascade a decoder and an encoder directly. In this architecture, the incoming source video stream is fully decoded, and then the decoded video is re-encoded into the target video stream with desirable bitrate or format. It is computationally very expensive, but often used way of video transcoding.

A more efficient solution to perform conversion between video bitstreams of the same standard - homogeneous transcoding - is open-loop transcoding. In an open-loop transcoder the process of video coding is reversed until the quantization step, a new quantizer value is calculated for lower bitrate, then the DCT coefficients requantized with this new quantizer value, and the rest of the video coding process is executed again with the new DCT coefficient values. Because of the higher quantizer step, the amount of information contained in each picture will be lower, which means lower bitrate for the entire video stream. Open-loop transcoders are computationally efficient, since they operate directly on the DCT coefficients. However, they suffer from the drift problem [3].

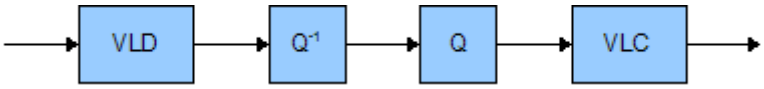


Figure 1: Open-loop transcoder.

The drift problem is explained as follows. A video picture is predicted from its reference pictures and only the prediction errors are coded. For the decoder to work properly, the reference pictures reconstructed and stored in the decoder predictor must be same as those in the encoder predictor. The open-loop transcoders change the prediction errors and, therefore, make the reference pictures in the decoder predictor different from those in the encoder predictor. The differences accumulate and cause the video quality to deteriorate with time until an intrapicture is reached. The error accumulation caused by the encoder/decoder predictor mismatch is called drift and it may cause severe degradation to the video quality [4], [5].

In MPEG-2 video compression, Intra-coded frames (I frames) are encoded without reference frame, MC is not needed in encoding I frames, so the transcoding of I frames is not subject to the drift. Bi-directionally predictive coded frames (B frames) are not used for predicting future frames [6]. Therefore, the transcoding of B frames does not contribute to the propagation and accumulation of the drift. The drift error is only caused by the transcoding operation of predictively coded frames (P frames), because they are used as further reference for prediction and can accumulate through a GOP (Group Of

Pictures). The quality deterioration gradually increases until the next I frame refreshes the video scene, so this error is also called “breathing”.

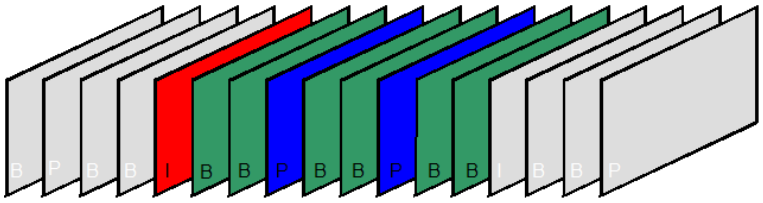


Figure 2: Picture coding types in MPEG-2.

However, the degree of the video quality degradation caused by the drift varies with architectures. In addition, the drift will be terminated by an intrapicture. In the applications where the number of coded pictures between two consecutive intrapictures is small and the quality degradation caused by the drift is acceptable, these architectures, although not drift free, can still be quite useful due to the potentially lower cost in terms of computation and required frame memory.

3. Rate Control

The goal of rate control in video coding and transcoding is to achieve a target bitrate with good and consistent visual quality. Rate control is responsible for maintaining consistent video quality while satisfying bandwidth, delay, and memory constraints by determining picture quantization parameters. The particular picture quality and rate of an MPEG-2 encoder is achieved by selecting a specific quantizer scale for each macroblock in picture. This value is calculated for each macroblock. A picture global quantization scale is the average of the macroblock quantizer scale values in that picture.

In the widely used MPEG-2 Test Model 5 (TM5) [7], [8], a picture complexity measure characterizes the difficulty in coding a picture, so that the target number of bits for coding that picture is proportional to its complexity. This complexity measure can be computed from the picture's spatial properties. Measurements have shown that this is actually a fairly accurate model for a large variety of scenes.

$$b_i = f(Q_i) = \frac{c_i}{Q_i}, \tag{1}$$

where b_i refers to the target bits per picture, Q_i is the picture global quantization scale, c_i is the picture complexity measure.

However, decoded images are not available in an open-loop transcoder, so complexity measure can't be calculated in a similar way to encoders. In transcoding, it is yet possible to compute the frame complexities from the input bit-stream, since the quantization step sizes and the number of bits information per picture are available, and then we can use these complexities for the bit allocation in transcoding. From equation (1), we could assume that:

$$\frac{b_{ti}}{b_{oi}} = f\left(\frac{Q_{oi}}{Q_{ti}}\right) = \alpha\left(\frac{Q_{oi}}{Q_{ti}}\right) + \beta\,, \tag{2}$$

where b_{ti} is the target bits for transcoded picture, b_{oi} is the original size of the current picture in bits, Q_{oi} is the picture global quantization scale of the original picture, Q_{ti} is the picture global quantization scale for transcoded picture, and α and β are parameters of the linear equation. We assume that the value above can be approximated by a linear equation.

Since a video sequence consist of successive video frames, the bitrate of the sequence depends on the size of the successive pictures, so we can say that:

$$\frac{b_{ti}}{b_{oi}} = \frac{R_a}{R_o}\,, \tag{3}$$

where R_a is the average output bitrate we would like to achieve, R_o is the input bitrate of the original video sequence.

Since different types of pictures have different sizes and in a group of pictures all picture types can be found, starting with an I picture, usually the GOP size, or GOP bitrate is specified in an encoder.

$$R_{GOP} = \frac{\sum_{n=1}^{N_{GOP}} b_n}{TN_{GOP}}\,, \tag{4}$$

where R_{GOP} is the bitrate of the GOP, N_{GOP} is the number of frames in a GOP, b_n is the size of the single pictures and T is the frame time.

A rate-control algorithm usually needs to know the GOP configuration. However, in transcoding, the output GOP configuration is often determined by the input one, since transcoding typically does not change the frame coding types in order to keep complexity low. In real-time transcoding, the input GOP configuration is usually unknown to the transcoder.

To control the output bitrate and the size of the single pictures, our rate control algorithm needs to know the occurrence rate of different pictures types. The input video stream has to be analyzed as long as the frame coding type rate can be calculated. As GOP size is variable it means several GOP times.

To set the output bitrate, the picture global quantization scale of the currently transcoded picture has to be calculated from the current picture's known input quantization scale.

$$Q_{ti} = mQ_{oi} \ , \tag{5}$$

where Q_{oi} is the current picture's input global quantization scale, Q_{ti} is the global quantization scale of the currently transcoded picture, and m is a multiplier that controls the bitrate reduction. Using equations (2), (3) and (5), we can write:

$$\frac{R_a}{R_o} = \alpha\left(\frac{1}{m}\right) + \beta \ . \tag{6}$$

In equation (6), there is m that can be expressed as:

$$m = \frac{\alpha}{\left(\frac{R_a}{R_o}\right) - \beta}, \tag{7}$$

where R_a is the output average bitrate we would like to achieve, R_o is the input bitrate of the original video sequence, and α and β are parameters from equation (2).

In our previous calculations we did not take into account the real output bitrate, only the output bitrate that should be achieved. Similarly, we did not take into account that values of the used parameters are dependent on the picture characteristics. So we have to examine the effective output bitrate and correct the calculated parameter values:

$$e = \frac{R_t - R_a}{R_a} \ , \tag{8}$$

where R_t is the effective output bitrate, R_a is the output average bitrate that should be achieved and e is a weighted difference from the desired output bitrate. With this weighted difference value used as an error signal, we create a feedback to influence the rate control algorithm and the new quantizer value.

From the above and with the modification of equation (7), we get the final equation which contains all the parameters from equations above and gives the m multiplier value for the actual picture:

$$m = \frac{\alpha(1+e)}{\left(\frac{R_a}{R_o}\right) - \beta}, \tag{9}$$

where R_a is the output average bitrate that should be achieved, R_o is the input bitrate of the original video sequence, and e is the weighted error signal. Only the definition of α and β remains.

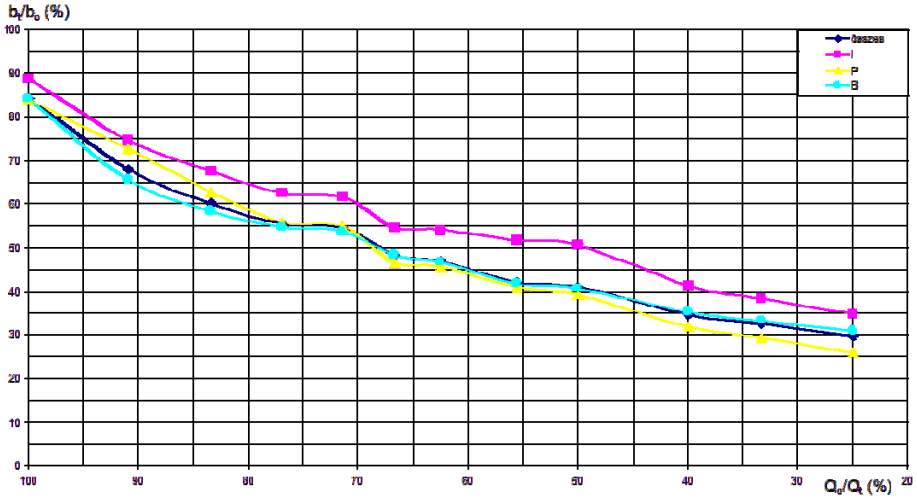


Figure 3: Picture size and picture global quantization scale.

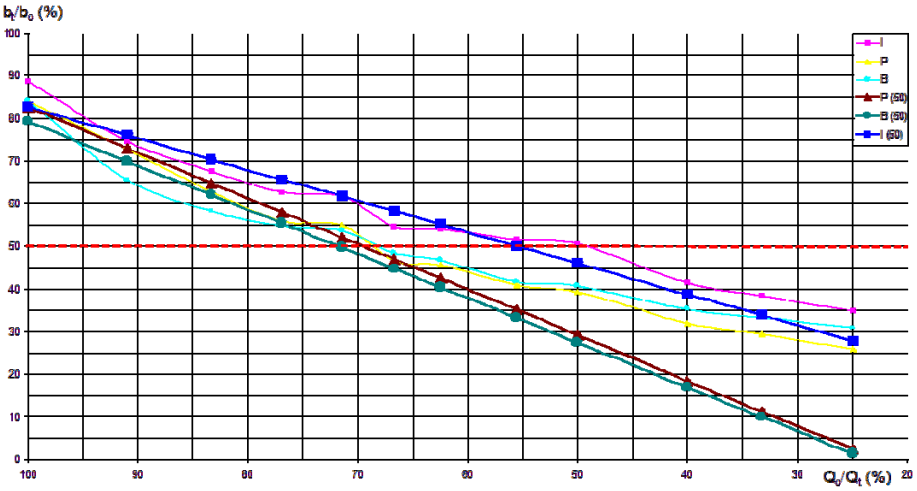


Figure 4: Picture size and linear approximation of picture global quantization scale.

Experiments were carried out to determine the value of α and β. We analyzed the average ratio between the number of bits in corresponding input

and output images, with different *m* values in different video sequences. The experimental results are shown in *Fig. 3* for one of the used video sequence (Mill).

It can be seen from the experimental results, that the ratio can be approximated by a linear function (2) between the number of bits in corresponding input and output images, and that α and β parameter values are different for different picture types. *Fig. 4* shows the lines of a linear approximation with a further assumption that – in general –, the degree of bitrate reduction will not exceed 50%. To achieve a more accurate approximation on that section only points belonging to ratios above 50% are taken into consideration.

4. Experimental results

We analyzed a transcoder realization based on the concept above with different video streams.

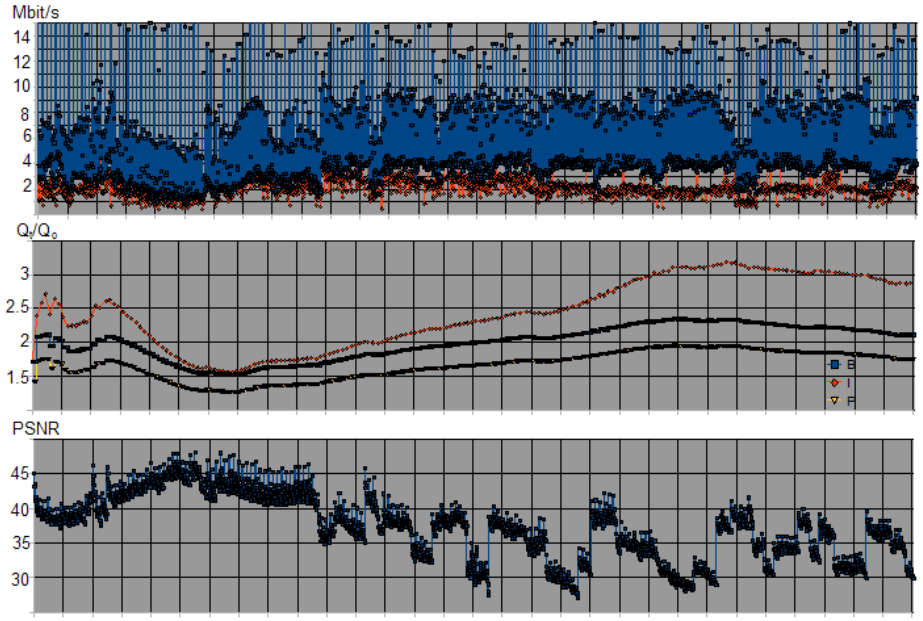


Figure 5: 3,000 frame, bitrate, value of *m* and picture quality (PSNR (dB)).

Results in *Fig. 5* are from a scene of 3,000 pictures from a VBR (Variable Bit Rate) live television broadcast. The input average bitrate is 4.5 Mbit/s, the

output average bitrate is set to 2Mbit/s, and it is VBR also. *m* multiplier is shown as changing to keep the set output bitrate.

Fig. 6 shows an enlarged part of *Fig. 5*, it is a scene of 300 pictures. The typical open-loop transcoder error called "breathing" can be seen on the picture quality graph.

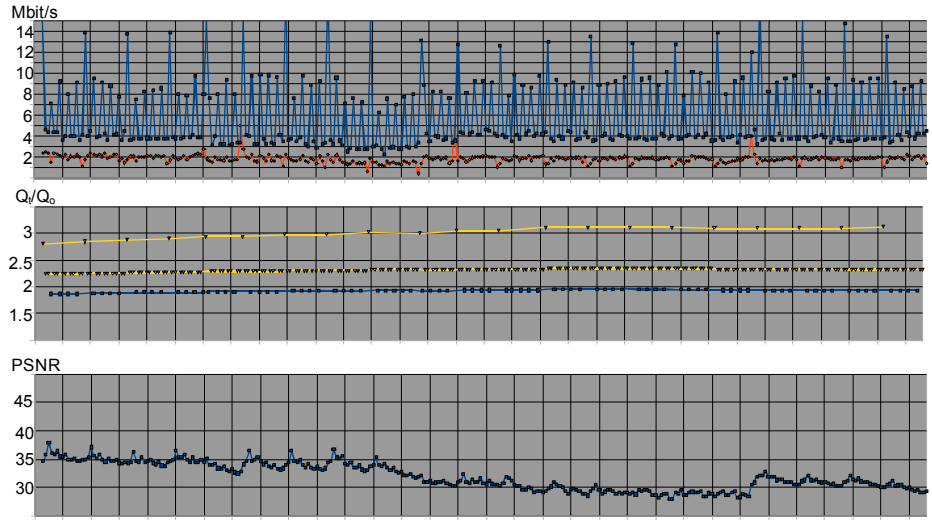


Figure 6: 300 frame, bitrate, value of *m* and picture quality (PSNR (dB)).

References

[1] Ahmad, I., Wei, X., Sun, Y., Zhang, Y.-Q., "Video transcoding: an overview of various techniques and research issues", *IEEE Trans. on Multimedia,* Vol. 7, Issue 5, pp. 793-804, Oct. 2005.

[2] Tudor, P.N., and Werner, O. H., BBC R&D "Real-time transcoding of MPEG-2 video bit streams" *IBC'97 Amsterdam*, pp. 286-301, 1997.

[3] Lastein, L., and Reul, B., "Transrating of MPEG-2 streams - various applications and different techniques", *BarcoNet - a Scientific Atlanta Company,* Denmark.

[4] Xin, J., Lin, C.-W., Sun, M.-T., "Digital Video Transcoding", *Proceedings of the IEEE*, Vol. 93, Issue 1, pp. 84 – 97, Jan. 2005.

[5] Assuncao, P. A. A.; Ghanbari, M., "A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG-2 bit streams", *IEEE Trans. on Circuits and Systems for Video Technology,* Vol. 8, Issue 8, pp. 953 – 967, Dec 1998.

[6] Int. Standards Org./ Int. Electrotech. Comm., (ISO/IEC), "Information technology - Generic coding of moving pictures and associated audio information: Video (MPEG-2 video)", *ISO/IEC 13818-2, 2nd ed*., 2000.

[7] "ISO/IEC-JTC1/SC29/WG11, Test model 5, MPEG93/457", Apr. 1993.

[8] Westerink, P. H., Rajagopalan, R., Gonzales, C. A., "Two-pass MPEG-2 variable-bitrate encoding", *IBM J. Res. Develop.,* Vol. 43, No. 4, July 1999.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 133-142

Subjective Video Quality Measurements of Digital Television Streams with Various Bitrates

Dénes DALMI[1], Tihamér ÁDÁM[2], Bence FORMANEK[3]

[1,2] Department of Automation, Faculty of Mechanical Engineering and Information Science, University of Miskolc, Miskolc, Hungary,
e-mail: [1]daldenisz@gmail.com, [2]adam@mazsola.iit.uni-miskolc.hu
[3] CableWorld Kft., Budapest, Hungary,
e-mail: formanek.bence@cableworld.hu

Manuscript received March 15, 2009; revised June 10, 2009.

Abstract: This paper first presents the most important standardized subjective quality assessment methods described in the ITU-R BT.500 recommendation. We briefly summarise why these subjective tests are so important. Finally, we discuss the implementation of the new subjective video quality measurement related to the impaired digital quality television programs. Our aim is to improve these subjective picture quality assessment methods to get sophisticated results, which correlate better with the objective picture quality test results. We would like to develop some objective picture quality measurements in the future.

Keywords: Subjective quality, Objective quality, Statistical multiplexing, Transport stream

Introduction

For the past few years we have dealt with subjective and objective picture quality measurements of digital television streams in the Digital Television Laboratory of the Department of Automation. After we had analysed the results of our subjective tests and drawn the conclusions, we started new subjective quality measurements, which focus on the video quality of the digital television streams, so-called transport streams having different bitrates.

Compression methods for digital television use different compression algorithms. Quality measurements are used to find the best compression method. There are two main categories of comparison methods: the objective video quality evaluation method based on mathematical calculations and the

subjective video quality evaluation methods based on tests performed by the audience.

Digital television streams are compressed according to the MPEG-2 or MPEG-4 standards. Nowadays digital television broadcasting systems often use statistical multiplexers. In statistical multiplexing, the communication channel is divided into an appropriate number of variable bitrate digital channels or data streams. Our goal is to determine the lowest bitrate, which has still acceptable quality. This bitrate would be used in statistical multiplexers as minimum bit-rate. Consequently, we use these quality measurements in order to find the compression parameters, which still result in acceptable video quality.

1. Subjective Television Picture Quality Assessment Methods

In this section we would like to introduce the most common subjective quality assessment methods of the digital television picture [1].

International recommendations for subjective quality assessment of television picture consist of specifications how to perform many different types of subjective tests. Subjective assessment methods are used to establish the performance of television systems. Measurements are therefore applied, which more directly anticipate the reactions of those who might view the tested systems. In this regard, it is understood that it may not be possible to fully characterize the system performance by objective means. Consequently, it is necessary to supplement objective measurements with subjective measurements.

In the course of a typical subjective quality test, a number of non-expert observers are selected, tested for their visual capabilities, shown a series of test scenes for about 10 to 30 minutes in a controlled environment and asked to score the quality of the scenes in one of a variety of manners.

In general, there are two types of subjective assessments. First, there are assessments that bring about the performance of systems under optimum conditions. These are usually called quality assessments. Second, there are assessments that create the ability of systems to retain quality under non-optimum conditions associated with the transmission or emission called impairment assessments. Some of these test methods are double-stimulus where viewers rate the quality or the change in quality between two video streams (reference and impaired). Others are single-stimulus where viewers rate the quality of just one video stream (the impaired). These methods will be later described.

In a modern television system, however, the picture quality is not a constant over time due to the compression streams. In the case of statistical multiplexing, the picture quality is a function of the complexity of the program material and the continuous operation of the transmission system. The selection of the

assessment method is affected by a number of procedural elements. These are the viewing conditions, the choice of observers, the scaling method to score the opinions, the reference conditions, the signal sources for the test scenes, the timing of the presentation of the various test scenes, the selection of a range of test scenes and the analysis of the resulting scores.

A description of the various subjective measurement methods provides some insight in the following sections.

1.1 Double-stimulus Impairment Scale Method

Double-stimulus Impairment Scale (DSIS) is a subjective assessment method when observers are shown multiple reference scenes and degraded scene pairs. The reference scene is always shown at first. Scoring is on an overall impression scale of impairment.

Table 1: Five-grade scale recommended by ITU

Five-grade scale	
Quality	**Impairment**
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

This scale is commonly known as the 5-point scale, where 5 equals with the imperceptible level of impairment and 1 shows the very annoying level as it is shown in *Table 1*.

1.2 Double-stimulus Continuous Quality-scale Method

In case of the Double-stimulus Continuous Quality-scale (DSCQS) method, observers are shown multiple sequence pairs with the reference and degraded sequences randomly first. Scoring is on a continuous quality scale from excellent to bad where each sequence of the pair is separately rated but in reference to the other sequence in the pair. Analysis is based on the difference in rating for each pair rather than the absolute values [2].

1.3 Single-stimulus Methods

Multiple separate scenes are shown in the Single-stimulus methods. There are two approaches: SS with no repetition of test scenes and SSMR where the test scenes are repeated multiple times. Three different scoring methods are

used. Adjectival scoring method has a 5-grade impairment scale, however half-grades may be allowed. Numerical scoring method has an 11-grade numerical scale, useful if a reference is not available. And finally there is a Non-categorical scoring, where assessors can score in a continuous scale with no numbers or a large range.

1.4 Stimulus-comparison Method

Stimulus-comparison method is usually accomplished with two well matched monitors but may be done with one. The differences between sequence pairs are scored in two different ways: Adjectival scale is a 7-grade, +3 to -3 scale labelled: much better, better, slightly better, the same, slightly worse, worse, and much worse, while Non-categorical is a continuous scale with no numbers or a relation number either in absolute terms or related to a standard pair.

1.5 Single Stimulus Continuous Quality Evaluation

Single Stimulus Continuous Quality Evaluation (SSCQE) is performed with a program, as opposed to separate test scenes, which is continuously evaluated over a long period of 10 to 20 minutes. Data is taken from a continuous scale every few seconds. Scoring is a distribution of the amount of time a particular score is given. This method relates well to the time variant qualities of new compressed systems. However, it tends to have a significant content of program quality in addition to the picture quality [4].

2. Statistical Multiplexing

The flexibility of the MPEG-2 coding system provides the opportunity to broadcast digital television streams, which have more or less bitrates. Everybody knows that the picture contains more information and has better quality when the rate of the stream, which transmits the compressed picture, is higher. In case of still or slowly moving picture sequences, which do not contain fine details, there is a limit, above which there is no use increasing the data rate, the picture, which has good quality, cannot be better at the receiver side. The change of the picture content and the moving of picture elements increase the amount of information to be transfer. Consequently, to observe the video quality, the data rate must be raised.

The creation of data rate depending on the picture content only makes sense when we can utilize the unused data rate range. In different transmission networks, where more TV programmes can be simultaneously transmitted, in

the spaces, which become vacant, one or more TV programmes can be delivered if we can control the resulting data rate.

Statistical multiplexing means that at transmitter site we compress the data stream with content-dependent data rate; however, we should meet the requirements that the resulting data rate cannot be higher than a predefined value. It is also important to determine a predefined order with which we ensure how much data rate will be allocated to the given programme in case of a large bitrate demand at the same time [3].

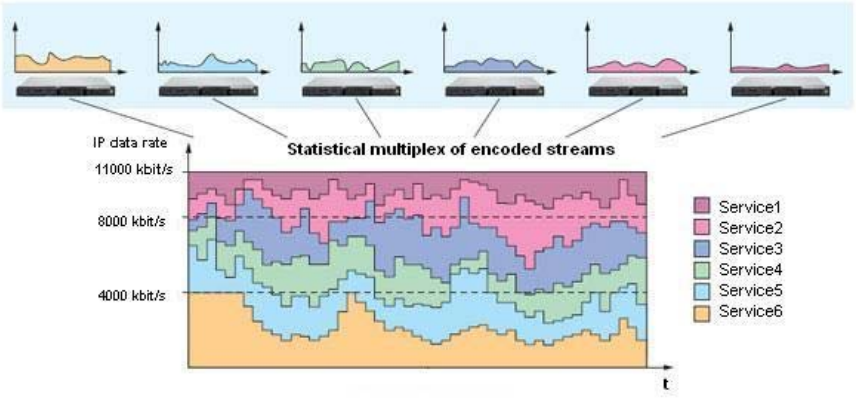


Figure 1: Statistical multiplexing.

Fig. 1. shows how the statistical multiplex works, so the digital television streams, which are coming from different locations (e.g. studios) with variable bitrates are added in one statistical multiplex stream.

With subjective quality measurements of digital TV streams, the minimum level of bitrate and other coding parameters, such as GOP (Group of Pictures) size and structure, as well as video picture parameters like brightness, contrast, saturation, can be determined. Nowadays there is a significant demand for these subjective results.

3. Subjective Video Quality Measurements

In this section we would like to describe our previous subjective picture quality measurements, and then we would like to go into details about our new measurements.

3.1 Short Presentation of Previous Quality Tests

We have previously executed three different types of subjective picture quality tests of digital television pictures coming from different digital television channels. We used a wide screen LCD television for the experiment,

whose screen could be separated into two parts. We chose three different digital television channels: satellite, cable and terrestrial. We selected three different programs: m2, Duna and Autonómia, which can be freely received in Hungary. The observers were undergraduates and one test session consisted of 5-15 of them. In the first test, observers rated the still pictures one after the other. In the second one, picture sequences were displayed in the two separate screens, so students had to evaluate the picture quality simultaneously. Finally, in the last test, observers assessed the quality of short motion picture sequences.

The evaluation was created by taking into account three aspects: sharpness, naturalness and subjective order. Therefore, observers had to determine an order between A and B pictures. They could note the results in an evaluation form. Test sessions took about 20-30 minutes. One test session comprised 8-12 pairs of 10-second pictures, covered the possible combination of different sources, such as satellite vs. cable. Between pictures there was a 10-second interval for the evaluation. Before the test pictures there was a mid-grey picture as mentioned in the ITU standard. We evaluated the test results by counting the votes of the observers in the different categories. In the serial subjective test of still pictures, we collected 216 votes, according to which the cable system got most of the votes in each category. In the serial test of motion pictures, we obtained a varied result, from the 243 votes gathered, the terrestrial system dominated in the sharpness category, while the satellite system got most of the votes in the naturalness and the subjective order categories [5].

Drawing the conclusions, we can make some important remarks. First of all, we should create some teaching methods for the video assessment, so that the non-expert observers could prepare for voting the quality. It is very important to teach the observers what they should pay attention before the real test, because it is really influence the test results. The experimenter should explain and demonstrate the evaluation categories (naturalness, sharpness, saturation, hue, etc.), the typical errors, which can occur in the digital video streams, and of course the essential information about the subjective quality assessment (number of test sequences, the duration of the voting period, the voting scale, etc.). In our opinion, by using a well-implemented teaching method, the fidelity of the subjective quality assessment can be improved.

Another important point is to select and record the test material in an appropriate way. In our previous subjective quality measurements it was a serious problem, that the test sequences were recorded after the error correction on the receiver side and not at the end of the transmission channel before the error correction. In the new subjective quality assessment, it was also a difficult task how to record test samples with various bitrates. We provide the related information in the following section.

We should also consider the laboratory circumstances (the distance between the screen and the observers, the resolution and other parameters of the television set, etc.). The ITU recommendation has good criteria to establish the appropriate laboratory environment; however, it has financial implication.

Finally, we should find a better way to record the votes of the observers, because so far they have filled a voting form. We had to evaluate thousands of voting papers, which resulted in mistakes. Consequently, a subjective quality assessment application is developed in order to help our work.

3.2 New Subjective Quality Measurement

As previously mentioned, our purpose is to conduct some subjective video quality tests of digital television streams, which have various bitrates.

3.2.1 Subjective Quality Assessment Supporter Application

For these measurements we have developed an application in Java environment, which provides a graphical interface in order to easily assess the digital television video.

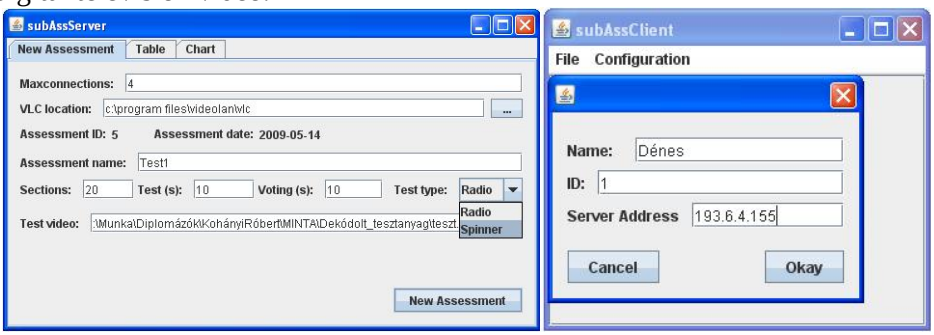


Figure 2: Subjective quality assessment software.

The program has two parts: the server and the client, which can be seen in *Figure 2*. The experimenter, who conducts the measurement, can configure or customize the subjective quality test on the *New Assessment* tab in the server software. First, the *Maxconnections* field has to be set, which determines the number of observers. Then, the experimenter should give the path of the VLC location. If it is well configured, then after the start of the new assessment, the VLC media player will display the test sequences. The assessment name and date is automatically set by the program. In the following steps the experimenter should give the name of the assessment, set the number of sections in the test session, configure the duration of one test sequence and the voting period in seconds and select the type of the test scale, which can be a 5-grade scale

recommended by ITU as it is shown in *Table 1.* or a spinner, which is a 100-grade continuous scale. Finally, the path of the test material has to be set.

The observers should run the client program and set some parameters, such as the name, the unique ID and the IP of the computer on which the server application runs.

When the experimenter starts the measurement, which can be automatic or manual, the voting screen will automatically appear on the client screen and the observers will have a defined amount of time to score the quality. The client software sends the scores to the server application, which stores them into its database. When the subjective measurement is finished, the experimenter can evaluate the results in a table or in a chart. The table contains the assessment ID, the assessment name and date, the assessor ID and name, the section number and the quality score. With SQL commands, the experimenter can create some queries in order to filter the huge amount of data. In the chart, the results of a given assessment can be seen, where the two axes are the number of sections and the mean value of the scores voted by the observers.

3.2.2 Recording the Test Material

Our first task was to record digital television video samples, which have different bitrates. *Fig.3.* presents the environment, how we recorded the test material.

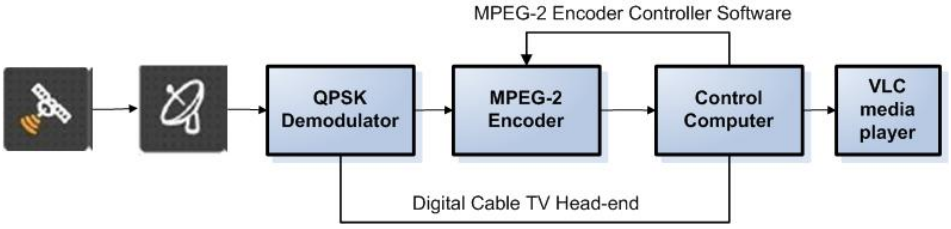


Figure 3: Environment for Recording the Test Material.

In the Digital Television Laboratory we used the Digital Cable TV Head-end, which contains special hardware devices developed by CableWorld Ltd. The QPSK demodulator is used to receive the digital transport streams broadcasted via satellite channel. The demodulated transport stream is then sent to the MPEG-2 Encoder. With the MPEG-2 Encoder Controller application running on the Control Computer, the coding parameters and the bitrates of the transport stream could be configured. In the final step, this encoded transport stream was displayed with the VLC media player. We used this media player to record video samples.

The problem was that we could not record test samples with various bitrates continuously; it was the fault of the VLC media player. Therefore, we recorded

10-second video samples and concatenated them into one test video sequence, which could be later used for the subjective quality measurements. However, we have not found appropriate MPEG-2 editor software yet, with which we can concatenate the splitted sections without re-encoding them. So it is a problem, which needs to be solved in the future.

3.2.3 Presentation of the Subjective Quality Assessment and the Result

We established a quality assessment environment in our laboratory. We created a computer network with 9-12 personnel and one server computers. Observers used the personal computers to run the client application. On the server machine the experimenter run the server application and conducted the subjective quality test. One test session was taken about 10-20 minutes, because the observers were needed to concentrate hard under the quality assessment.

Table 2: Five-grade scale recommended by ITU

Seq. N.	Bitrate (kbps)	1. Measurement (0-5)	2. Measurement (0-100)
1.	8000	2.75	39.75
2.	992	1.25	4.75
3.	1504	3.75	51.50
4.	4000	4.50	73.25
5.	1104	1.50	8.25
6.	1600	2.50	39.25
7.	2608	5.00	87.75
8.	3504	4.25	79.75
9.	3008	3.75	69
10.	2800	4.50	67.75
11.	1904	3.50	33.50
12.	1200	2.25	21
13.	6000	3.50	76
14.	1312	1.00	7.75
15.	4512	4.00	67.75
16.	1408	2.25	22.25
17.	2400	3.25	65
18.	5008	4.00	73
19.	2000	4.25	75.50

So far we have only a few number of test result as described in *Table 2.*We used a test material included 19 sections with different bitrates. In the first and the second measurements the mean of the quality scores can be seen. The difference between the two measurements is the voting scale, which was used for the test. It can be seen that the video sequence, which has higher bitrate, had got better quality scores, but there are discrepancies in the test results. It is

important to mention that this result is not representative, because the number of assessors, who have already taken part in our assessment, is less than 10.

To give a significant result we need to repeat this measurement with a large number of observers. According to our assumption, the lowest bitrate, which has still acceptable quality, is about 1500 Kbit/s. However, it will be our future work to verify it.

4. Conclusion

In this paper we have dealt with subjective quality assessments. We have introduced different assessment methods that we would like to apply for future measurements. Then, we have described our previous subjective quality assessment tests and listed some points in which we could improve. Finally, we have presented a new subjective video quality measurement of digital television streams in order to specify the minimum bitrate with an adequate quality. We have had only assumptions for the exact value of this bitrate; however we collected some useful experiences. We will have to solve some problems in the future, e.g. to create test materials in an appropriate way, to develop a well-applicable teaching method, etc.

Acknowledgements

We would like to say thank you to all employees of CableWorld Ltd. and the members of the Automation Department to help our works. Finally, special thanks to Prof. György Lajtha and Mihály Szolokai to contribute to our work with valuable advice.

References

[1] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures", *ITU-R Recommendation BT. 500-11*, Geneva, Switzerland, 2002, pp. 2-24.

[2] Veres, P., "Digitális adatjelek átvitele és kiértékelése", in *CableWorld hírek (CableWorld Kft. technikai magazinja)*, Vol. 11, Budapest, 1999.

[3] Zigó, J., "A statisztikus multiplexelés, és az MPEG-2 adatsebesség csökkentése", in *CableWorld hírek (CableWorld Kft. technikai magazinja)*, Vol. 38, Budapest, 2008.

[4] Dalmi, D. "Subjective assessment of picture quality of different digital television channels", in *6th International Conference of PhD Students*, Pécs, 2007, pp. 25-30.

[5] Dalmi, D., Ádám, T., "Subjective and objective picture quality test of digital television programs", in *9th International Carpathian Control Conference*, Sinaia, 2008, pp. 111-114.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 143-154

Challenges in a Web-enhanced Personalised IPTV Service

Sándor SZÉKELY [1], Tamás SZÁSZ [2],
Zoltán SZAPPANYOS [2], Zsolt TÓFALVI [2]

[1] Nokia Siemens Networks, Munich, Germany,
e-mail: sandor.szekely@nsn.com
[2] Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tîrgu Mureş, Romania,
e-mail: szasz.tamas@gmail.com; szappanyos.zoli@gmail.com; tofalvi.zsolt@gmail.com

Manuscript received March 15, 2009; revised May 28, 2009.

Abstract: Internet protocol television (IPTV), one of the most emerging services, offers multimedia streaming services with security, reliability, and relevant quality of service (QoS) / quality of experience (QoE). It provides added values to all the involved players including customers and also brings technical and business challenges to those players. For IPTV services, we expect to adopt the managed network environment for high quality and the Web technologies for personalization to meet the customer's necessity. Web can provide an open, flexible, and agile platform. Therefore, in this paper, we propose personalized IPTV services based on Web-enhanced open platform and present the functional architecture. Technical issues for deploying the proposed services using Web are also provided. The objective of this paper is to analyze the critical architectural and design issues for developing an attractive, high-quality, viable and feasible model for personalised service.

Keywords: internet protocol television (IPTV), web-based television (WebTV), Electronic Programming Guide (EPG), metadata functionality, recommendation engine (RE)

1. Introduction

Currently, digital television is gradually replacing analogue TV. Although these digital TV services can be delivered via various broadcast networks (e.g., terrestrial, cable, satellite), Internet Protocol TV (IPTV) over broadband telecommunication networks offers much more than traditional broadcast TV. It can improve the quality that users experience with this linear programming TV service, but it also paves the way for new TV services, such as video-on-demand, time-shifted TV, and network personal video recorder services, because of its integral return channel and the ability to address individual users

[1]. IPTV service is considered as the emerging application that has a great potential to generate new revenues for contents and service providers.

In the last few years, we have witnessed a rapid growth of multimedia content delivery across the networks. Peer-to-Peer (P2P) networks and user generated content (UGC) are considered as one of the most suitable targeted infrastructure for supporting real time streaming and has played vital role in this growth. One of the major challenges of this approach is to reach the same quality of service of traditional television and commercial IPTV by employing only best effort network layer services. This service gives more choice to the end users to consume TV programs, on-demand content and UGC in a personalized way and beyond any geographical constraints [2], [3].

IPTV represents a solution for interactive television-like services. It combines streamed video, Web services, and eventually voice services. Personalized IPTV will be a key solution for value added services over the internet and the emergence of Web and next-generation networks (NGN) will provide open platforms and new business models for IPTV services in the future [4], [5].

The IPTV service stays in the centre of interest and is globally deployed. Web-based IPTV service is one scenario for efficient delivery of service. One can serve IPTV service through the internet, and IPTV subscribers can connect the service using the Web browser. Here, one can see that the existing Terrestrial / Cable / Satellite TV contents can easily deliver through the internet by integration of Web and IP data services (e.g. Joost, Zattoo). But the number of output channels to the subscribers is less than the number of input channels from existing broadcaster when existing broadcasting services are transmitted through the internet. In general, the output channel is allocated in the sequence required by the subscriber [6].

Current standardisation activities are presented in [7] for Web-based metadata applications, especially those of the ITU-T IPTV Focus Group, which is interested in service requirements, architecture and functionality to provide IPTV service. Among others, metadata functionality and Web technology are necessary to support interactive data service. Web technology is moving towards a so called Web 2.0, where one of the major features is to support more interactive capability.

The objective of this article is to present the design of a web-enhancement of an IPTV model, establish the associated framework for such IPTV services delivery, and contribute to discussions and research activities. At the same time, the conclusions and the proposed model will serve as a framework for future efforts in our laboratory test environment. Chapter 2 presents our modelling approach, explains the selected architecture, gives implementation details and describes the reasoning behind the selected programming techniques. Chapter 3

gives further details on the methodology of the implemented recommendation engine. In Chapter 4 we discuss the challenges during design and implementation phase. Our future plans are revealed in Chapter 5. Finally we draw our conclusions and provide a list of recent reference works.

2. Architecture of the web-enhancement of an IPTV service

At the beginning of our investigations we have analyzed tens of web-based applications, like youTube, joost, Zattoo, audiTv, videa, mySpace, flickr, hirTV, mtv, DunaTV, etc. We have looked for fancy design ideas, creative features and intuitive navigation with a clear target to create something new. Beside this, we have run speed tests and taken into account the usability, simplicity, platform independency, multiple tools and many features that web-based systems provided. Based on this learning, we started to design our own model. For example, we decided that the clients do not need to download any additional software when they open our web-based application, and there should not be any need for additional installation and configuration. A well-known WebTV application that started with downloadable client software is Joost. After a while they had to abandon the idea of using downloadable player and switched to an all web-based solution. But Joost is not the only one who realised that the future is a web-based solution. Others who had not switch to web-based applications failed or they are about to fail or simply they are not enjoying popularity.

For the design of our application, the main goal was to structure it in modules which can be easily maintained and extended in the future (see *Fig. 1*). Besides that, the optimization process will be easier too, namely because a well structured and clean code is easier to be optimized.

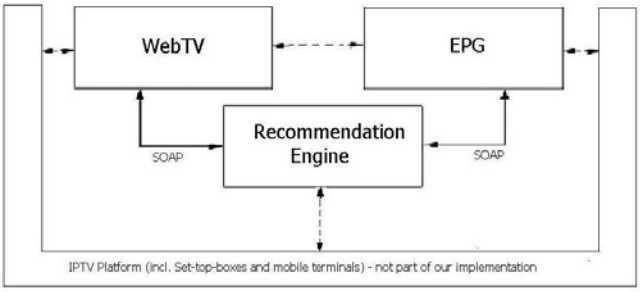


Figure 1: The basic modules of our implementation.

Our system consists of three main modules which are to be integrated into an IPTV platform in the future. However, currently these modules work stand-alone and co-exist without any background IPTV platform. The WebTV module

is responsible for displaying the aimed content on a PC or a TV set with integrated internet access. The consumer shall search for a specific video content or linear TV program. To ease the selection of a TV channel and that of a dedicated program within the channel, the EPG module was created. We have analysed the available EPG sites on the web, like tvtv.de, zingzing.co.uk, programy-tv.cz, port.ro, dunatv.hu, neuf.fr, then we have tried to take the best ideas out of them and implemented it in our own way. The third module is a so called recommendation engine (RE). Our aim is that the consumer after watching a movie or a TV program will be provided with a recommendation to watch another movie or program of the same category or with the same actor, or from the same author, etc. Our output is based on the personal profile of consuming movies and TV programs, while this profile will be updated continuously. Our recommendation engine is presented in more details in Chapter 3. *Fig. 2* presents more details on the modularity of our solution. The most important sub-modules, programming languages and interfaces are also depicted.

Without being subjective, we realized that one of the best web-based video-players is that of Adobe Flash, which is also easy to embed into our application. However, it is not optimal to implement the whole application in Flash technology. The main reasons are: long loading-time of the whole site, relatively big amount of data required, lack of modularity of the site, more time needed for development and difficulties in the personalization of the URL site. Presentation of text-based frames in Flash is difficult; copying text is simply not possible. In contrary, an html-based site performs better in all above examples. Therefore, it is straight-forward to decide for a flash player embedded into an html-based presentation layer.

Adobe ***Flash*** is a multimedia platform, Flash is commonly used to create animation, advertisements, and various web page components, to integrate video into web pages, and more recently, to develop rich Internet applications. Several software products, systems, and devices are able to create or display Flash content, including Adobe Flash Player, which is available for most common web browsers.

Ajax (Asynchronous JavaScript and XML), is a group of interrelated web development techniques used to create interactive web applications or rich Internet applications. With Ajax, web applications can retrieve data from the server asynchronously in the background without interfering with the display and behaviour of the existing page. The use of Ajax has led to an increase in interactive animation on web pages. Data is retrieved using the XMLHttpRequest object or through the use of Remote Scripting in browsers that do not support it. Despite the name, the use of JavaScript and XML is not actually required, nor do the requests need to be asynchronous.

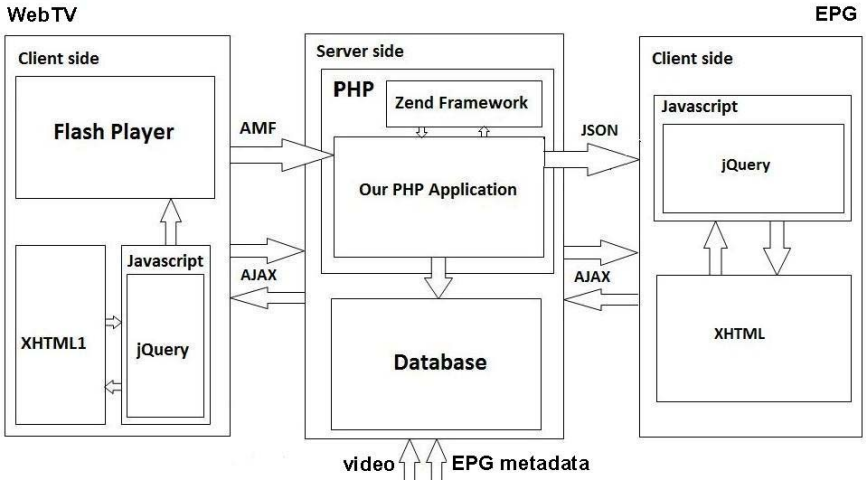


Figure 2: Sub-modules and interfaces of our WebTV and EPG modules.

XML (Extensible Mark-up Language) is a general-purpose specification for creating custom mark-up languages. It is classified as an extensible language, because it allows the user to define the mark-up elements. XML's purpose is to aid information systems in sharing structured data, especially via the Internet, to encode documents, and to serialize data; in the last context, it compares with text-based serialization languages such as JSON.

JSON (JavaScript Object Notation) is a lightweight computer data interchange format. It is a text-based, human-readable format for representing simple data structures and associative arrays (called objects).

PHP is a widely-used general-purpose scripting language that is especially suited for web development and can be embedded into HTML. It generally runs on a web server, taking PHP code as its input and creating web pages as output. It can be deployed on most web servers and on almost every operating system and platform free of charge.

Zend Framework (ZF) is an open source, object-oriented web application framework implemented in PHP 5. ZF is a use-at-will framework. There is no single development paradigm or pattern that all Zend Framework users must follow, although ZF does provide components for the Table Data Gateway, and Row Data Gateway design patterns. Zend Framework provides individual components for many other common requirements in web application development. Zend Framework also seeks to promote web development best practices in the PHP community; conventions are not as commonly used in ZF as in many other frameworks [8].

JavaScript is a scripting language used to enable programmatic access to objects within other applications. It is primarily used in the form of client-side JavaScript for the development of dynamic websites. JavaScript, despite the name, is essentially unrelated to the Java programming language even though the two do have superficial similarities. Both languages use syntaxes influenced by that of C syntax, and JavaScript copies many Java names and naming conventions. ***jQuery*** is a lightweight JavaScript library that emphasizes interaction between JavaScript and HTML [9].

Action Message Format or ***AMF*** is a binary format based loosely on the Simple Object Access Protocol (SOAP). It is used primarily to exchange data between an Adobe Flash application and a database, using a Remote Procedure Call. Each AMF message contains a body which holds the error or response, which will be expressed as an ActionScript Object.

SOAP, originally defined as Simple Object Access Protocol, is a protocol specification for exchanging structured information in the implementation of Web Services in computer networks. It relies on Extensible Mark-up Language (XML) as its message format, and usually relies on other Application Layer protocols (most notably Remote Procedure Call (RPC) and HTTP) for message negotiation and transmission. SOAP can form the foundation layer of a web services protocol stack, providing a basic messaging framework upon which web services can be built.

An electronic program guide (***EPG***) is a digital guide to scheduled broadcast television or radio programs, typically displayed on-screen with functions allowing a viewer to navigate, select, and discover content by time, title, channel, genre, etc. by use of their remote control, mouse or a keyboard.

We have used PHP script language to generate dynamic html pages, as PHP is easy to connect with html presentation. Thus we can refresh in background parts of mixed web pages using Ajax technology. The communication takes place in the background and remains invisible for the consumer. The above technologies are all open, so we had the possibility to embed these frameworks in order to support our work. We have selected the Zend Framework on the server-side and the jQuery Framework on the client-side, as being the fastest, most robust and extendable among many frameworks. Usually there is a non-trivial problem to communicate between PHP and Flash, but the selected Zend Framework provides an elegant solution to this using AMF format.

Our implemented EPG allows a very easy navigation and selection of 100+ broadcast TV programs. Radio programs are currently not included. A sample implementation can be found in *Fig. 3*.

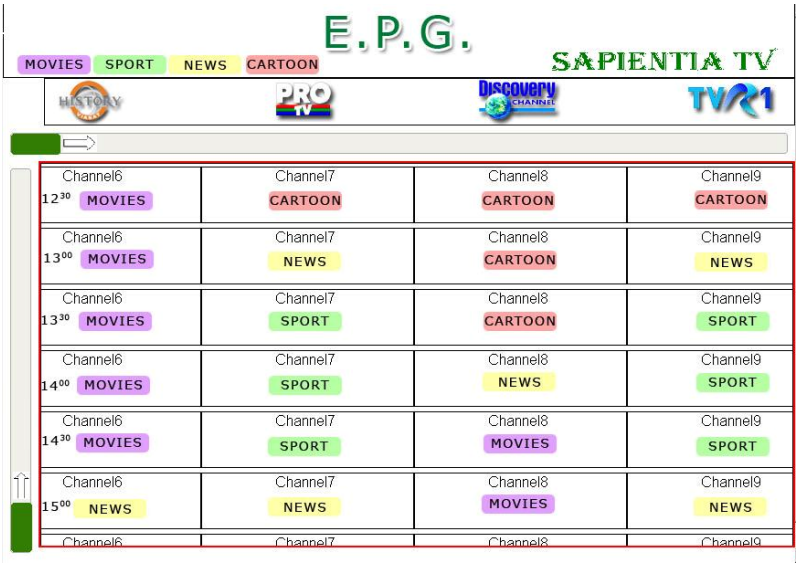


Figure 3: The EPG presentation allows easy navigation and selection (simplified).

All information about TV channels are stored in an XML file. This file is a Linux shell script, which is periodically updated with help of an external program. Data extraction from this XML file is processed in the server-side with help of the XPath expression that provides input to the PHP_simpleXML function. The server-side inserts the provided TV channels into the browser and generates a table that holds the daily channel information. The requested data arrives asynchronously to the client-side in JSON format, as reply to an XHTTP request generated by the client. The received data is processed by a JavaScript, and then displayed on the screen. The interaction with the browser is done by the same function. When the end-user navigates on the screen and moves the cursor, new data might be needed to replace old data. A JavaScript logic decides which data is necessary and if so, then generates new XHTTP requests to the server.

3. Recommendation Engine

The personalization and recommendation has more and more importance in today's TV consumption and web behaviour. More than 100 TV channels are a lot more than the consumer can easily manage; furthermore the internet is a huge information labyrinth, so our clear purpose is to make users' navigation and decision easier.

The manually filled profiles have risen seriously, but this is subjective and does not adapt to changing interests of users. There are two frequently used algorithms in today's recommendation systems: Slope one and Pearson Correlation [10]. These algorithms do not take into consideration the different kind of users and the human nature. A complete intelligent Web personalization system is generally based on Web usage data mining to discover useful knowledge about user access patterns, followed by a recommendation system to act on this knowledge in order to respond to the users' individual interest. The knowledge discovery component must discover distinct user profiles from Web usage data. Their unsupervised nature also avoids reliance on input parameters or prior knowledge about the number of profiles to be sought. A recent work [11], used item-to-item collaborative filtering as a recommendation strategy in Amazon.com. This approach matches each of the user's purchased and rated items to similar items then combines those similar items into a recommendation list. A similar-item table is built by finding items which customers trend to purchase together. Unfortunately, they do not present any empirical results, or sufficient details about the proposed technique. Moreover, it is not clear how this approach differs from standard frequent item-set based techniques.

There are three steps in the recommendation process: personalization, making clusters and decision-making with Fuzzy logic.

A. Personalization

This step is the automatic identification of the user profiles and updating these profiles via training with the newly mined information. At a first step it is better to use predefined user profiles. After the system collects enough information (e.g., about user's navigation activities and viewed content), we use clustering to give an updated description of the user's interest (e.g. in *Fig. 4*).

We implemented a hierarchical version of a robust genetic clustering approach. This process can be executed offline and periodically.

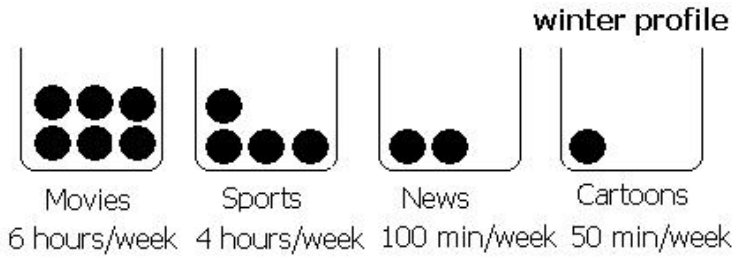


Figure 4: A sample viewing profile of a family member.

B. Making clusters

Our selected clustering algorithm is a so called Hierarchical Unsupervised Niche Clustering Algorithm (H-UNC). The basic steps of the algorithm are as follows:

- Encode binary session vectors;
- Set current resolution Level L = 1;
- Start by applying UNC to entire data set w/ small population size;
- Repeat recursively until cluster cardinality or scale become too small
  {
  - Increment resolution level: L = L + 1;
  - For each parent cluster found at Level (L-1)
    Reapply UNC only on data subset assigned to this parent cluster to
  - Extract more child clusters at higher resolution (L > 1)
  }

This clustering algorithm needs to execute in long periods (monthly or 3 monthly).

C. Decision making

Fuzzy approximate reasoning is an inference procedure that derives its conclusions from fuzzy rules and known facts. Generalized Modus Ponens is a generalization of modus-ponens which is the basic rule of inference in traditional two-valued logic. In two valued logic an implication (A _B) is used to infer the truth of proposition B from the truth of proposition A, and these truths can only be absolute (i.e., true or false). Generalized Modus Ponens performs logical implication in an approximate manner to deal with uncertainties in the facts (input) and in the implication itself. It is formalized as follows:

Fuzzy Rule: If x is A then y is B. Fact: x is A' → Conclusion: y is B' (1)

Where A and B are linguistic variables defined by fuzzy sets on the universes of discourse X and Y, respectively. A fuzzy If-then rule can be defined as a binary fuzzy relation R on the product space X × Y. The relation matrix R, which encodes the fuzzy implication, can be considered as a fuzzy set with a two-dimensional membership function: $\mu R(x,y) = f(\mu A(x), \mu B(y))$ that maps each element in X × Y to a membership grade in [0,1]. Using the compositional rule of inference, we can formulate the inference procedure in fuzzy reasoning as follows:

B' = A' ~ R (2) (2)

where ~ denotes a fuzzy composition operator, consisting of a conjunction, followed by a disjunction.

In the context of personalization, X denotes the space of profiles, Y denotes the space of visited contents, R is a relation that maps profiles in X to contents in Y with varying degrees of relevance. The rows of R can be defined by the relevance weights/components of the profile vectors. Input fact A' is a fuzzy set defined on X, thus naturally consisting of the memberships of a given session in each one of the profiles in X. Output B', the composition of A' and R is a fuzzy set defined on the set of contents, Y. This is the conclusion of the inference procedure, and compared to the original user session that was limited to a crisp set of contents, B' represents the possibility that each content on our website (or later on the TV screen) is of relevance to the current user as inferred via relation matrix R.

For better results we should train the membership-functions with a multilevel neural network. The "learning" data is the currently mined content information about the user, in this case the profile information we are storing in the weights of input membership-functions.

4. Challenges during design and implementation

Browser compatibility: Today we are over-flooded with a lot of different browsers on the market. Competition is good, but the disadvantage is that not all browsers support all the frameworks we are using, thus we have the challenge to test our solution with as many browsers as possible and provide compatibility. The jQuery framework helps a lot, but is still not so straight forward.

Common database: We faced some problems by database access from different modules. A common database offers advantages in many senses, but brings also a lot of disadvantages into the system, e.g., the concurrent access.

Modularity: We have decomposed our system into modules, aiming to define independent entities as much as possible. This is not so trivial. Due to security reasons the different modules cannot access the code of the other ones, thus we introduced the SOAP interface to realize the communication between modules.

Security issues: Due to the fact that parts of our source code are visible for smart users, it is extremely important to isolate the JavaScript on the client side from the rest of the modules. Another security threat could be any unverified input data from end users, therefore we strictly verify every input data before running, and this control is well supported by the Zeng Framework.

Concurrent usage: In case that our system is used by many users at the same time, we are challenged to run through an optimization process in the

current configuration. The availability of necessary external bandwidth is the problem no.1. As we already use the most modern and efficient encoding technology (H.264), it only remains to ensure more capacity. Suppose that we do not have a bandwidth problem, then our architecture could serve thousands concurrent users. This limits of course the usage of our solution in commercial systems, which can be easily solved by proper engineering (e.g. more HW).

Storage: Last but not least the storage space and its read/write speed is also a challenging problem in our solution. We did not go for expensive commercial solution, but selected a scalable file-server.

5. Next steps

At the moment we are still in the implementation phase, but our roadmap contains a couple of new elements. We aim to extend our PC-client based solution towards set-top-box based TV consumption and mobile handset. An important aspect will be adapting the display resolution to TV screen and mobile screen. Another aspect is derived from the consumer behaviour of leaning backward when watching TV, thus we need to adapt the size of the text which appears on the screen.

A third challenge is the integration of the remote controller into our application. We have started to create a new module which processes the necessary programs to interface with the remote controller. Basic functions are already available (like volume control). Exporting our application to mobile screens basically introduces similar problems as with the TV set, except the video resolution has to be reduced. The biggest challenge however is the unavailability of the right software for mobile handsets or the integration of our application into a handset, which does not necessarily provide or support open interfaces. A first approach is the use of Adobe flash lite, which is already implemented in a reduced number of mobile terminals.

Currently our web architecture supports only a video-on-demand solution, but we plan to integrate live TV sessions and IP cameras as well. Last but not least we also plan to introduce further features to support instinct navigation on the screen.

6. Conclusions

The IPTV service can deliver TV programs anytime anywhere. IPTV supports both broadcast and unicast services like Live-TV and Video-on-Demand. This paper identified the challenges in delivering web-enhanced IPTV and proposed a framework to provide solutions to those challenges. We have built a prototype of the system and demonstrated its flexible features, integrated

EPG, remote controller and recommendation engine. We are firmly convinced that our new EPG design is one of the most competitive designs on the market.

We are living in the content-centric world. The user experience of this new media is thought as a key factor for the success of an IPTV service. We are aware that still a lot of innovation is necessary in order to win the battle on "the last millimetres", namely between the consumer eyes and his brain.

Acknowledgements

The authors would like to express their gratitude to the Department of Electrical Engineering at the Sapientia University for providing the necessary hardware in order to support the ongoing research work.

References

[1] Degrande, N., Laevens, K., De Vleeschauwer, D., Sharpe, R., "Increasing the user perceived quality for IPTV services", *IEEE Communications Magazine*, Volume 46, Issue 2, pp.94–100, February 2008.

[2] Manzato, D., da Fonseca, N., "Peer-to-Peer IPTV Services", *IEEE GLOBECOM Workshops*'08, 30 Nov.–4 Dec. 2008, pp.1-6.

[3] Mushtaq, M., Ahmed, T., "P2P-based mobile IPTV: Challenges and opportunities", *IEEE/ACS Int. Conf. on Computer Systems and Applications, AICCSA'08*, 31 March - 4 April 2008, pp.975-980.

[4] Volk, M., Guna, J., Kos, A., Bester, J., "IPTV Systems, Standards and Architectures: Part II - Quality-Assured Provisioning of IPTV Services within the NGN Environment", *IEEE Communications Magazine*, Volume 46, Issue 5, pp.118–126, May 2008.

[5] Gyu M.L., Jun K.C., "Personalized IPTV Services using Web-based Open Platform in NGN", *IEEE Global Telecommunications Conference, GLOBECOM'08*, 30 Nov. - 4 Dec. 2008, pp.1–5.

[6] Hongnyun K., Junkyun C., Sanghyun C., "Flexible Channel Allocation Algorithm for Web-based IPTV Service", *10th Int. Conf. on Advanced Communication Technology, ICACT'08*, Volume 3, 17-20 Feb. 2008, pp.1544–1547.

[7] Sunghan K., Lee, S.Y., "Web Technology and Standardization for Web 2.0 based IPTV Service", *10th Int. Conf. on Advanced Communication Technology, ICACT'08*, Volume 3, 17-20 Feb. 2008, pp.1751–1754.

[8] Allen R., Lo N., Brown S., "Zend Framework in Action", *Hanning Publ*., available at Amazon.com, 2009.

[9] Chaffer J., Swedberg K., Resig J., "Learning jQuery 1.3: Better Interaction Design and Web Development with Simple JavaScript Techniques", *Packt Publ.,* available at Amazon.com, February 2009.

[10] Klir, G. J., Yuan, B., "Fuzzy Sets and Fuzzy Logic", *Prentice Hall Publ.*, 1995.

[11] Nasraoui, O., Petenes, C., "An Intelligent Web Recommendation Engine Based on Fuzzy Approximate Reasoning", University of Memphis, 2008.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 155-161

Adapted Discrete Wavelet Function Design for ECG Signal Analysis

Zoltán GERMÁN-SALLÓ

Department of Electrical Engineering, Faculty of Engineering,
"Petru Maior" University of Tîrgu Mureş, Tîrgu Mureş, Romania,
e-mail: zgerman@engineering.upm.ro

Manuscript received March 15, 2009; revised July 06, 2009.

Abstract: The main task in wavelet analysis (decomposition and reconstruction) is to find a good wavelet function (mother wavelet) to perform an optimal decomposition. The goal of most wavelet researches is to create a set of basis functions and transforms that will give an informative, efficient and useful description of a signal. It is better if the wavelet function is adapted to the signal, because the computational costs can be reduced and more accurate analysis can be obtained. This paper presents a discrete wavelet function synthesizer, which starts from an arbitrary, discretized sequence, to obtain the reconstruction and decomposition filters. The pursued criterion (expected result) is to minimize the reconstruction error between a first or second order approximation and the original signal.

Keywords: wavelet analysis, decomposition and reconstruction filter banks, multiresolution analysis, discrete wavelet transform

1. Introduction

The analysis of an ECG signal has been used as a diagnostic tool to provide information on the functions of the heart. The ECG is the graphical representation of variation in time of a potential difference between two points on a human body surface as a result of the activity of the heart. The wavelet transform is a recently developed signal processing technique, created to overcome the limits of the classical Fourier analysis, to deal with non-stationary signals like biomedical signals. The wavelet transform of a signal is calculated by taking the convolutive product between the biological signal and basis functions, measuring the similarity between them. The result of this product is a

set of coefficients. This set of coefficients indicates how similar is the signal relative to the basis functions. In the case of wavelet analysis, the basis functions are scaled (stretched or compressed) and translated versions of the same prototype function, called the mother wavelet $\psi(t)$. Theoretical knowledge about mathematical backgrounds of wavelet transform can be found in [1], [2], [3]. This paper briefly introduces a new method, using softcomputing elements to synthesize new wavelet functions in order to have with them an optimal decomposition structure. The expected result is to minimize the reconstruction error between a first order approximation and the original signal.

2. Wavelet Decomposition and Reconstruction

The wavelet transform is a decomposition of the signal as a combination of a set of basis functions, obtained by means of scaling a and translation b of a mother wavelet $\psi(t)$. The continuous wavelet transform (CWT) uses the dilation and translation of the mother wavelet function ψ. The CWT of signal $x(t)$ is defined as [1]:

$$W_a x(b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} x(t) \cdot \overline{\psi}\left(\frac{t-b}{a}\right) dt . \tag{1}$$

where a is a scale factor which is proportional to the inverse of frequency and b is the translation parameter. The scale factor and the translation parameter can be discretized, the usual choice is to follow a dyadic grid for them. The transform is then called the (dyadic) discrete wavelet transform (DWT).

$$C(j,k) = \sum_{n \in Z} x(n) \cdot \psi_{j,k}(n),\ \psi_{j,k}(n) = 2^{-j/2} \cdot \psi\left(2^{-j} n - k\right) \tag{2}$$

For discrete time-signals, the dyadic discrete wavelet transform (DWT) is equivalent according to Mallat's algorithm [1] to an octave filter bank, and can be implemented as a cascade of identical filter cells (low-pass and high-pass finite impulse response(FIR) filters) as shown in *Fig. 1*.

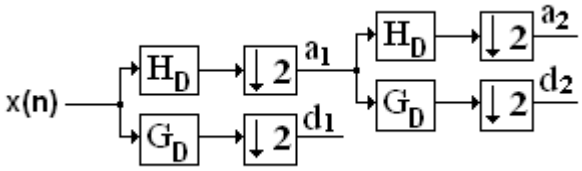


Figure1: FIR filter structure for dyadic scale decomposition.

The decomposition procedure starts with passing signal (discrete sequence) through a half band digital low-pass filter H_D with impulse response h(n). Filtering corresponds to the convolution of signal with the impulse response of the filter. A half band low-pass filter removes all frequencies that are above half of the highest frequency in the signal, but leaves the scale unchanged. Only the subsampling process changes the scale. In summary the low-pass filtering halves the resolution but leaves the scale unchanged. The signal is then subsampled by 2 since half of the number of samples is redundant. This operation doubles the scale (*Fig.1*). The operators H_D and G_D correspond to one stage in the wavelet decomposition, the spectrum of the signal is split in two equal parts, a low-pass (smoothed) and the high-pass part. The low-pass part can be split again and again until the number of bands created satisfies the computational demands. Thus, the discrete wavelet transformation can be summarized (after *j* stages) as

$$x \to \left(Gx, GHx, GH^2x, \ldots GH^{j-1}x, H^j x\right) \to \left(d_{j-1}, d_{j-2}, \ldots d_1, a_1\right) \qquad (3)$$

The output of the DWT consists of the remaining several times smoothed components, and all of the accumulated "detail" components [5].
The reconstruction procedure is similar to decomposition. The signal at every level is upsampled by two, passed through the synthesis (low-pass and high-pass) filters and then the filtered components are summed [4].

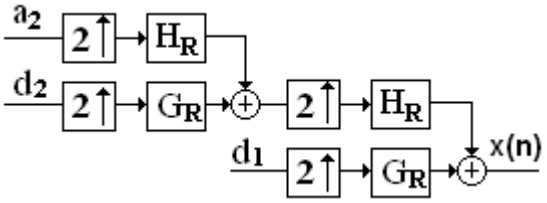


Figure 2: FIR filter structure for dyadic scale reconstruction.

The filters must satisfy certain requirements to enable perfect reconstruction from the two output signals after downsampling, and to yield an orthogonal underlying wavelet basis. To end up with a corresponding mother wavelet $\psi(t)$ having compact support, the filters (H_D,G_D,H_R,G_R) must be finite impulse response (FIR) filters [5]. All the filters are intimately related to the sequence $W = (w_n), n \in Z$ which defines the dilation (or refinement) relation

$$\frac{1}{2}\phi\left(\frac{x}{2}\right) = \sum_{n \in Z} w_n \phi(x - n). \tag{4}$$

If ϕ is compactly supported, the sequence $W = (w_n), n \in Z$ can be viewed as a filter, and from this we can define four FIR filters of length 2N organized as in *Fig. 3*. G_R and H_R are quadrature mirror filters (qmf) [3], H_D is obtained from H_R by flipping its coefficients. H_D and G_D are also quadrature mirror filters [3].

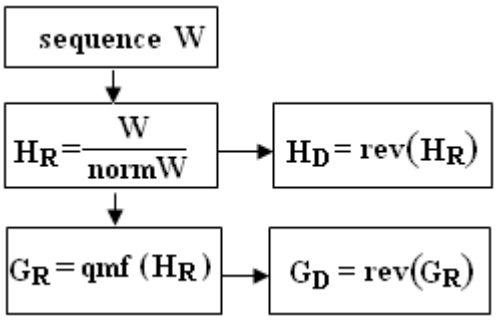


Figure 3: FIR filter synthesis according to the perfect reconstruction conditions.

3. Method and materials

The wavelet functions are obtained using an artificial neural network based function synthesizer. The basis function (wavelet) sequence is synthesized following the algorithm presented in *Fig. 4*. The main criteria for these filters are: low-pass FIR filter of *2N* length, with norm $\sqrt{2}$; the low-pass and high-pass structures are obtained from this arbitrary sequence.

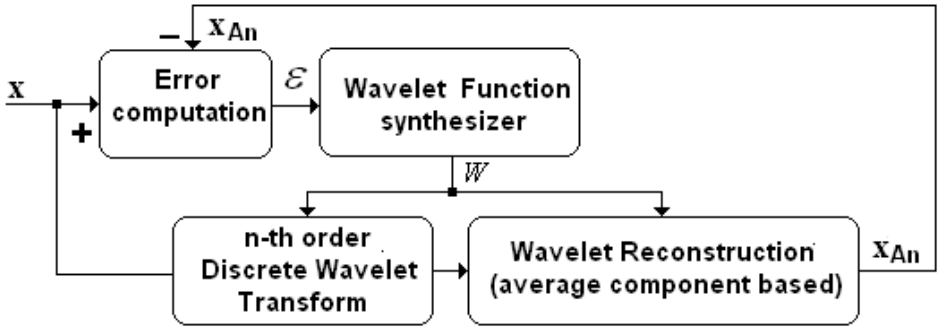


Figure 4: Function synthesizer.

The calculation of the error ε is performed comparing the reconstructed function (from first average components) with the signal. The approximation error is defined as a difference between the signal x and the function reconstructed from x_{An} average components only.

$$\varepsilon = \varepsilon_{wavelet} + \varepsilon_{norm} \tag{6}$$

$$\varepsilon_{wavelet} = \left\{X - IDWT\left[DWT\left(X\right)\right]\right\}^2 \tag{7}$$

$$\varepsilon_{norm} = \left(norm(W) - \sqrt{2}\right)^2 \tag{8}$$

The criterion-function was defined as:

$$w_i = w_i + \mu \frac{\delta\varepsilon}{\delta w_i}, \tag{9}$$

where μ is the learning rate and $\frac{\delta\varepsilon}{\delta w_i}$ is the variation of error. The used test signal is from MIT-BIH Arrhythmia Database.

4. Results

The test signal is from MIT-BIH database; we used only a short sequence and several existing wavelet functions. The resulted wavelet sequence is presented in *Fig. 4*, the analyzed signal and its average and detail components are presented in *Fig. 5*.

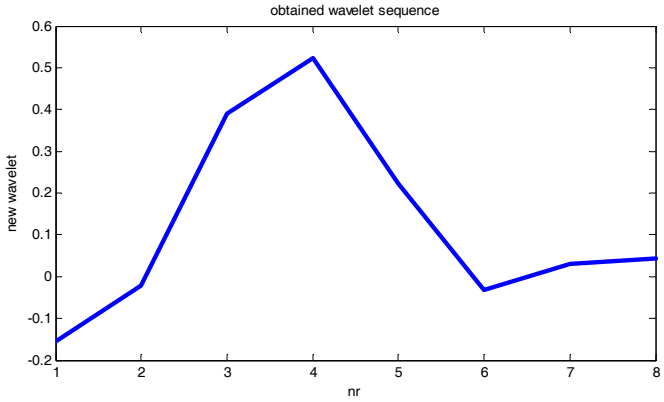


Figure 4: The resulted wavelet function.

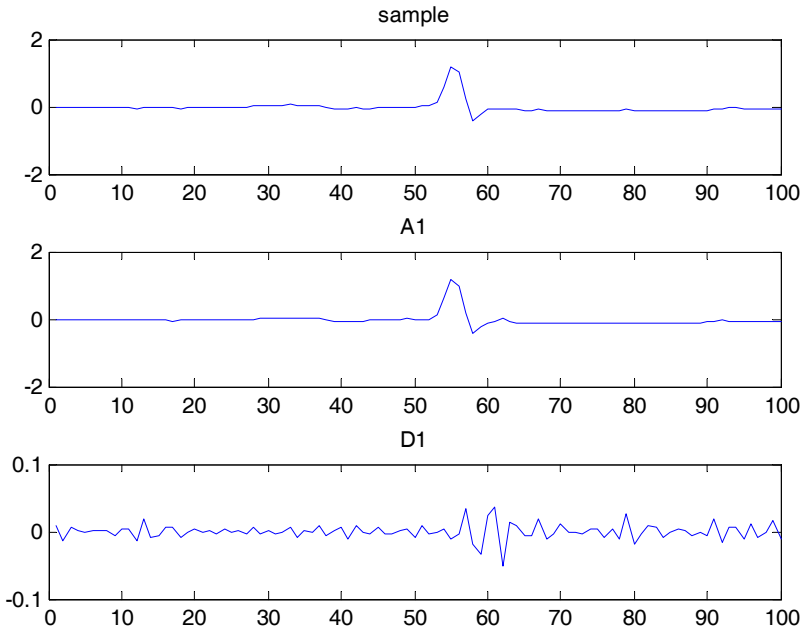


Figure 5: The analyzed signal sequence and its first average and detail components.

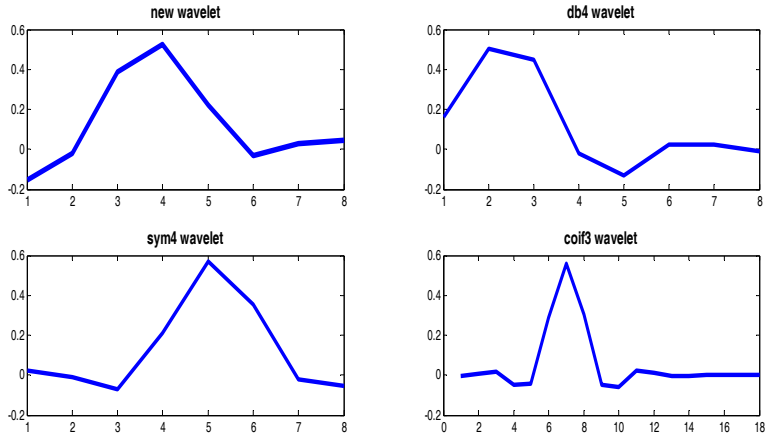


Figure 6: The obtained wavelet sequence compared to other similar functions

5. Conclusions

The main advantage of the presented method is that the analyzing discrete wavelet can be adapted to the signal. Using filter banks it's possible to obtain a discrete wavelet transform of a sequence without specifying any function. The obtained functions gave better or almost the same results in decomposition, reconstruction as the existing functions. Using adapted wavelet functions reduces the computational costs and gives a more accurate multiresolution analysis.

References

[1] Mallat, S.: A wavelet tour of signal processing Academic Press London 2001.
[2] Aldroubi, A., Unser, M.: Wavelets in Medicine and Biology. CRC Press New York 1996
[3] Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J-M.: WaveletToolbox. For Use with Matlab. User's Guide. Version 2.The MathWorks Inc 2000
[4] Coifman, R.R.: M.V Wickerhauser: Entropy-based algorithms for best basis selection IEEE Trans. on Inf. Theory, 1992, vol. 38, 2, pp. 713–718.
[5] Burrus, S. C., Gopinath, R..A., Guo, H., Introduction to wavelets and Wavelet Transforms
[6] Karel, JMH, Peeters, RLM, Westras, RL, moermans, KMS, Haddad, S.A.P., Serdijn, W.A.: Optimal wavelet design for cardiac signal processing. Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 163-174

The Mojette Transform Tool and Its Feasibility

Péter SZOBOSZLAI[1], Jan TURÁN[2],
József VÁSÁRHELYI[3], Péter SERFŐZŐ[4]

[1] Magyar Telekom, Budapest, Hungary,
e-mail: szoboszlai.peter@telekom.hu
[2] Department of Electronics and Multimedia Communications,
Technical University of Kosice, Košice, Slovak Republic,
e-mail: jan.turan@tuke.sk
[3] Department of Automation, University of Miskolc, Miskolc-Egyetemváros, Hungary,
e-mail: vajo@mazsola.iit.uni-miskolc.hu
[4] Ericsson Hungary Ltd, Budapest, Hungary, e-mail: peter.serf.z@ericsson.com

Manuscript received March 15, 2009; revised June 10, 2009.

Abstract: The Mojette Transformation Tool (MTTool) is an implementation of the Direct Mojette transform and its inverse in Net environment. In contrast with the hardware development (MoTIMoT) [1], the software development provides us both an endless possibility of different variations of the Mojette Transform in a shorter time frame and lower costs. Tests with such a tool are much easier and it is also better for demonstration and training purposes. This paper tries to outline how the MTTool could be helpful for further developments both in software and hardware development.

Keywords: Mojette Transform, MoTIMoT, MTTool, performance test, image processing, software development.

1. Introduction

The Mojette Transform (MT) originates from France where J-P. Guédon referred to an old French class of white beans, which were used to teach children computing basics of arithmetic with simple addition and subtraction. He named it after the analogy of beans and bins. Bins contain the sum of pixel values of the respective projection line [2]. There are several different variations of MT applications nowadays which are used in different areas, such as tomography [3], internet distributed data bases [4], encoding, multimedia error correction [5], or The Mojette Transform Tool (MTTool), which was created for testing purposes. Moreover, it can be used for demonstrations and training purposes as well.

Although the MTTool development has not been finished yet, we have already gained much experience with it, and we can see how it may become more helpful for further projects both in software and hardware development. So the main purpose to build such an environment is that with its help we could try to compare MT software version with the hardware one. Possible application of the SW and the HW can be a surveillance system, where the captured and transformed data is stored on different servers for security reasons. From one transformed data the recorded data cannot be restored and loosing connection to one storage server is not affecting the restoration of the requested data.

2. Mojette and Inverse Mojette Transform

Mojette Transform: The main idea behind the Mojette transformation (similarly to the Radon transformation) is to calculate a group of projections on an image block [6]. The Mojette transform (MOT) (see [7], [8] and [9]) projects the original digital 2D image:

$$F = \left\{ F\left(i, j\right); i = 1, ..., N; j = 1, ..., M \right\} \tag{1}$$

onto a set of K discrete 1D projections with:

$$M = \left\{ M_k\left(1\right); k = 1, ..., K; 1 = 1, ..., 1_K \right\}. \tag{2}$$

MOT is an exact discrete Radon transform defined for a set $S = \{(p_k, q_k), k = 1, ..., K\}$ specific projections angles:

$$M_K\left(l\right) = proj\left(p_k, q_k, b_l\right) = \sum_{(i,j)\in L} F\left(i, j\right)\delta\left(b_l - iq_k - jp_k\right), \tag{3}$$

where $proj\ (p_k, q_k, b_l)$ defines the projection lines p_k, q_k, $\delta(x)$ is the Dirac delta with the form:

$$\delta\left(x\right) = \begin{cases} 1, if\_x = 0 \\ 0, if\_x = 1 \end{cases} \tag{4}$$

and

$$L = \left\{ (i, j); b_l - iq_k - jp_k = 0 \right\} \tag{5}$$

is a digital bin in the direction θ_k and on set b_l.

So the projection operator sums up all pixels values whose centers are intersected by the discrete projection line l. The restriction of angle θ_k leads both

to a different sampling and a different number of bins in each projection (p_k, q_k). For a projection defined by θ_i, the number of bins n_i can be calculated by:

$$n_i = \left(N-1\right)\left|p_i\right| + \left(M-1\right)\left|q_i\right| + 1 \tag{6}$$

The direct MOT is depicted in *Figure 1* for a 4x4 pixel image. The set of three directions *S={(-1,2),(1,1),(0,-1)}* results in 20 bins.

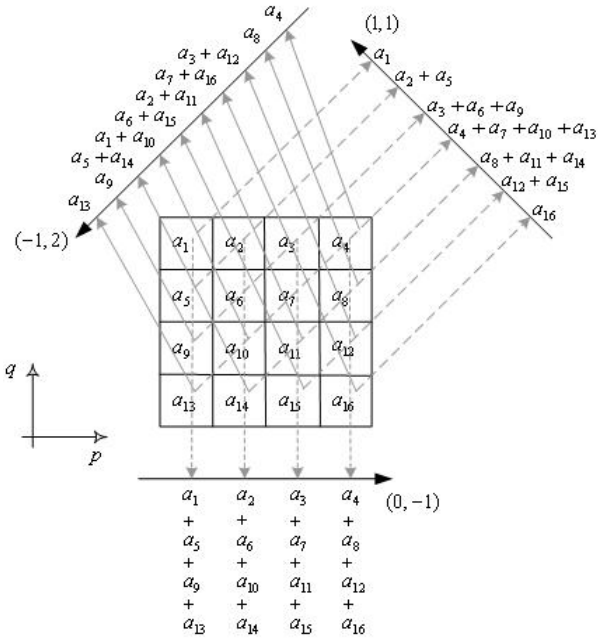


Figure 1: The set of three projections computed from a 4x4 image.

The MT can be performed by direct addition of the image pixel values in grey scale images, and for bitmaps we can add the different bitmap color table values.

Inverse Mojette Transform: The basic principle of the inverse Mojette transform is the following. We start the image reconstruction with bins corresponding to a single pixel summation. This reconstructed pixel value is then subtracted from the other projections and the process is iterated for the N^2-*1* pixels: the image is then completely decoded. In the case of a 4x4 pixel image reconstruction, if the directions of the MT sets are *S={(-1,2),(1,1),(0,-1)}*, then the minimum number of subtractions needed is 10, from the 20 bins. So should

it happen to lose some of the bins we could still reconstruct the image due to the redundancy of the MT.

3. Mojette Transform in MTTool

In MTTool the implementation of the MT was applied in three different ways. This is due to the fact that this application is still under development and the three different ways were constructed not at the same time, but in the previous years.

Table 1: MT implementation and its main differences

Nr.	Image Format	Projections	MT and Inverse MT
1	PGM	p={1,-1,3,-3}; q={quarter of the image size}	addition and subtraction
2	BMP	p={2,-2}; q={1} and p={3,-3,2}; q={1}	addition and subtraction
3	BMP	p={2,-2}; q={1} and p={3,-3,2}; q={1}	Matrix

The First Version: In the initial release one of the hardest decisions was to declare some rules, which had to be both flexible and at the same time not very complex. We had to declare the image sizes we had to work later with, and to look for a useful relationship between the picture size and the vectors we use in the MT, Inverse Mojette Transform (IMT). Considering several different file sizes, it was clear the smallest image size which can be used in real system is the 256 x 256 so, we decided to take the picture size $2^n x 2^n$, where n is equal to 8 and 9, but can be changed easily later on. So the transformable picture size are 256 x 256 and 512 x 512. In the Picture Preview we can open and display any kind of PGM or BMP file irrespective of the picture size, but some of the images are increased or decreased to fit on the screen.

Table 2: Image display in Picture Preview

Original size	Displayed size	Ratio
1600 x 1200	400 x 300	0,25
1599 x 1199	799 x 599	0,5
1024 x 768	512 x 384	0,5
Height < 1024	Height +180	Other

After checking the restrictions, the first step in the MT is to make a vector from the pixels of the image. When following a simple rule (*1*, 2^n *x* 2^n), it is easy to define the size of this vector. If *n=8,* this result in the vector (*1, 65536*), in which every line contains a pixel value from the picture. Because the PGM picture is a 256 greyscale image, a PGM file contains pixel values only from 0 to 255. In case of a BMP image, we could make it three times because of the different bitmap color table values.

In the second step we make the Mojette Transformation. The vector *p* is predefined for the four projection directions and the *q* vector has the same value in each case (quarter size of the 2^n *x* 2^n image). We generate four files for the four different projections, which are the following:

- originalfilename.pgm.moj1 (*1, q*)
- originalfilename.pgm.moj2 (*-1, q*)
- originalfilename.pgm.moj3 (*3, q*)
- originalfilename.pgm.moj4 (*-3, q*).

From the existing MT files (moj1, … moj4), we get the original PGM picture with the IMT. In this case all of the four Mojette Transformed files are needed to rebuild the original image without any errors at all. If any of the Mojette Transform files is defect or incomplete, the Inverse Mojette Transform will not give back the original image. Each of the four files contains a vector described above. The next step of the IMT is to read the first and last vectors of the third and fourth MT files and put them in their place. So we have in all four corners of the picture the valid pixel values filled up. See step 1, 2, 3 and 4 on the following figure:

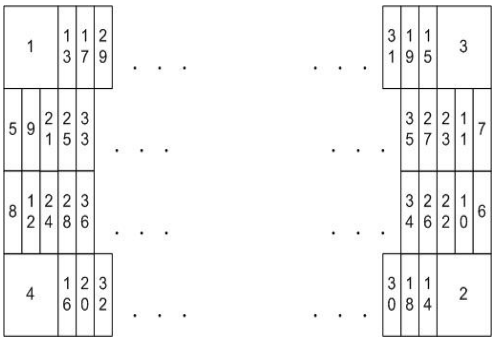


Figure 2: First 30 steps of the IMT.

After recreating the pixel values, we only need to add the new header for the file and the restoration of the original image is already performed.

The Second and Third Version: These solutions differ from the previous one in such a way that these are applied on BMP images and in these cases we perform the MT and IMT on the three different bitmap color tables. We use the same algorithm for the three different color maps and collecting the bins into 3 separate files which differ in their extensions and of course in their content. On the bitmap images we use the directions $S_1=\{(2,1),(-2,1)\}$ and $S_2=\{(3,1),(-3,1),(2,1)\}$ for the block sizes 4 and 8. Although the MT is also prepared for the block size 16 and 32, the implementation of the IMT isn't done yet. In the second version, we use simple addition and subtraction – different from the one mentioned in the first version –, since here we have block sizes 4 and 8 and there we perform the MT and IMT on the whole image at once and not step by step. In the third version, instead of addition and subtraction, we use matrices for the MT and IMT on the above mentioned block sizes. The MT with matrices is implemented in the following way, where b_i is the bin resulted from the following equation:

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \\ b_{10} \\ b_{15} \\ b_{16} \\ b_{17} \\ b_{18} \\ b_{19} \\ b_{20} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \\ a_{11} \\ a_{12} \\ a_{13} \\ a_{14} \\ a_{15} \\ a_{16} \end{bmatrix} = \begin{bmatrix} 10 \\ 123 \\ 37 \\ 137 \\ 254 \\ 319 \\ 433 \\ 68 \\ 6 \\ 234 \\ 125 \\ 267 \\ 312 \\ 8 \\ 45 \\ 178 \end{bmatrix} \quad (6)$$

The inverse matrix for the previous example (for the 4x4 matrix size) is implemented as it is shown in the next equation, where a_i stands for the original values of the matrix:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \\ a_{11} \\ a_{12} \\ a_{13} \\ a_{14} \\ a_{15} \\ a_{16} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \\ b_{10} \\ b_{15} \\ b_{16} \\ b_{17} \\ b_{18} \\ b_{19} \\ b_{20} \end{bmatrix} = \begin{bmatrix} 10 \\ 123 \\ 25 \\ 35 \\ 12 \\ 102 \\ 252 \\ 241 \\ 2 \\ 78 \\ 255 \\ 23 \\ 178 \\ 45 \\ 6 \\ 234 \end{bmatrix} \quad (7)$$

4. MoTIMoT implementation

The ‘MoTIMoT’ co-processor denotes the hardware implementation of the direct MT and IMT as co-processing elements of an embedded processor. The advantage in using FPGAs for this is the flexibility of the development tools, the hardware-software co-design solutions, and the compact hardware in the loop simulation. The embedded reconfigurable hardware is based on Xilinx ML310 board [10].

Starting from a 256x256 pixel size greyscale image, this requires 64KB memory. In order to process all the projection lines in parallel (in the case of the MT), one needs as many 64KB sized memory blocks (i.e. Block RAM) as many projection lines we have for this image size. The bin vectors are stored in the external memory in a so-called ‘MT memory file’. The division in slices of the original image is motivated by the fact that this can be corrupted during the transmission of the MT file. The image can not be reconstructed without the damaged area from the corrupted MT file. While applying the MT to slices, the effect of the MT corruption is diminished. Similarly, the memory necessary to calculate the MT and IMT is smaller in the case of slices. The whole image process would result in a need for reconfiguration in order to calculate all the projections, because 256KB are needed to load the 256x256 image in the embedded memory only for 4 projection lines (the XC2VP30 has 1.7Mb Block RAM ≈ 212KB). For this reason the 256x256 pixel image is divided in 4 slices (128x128).

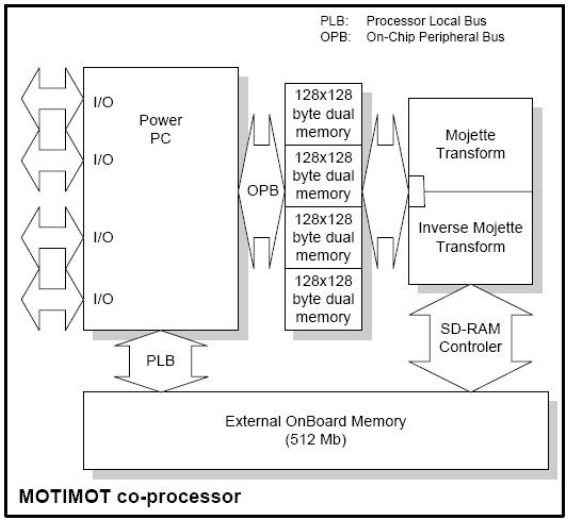


Figure 3: MoTIMoT co-Processor Block Scheme [1].

The main parts of the image processing systems are: the PowerPC as the main processing unit, the MoT unit and the IMoT unit. Both MoT and IMoT processors are connected to the PowerPC via the internal PLB bus, because their work runs under the main processor control.

While processing the IMT (for the same image size and projection lines) one needs to read the MT memory file from the external memory and to reconstruct the image. The bin size means the maximum pixel numbers contained by a bin and also defines the number of bits needed for the unary bins of the unary MT file. Bins containing only one pixel value are placed to their corresponding position during the back projection (IMT). These pixel values contained by other bins (in other projections) as well and which bin values have to be decreased with the current pixel value. To calculate the positions of the bins in the projections and to calculate the correspondence in the image of a single pixel bin we need the unary image. Thus when the value of a single pixel bin is substituted and the other bin values are decreased, the changes have to be validated as well in the unary MT file. The unary MT file contains unary bins. These bins contain not only the current pixel number included in the bin in the MT file, but the corresponding position in the image of these pixels too.

5. Experiments and results with MTTool and MoTIMoT

MTTool: We can decrease the size of any vectors which are created from the projections of MT with the built in ZIP and Huffman coding opportunities. The Huffman lossless encoding and decoding algorithm was chosen due to its binary block encoding attribute and not because of its compression capability. Good data compression can be achieved with Zip and Unzip, which are also implemented. The possibility of time measuring with simple tools, such as labels or easily generated text files which include the test results, can give us a good insight into the MT and IMT. From these results we can estimate and predict the consumed time on hardware implementation and its cost as well.

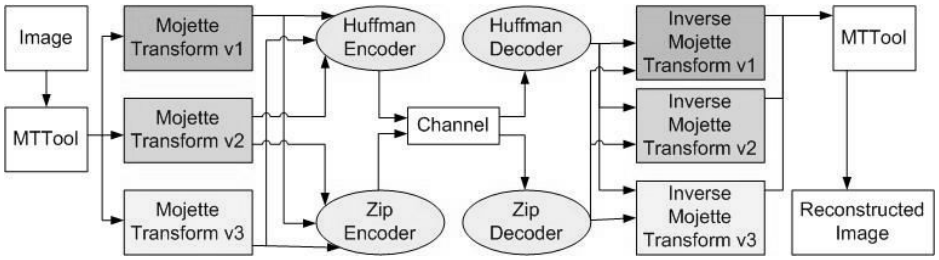


Figure 4: Logical system architecture of the MTTool.

The time measurement was applied on three different images with three different image sizes and with three different periods. The images were black and white PGM files with pixel values of 0 and 255 and the LENA.PGM. The first test ran only once, after which the second test ran for 6 times in a row, and the last test ran 24 times. Each test was performed with sizes of 16x16, 32x32 and 512x512. The results of the two smallest image sizes are nearly identical, and the results were nearly always under 20 milliseconds for MT and IMT, but we could see the following difference regarding the 512x512 image size:

Table 3: Test result of the MT and IMT with the first version

IMAGE	Black (512x512)		White (512x512)		Lena (512x512)	
	Minute: Second: Millisecond	MT and IMT in Millisecond	Minute: Second: Millisecond	MT and IMT in Millisecond	Minute: Second: Millisecond	MT and IMT in Millisecond
MT start	57:14:277		3:45:510		21:36:79	
MT end IMT start	57:15:439	1162	3:47:403	1893	21:37:762	1683
IMT end	57:15:910	471	3:47:964	561	21:38:303	541
MT start	57:22:259		4:0:822		21:49:749	
MT end IMT start	57:23:411	1152	4:2:555	1733	21:51:391	1642
IMT end	57:23:891	480	4:3:105	550	21:51:932	541

From this table we can see that the difference between black and white images is more than 50 percent, when it comes to the MT, and only 20 percent when we apply the IMT on the Mojette files. For a real time video surveillance application which should capture at least 25 image per second (PAL) this result is not enough.

MOTIMOT: The simulations were made on PC hardware environment using a portable greymap 256x256 image (Lena) without transmission and no bit-corrupted errors, just as in the MTTool. The simulation proved the correctness of the implemented algorithms and the functionality of the proposed hardware. In both implementation of the MT and IMT the images were restored with only small differences. The original creation date of the image is replaced with the date when the IMT was performed. All the header information is cut and isn't restored later on. Restoration of the image includes only the pixel values of the original image. We also attach a new automatically generated header to these pixels, so the restored image pixel values are exactly the same as the pixel values in the original image. Therefore any information included in the image itself such as watermark, time and date etc. can be restored later on easily.

6. Conclusion

The paper outlines the different ways, how the Mojette Transform is currently implemented in the MTTool and also gives an insight how the hardware implementation of Mojette and inverse transformation in the embedded system using FPGA has been done. The original contribution is the calculation of the dimension of Mojette memory file, the definition and analysis of the hardware structure as a whole with the simulation results. Future work is needed both in the software and hardware versions. In the software version (MTTool) more tests should be performed to get more accurate results, and by comparing them to the results of the hardware, we should find an optimal way to perform the Mojette Transform. In the hardware version (MoTIMoT), finalizing the implementation of the co-processors as a whole with run-time reconfiguration is needed.

Acknowledgements

The authors gratefully acknowledge the donations of Xilinx Inc. and Celoxica Inc., which made it possible to start this research.

Thanks for Ferenc Nagy who offered the necessary space and time for our work.

References

[1] Serfőző, P., Vásárhelyi, J., "Development work of a Mojette transform based hardware codec for distributed database systems", in *Proceedings of 8th Interational Caroathian Control Conference ICCC2007, Strebske Pleso, Slovakia,* 2007 May 24-27, pp. 631-635.

[2] Guédon, J.-P., Normand, N., "The Mojette transform: The first ten years", in *Proceedings of DGCI 2005*, LNCS 3429, 2005, pp. 79-91.

[3] Guédon, J.-P., Normand, N., "Spline Mojette transform application in tomography and communication", in *EUSIPCO*, Sep. 2002.

[4] Guédon, J.-P., Parrein, B., Normand, N., "Internet distributed image databases", *Int. Comp. Aided Eng.*, Vol. 8, pp. 205–214, 2001.

[5] Parrein, B., Normand, N., Guédon, J.-P., "Multimedia forward error correcting codes for wireless LAN", *Annals of Telecommunications (3-4),* pp. 448-463, March-April, 2003.

[6] Normand, N., Guédon, J.-P., "La transformee Mojette: une representation recordante pour l'image", *Comptes Rendus Academie des Sciences de Paris, Theoretical Comp. SCI. Section*, 1998, pp. 124–127.

[7] Katz, M., "Questions of uniqueness and resolution in reconstruction from projections", Springer Verlag, Berlin, 1977.

[8] Autrusseau, F., Guédon, J.-P., "Image watermarking for copyright protection and data hiding via the Mojette transform", in *Proceedings of SPIE*, Vol. 4675, 2002, pp. 378–386.
[9] Turán, J., Ovsenik, L., Benca, M., Turán, J. Jr., "Implementation of CT and IHT processors for invariant object recognition system", *Radioengineering*, Vol. 13, No. 4, pp. 65-71, Dec. 2004.
[10] Xilinx, ML310 User Guide, pp. 73, http://xilinx.com.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 175-186

Assessment of Building Classifiers for Face Detection

Szidónia LEFKOVITS

Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tîrgu Mureş,
e-mail: lefko_cd@yahoo.com

Manuscript received March 15, 2009; revised May 26, 2009.

Abstract: Nowadays an increasing number of applications require fast and reliable object detection systems. The most efficient system presented in the publications of Viola-Jones object detection framework and the open source implementation of their ideas creates a solid baseline for future detectors. This approach has been extensively used in Computer Vision research, particularly for detecting faces and facial features. The OpenCV community shares a collection of such classifiers. The analyses of such public classifiers define the basis of future work in the object detection domain. In this paper the performance of cascade classifiers is analyzed. A series of ambiguities concerning the teaching process is also presented together with a few proposals how to solve them. It has been tried to discover and overtake the limitations of the OpenCV implementation and use it to create the author's own classifier. Finally, an algorithm is proposed to get 10^{-5} false alarm rate.

Keywords: face detection, AdaBoost, Haar features, training data set, supervised learning.

1. Introduction

Face detection and recognition has become an increasingly researched area. The Viola and Jones method for face detection [1], [2], [3], [4] is an especially successful method, as it has a very low false positive rate. It can detect faces in real time and yet is very flexible in the sense that it can be trained for different level of computational complexity, speed and detection rate suitable for specific applications. The implementation offered by Intel in the OpenCV application made this algorithm more attractive. It is highly desirable to use this versatile method for anyone who might want to make research in this area. The OpenCV application has a poor tutorial in reference to the creation of classifiers. To exceed this, a lot of authors published their own experience on the internet [5], [6]. One can find a lot of comments, experiences and useful functionalities in order to create the training data set.

2. The facial detection system

The used detection system is a combination of geometrically-based and image-based methods. It is geometrical, because it uses general features of human faces: position of particular features among which the eyes, the nose and the mouth. These features are the selected Haar functions, but the selection is based on a statistical learning, which uses a training data set to build the face model.

2.1. The AdaBoost Algorithm

The AdaBoost algorithm is proposed by Freund and Shapire [7]. It constructs an ensemble of classifiers and uses a voting mechanism for the classification. In a wide variety of classification problems, their weighting scheme and final classifier merge have proven to be an efficient method for reducing bias and variance, and improving misclassification rates.

The idea of boosting is to use the weak classifier to form a highly accurate prediction rule by calling the weak classifier repeatedly on different distributions over the training examples. Initially, all the weights are set equally, but each round the weights of incorrectly classified examples are increased so that the images, which were poorly predicted by the previous classifier, will receive greater weight on the next iteration.

The most important theoretical propriety of AdaBoost concerns in its ability to reduce the training error. The AdaBoost converts a set of weak classifiers into a strong learning algorithm, which can generate an arbitrarily low error rate.

2.2. Haar functions

Many descriptive features could be used to train a classifier by boosting. In face detection, the feature based method seems to be quite efficient. The Haar wavelets are naturally set basis functions, which compute the difference of intensity in neighboring regions [3]. The value of the Haar function is the measure of likeliness between the specified region of an image and the definition of the Haar function.

Significant is the very fast evaluation of this function by using a new image representation called Integral Image. Another important property is the fact that the value of a Haar function is the same if the picture is reduced by a factor, or the Haar function is increased by the same factor. This property decreases more the evaluation time.

A corresponding weak classifier can be built from each Haar function. For this we need to determine the optimal threshold for each function. This

optimum is reached when the number of misclassified examples is the lowest weak classifier $h_j(x)$ consists of a feature $f_j(x)$, of a threshold value θj and a parity p_j to indicate the direction of inequality:

$$h_j(x) = \begin{cases} 1 \ ,if & p_j \cdot f_j(x) > p_j \cdot \theta_j \quad x - face - image \\ 0 \ ,if & p_j \cdot f_j(x) < p_j \cdot \theta \ x - non - face - image \end{cases} \tag{1}$$

2.2. The monolithic classifier

The monolithic classifier for face detection is built by using the AdaBoost algorithm and the weak classifier based on Haar functions [3]. We also need a set of training examples consisting of all significant human faces (5,000) and various non face images (10,000) for the learning process. The first step is to build the classifier-image table, which contains the value of each classifier for every image of the training set. The table is quite large, and we have to use each value of it to compute the weighted error for each round and for each weak classifier. The minimum error determines the weak classifier for one round and the weight for it in the final classifier. Thus, this error modifies the weight of each picture. While the weight of the misclassified pictures increases, the weight of the well-classified ones decreases. The algorithm cycle stops when one of the learning conditions is satisfied; these conditions can be:

- the maximum number of T cycle.
- the error condition for the weak classifier
- the desired performances are reached.

2.3. The cascade classifier

The response time of a monolithic classifier is the same for every image. In order to reduce this time, it is necessary to evaluate one part of the images with few weak classifiers and evaluate only complex images with the whole classifier. The idea is to reject rapidly the most part of negative images. The cascade design process is driven by a set of detection performances [2]. If each stage classifier is taught for low performances (f<0.5, false detection rate/stage and d>0.999, hit rate/stage), then the whole cascade will have the same performances as a monolithic classifier, but 10 times faster.

Each stage is taught with the remaining images from previous stage. The stop condition of the learning process is given by the reached performance. That is why it is necessary to measure this performance with a validation set of images.

If we build a classifier with 20 stages, each with the above performances, then we will obtain both the global false positive error rate $F < f^{20} = 0.5^{20} = 9.6 \cdot 10^{-7}$, and the error detection rate $D \geq (1-d)^{20} = (1-0.001)^{20} = 0.98$ (*Fig. 1*).

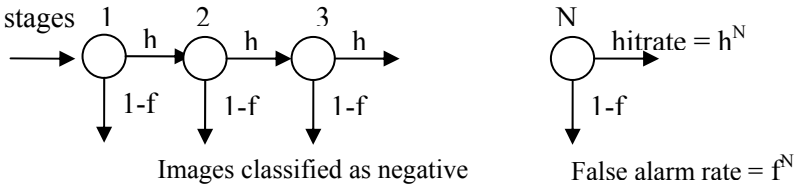


Figure 1: Cascade classifier with N stages [3].

To obtain the given performances, the global threshold has to be modified until the current cascade has a detection rate of at least d_k. This task is reachable by decreasing the threshold; then the detection rate increases, and the false detection rate increases, too. If this parameter is too big and does not fit the learning conditions, the algorithm takes one other weak classifier in the current stage.

The overall training process involves two types of tradeoffs. In most cases, classifiers with more features will achieve higher detection rates and lower false positive rates. At the same time classifiers with more features require more time to compute.

3. Building a classifier

This section presents the different results obtained by the face detectors that have been developed. A few results of various authors will be discussed and interpreted and then the conclusions drawn regarding our experimental results.

Only a few classifiers can be found which work with this principle [8] in the public domain. The OpenCV community shares their collection of classifiers (see *Table. 1*). Only a few authors, disclosed their classifiers, but none of them published their training sets and training methodology. In this domain there are several unsuccessful attempts to train classifiers with the presented algorithm.

3.1 Building the training set

We propose to create a classifier for face detection with the cvHaarTraining program. It seems to be easy enough to follow the procedures described in

OpenCV tutorial. In fact, there are a lot of problems to solve in order to obtain an efficient classifier.

There are a lot of basic questions with the training set.

1. What is the best input pattern size?
2. How to crop the face images?
3. What is the background image size?
4. Which are the significant images?
5. What is the necessary training set size?
6. How to teach the classifier to obtain 5×10^{-6} false alarm rate?

These questions are further debated.

1. The input size of the image determines the number of used features in the learning process. For a pattern of 24×24 pixels size, there are 84848 (BASE) features in the basic set and 111360 (CORE) in the extended set, and 138694(ALL) the entire set features to evaluate.

Larger images are more detailed and need more memory and more feature to evaluate. That means larger feature-image table. Experimental analysis can conclude the size of image pattern depending on each application. According to the experiments [3], the images' pattern size 24×24 is the best in face detection, because it has the lowest false alarm rate at the same hit rate.

According to our experimental results, the optimal pattern size is 18×24. We concluded that the optimal pattern size depends on the variety of the data base used for training. Other approaches of face detectors have obtained other optimal dimensions for the training image pattern.

2. There are a lot of possibilities to crop face images.

a. cropping only the significant part of the images

We can define the lower and upper boundary of the face by adding the distance between the mouth and the nose to the height of the eye-line, and subtract from the height of mouth-line the same distance. We define the left-right margins by adding the distance between the eyes to the right margin of the right eye and subtract it from the left margin of the left eye [5].

b. cropping images that include extra visual information, such as contours of the chin and cheeks and the hair line. It seems that additional information in larger sub-windows can be used to reject non-faces earlier in the detection cascades [3]. The second case included additional information of the faces. The evaluable features are also more numerous, which make the detection process more accurate.

3. The size of the background image does not seem to be so important, it is never explicitly specified. It can be taken to be the same size as the positive pattern size, namely 24×24 pixels.

The OpenCV HaarTraining program can read background images of any size and it crops from this various number of backgrounds by shifting the cropping

window through the whole image by a step of width/2 and height/2. It takes the specified number of backgrounds from the same given images each time. Smaller images with different characteristics are recommended to crop as backgrounds. There are two possibilities to create your own classifier with OpenCV. One way is to create the whole classifier containing more stages at once. Probably, in this case the background images are filtered in each step and only the false alarm images are used in the further stages. The other way is to create each stage separately. One drawback of the OpenCV training process is the building of the background set. In this case, one should choose backgrounds randomly, in an other way than OpenCV, which uses the same image table for all the stages created separately. In their application, the background training set contains the same aggregation of images in the same order for each stage. This is the reason we propose using different backgrounds in each stage.

4. The image-based learning method needs a number of significant positive and negative images. At this step, one should have a methodology to choose only the significant images. Practice proves that an amount of 5,000 face images would be enough, but the difficult question is the number of backgrounds. The positive images are usually cropped manually and each is verified by a human operator. To increase the insensibility of the built classifier, many images can be generated from one image by applying distortion functions (randomly little rotation, translation and resizing). This little variation can be applied to the whole image set in order to multiply the number of used images. The images of our own database were collected from public marked face databases FERET and Yale (about 1,800 pictures), and these were completed by a self marked cropped studio images (about 1,100).

The background images are generated automatically in general randomly from a set of images. The backgrounds were downloaded randomly from the internet, and besides we used the Corel Draw image set. To increase the variety and cardinality of them, several random operations were performed: rotation, translation, resizing. Because of the large variety of background the selection of significant patterns is a difficult task. One idea is to take the images filtered by the existing stages of the classifier as significant background pattern.

5. The training set size determines the learning time. Leinhart proposed a training set with 5,000 positive and 3,000 negative images [1]. Viola and Jones built their classifier with 4,916 faces and 10,000 non-faces selected randomly from a set of 9,500 images which did not contain faces [3].

Our first experience had 3,000 positive images and 27,000 negative ones. The scanning Windows sizes were 18x18, which contained 33,000 Haar features from the basic set. The dimension of feature-image table is 10^9, and used 1GB memory. The learning program reported 30s for the selection of one Haar feature. The computer we used was an Intel Core 2 Duo, CPU E460

frequency 2.4GHz, Gigabyte motherboard FSB 1333MHz, Dual Channel DDR2 800 and 2GB of RAM. It needs about 30,000 seconds, more than 8 hours, in this the selection of 1,000 features.

The functionality principle of the detector is to scan the image at multiple scales and locations. Good results are obtained by using a scale factor of 1.3 and a step size of s=1 pixel. With this scanning parameters, one image of 320×240 pixels size has 130,000 sub-windows of 24×24 pixels. The processing of the high number of sub-windows, suppose a false alarm rate lower than 10^{-5}.

6. The question is how to teach the classifier to obtain a 5×10^{-6} false alarm rate. In order to achieve this performance, more million different background pictures are needed. [4]

If we use 2 million background pictures, the processing time of one feature selection increases dramatically. The feature-image table needs a huge amount of memory space which exceeds the usable RAM memories. This limitation can be solved by the usage of virtual memory created on HDD. Access time to virtual memory rises computation time. It will become 100 times longer, so it will take at least one month to build a classifier containing 1,000 features.

The solution is probably behind a methodology of choosing the background images. I propose to teach each stage of classifiers the same amount of 5,000 positive images and a number of 10,000 background images filtered by the previous stages. A simple program is needed, which randomly crops background patterns from a specified set of images. The background patterns are taken for each step from a different set of given images, thus the needed number of pictures will increase each time, because one needs 10,000 remaining images after the filtration by previous stages. Supposedly, this is a way to get more and more specialized stages. The process ends when we cannot get more significant images or the time of choosing background patterns increases over a given limit (Fig. 2).

The inner block cycle executes the selection of background images until the *nmax* cycle limit is reached, namely the desired number of non-face images. The outer block represents the training process with the face and previously chosen non-face images. This cycle ends if we achieve a given number of stages or the set of backgrounds is not sufficient any more.

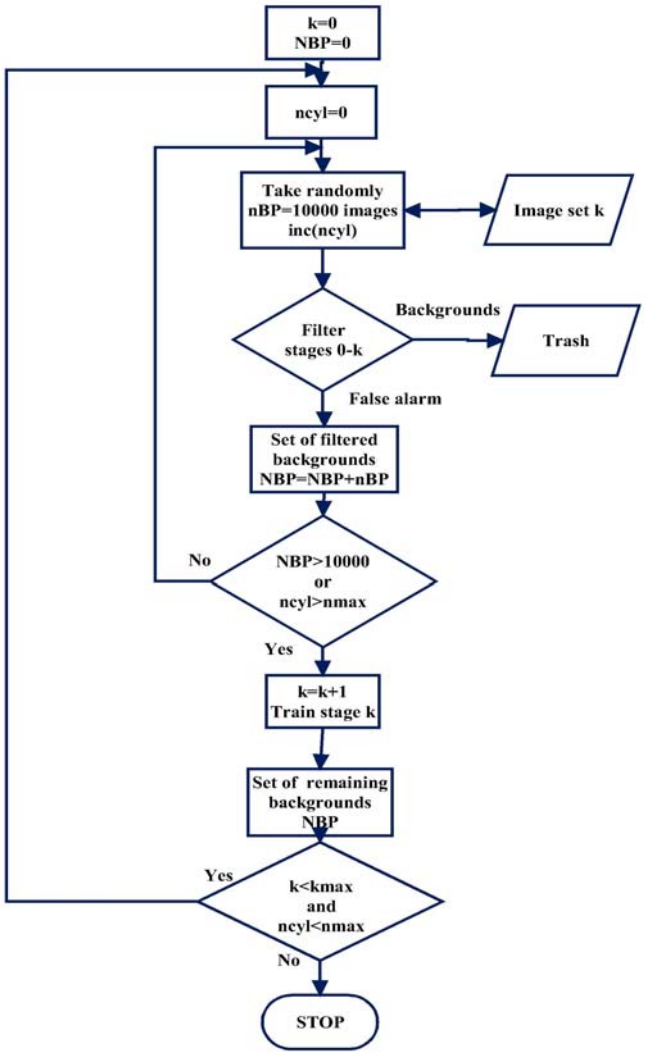


Figure 2: Significant background generator.

3.2 Performance analysis – Reported experiments

The classifier structure had to be deduced from the available public classifiers stored in xml files. Owing to this, a cascade classifier consists of several stages and a stage threshold. On one stage, the weak classifiers have to be cumulated according to the set hit and false detection rate. Thus on the first stages only few features have to be used with low error rate and on the further stages more, less efficient classifiers should be combined.

According to Viola and Jones's writing:

"Training time for entire 32 layers was on the order of weeks on a single 466MHz AlphaStation XP800. During this laborious training process several improvements to the learning algorithm were discovered. This improvements which will be described elsewhere, yield to 100 fold decrease in training time."[3] These improvements were never published and remain the secret of the learning process.

The following can be concluded by analyzing the published classifiers: the number of used stages of a classifier varies between 16 and 46, the mean value is about 22-23 stages. The first stages contain 2 - 10 features, whereas the last stages contain 100-200 features. It can be asserted that a good classifier should have more than 1,000 features (see Table 1). The table contains the measured parameters of public classifiers FD (frontal default), FA1 (frontal alternate 1), FAT (frontal alternate tree), FA2 (frontal alternate 2) and of my own classifiers (Class_04, Class_05, Class_06).

Table 1: Measured performances of classifiers

	Detection	Missed	False	Stages	Features	DetTime(s)
FD	344	24	192	25	2913	16.77
FA1	337	31	106	22	2135	20.47
FAT	325	43	61	47	8468	18.99
FA2	344	24	143	20	1047	17.73
Class_04	293	75	666	6	677	17.53
Class_05	329	39	1190	12	1632	26.58
Class_06	279	90	150	16	1319	17.87

Most authors created their classifiers and published their results on face detection in tables containing the performance parameters, usually a representation of the detection rate and the amount of false detections. A table, that is, one measurement, represents only a single point on the ROC (Receiver Operating Characteristic) curve. In order to be able to compare the detectors, we

need the ROC that represents the variation of the detection rate depending on the number of false alarms [8]. The ROC curves would be very helpful if the publishers appended the used database and the measure methodology. In their absence, we took some measurements on each stage, and based on these results we draw up the ROC curve for classifiers. There are two types of detectors: the first one has a good (90%) hit rate, which implies a huge amount of false detections. The second has much less false alarms, but the detection rate and detection time also decreases.

The authors found their optimum between these contradictory requirements. If each stage is taught separately in the learning process, then the threshold of each stage can be modified by the required performance values. Referring to this, we could observe the difference between theory and OpenCV implementation. OpenCV implementation only uses the training data set for testing performances. But the theoretical algorithm requires the result of the performance on the test data set and, accordingly, modifies the stage threshold. Depending on the result of the modified stage threshold, one decides to continue the learning process until the required parameters are reached.

In order to evaluate and compare detectors, we used the performance evaluation of OpenCV, which resulted in as follows.

- A table containing the tested pictures and the number of hits and false alarms. The last row is the sum of the results, out of which we can calculate the necessary percentage for the ROC.
- A second table contains the points of the ROC (not suitable for comparing classifiers).

The face detected regions are directly marked on tested images for visual performance analysis.

One needs a test data set in order to test detector performance. Fortunately, this is available on the CMU's (Carnegie Mellon University) internet sites [9]. The description of data set does not correspond to the request of OpenCV program; consequently, we modified it according to the face positions. This set contains 105 images with 368 faces. The performance results are given in Table 1.

Based on the fact that the detection rate of the next stage is greater or equal to the previous stage rate and only the false detection rate decreases (i.e. every stage corrects the number of false images), we propose to build our own stage by stage measured ROC curve for classifiers comparison (*Table 2* and *Fig. 3*). We consider that this curve is more suitable for the comparison of classifiers.

We can conclude from the mentioned tables (*Table 2*), that our classifier presents a lower false detection rate, beginning form the earlier stages. This result is due to the proposed algorithm.

Table: 2 Stage by stage measurements for FA2 and Class_05

stage no.	10	11	12	13	14	15	16	17	18	19	20
hit rate	91,84	94,29	96,1	95,9	96,1	95,6	94,5	94	93,4	93,75	94,3
no.false det	4458	3412	2805	2157	1685	1199	832	596	346	219	143

stage no.	6	7	8	9	10	11	12	13	14	15	16
hit rate	85	89,4	89,9	90,4	88,85	87,8	85,32	83,96	80,16	75,8	75,5
no.false det	4661	3588	2646	2230	1722	1363	1188	846	390	168	150

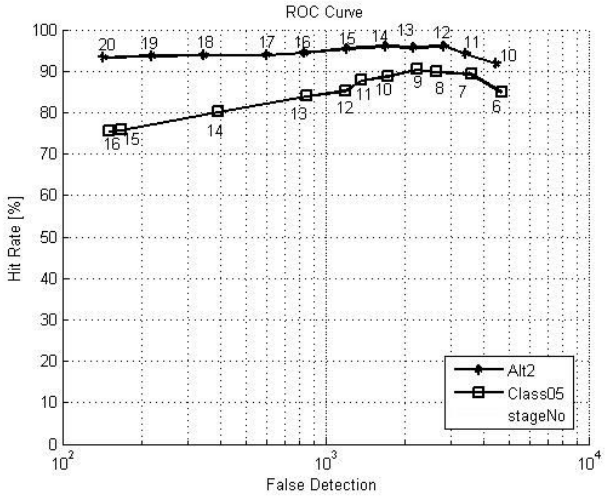


Figure 3: Stage-by-stage ROC curve.

Theoretically, the ROC has to be a decreasing curve. In the early stages the curve increases because of the numerous false alarms. So this modifies the center of the resulting detected object.

But it is evident that the detection rate is less than the value of the known best classifiers. The maximum hit rate (95%, respectively 90%) shows the properness of the face database used. This is due to the prepared pictures which do not contain sufficient various faces. They predominantly present young European people without beard or moustache, and very few of them wear glasses. Thus, without these lacks, the detection rate would be also over 90%, closer to the best known results.

In conclusion, today's known most efficient classifiers for frontal face detection are those of OpenCV source, which were trained by Leinhart, Kuranov and Pisarevsky [1].

4. Conclusion

In several cases, under more natural conditions, our classifier presents fewer false alarms at the same detection rate than the known ones. It is necessary to enlarge the face database and, simultaneously, create an algorithm for detection of significant faces, in order to eliminate the most part of the previously supervised human-classification. The efficiency of the detector depends on the training data set and the used methodology, but this remains the secret of the authors. The building of the training data set is very laborious and drudgery. The inconveniences of the OpenCV program can be avoided by the ability and knowledge of the user.

References

[1] Leinhart, R., Kuranov, A., Pisarevsky, V., "Empirical analysis of detection cascades of boosted classifiers for rapid object detection", *25th DAGM Pattern Recognition Symposium*, 2003.

[2] Leinhart, R., Liang, L., Kuranov, A., "A detector tree of boosted classifiers for real-time object detection and tracking", in *Proceedings of International Conference on Multimedia and Expo (ICME'03)*, 2003, Vol. 2, pp. 277-280.

[3] Viola, P., Jones, M., "Robust real-time object detection", *International Journal of Computer Vision*, Vol. 57, No. 2, pp. 137-154, 2004.

[4] Viola**,** P., Jones, M., "Fast multi-view face detection", *Mitsubishi Electric Research Laboratories*, Cambridge, Technical Report TR2003-096, July 15, 2003.

[5] Naotoshi Seo "Tutorial: OpenCV haartraining" http://note.sonots.com/SciSoftware/haartraining.html.

[6] Adolf, F. "How to build a cascade of boosted classifiers based on Haar-like features". http://robotik.inflomatik.info/other/opencv/OpenCV_ObjectDetection_HowTo.pdf

[7] Freund, Y. Y., Schapire, R. E., "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, Vol. 55, pp. 119-139, Art. No. SS971504, 1997.

[8] Castrillón-Santana, M., Déniz-Suárez, O., Antón-Canalís, L., and Lorenzo-Navarro, J., "Face and facial feature detection evaluation - performance evaluation of public domain Haar detectors for face and facial feature detection", in *Proceedings of 3rd International Conference on Computer Vision Theory and Applications (VISAPP'2008)*, 2008, pp. 167-172.

[9] CMU/VACS image data base: http://www.cs.cmu.edu/~cil/v-images.html.

[10] Huang, C., Wu, B., Ai, H., and Lao, S., "Omni-directional face detection based on real AdaBoost", in *Proceedings of International Conference on Image Processing*, 2004, Vol. 1, pp. 593-596.

[11] http://www.intel.com/research/mrl/research/opencv.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 187-197

Text Conditioning and Statistical Language Modeling Aspects for Romanian Language

József DOMOKOS[1,2], Gavril TODEREAN[2]

[1] Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tîrgu Mureș, Romania,
e-mail: domi@ms.sapientia.ro
[2] Department of Communications, Faculty of Electronics,
Telecommunications and Information Technology,
Technical University of Cluj-Napoca, Cluj-Napoca, Romania,
e-mail: toderean@pro3soft.ro

Manuscript received March 30, 2009; revised June 30, 2009.

Abstract: In this paper we present a synthesis of the theoretical fundamentals and some practical aspects of statistical (n-gram) language modeling which is a main part of a large vocabulary statistical speech recognition system. There are presented the unigram, bigram and trigram language models as well as the add one, Witten-Bell and Good-Turing estimator based Katz back-off smoothing algorithms. The perplexity measure of a language model used for evaluation is also described.
The practical experiments were made on Romanian Constitution corpus. Text normalization steps before the language model generation are also presented. The results are ARPA-MIT format language models for Romanian language. The models were tested and compared using perplexity measure.
Finally some conclusions are drawn based on the experimental results.

Keywords: Romanian statistical language modeling, natural language processing, text conditioning, ARPA-MIT language model format, n-gram language model, smoothing, perplexity.

1. Introduction

Statistical speech recognition is based on Hidden Markov Models (HMMs). Such a system, depicted in *Fig. 1*, is built using multiple chained HMMs for acoustic modeling and language modeling [1], [2], [3].

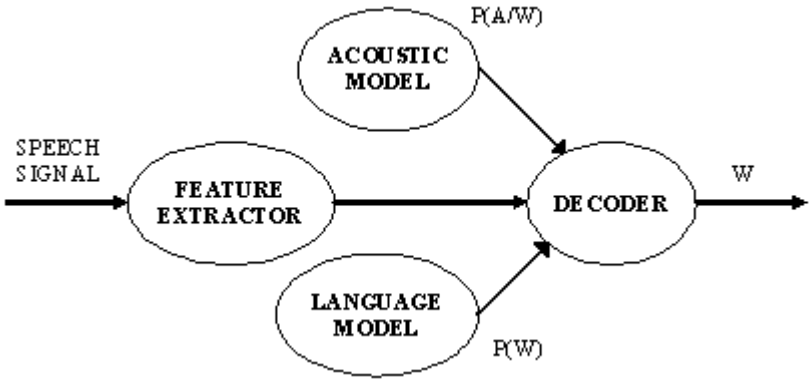


Figure 1: Statistical speech recognition system architecture.

The system presented in *Figure 1*, can be described mathematically as follows [1][4]: we have a set of acoustic vectors $A = \{a_1, a_2, ..., a_n\}$ and we are searching the most probable word sequence: $W^* = \{w_1, w_2, ..., w_m\}$.

$$W^* = \arg\max_w \{P(W \mid A)\} \tag{1}$$

Using the Bayes formula, we can transcribe (1) as follows:

$$W^* = \arg\max_w \left\{ \frac{P(A \mid W) \cdot P(W)}{P(A)} \right\} \tag{2}$$

We know, that probability of acoustic vector *P(A)* is constant, and we have:

$$W^* = \arg\max_w \{P(A \mid W) \cdot P(W)\} \tag{3}$$

In (3) we can distinguish [4], [5], [6], [7]:

- *P(W)* – the language model;
- *P(A|W)* – the acoustic model.

The acoustic modeling part of the speech recognition system can be developed using HMMs, Gaussian Mixture Models (GMMs) or Artificial Neural Networks (ANNs).

The language modeling part of the system can be one of the following [2]:

- statistical language model;
- context-free grammar (CFG);
- Probabilistic context-free grammar (PCFG).

In this paper we want to present the statistical n-gram type language model which is the most powerful and the most widely used one, and we want to create Romanian language models in ARPA-MIT [8] standard format for large vocabulary continuous speech recognition systems.

We have developed so far the feature extractor module and also the acoustic modeling part (using artificial neural networks) of an automatic speech recognition system.

2. Statistical language modeling

The speech can be considered a stochastic process and every linguistic unit (phoneme, syllabus, or word) can be considered a random variable with a random probability distribution. If we are talking at word level, the n-gram language models try to estimate the probability of the next word based on the history (the last *n-1* preceding words)) [1], [4], [6].
The language model try to estimate the probability of word sequence: $w_1^n = (w_1, w_2, \ldots, w_n)$, which is:

$$p(w_1^n) = p(w_1) \cdot p(w_2 \mid w_1) \cdot p(w_3 \mid w_1^2) \ldots \cdot p(w_n \mid w_1^{n-1}) \tag{4a}$$

$$p(w_1^n) = p(w_1) \cdot \prod_{k=1}^{n} p(w_k \mid w_1^{k-1}) \tag{4b}$$

Using Markov assumption, the history can be reduced to the last *n-1* words, and we have:

$$p(w_k \mid w_1^{k-1}) \approx p(w_k \mid w_{k-n+1}^{k-1}) \tag{5}$$

Even (5) is hard to compute for $n > 3$ because we need a large training corpus to properly evaluate the probabilities.

For $n = 1 \ldots 3$ we have:

- Unigram language model ($n=1$);
- Bigram language model ($n=2$);
- Trigram language model ($n=3$).

A. Unigram language model

The unigram language model considers all words independent. This means that no history information is involved.

$$P(w_k \mid w_1^{k-1}) \approx P(w_k) \tag{6}$$

If we use (4), the probability estimation for the unigram model will be:

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k) \tag{7}$$

B. Bigram language model

Bigram language model takes into consideration one word for history.

$$P(w_k \mid w_1^{k-1}) \approx P(w_k \mid w_{k-1}) \tag{8}$$

If we substitute (8) into (4), we have the probability estimation formula for bigram language model:

$$p(w_1^n) = p(w_1) \cdot \prod_{k=2}^{n} p(w_k \mid w_{k-1}) \tag{9}$$

C. Trigram language model

Trigram language model uses a two-word history.

$$P(w_k \mid w_1^{k-1}) \approx P(w_k \mid w_{k-1}, w_{k-2}) = P(w_k \mid w_{k-2}^{k-1}) \tag{10}$$

The probability estimation formula is given by (11).

$$p(w_1^n) = p(w_1) \cdot p(w_2 \mid w_1) \cdot \prod_{k=3}^{n} p(w_k \mid w_{k-1}, w_{k-2}) \tag{11}$$

3. Probability estimation and smoothing

The probabilities for (7), (9), and (11) can be simply calculated using MLE (Maximum Likelihood Expectation) algorithm [2]. Thus we have the following MLE estimators for unigram, bigram and trigram language models:

$$p(w_k) = \frac{n_k}{N}, \tag{12}$$

$$p(w_k \mid w_{k-1}) \approx \frac{Nr.(w_{k-1}, w_k)}{Nr.(w_{k-1})}, \tag{13}$$

$$p(w_k \mid w_{k-1}, w_{k-2}) \approx \frac{Nr.(w_{k-2}, w_{k-1}, w_k)}{Nr.(w_{k-2}, w_{k-1})}, \tag{14}$$

Where:

n_k – is the number of occurrences of word w_k;

N – is the total number of words in training corpus;

Nr. (...) – is the number of occurrences of a specific word sequence.

These probabilities calculated using MLE algorithm do not provide useful results. In order to be able to use the probabilities in language modeling experiments, they must be smoothed [4], [6], [7], [9].

Smoothing means that a probability mass is retained from high probabilities to be reallocated to zero or small probability values. There are a lot of useful smoothing techniques [4], [9], [10]:

- Add one or Laplace smoothing;
- Good-Turing estimator;

- Back-off or Katz smoothing;
- Kneser - Ney smoothing;
- Jelinek - Mercer smoothing or interpolation.

For practical experiments we used Good - Turing estimator and back-off smoothing. We had also implemented the add-one and Witten-Bell smoothing algorithms in Microsoft Visual Studio environment using C++ programming language.

A. Good – Turing estimator

The Good-Turing estimator comes from biology where it was used for species estimation. The general form of the estimator is [4]:

$$P(X) = \frac{r^*}{N}, \quad where, r^* = (r+1)\cdot\frac{E(N_{r+1})}{E(N_r)} \tag{15}$$

In (15) we have the following notations:

r is the number of occurrences of word X;

N_r is the number of words which occurs exactly r times in the training corpus;

N is the total number of words from the training corpus;

E is an estimation function for N_r;

r^* is the adjusted number of occurrences;

The total value of probability calculated using Good-Turing estimator is always smaller than *1*. The remaining probability mass is reallocated to the unseen words from the vocabulary. The simplest way to choose the estimation function E [5] is presented in (16).

$$\frac{E(n+1)}{E(n)} = \frac{n}{n+1}\cdot(1-\frac{E(1)}{N}) \tag{16}$$

B. Katz back – off smoothing

Back-off smoothing was firstly introduced by Katz. He showed that MLE estimation of probabilities is good enough if the number of occurrences of a word is bigger than a threshold value $K = 6$ [4].

All the probabilities for n-gram word sequences which have an occurrence number between *0* and K will be smoothed using Good-Turing estimator to save probability mass for unseen word sequences. If a word sequence has zero occurrences we try to estimate its probability using the inferior *(n-1)*-gram model. If the occurrence is still zero for this inferior model we continue to back-off to a lower model. Finally if we reach the unigram model, we have the relative frequency of a word bigger than zero.

For a trigram back-off model we have the fallowing relations:

$$\hat{p}(w_3 \mid w_1, w_2) = \begin{cases} f(w_3 \mid w_1, w_2), & \text{if nr. } (w_1, w_2, w_3) \geq K \\ \alpha \cdot \mathrm{Q_T}(w_3 \mid w_1, w_2), & \text{if } 0 < \text{nr.}(w_1, w_2, w_3) < K \\ \beta \cdot \hat{p}(w_3 \mid w_2), & \text{else} \end{cases} \tag{17}$$

$$\hat{p}(w_3 \mid w_2) = \begin{cases} f(w_3 \mid w_2), & \text{if nr. } (w_2, w_3) \geq L \\ \alpha \cdot \mathrm{Q_T}(w_3 \mid w_2), & \text{if } 0 < \text{nr.}(w_2, w_3) < L \\ \beta \cdot f(w_3), & \text{else} \end{cases} \tag{18}$$

4. Language model evaluation

Language model evaluation can be done in different ways, for instance using [4][6]:

- random sentence generation;
- words reordering in sentences;
- perplexity;
- integration in an existing speech recognition system.

In our experiments we used perplexity to measure the quality of language models. Perplexity is the most used measurement for language model evaluation.

Perplexity can be defined using entropy from information theory. For a random variable $X=\{x_1, x_2, \ldots, x_N\}$, the entropy can be defined:

$$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x) \tag{19}$$

Instead of entropy, we use the entropy rate calculated as follows:

$$\frac{1}{N} H(w_1^n) = -\frac{1}{N} \sum_{w \in V} p(w_1^n) \cdot \log_2 p(w_1^n) \tag{20}$$

For a real language we should consider infinitely long word sequences:

$$\frac{1}{N} H(w_1^n) = \lim_{x \to \infty} -\frac{1}{N} \sum_{w \in V} p(w_1^n) \cdot \log_2 p(w_1^n) \tag{21}$$

Using the Shannon – McMillan - Breiman theorem, if the language is stationary and ergodic (which is true for the natural languages), the above formula can be simplified:

$$H(L) = \lim_{x \to \infty} \left(-\frac{1}{N} \log_2 p(w_1^n) \right) \tag{22}$$

Finally we use a large training corpus to estimate probabilities p* and we have the *logprob* value instead of the entropy rate:

$$LP = -\frac{1}{N}\log_2 p^*(w_1^n) \tag{23}$$

Perplexity is defined as:

$$PP = 2^{LP} \tag{24}$$

5. Text conditioning

Collecting sufficient language model training data for good speech recognition performance in a new domain is often difficult. There are some text corpora for Romanian, which can be used for language modeling, but they are not normalized. This chapter presents the text normalization steps which are used to make these data more suitable for language model training.

Text is unlike speech in a variety of ways. For example, a written text may also include numbers, abbreviations, acronyms, punctuation, and other "non-standard words" (NSWs) which are not written in their spoken form. In order to effectively use this text for language modeling, these items must be converted to their spoken forms. This process has been referred to as text conditioning or normalization and is often used in text-to-speech systems.

State of the art language modeling tools like HLMTools [8], SRI LM [11] or CMU LM [12] do not provide professional text conditioning tools. A set of text conditioning tools are available from the Linguistic Data Consortium (LDC). A more systematic approach to the NSW normalization problem is referred to here as the NSW tools [13], [14]. These tools perform text normalization using a set of ad-hoc rules, converting numerals to words and expanding abbreviations listed in a table. They also use models trained on data from several categories. The NSW tools perform well in a variety of domains, unlike the LDC tools which were developed for business news.

Text normalization for Romanian is a hard process because of the diacritic characters as well.

Our system performs the following basic conditionings:

- it segments the text into sentences on the basis of punctuation, marking with <s> and </s> tags the beginning and the end of the sentences and puts only one sentence per line;
- eliminates most punctuation symbols;
- converts numbers into words;
- converts the whole text to uppercase;
- deletes all empty lines;
- eliminates redundant white spaces;
- replaces diacritic characters;

6. Experimental results

The built language models are based on Romanian Constitution text corpus. This little corpus contains 9936 words including both the train part and the test part (a total number of words). We count n-grams up to $n = 4$, however the corpus size does not allow us to compute valuable trigram and four-gram probabilities as you can see from perplexity results.

In *Table 1* the four-grams with the greatest frequency of appearance can be seen.

Table 2 presents the most probable 15 words from Romanian Constitution corpus.

Table 1. Most frequent four-grams in Romanian Constitution corpus

Word 1	Word 2	Word 3	Word 4	Number of appearance
</S>	<S>	DREPTUL	LA	27
DE	LEGE	</S>	<S>	16
SE	STABILESC	PRIN	LEGE	12
</S>	<S>	DREPTUL	DE	11
PRIN	LEGE	ORGANICA	</S>	10
LEGE	ORGANICA	</S>	<S>	10
</S>	<S>	CAMERA	DEPUTATILOR	9
CONDITIILE	LEGII	</S>	<S>	8
IN	CONDITIILE	LEGII	</S>	8
CAMERA	DEPUTATILOR	SI	SENATUL	8

The total number of distinct words in corpus is 1928, grouped in 718 sentences. 963 of them had more than one appearance.

We generated a dictionary from the most probable 963 words from the corpus (in fact these words appear more than once in the training corpus), and then we mapped all the other words into an unknown word class. We than generated the unigram, bigram, and trigram language models with Katz cut-off based on the corpus. The built language models were stored in ARPA MIT standard format.

Language model evaluation was made using perplexity measure for the three models. The perplexity results of the created models using the 963 - word dictionary, are presented in *Table 3*.

We made a second experiment with a smaller dictionary, containing only the words with appearance greater than 2 (626 words). The perplexity results of

our second experiment using the dictionary with 626 words are synthesized in *Table 4*.

Table 2. The most probable 15 words from Romanian Constitution

Word	Appearance	Word	Appearance	Word	Appearance
</S>	718	A	195	AL	71
<S>	718	LA	151	DREPTUL	70
DE	470	SE	128	PRIN	69
SI	405	SAU	116	PENTRU	68
IN	287	ESTE	82	CU	66

Table 1. Perplexity results for Romanian Constitution corpus using a 963-word dictionary.

Model	Perplexity
Unigram	559.74
Bigram	397.37
Trigram	419.52

Table 2. Perplexity results for Romanian Constitution corpus using a 626-word dictionary.

Model	Perplexity
Unigram	509.64
Bigram	332.24
Trigram	419.23

7. Conclusions and future works

We can see from *Table 2* that the most frequent words in Romanian Constitution corpus are the prepositions: *"de", "şi"," în", "a", "la", "se", "sau", "al", "prin"*, *"cu"* etc. The verb *"este"(to be)* has surprisingly high frequency of appearance, and because of the text corpus domain we also have the noun *"dreptul"(law)* in the first 15 words in order of frequency of appearance.

The sentence start tags and sentence end tags are also present in the four-grams and in the most frequent word list because of the short sentences of the corpus.

The best results are achieved using bigram model. The trigram model cannot improve the results because there is insufficient data for training the model.

We can see from the results that if the number of words increases, the perplexity of the model increases too, and the model has weaker quality. The

models built by eliminating the words with occurrences smaller than a threshold are simpler and perform better. This threshold can be experimentally settled. This technique is called in the literature n-gram pruning [1], [2].

The words with single occurrence do not improve the model quality, they will raise perplexity and they should be eliminated from the vocabulary. We have drawn the same conclusion in our previous work [1] based on the Susanne Corpus.

In conclusion the used n-gram model dimension should be chosen considering the amount of training data available. This little corpus is good enough for preliminary testing of a text conditioning and language modeling tool and we can train well enough just unigram and bigram language models.

As future work, we would like to improve the text conditioning tool with diacritic restoration feature, and try to generate language models based upon a much bigger training corpus of Romanian journal articles and to implement the state of the art Kneser - Ney smoothing algorithm [4],[8],[10],[11].

In order to compare our language modeling tools with others we will try to use open source language modeling toolkits (e.g., CMU-LM [9], SRILM [7]) on the same corpora.

We also try to collect a larger Romanian text corpus based on WEB resources.

Acknowledgements

This research work is supported by the Institute for Research Programs of the Sapientia University.

References

[1] Jelinek, F., "Statistical Methods for speech recognition", *The MIT Press*, 2001.
[2] Becchetti, C., Ricotti, L. P., "Speech recognition. Theory and C++ implementations", *John Wiley & sons*, 1999.
[3] Domokos, J., Toderean, G., Buza, O., "Statistical Language modeling on Susanne corpus", *IEEE International Conference COMMUNICATIONS 2008, Proceedings*, pp. 69-72, 2008.
[4] Jurafsky, D., Martin, J. H., "Speech and language processing. An introduction to Natural language Processing", *Computational Linguistics and Speech Recognition, Prentice Hall*, 2000.
[5] Huang, X., Acero, A., Hon, H., "Spoken Language Processing. A Guide to Theory, Algorithm & System Development", *Prentice Hall*, 2001.
[6] Manning, C., Heinrich, S., "Foundations of statistical language processing", *The MIT Press*, 1999.
[7] Rosenfeld, R., "Two decades of statistical language modeling: where do we go from here?", *Proceedings of the IEEE*, Vol. 88, pp. 1270 – 1278, 2000.

[8] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., "The HTK Book", *Cambridge University Engineering Department*, 2005.

[9] Chen, S., Goodman J., "An Empirical Study of Smoothing Techniques for Language Modeling", *Harvard Computer Science Technical report TR-10-98*, 1998.

[10] Goodman, J., Gao, J., "Language model size reduction by pruning and clustering", ICSLP-2000, *International Conference on Spoken Language Processing*, Beijing, 2000.

[11] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", *Proceedings of International Conference on Spoken Language Processing*, Denver, Colorado, 2002.

[12] Clarkson, P. R., Rosenfeld, R., "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proceedings ESCA Eurospeech*, 1997.

[13] Paul, D. B., Baker, J. M., **"**The design for the wall street journal-based CSR corpus"**,** *Workshop on Speech and Natural Language, Proceedings*, pp. 357 – 362, 1992.

[14] Schwarm, S., Ostendorf, M., **"**Text normalization with varied data sources for conversational speech language modeling"**,** *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '02*, Vol. 1, pp. 789 – 792, 2002.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 199-206

Past, Present and Future of Teaching Mechatronics at the Faculty of Mechanical Engineering and Informatics of the University of Miskolc

Gyula PATKÓ[1], Ádám DÖBRÖCZÖNI[2], Endre JAKAB[3]

[1]Department of Machine Tools, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Miskolc, Hungary,
e-mail: patko@uni-miskolc.hu
[2]Department of Machine and Product Design, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Miskolc, Hungary,
e-mail: machda@uni-miskolc.hu
[3]Robert Bosch Department of Mechatronics, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Miskolc, Hungary,
e-mail: jakab.endre@uni-miskolc.hu

Manuscript received March 15, 2009; revised April 15, 2009.

Abstract: The paper presents the development of and changes in teaching mechatronics at the Faculty of Mechanical Engineering and Informatics of the University of Miskolc starting in the 1960s. It is easy to see how fast the University responded to the technical changes in the world, what directions it followed and how it became involved in solving industrial problems of a mechatronic type, and how all these affected engineering education and training. The paper presents a novel example of cooperation between industry and higher education, amounting to an innovative solution. Furthermore, it will also display the extensive opportunities for evolution of a specialist field.

1. Introduction

There are several definitions known about the discipline of *mechatronics*. Each of them includes that it is an interdisciplinary field of science, where synergic integration of mechanical and electronic systems and information technology is achieved.

In the stiff competition of a globalised world, extraordinary intellectual and material resources are concentrated in order to conquer markets, new methods and technologies are applied in developments with greatly reduced cycle times. Mechatronic devices have also entered everyday life, for which a range of examples can be mentioned.

In its directives, the Design Methodology for Mechatronic Systems VDI 2206 describes mechatronics as the resultant of the engineering sciences above, shown in *Fig. 1*.

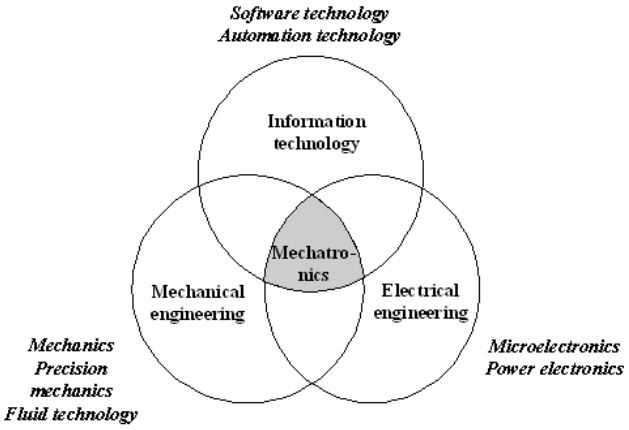


Figure 1: VDI 2206.

2. Establishment of mechatronic engineering programmes in Hungary

In accordance with the new Act on Higher Education, the introduction of the two-stage system, based on the preliminaries, made it possible to establish the BSc programme and the MSc programme in Mechatronic Engineering, which had not existed previously. In the course of the establishment, higher education institutions were also interested. Having previous experience, they elaborated and submitted the documents for establishing and introducing the programmes in a consortium partnership. The documents are available on the web page of the Hungarian Accreditation Committee (Magyar Akkreditációs Bizottság, MAB): http://www.mab.hu/a_tajekoztatok.html. The detailed programme and degree requirements („Képzési és kimeneti követelmények", KKK) for the undergraduate and master programmes are available on the web page of the Hungarian Ministry of Education and Culture (OKM): http://www.okm.gov.hu/main.php.

For comparison, the table below shows the credit point requirements of the "Fields of knowledge definitive for the qualification" for the programmes in mechatronic engineering and the data for the programmes in mechatronic engineering accredited at the University of Miskolc, Faculty of Mechanical

Engineering and Informatics (GEK). The KKK also specifies which topics have to be covered by the fields of knowledge. Within the number of admissions permitted, only graduates of the undergraduate programme in mechatronic engineering can be admitted to the MSc programme with full recognition of credit value. For the other programmes that may be taken into account (mechanical engineering, transport engineering, electrical engineering, engineering informatics, mechanical engineering for agriculture and food industry, and energetics), the admission conditions into the MSc programme can be found on the web page of the above-mentioned Ministry.

According to the latest data, there are nine higher education institutions in Hungary offering accredited programmes in mechatronic engineering. Both BSc and MSc programmes are offered by the following institutions: Budapesti Műszaki és Gazdaságtudományi Egyetem (BME), Budapesti Műszaki Főiskola (BMF), Miskolci Egyetem (ME), Széchenyi István Egyetem (in Győr), Szent István Egyetem (in Gödöllő). BSc programmes are offered by the following institutions: Debreceni Egyetem, Pannon Egyetem (in Veszprém), Pécsi Tudományegyetem, and Szegedi Tudományegyetem.

Table 1: Credit point requirements.

Fields of knowledge	**BSc, 7 semesters**		**MSc, 4 semesters**	
	MAB (cr)	**GEK (cr)**	**MAB (cr)**	**GEK (cr)**
Fundamental knowledge in science	40-50	47	26-36	27
Knowledge in economics and humanities	16-30	16	10-16	10
Professional core material	70-103	96	20-36	30
Differentiated professional knowledge, (degree work)	min 40	51	46-60	53
	-	(15)	(30)	(30)
Total credit points	**210**	**210**	**120**	**120**

3. Background of the programmes in mechatronic engineering established at the University of Miskolc

The Faculty of Mechanical Engineering (today Faculty of Mechanical Engineering and Informatics) at the University of Miskolc was established in 1949 with the purpose of training mechanical engineers. The 60th anniversary of the establishment is celebrated this year. The initial focus of the programmes included machine tool design and production engineering. This focus has been widened by now, meeting various demands. Currently, the Faculty has nine accredited BSc programmes (Energetics, Mechanical Engineering, Industrial Product and Design Engineering, Mechatronic Engineering, Industrial Management, Electrical Engineering, Economic Informatics, Engineering Informatics, Informatics, and Programme Informatics**)** and four MSc programmes (Mechanical Engineering, Mechatronic Engineering, Energetics, and Engineering Informatics), and the accreditation of the MSc programme in Logistics Engineering is under way. It shows the great number of new programmes which have emerged from the originally uniform training; all of them are built on the foundations created by the programme in mechanical engineering offered by the Faculty.

The roots of introducing the BSc programme in mechatronic engineering go back to machine tools and their automation at the Faculty of Mechanical Engineering and Informatics of the University of Miskolc. The programme for specialist engineers in *Machine Tool Automation* began in 1966 as a postgraduate programme and extended over five years. The full-time university programme in *Machine Tool Design* was also introduced in 1966, which was divided into two specialised programmes: *Machine Tool Design* and *Machine Tool Automation* in 1972**.** The latter widely applied the achievements of flexible automation in the courses on machine tools and robots, representing essentially mechatronics.

The technical-technological development and restructuring of the Hungarian industry encouraged education and research to introduce changes. The Faculty became involved in mechatronics, which emerged as a new discipline, with the help of substantial Tempus, PHARE-ESZA and HEFOP projects. These included exchange of experience, laboratory infrastructure development as well as writing textbooks.

The full-time university and college programmes of the *Specialisation in Mechatronics* appeared at the Faculty of Mechanical Engineering in the curricula valid from the academic year 1993/1994, at the same time when they appeared at BME, and they are to be phased out in 2009. These specialisations were included in the professional responsibilities of the Department of Machine Tools founded in 1963, but in close cooperation with the Departments of

Electrotechnique-Electronics and Automation, which were professionally responsible for the university programme of the *Specialisation in Electronics and Automation* in addition to college programmes in electrical engineering.

The results and achievements in mechatronics research appeared in the two Doctoral Schools of the Faculty of Mechanical Engineering and Informatics. The departments of the Faculty achieved highly significant results related to mechatronics in the research and development work commissioned by the industry.

4. Introduction of the programme in Mechatronic Engineering and establishment of the Robert Bosch Department of Mechatronics at the University of Miskolc

The Faculty of Mechanical Engineering and Informatics introduced the BSc programme of *Mechatronic Engineering* in the academic year 2007/2008 with 29 students. It should be noted that the average admission score of the students was among the highest at the Faculty. In 2008 the MSc programme in *Mechatronic Engineering* was also accredited. The programmes belong to the professional responsibilities of the Robert Bosch Department of Mechatronics and the Department of Machine Tools.

The curriculum in the BSc programme includes a single specialisation, Engineering Mechatronics, after two years of studies, and in the MSc programme it includes the specialisation, Mechatronics of Production Tools, after one year of studies. These rely on the achievements of previous programmes and are also aimed at a section of the industry that may be a significant employer of the mechatronics engineers graduating here. In elaborating the curricula and course syllabuses the experts in the major factories in the region were consulted.

Professors of the Duisburg-Essen University, who are involved in the programme as visiting professors, have performed a considerable role contributing to the establishment and promotion of the Department.

In consortium with the Széchenyi István University significant infrastructure and electronic learning material development was performed, also in relation to mechatronics, in the framework of the projects HEFOP-3.3.1-P.-2004-09-0102/1.0. The relevant material can be found at http://www.gepesz.uni-miskolc.hu/.

In order to establish and develop programmes in mechatronic engineering, on July 1, 2005 the Robert Bosch Department of Mechatronics was established at the Faculty of Mechanical Engineering and Informatics of the University of Miskolc *with support by the executive management of Bosch and the Bosch factories in the region.* The Department operated as an enterprise for three

years. The objective of the cooperation between the factories and the University is: *to apply and expand the technical and scientific knowledge in the research, teaching and wide-ranging application of mechatronics, to provide practice-oriented academic programmes and to meet the demand of the factories for engineers.*

The University of Miskolc was pleased to accommodate the first department to be financed by companies since World War II and took over its operation on July 1, 2008. The University also granted substantial funds for the project. The example, although not in the same structure, has been followed by several higher education institutions.

The Department has been funded and supported through the *professional training contributions* and the *innovation contributions* paid by companies. In the latter framework, the Department has completed several R+D projects with the involvement of the staff of the Faculty.

The practice-oriented training of the students of the programme in Mechatronic Engineering is supported by the mechatronics laboratories: hydraulics-pneumatics, PLC, drive technology, sensor technology and mechatronics systems (*Fig. 2*) constructed using the professional training contributions.

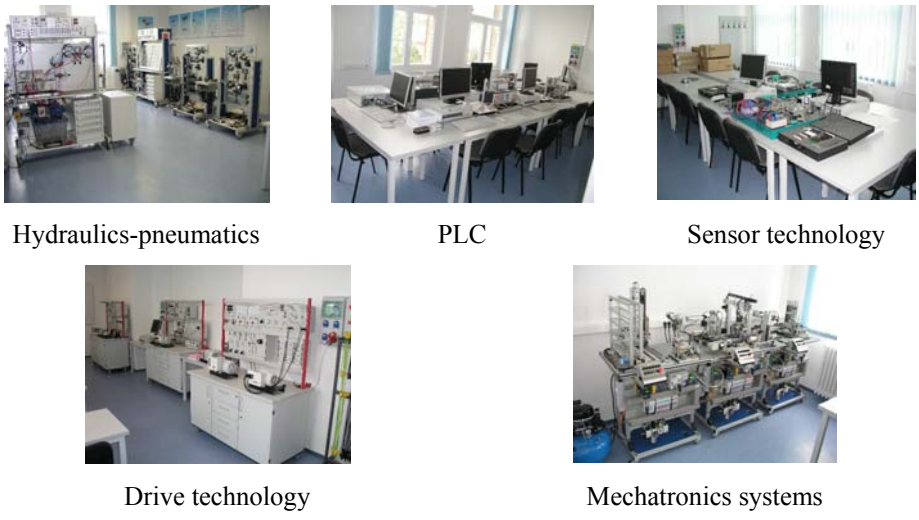


Figure 2: Laboratories of the Robert Bosch Department of Mechatronics.

5. Future of programmes in mechatronics

The development of disciplines giving the basis of mechatronics is unbroken, and the new achievements emerge in products of an increasing range. The pursuit of this integrated science requires students who are both inherently and willingly suitable for receiving complexity and for absorption in one of the fields of mechatronics as well as cooperation.

The particular characteristics and requirements of development in the field of mechatronics also have to be given an increasing emphasis in the academic work, and this has been partially formulated in the directives. Some of them are listed below without striving for completeness:

- accurate formulation of the task (requirement), application of various design methods and development tools (QFD, FMEA, etc.), virtual design, modelling and simulation tools,
- Concurrent Engineering,
- application of a wide range of tools of information and electric technology,
- acceleration of the innovation process, increasing intelligence, and use of techniques such as modularity, change of technology, functional and spatial combination, prototyping, testing,
- possibility of access to databases and interfaces irrespective of location and time,
- systemic thinking,
- cooperation in team work, including the evolution of individual capacities, etc.

In addition, it is indispensable to get to know and use further tools, methods and fields, such as logistic processes, quality control, production systems, production control, reliability, operation and maintenance.

The above requirements clearly show that part of the required knowledge meeting the features of the field for solving the industrial tasks in mechatronics can only be obtained in the BSc and MSc programmes, therefore advanced training programmes are of great importance. An active environment is needed for the abilities to develop and in order to achieve the objectives of the programmes *cooperation between the industry and the university* is essential and is already evident in several areas.

Besides theoretical and practical education, factories manufacturing mechatronic devices or using them in production play a considerable role. In the factories students or trainees on industrial placement can familiarise themselves with the various techniques, they may be given project assignments, topics for degree work and they can be involved in programs and tenders. It is to be noted here that the competitions *'Pneumobil 2009'* and *'Elektromobil 2009'* of the Bosch factories in Hungary attracted student teams from a number of higher

education institutions. The addresses http://www.pneumobil.hu and http://www.olh.hu:80/bosch/cms/elektromobil give all the details of the call. The large number of the teams participating shows that students need such challenges requiring independent and creative work.

Integrating industrial experience and knowledge into education and academic programmes is a very significant factor. One efficient way of doing so is involvement in industrial research and development work, getting students involved and integrating the experience in the academic work. Another important element is what higher education in the developed countries has been using for a long time, namely involving industrial experts in education. Regarding mechatronics, doctoral programmes have made it possible to award higher qualifications to experts.

6. Summary

One may wonder about the perspectives of the programme in mechatronic engineering in the current circumstances and the job opportunities for the graduated engineers. The answers are that this period particularly requires experts who are able to develop and operate intelligent, innovative products and instruments. Another major objective of the programme is to train engineers, who – due to their well-founded, comprehensive knowledge – can offer flexible solutions in changing conditions. The industry offers opportunities for them to find jobs in wide professional fields, for the knowledge they have acquired can be utilised anywhere in the world.

7. Acknowledgements

The authors wish to express their thanks to all the colleagues who have contributed to the foundations for the establishment and introduction of the programme in mechatronic engineering and are currently involved in teaching the courses.

References

[1] Curricula of the Faculty of Mechanical Engineering and Informatics of the University of Miskolc: http://www.gepesz.uni-miskolc.hu/oktatas/
[2] ProjectHEFOP-3.3.1-P.-2004-09-0102/1.0
[3] VDI 2206
[4] R. Isermann,: Mechatronische Systeme – Grundlagen, Berlin: Springer Verlag, 2008

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 207-213

Phase Transformations in the Heat Treated and Untreated Zn-Al Alloys

Béla VARGA[1], Ioan SZÁVA[2]

[1]Department of Technologic Equipment and Materials Science, Faculty of Materials Science,
"Transilvania" University of Braşov, Braşov, Romania,
e-mail: varga.b@unitbv.ro
[2]Department of Strength of Materials and Vibrations, Faculty of Mechanical Engineering,
"Transilvania" University of Braşov, Braşov, Romania,
e-mail: janoska@clicknet.ro

Manuscript received March 15, 2009; revised April 30, 2009.

Abstract: Microstructure changes and phase transformations of Zn-Al based alloys have been systematically studied, using XRD, SEM and TEM techniques. The paper presents the results of experimental research concerning the eutectoid transformation in the Zn-Al system. The paper focuses on the determination of the activation energy for the eutectoid transformation in the binary Zn- (4, 8, 12, 22, 27) % Al system, using the values of the temperatures corresponding to the peaks on the derivatives of the dilatation curves.

Keywords: Zn-Al alloys, eutectoid reaction, dilatometric analysis, activation energy.

1. Introduction

Over the last years the category of industrial alloys of the Zn-Al system has been extended by the standardizing of compositions with 8, 12, 22, 27 and 40% aluminium, respectively. Their properties have been thoroughly researched and documented [1]. Alloys of the Zn-Al category have excellent castability and good wear and friction strength. The disadvantage of such materials is their instability in time. In order to avoid this disadvantage Zn-Al alloys are exposed to heat treatments designed to contribute to an increased structural stability during deployment [2], [3].
At the same time compositions with 18 – 40% aluminium have super-plastic properties [4], [5]. The heat treatments applied to these alloys, as well as super-plasticity are based on structural transformations in the solid phase. The thermodynamics of the process is of highest relevance to an analysis of the structural transformations. The paper approaches the study of the eutectoid

transformation for binary compositions with 4, 8, 12, 22 and 27 % aluminium, respectively. During cooling the eutectoid transformation occurs at a temperature of 278 ºC according to the reaction:

$$\beta \rightarrow \eta + \alpha,$$

while during heating the reaction unfolds in the opposite direction.
The activation energies for heating and cooling were determined by dilatometric analysis.

2. Experimental determinations

Primary metals were used to melt of Zn-(4, 8, 12, 22, 27) % Al compositions. Melting was achieved in an electric furnace with silit bar heaters (electrical resistances) in a graphite crucible. The over-heating temperature prior to casting was of 100 °C. Billets of 80x200x10 mm were cast in steel ingot moulds.

Test pieces were machined from the cast billets, to be used for determining chemical composition and hardness for the structural analyses conducted on heat treated and untreated alloys, as well as for dilatometric analysis. The applied heat treatment consisted of heating to 350 °C over 4 days (96 hours) followed by furnace cooling. Structure analysis took into account Presnyakov's thermal equilibrium diagram, *Fig. 1*.

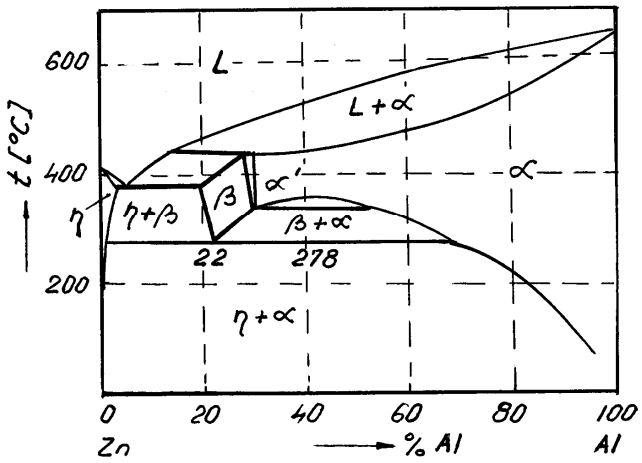


Figure 1: Zn-Al thermal equilibrium diagram.

Figure 2 shows obtained significant microstructures observed by means of a Nikon microscope for increase of 200 – 1000 times.

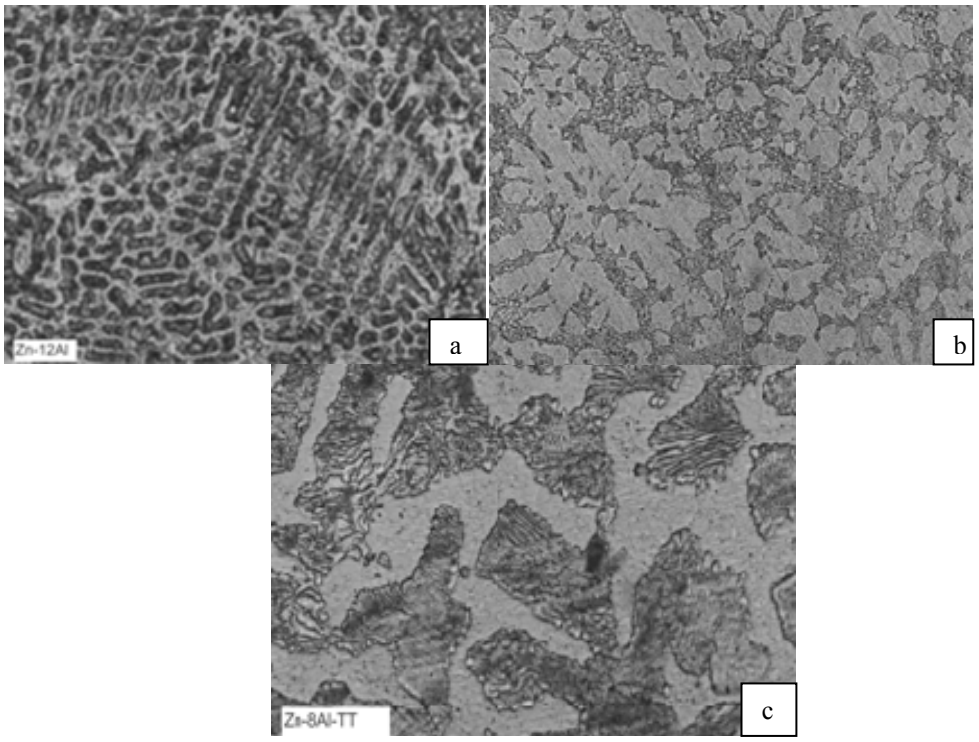


Figure 2: Structure of Zn-Al alloys
a.) Zn-12%Al, cast, x500 b.) Zn-22%Al, cast, x500 c.) Zn-8 %Al heat treated, x1000.

In alloys of hypoeutectoid composition a dendritic structure can be observed, consisting of η phase in the hypoeutectic alloy and phase α in the case of hypereutectic alloys, embedded in the η + α. eutectic. In the test sample of eutectoid composition reveal the effect of the peritectic reaction can be observed from the location of the eutectoid at the margins of the α solid solution dendrites. It has to be remarked that achieving the equilibrium structure requires long term heat treatment, as only under such conditions is the structure is in accordance with the nature and proportion of phases indicated by the thermal equilibrium diagram.

For dilatometric analysis round test pieces of 6 mm diameter and 15 – 17 mm length were machined. Dilatometric analysis was conducted by means of a LINSEIS, L75/230 device, *Fig. 3*.

Phase transformations in solid state can be studied with a dilatometer only when volume variations during transformation are also involved. Any heated metallic body dilates according to the equation:

Figure 3: LINSEIS, L75/230 dilatometer.

$$L_t = L_o\,(1 + \alpha.t). \tag{1}$$

The variation in length is computed by the equation:

$$\Delta L_{Thermal} = L_t - L_0 = L_o\,(1 + \alpha.t). \tag{2}$$

In case the metal (alloy) presents phase transformations in solid state, as the eutectoid one, a variation in length determined by the phase transformation is added to the one determined by temperature increase:

$$\Delta L_{Total} = \Delta L_{Thermal} + \Delta L_{Phase} \tag{3}$$

Variations in length determined by phase transformations are more visible on the derivatives of the dilatation curves, *Fig. 4*.
The value of temperature corresponding to the peaks on the dilatation curves derivative was used to determine the activation energy in the case of the eutectoid transformation in the Zn-Al system.

The activation energy (E_a) is calculated by Kissinger's equation written in the form below:

$$\ln \frac{v}{T_m^2} = -\frac{E_a}{R \cdot T} + M \tag{4}$$

where: T - temperature;

v – heating rate in ˚C/s

R – constant of gases, R= 8.3144 [J/mol.K];

M - constant;

T_m – maximum point on the dilatation curves derivatives, in degrees Kelvin.

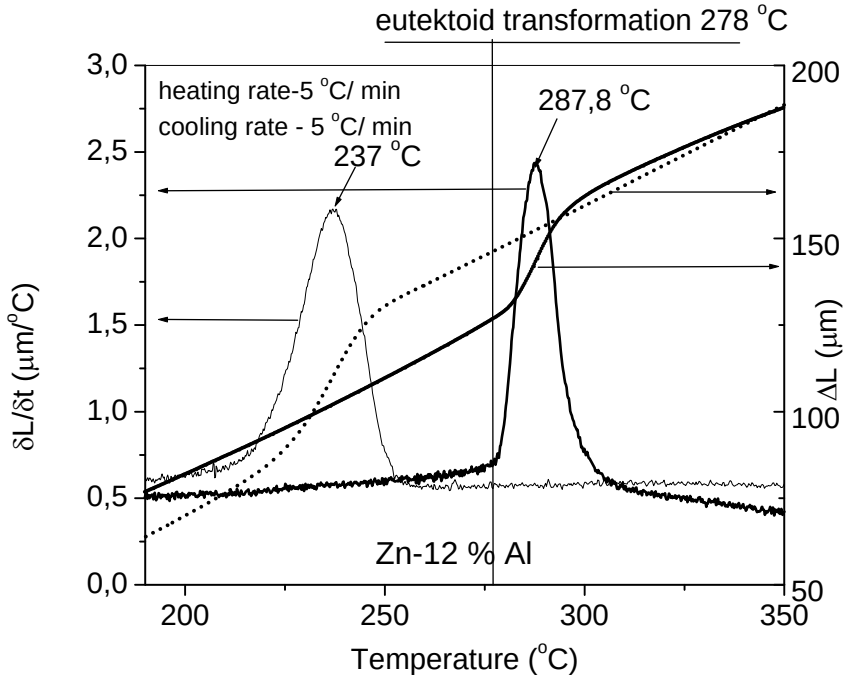


Figure 4: Dilatation curves for Zn-Al 12 % composition and their derivatives.

Figures 5 and *6* show the diagrams for determining of the activation energy of the eutectoid transformation during heating and cooling, respectively, for a Zn-Al 12 % composition.

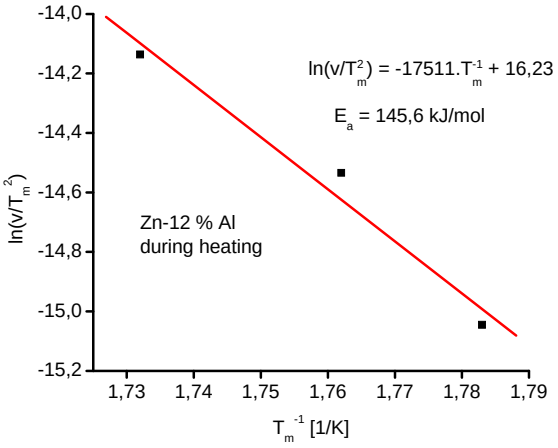


Figure 5: Determination of the activation energy during heating of the Zn-12 %Al %alloy.

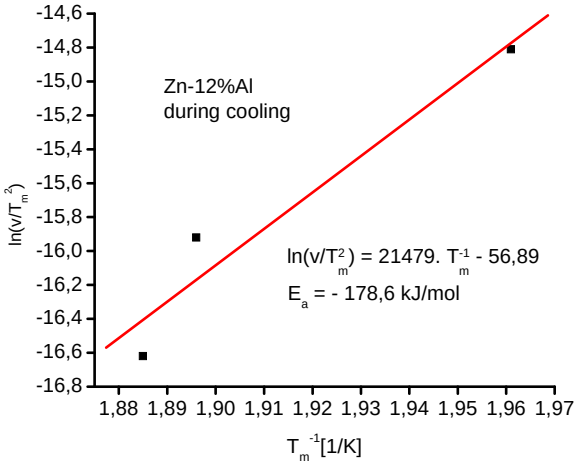


Figure 6: Determination of the activation energy during cooling of the Zn-12%Al alloy.

The diagram in *Fig. 7* presents the variation of the activation energy versus zinc concentration in the alloy, during heating and cooling, respectively. Relationships presented in the chart in *Figure 7* are valid for hypoeutectic compositions.

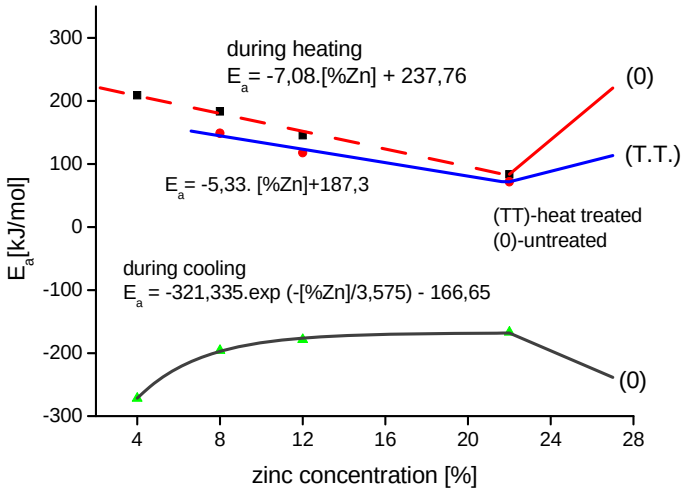


Figure 7: Variation of the activation energy versus zinc concentration during heating and cooling, respectively.

3. Conclusions

The degree of generating of the peritectic transformation significantly influences the structure of alloys at environment temperature.

In transformation processes during both heating and cooling the increase of zinc contents determines a decrease of the value of the activation energy.

The variation of activation energy depending on zinc or eutectoid concentration in the structure of the alloy is governed by different laws of constituent germination for the cases of heating and cooling of the alloy, respectively.

References

[1] http://en.wikipedia.org/wiki/ZAMAK, 21/12/2008 20.43.

[2] Savas, M. A., Altintas, S., "The microstructural control of cast and mechanical properties of zinc-aluminium alloys", *J. Materials Science,Vol. 28*, pp. 1775-1780, 1993.

[3] Xu, X. L., Yu, Z. W., Ji, S. J., Sun, J. C., Hei, Z. K., "Differential Scanning calorimetry and X-ray diffraction studies on aging behavior of Zn-Al alloys", *Acta Metallurgica Sinica*, Vol. 14, No. 2, pp. 109-114, Apr. 2001.

[4] Zhu, Y. H., " Phase transformations of eutectoid Zn-Al alloys", *J. Materials Science,* Vol. 36, pp. 3973-3980, 2001.

[5] Zhu, Y. H., Chan, K. C., Pang, G. K. H., Yue, T. M., Lee, W. B., "Structural Changes of α Phase in Furnace Cooled Eutectoid Zn-Al Based Alloy", *J. Mater. Sci. Technol.*, Vol. 23 No.3, pp. 347-352, 2007.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 1 (2009) 215-226

Parameters and Models of the Vehicle Thermal Comfort

Radu MUSAT[1], Elena HELEREA[1]

[1] Department of Electrical Engineering,
Faculty of Electrical Engineering and Computer Science,
"Transilvania" University of Braşov, Braşov, Romania,
e-mail: r_musat@yahoo.com, helerea@unitbv.ro

Manuscript received May 30, 2009; revised June 30, 2009.

Abstract: Nowadays, efforts are being made to estimate the thermal comfort in vehicles by measuring each environmental parameter - air temperature, air humidity, mean radiant temperature, air velocity, human activity and clothing insulation. An optimum level of comfort in the vehicle is obtained only by using an automatic air conditioning and climate control system. The paper focuses on the analysis of the vehicle thermal comfort parameters in order to improve the measurement methods and to establish the optimum thermal comfort inside a vehicle. The paper also describes two thermal comfort models used to estimate thermal comfort inside the vehicles.

Keywords: Thermal comfort, measurement, parameters, air conditioning, vehicle.

1. Introduction

Over the last years, with the trends of reducing costs and carrying weight, the interest in ensuring an optimal efficiency for vehicles has increased in a large sense (comfort, dynamicity, performances and energy efficiency).

Construction of vehicles developed from simplistic to modern, integrating the state-of-the-art technologies, organized on functional and aesthetic criteria, which ensure the passengers' comfort, ergonomics and safety.

Thermal comfort in vehicles represents a subjective sensation of heat balance that occurs in the human body when environmental parameters - *air temperature, air humidity, radiant temperature, air velocity, human level activity and clothing insulation* - are in a range of well-defined values [1].

ASHRAE Standard 55 defines thermal comfort as "that state of mind which expresses satisfaction with the thermal environment" [2].

As Parson observed in his studies [3], thermal comfort is influenced by a combination of physical, physiological and psychological factors. Some factors include solar radiation and glazing, inside and outside colours, the size of the vehicle, the clothing type of the passengers and passenger capacity of the vehicle cabin [4]. In a vehicle environment each passenger, regardless of their size, can affect the thermal environment inside a vehicle [5-9].

Thermal comfort is achieved (i) by ensuring temperatures of 20°C ÷ 22°C, as a result of air temperature, delimitation areas, humidity and air velocity in accordance with the activity level and clothing insulation of the occupants, (ii) by avoiding situations such as the occupants coming into contact with very cold or very hot surfaces, (iii) by avoiding air currents. These requirements must be met throughout the entire year, both summer and wintertime.

The research of thermal comfort has been in progress for many years. The study of thermal comfort in vehicles was developed from basic thermal comfort research and applied work related to factories and buildings [10]. The first research in vehicles dealt mainly with agricultural vehicles and public transport systems such as subways, trains and buses [11]. Achieving a thermally comfortable vehicle environment has become an issue of main importance.

The paper presents an analysis of the vehicle thermal comfort parameters and describes two thermal comfort models (Fanger's model and thermal manikin model) used to estimate the thermal comfort inside vehicles.

2. Environmental parameters of the vehicle

Very few articles have explicitly defined the differences between vehicle and building environment. However, there are researchers who dealt with the vehicle environmental parameters and their measurement methods [5-9], [11].

ISO 7726 standard describes some methods for measuring physical qualities related to thermal comfort parameters [12]. Whereas the tendency in measuring thermal comfort has been towards using individual instruments (e.g. thermocouples, globe thermometers, net radiometers, hotwire anemometers, hygrometers etc.) to measure single parameters in buildings, the automotive research has adopted a different approach, mainly due to the small available working space and the dynamic driving tests that are required when making thermal measurements. Installing large amounts of equipment in vehicle cabins is time-consuming and presents difficulties when all parameters have to be measured in the same position. Using a transducer that measures the combined effect of all environmental parameters will make the evaluation very efficient.

In his studies, Temming observed that the thermal environment in a vehicle cabin is very complex and thus difficult to evaluate. These difficulties are due to the influence of convective, radiative and conductive heat exchange created by

external thermal loads, the internal heating and by air conditioning and ventilation system [13].

The usual method to evaluate the thermal comfort parameters in vehicles is to use sensors to measure the air temperature at the level of the head and feet. The main purpose of such measurements is to determine how quickly the temperature will increase or decrease in a cold or warm vehicle cabin, to study the difference between the temperature at the feet and head level and to establish when the temperature reaches the thermal comfort level. However, using this method, only one of the needed parameters that concern the thermal comfort sensation is measured. By measuring only the air temperature, any influence of the air velocity and radiation (cold or hot) are neglected and the measurements might lead to false conclusions. This fact appears more often in vehicles than buildings, because the air conditioning system can create high local air velocities [14].

Nowadays, efforts are being made to estimate the thermal comfort in vehicle environments by measuring each environment parameter - air temperature, air humidity, mean radiant temperature, air velocity, human activity and clothing insulation. There is a great inter-correlation among these parameters. That is why the values recommended in standards are in well defined ranges. The thermal comfort can be obtained by correlating all these parameters.

Air temperature

The optimal value for the inside temperature is a function of the season time. During wintertime the optimal inside temperature adopted is $\theta_i = +22^oC$; during summertime different values for inside temperature are indicated in the literature.

Temming underlines that air temperature zones inside a vehicle are not homogenous. Whereas the air temperature in buildings generally increases with height from the floor to ceiling, this fact is not acceptable in vehicles. In vehicles, the air temperature at the ankle level is expected to be higher than at head level.

ASHRAE Standard 55 prescribes 3ºC for the vertical air temperature difference between head and ankle level [2]. Other studies set this limit up to 6ºC [15].

Moreover, the air temperature depends upon the “class” of the vehicle. A larger vehicle with leather upholstery during warm-up conditions may have an entirely different air temperature than a small economy-class vehicle during the same driving conditions.

The inside temperature is measured using temperature sensors. The recorded temperature values are between the values of the air temperature and the values

of the mean radiant temperature. In order to reduce the error introduced by the solar radiations, the temperature sensor must be as small as possible. The purpose of using appropriate temperature sensors is to see how quickly the temperature will increase or decrease in a cold or warm vehicle cabin and to measure the difference between the temperature at the level of head and feet.

Air velocity

Air velocity inside the vehicle usually has reduced values, ranging between 0.1 and 0.4 m/s. The maximum air velocity allowed inside a vehicle is considered a function of the air temperature determined by the convection heat exchange between the human body and environment. Due to air velocity fluctuations, the measurements must be carried out over a period of 3 ÷ 5 minutes to obtain a reasonable average value. When a model is developed, the air velocity value is neglected because it has a reduced value.

Air flow sensation is subjective and varies according to the person's sensitivity (some parts of the body are more sensitive, e.g. nape). The appearance of air currents is mostly due to the untight environment and to the air flow of the air conditioning system. The air flow coming through an open window increases the air velocities and the thermal discomfort as well.

Inside the vehicle, the air flow can only be directed to smaller sections because of a reduced volume (as opposed to buildings). The heated air should be directed toward the bottom half of the occupant's body and the cool air should be directed toward the upper half [16].

The studies of many researchers show that in a warm environment, higher air flow could provide a thermal comfort [17-19].

Figure 1 shows the correlation between the air velocity limit and the inside air temperature. As it can be seen in this figure the limits of air velocity values increase at high air temperature values.

The air flow sensation appears above the air velocity curve. The air flow sensation is subjective and affects mostly the back of the passengers' neck. Moreover, air flow sensation depends on the body's thermal state [20].

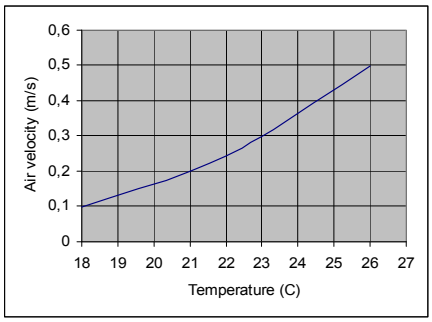


Figure 1: The air flow sensation curve [20].

Mean radiant temperature

The mean radiant temperature (MRT) is the uniform surface temperature of an imaginary black enclosure in which an occupant would exchange the same amount of radiant heat as in the actual non-uniform space [2]. MRT represents the mean temperature of all the objects surrounding the body. MRT will be positive when surrounding objects are warmer than the average skin temperature and negative when they are colder. MRT governs human energy balance and human body heat losses, especially on hot sunny days [21].

Mean radiant temperature, θ_m, is obtained if surface S_i and temperature θ_i are known for every construction element (e.g. door panels, dashboard) delimitating the passenger's area.

Mean radiant temperature is calculated by using the following formula:

$$\theta_m = \frac{\sum_{1}^{n} S_i \cdot \theta_i}{\sum_{1}^{n} S_i} \tag{1}$$

The mean radiant temperature can be determined if temperature and position of every construction element (e.g. door panels, dashboard) around the passenger inside the vehicle are known.

Relative humidity

ASHRAE Standard 55 defines relative humidity as the ratio of the partial pressure of water vapour in a gaseous mixture of air and water vapour to the saturated vapour pressure of water at a prescribed temperature [2].

Relative humidity is measured in only one place inside the vehicle because the pressure of the water vapour is uniform in the entire vehicle. The human body is sensitive to air humidity changes. The thermal comfort sensation is optimal when the relative humidity value is about 50%.

Temming observes in his studies that humidity plays a minor role. However, relative air humidity is correlated with inside temperature. These two parameters influence the thermal comfort of the passengers and they are the main parameters of the air conditioning system [21].

Figure 2 shows the correlation between temperature variation and the relative air humidity. As it can be seen in this figure, the relative air humidity increases when temperature decreases. A high relative humidity (over 70%) causes a sultry weather sensation increasing the discomfort level and can lead to problems of condensation, such as misting of windshields and shorting of electrical components. A low relative humidity (under 30%) causes a dry sensation, which can irritate the passenger's bronchial ways.

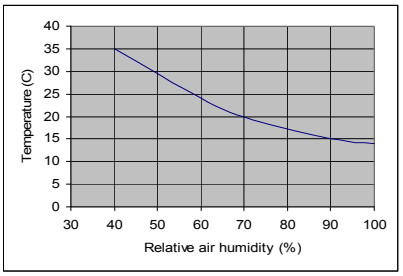


Figure 2: Correlation between temperature and relative air humidity [21].

The recommended values for inside temperature and air humidity in correlation with the outside temperature are given in *Table 1*.

Table 1: Inside temperature and air humidity as a function of outside temperature [22].

Outside temperature [˚C]			Winter	Summer			
			till +20	+20	+25	+30	+32
Inside temperature [˚C]			22	22	23	25	26
Relative humidity	%	Min	35	-	-	-	-
		Max	70	70	65	60	55

Human activity level and clothing insulation

The equivalent temperature, θ_{eq}, is calculated with formula [22]:

$$\theta_{eq} = A \cdot \theta_i + (1 - A) \cdot \theta_m \tag{2}$$

where: θ_i – inside temperature;
θ_m – mean radiant temperature;
A – weight factor (*Table 2*).

Table 2: Weight factor values at different air velocities values [22].

Inside air velocity, v [m/s]	< 0,2	0,2……0,6	0,7……1
Weight factor, *A*	0,5	0,6	0,7

Figure 3 shows the equivalent temperature curves of thermal comfort as a function of the human activity level, q_o, (1 met=58.2 W/m^2) and clothing insulation (1 clo=0.155 m^2·K/W). The diagram was created for relative air humidity of 50% and for inside air velocity of v_a = 0 m/s, if human activities $q_0 \leq 1$ met; for relative air humidity of 50% and for inside air velocity of $v_a = 0.3 \cdot (q_0 - 1)$ if human activities $q_0 > 1$ met.

The surface temperature of the human body is an average temperature, as people have different skin temperatures for different parts of the body. Clothing insulation increases together with the temperature increase due to the lower difference between the air temperature and the human body's surface temperature. *Figure 4* shows the skin temperature corresponding to different parts of the human body versus the inside air temperature. As it can be seen in this figure, the temperature at the feet level is lower than the temperature at the head level. These temperature differences influence the thermal comfort of the passenger.

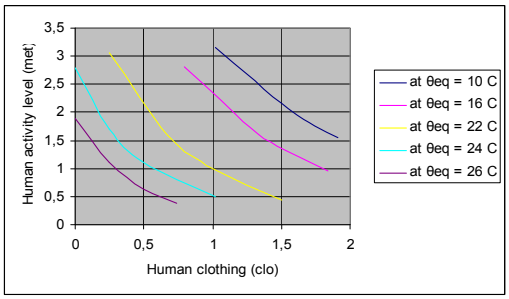


Figure 3: Thermal comfort limits for equivalent temperature as a function of human activity level and human clothing [23-25].

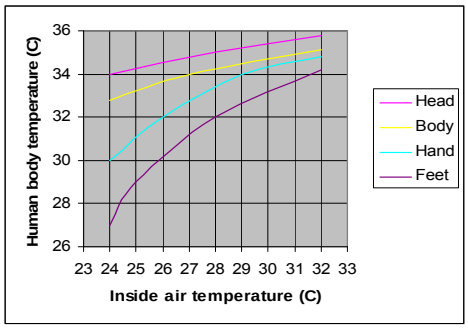


Figure 4: Skin temperature of different human body parts versus inside air temperature [23-25].

Several supporting standards for clothing insulation (ISO 9920) and metabolic rate (ISO 8996) are available [26-27].

When all these parameters are measured, their combined effect on the occupants of the vehicle can be determined.

In order to obtain an increased thermal comfort of the passengers, the parameters analyzed above are used to design the air conditioning system.

3. Thermal comfort models

Vehicle thermal comfort has been modelled using a combination of mathematical and statistical relationships.

Different models of thermal comfort have been also developed that can be used to predict subjective comfort assessment. The models are usually based on six parameters - air temperature, air humidity, mean radiant temperature, air velocity, human activity and clothing insulation [28]. Some models have been validated as a result of research on human subjects.

There are two significant models that can be used to predict thermal comfort and to estimate the environmental parameters of the vehicles (a mathematical model - Fanger's model and a physical model using thermal manikins). These models are the basis for designing and improving the air conditioning system.

Fanger's model

The most notable researcher on the thermal comfort analysis was P.O. Fanger [29]. Fanger's model suggests that thermal comfort can be predicted if the values of all six environment parameters are known. According to Fanger,

the thermal comfort is analyzed by PMV (*Predicted Mean Vote*), and thermal discomfort can be analyzed by PPD (*Predicted Percentage Dissatisfied*).

A vehicle represents a "moderate" thermal environment described by Fanger's equation. The equations that led Fanger to develop the concept of PMV and PPD are based on the physiological processes that underlie human heat balance. The interaction between the human body and the environment is described by the heat balance equation between thermal heat developed by metabolism in the human body and the heat transferred through convection, conduction, radiation and evaporation [30].

The PMV is based on the subjective seven-step scale [31]. The value of the PMV index has a range from -3 to +3, corresponding to human sensations from cold to hot, respectively, where the null value of the PMV index means neutral. The PMV evaluation method treats the entire body as one object. It does not distinguish between different parts of the body. If one side is warm and the other cold, the PMV model would calculate a zero thermal load and therefore yield a neutral thermal sensation (PMV=0). It is noted that the optimum value for thermal comfort (PPD is 5% and PMV is 0) can be obtained only with automatic air-conditioning systems.

The model was developed based on data from uniform thermal environments. Because it only calculates the heat transfer for the entire body, it cannot predict local discomfort. Clothing is assumed to cover the entire body uniformly which results in one skin temperature across the entire body.

The PMV model is a basis for most current standards prescribing methods for evaluating thermal comfort in vehicles [2], [26], [32].

Fanger model has limitations related to: (i) thermal steady state or dynamic state, (ii) distinction between local and whole-body thermal comfort, (iii) environmental particularities of the vehicle. The PMV model depends on the context and is more accurate in vehicles with air-conditioning systems than in the ones with natural ventilation, because of the influence of outside temperature. The inadequate measurements of the thermal insulation of clothing and the metabolic rate will reduce the accuracy of the PMV index. Other limitations are related to the local effects of asymmetrical conditions or local air movement around the face of occupants.

Thermal Manikins

Measurement and assessment of thermal climate using a thermal manikin makes it possible to evaluate the best solution for thermal control. It can also be used to measure clothing and chair insulation.

The first thermal manikin was introduced in 1985 by Wyon [33] and after that other manikins have been developed [34-36].

Thermal manikins are currently available but they are used primarily for measuring the thermal insulation condition of the clothing. The problem is that the manikin does not respond to the environment in the way the human body does. Most current manikins do not possess a sweating capability and hence only sense dry heat transfer. Evaporative cooling is a critical and often used component of the thermoregulatory system of the body. A thermal manikin should possess this capability in order to accurately simulate the response of the body in all thermal environments.

A thermal manikin that possesses a high degree of sensory spatial resolution, local thermoregulatory responses including sweating, a fast time response and a feedback loop to continuously react and adjust to a thermal environment like a human has never been developed. An advanced thermal manikin with these capabilities would help industry develop more effective and energy efficient climate control systems for transportation environments, or others where transient and extremely non-uniform thermal environments exist [37].

A thermal manikin needs to have the following properties in order to accurately simulate the human body: correct body shape and size; control of heat emission; control of the distribution of heat across the skin surface; emission of the skin; control of the distribution of perspiration across the skin surface; control of pose and movement and control of core [38].

So far, no manikin meeting all these criteria has been available. Depending on certain situations, for example, when wearing cold-weather gear in cold conditions, existing thermal manikins are limited to a uniform temperature distribution across the skin surface even though the human body's extremities experience large drops in skin temperature. This leads to an overestimation of heat loss from the extremities in the result obtained using the thermal manikin.

In order to improve the air conditioning system, measurements of local climate disturbances with a man-sized thermal manikin have to be made and have to be correlated with the thermal sensation experienced by subjects exposed to the same conditions. Criteria for acceptable climatic conditions can be defined in terms of quantities measured with the manikin.

Although there are some limitations, the thermal manikin model represents a quick, accurate and efficient model for evaluating the vehicle's thermal comfort.

3. Conclusion

The vehicle is characterized by a moderate thermal environment. This environment is defined by six thermal comfort parameters: air temperature, air humidity, mean radiant temperature, air velocity, human activity and clothing insulation. By measuring all these parameters, the combined effects on the

occupants of the vehicle can be calculated. Specific models for thermal comfort are analyzed and used to estimate the thermal comfort of the passengers.

Inside air temperature zones within the occupant space of vehicles are neither homogenous, nor desired to be homogenous. The inside temperature is correlated with inside relative air humidity. These two parameters influence the thermal comfort of the passengers and they are the main parameters of the air conditioning system. *Humidity fluctuations* play a minor effect if the values are in the range of 30% to 70%. *Mean radiant temperature* depends on the "class" (size and quality) of the vehicle and influence the passenger thermal comfort. *Air velocity fluctuations* are mostly due to the untight environment and to the air flow of the air conditioning system. In order to obtain an increased thermal comfort of the passengers, new measurement methods and efficient models must be developed.

References

[1] Sârbu, I., "Thermal comfort evaluation models", in *Theoretical considerations*, No. 2 (43), University of Timişoara, 2007.

[2] ***, ASHRAE American Society of Heating Refrigerating and Air Conditioning Engineers, Standard 55, "Thermal Environmental Conditions for Human Occupancy", Atlanta, 1992.

[3] Parsons, K. C., "Human Thermal Environments", *Taylor & Francis Inc.*, Bristol, 1993.

[4] Hymore, R. R., et al, "Development of a test procedure for quantifying performance benefits of solar control glazing on occupant comfort", in *SAE Technical Paper Series*, No. 910536, Warrendale, PA, 1991.

[5] Sayer, J. R., Traube, E. C., "Factors influencing visibility through motor vehicle windshields and windows: review of the literature", UMTRI-94-20, *University of Michigan Transportation Research Institute: Ann Arbor*, Michigan, 1994.

[6] Schacher, L., Adolphe, A., "Objective characterization of the thermal comfort of fabrics for car upholstery", *Proceedings*, *Niches in the World of Textiles World Conference of the Textile Institute,* Vol. 2, pp. 368-369, 1997.

[7] Shuster, A. A., "Heat comfort in passenger cars", *Tyazheloe Mashinostroenie*, No. 1, 1998.

[8] Madsen, T. L., et al., "New methods for evaluation of the thermal comfort in automotive vehicles", *ASHRAE Transactions*, Vol. 92, pp. 38-54, 1986.

[9] Gagge, A. P., Fobelets, A. P., "Standard predictive index of human response to the thermal environment", *ASHRAE Transactions*, Vol. 92, pp. 709-731, 1986.

[10] Parsons, K., "Human thermal environments – The effect of hot, moderate and cod environments on human health, comfort and performance", Second Edition, *Taylor&Francis Group*, London, 2002.

[11] Devonshire, J. M., Sayer, J. R., "The effects of infrared-reflective and antireflective glazing on thermal comfort", *University of Michigan*, March 2002.

[12] *** ISO 7726, "Ergonomics of the thermal environment — instruments for measuring Physical quantities", Geneva, 1998.

[13] Temming, J., "Comfort requirements for heating, ventilation and air conditioning in motor vehicle", in *International Conference on Ergonomics and Transport*, Wales, 1980.

[14] Huizenga, C., et al., "Window performance for human thermal comfort", in *National Fenestration Rating Council, University of California, Berkeley*, 2006.

[15] Zhang, H., et al., "Modeling thermal comfort in stratified environments", *Center for Environmental Design Research, University of California at Berkeley*, 2005.
[16] Olesen, B. W., Rosendahl, J., "Thermal comfort in trucks", in *SAE Technical Paper Series*, No. 905050, Warrendale, PA, pp. 349-355, 1990.
[17] Roles, F., et al, "The Effect of Air Movement and Temperature on the Thermal Sensations of Sedentary Man", in *ASHRAE Transactions*, No. 80 (1), pp. 101 – 119, 1974.
[18] Scheatzle, D., et al., "Extending the Summer Comfort Envelope with Ceiling, Fans in Hot, Arid Climates", in *ASHRAE Transactions*, No. 95 (1), pp.169 – 280, 1989.
[19] Tanabe, S., Kimura, K., "Thermal comfort requirements under hot and humid conditions", *Proc. - ASHRAE Far East Conf. on Air conditioning in hot climates, Singapore*, 1987.
[20] Fountain, M. E., Arens, E. A., "Air Movement and Thermal Comfort", *in ASHRAE Journal,* pp. 26 – 30, Aug. 1993.
[21] Temming, J., "Comfort requirements for heating, ventilation and air conditioning in motor vehicles", in *Human Factors in Transport Research vol. 2 – User Factors: Comfort, the Environment and Behaviour*, Ed. Oborne & Levis Academic Press, London, 1980.
[22] Rugh, J. P., Bharathan, D., "Predicting human thermal comfort in automobiles", in *Renewable Vehicle Thermal Management Systems Conference, Toronto*, May 2005.
[23] ***, ISO 14505-3, "Ergonomics of the thermal environment - Evaluation of thermal comfort using human subjects", Geneva, 2006.
[24] Tadakatsu, O., "Environmental ergonomics: The ergonomics of human comfort, health, and performance in the thermal environment", Elsevier Ergonomics Book Series, Vol. 3, 2005.
[25] ***, "Vehicle Thermal Management Systems – VTMS 6", *Institution of Mechanical Engineers, Ed. John Wiley and Sons*, 2003.
[26] ***, ISO DIS 9920 "Ergonomics of the thermal environment - Estimation of the thermal insulation and evaporative resistance of a clothing ensemble", Geneva, 2004.
[27] ***, ISO 8996 "Ergonomics of the thermal environment - Determination of metabolic rate", Geneva, 2004.
[28] Parsons, K. C., "Introduction to thermal comfort standards", in *Congress Moving Thermal Comfort Standards into the 21st Century Conference,* Vol. 34, No. 6, pp. 537-548, 2002.
[29] Fanger, P. O., "Thermal Comfort: Analysis and Applications in Environmental Engineering", *McGraw-Hill, Co.*, New York, 1970.
[30] Fanger, P.O." Thermal Comfort", *McGraw-Hill Co*., New York, 1973.
[31] Charles, K. E., "Fanger's Thermal Comfort and Draught Models", *Institute for Research in Construction*, Report RR-162, Ottawa, Oct. 2003.
[32] *** ISO 7730, "Moderate thermal environments—Determination of the PMV and PPD indices and specification of the conditions for thermal comfort", Geneva, 1994.
[33] Wyon, D. P., et al., "A new method for the detailed assessment of heat balances in vehicles–Volvo's thermal manikin", in *SAE Paper*, No. 850042, Warrendale, PA, 1985.
[34] Rolle, C. D., et al., "Real evaluation of thermal comfort in a car passenger compartment", in *SAE paper*, No. 925176, Warrendale, PA, pp. 69-73, 1992.
[35] Madsen, T. L., et al., "New methods for evaluation of the thermal comfort in automotive vehicles", in *ASHRAE Transactions*, Vol. 92, pp. 38-54, 1986.
[36] Matsunaga, K., et al., "Evaluation of comfort of thermal environment in vehicle occupant compartment", in *JSAE Review*, Vol. 18, pp. 57-82, 1997.
[37] McGuffin, R., Burke, R., "Human Thermal Comfort Model and Manikin", in *Renewable Energy Laboratory, Society of Automotive Engineers, Colorado*, 2002.
[38] Nielssen, O. H., "Comfort Climate Evaluation with Thermal Manikin Methods and Computer Simulation Models", in *Royal Institute of Technology, Stockholm*, Sweden, 2004.

ACKNOWLEDGEMENT

A majority of the papers was presented at the *Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (MACRo2009), Tîrgu Mureş, Romania, March 20-21, 2009*, organized by Sapientia University.

The Editors would like to acknowledge the contributions of all who coordinated and performed the double-blind peer reviews of manuscripts submitted to the journal. Besides the members of the Editorial Board, the following researchers made significant efforts to complete this process:

Review coordinators and peer reviewers:

Marian ALEXANDRU
László BAKÓ
András BORSOS
László DÁVID
András GULYÁS
Katalin GYÖRGY
Zalán HESZBERGER
András KAKUCS
Dénes Nimród KUTASI
László Ferenc MÁRTON
Lőrinc MÁRTON
László SZABÓ
László SZILÁGYI
Ferenc TÓTH
Tamás VAJDA

Linguistic review:

Attila IMRE

Acta Universitatis Sapientiae

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

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering

Sapientia University

Scientia Publishing House

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

Information for authors

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

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

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

The paper must be submitted both in MSWord document and PDF format. The submitted PDF document is used as reference. The camera-ready journal is prepared in PDF format by the editors. In order to reduce subsequent changes of aspect to minimum, an accurate formatting is required. The paper should be prepared on A4 paper (210 x 297 mm) and it must contain an abstract not exceeding 100 words.

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

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

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

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

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

Publication supported by