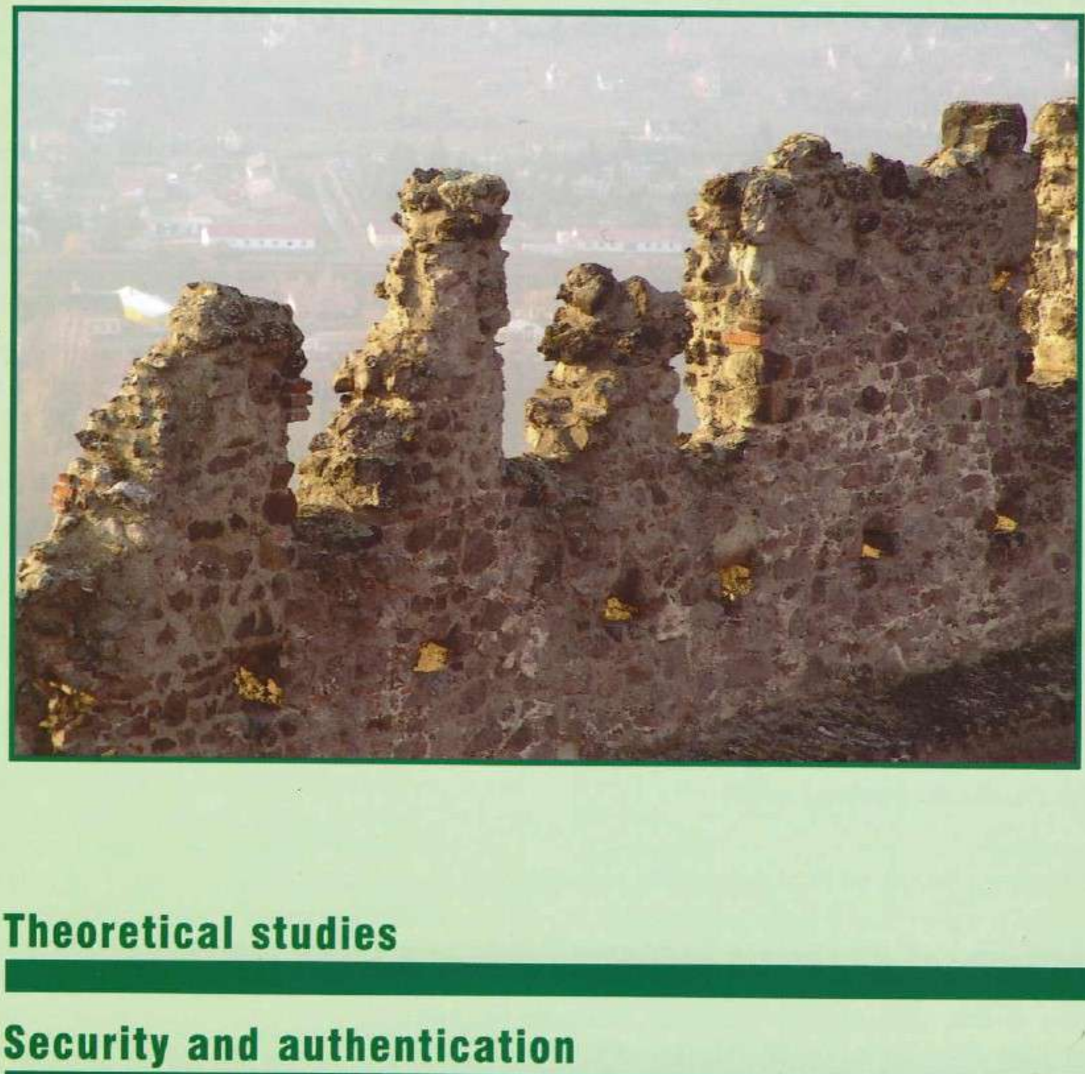
híradástechnika

VOLUME LIX.

2004/12

Selected Papers



Theoretical studies

Security and authentication

Management, Protocols and Transmission problems

Scientific Association for Infocommunications

Contents

THEORETICAL STUDIES

SECURITY AND AUTHENTICATION

MANAGEMENT

PROTOCOLS

TRANSMISSION PROBLEMS

Cover: The CASTLE OF VISEGRAD demonstrates that a solution can be succesful if it is based on an existing technical or natural creation.

Foreword
New results based on existing environment

lajtha.gyorgy@ln.matav.hu

Similarly to the earlier praxis we selected the papers from the last five Hungarian issues. To the proper selection it is necessary to define a clear concept. But this concept can be improved in every year, and novel directives can arise. Therefore primarily we describe what ideas led us to collect the papers for the English edition.

Generally the conventional way of selection is to choose the definitely high level papers. We wouldn't like to neglect this instruction. They can mostly interesting for our reader abroad. Namely our goal is to support the world-wide acknowledgment of our authors.

The second view point was to publish ideas which can influence the future trends in telecommunication. This type of research and development are usually only a specific part of a long-range process. A development can be only successful if it can make use of the existing equipments, methods and infrastructure. These is particularly important in services like telecommunications which are using a tremendous, valuable infrastructure. The utilization of the invested money is the precondition of the success of a novel method.

Last but not least we had to guess the professional use of a scientific result. If we estimate that the theoretically high level statement will have economic success then it has a high priority in the selection. Her we take in account whether the system can use the earlier investments and the new method should produce seamless co-operation with the existing network. The comfort of the user requires that the applicable future and existing subscriber terminals shouldn't differ.

In this issue of the journal the first three studies have nice theoretical importance. The first one is scrutinizing the properties of the speech. Its result is preparing and supporting the artificial speech research. The second one derives the solution of the Maxwell equations in the case impulse excitation. It can support the planning of digital radio systems. The author is using really high level mathematical instruments. The third paper is calculating the positioning errors using GPS.

The following two paper deals with authentication and security problems. Here the safest method is if we use one or more of the personal biometric identification. On other study describes the problem of anonymity accompanied with safety of transmitted information. The results are in close correlation with some practical tasks.

This two leading topics will be followed by two studies dealing with management problems. Networking is common in the next two contributions. One of them is discussing the problem of a special satellite test and control. In this case the round trip delay is more than some hour. Here sometimes local decision must be applied which will be followed later by the instructions coming from the terrestrial headquarter sent to the satellite. There is one study discussing the problem of optical communications and an other one the economic situation of telecom industry.

Finally we introduce some results achieved on multimedia and video transport services. We are not sure that they are the most actual problems, our purpose is only to demonstrate the results on technical fields of broad interest. They are representative example of the R&D results of the Hungarian Academic and industrial researchers.

Here we would like to draw the attention to the new tendencies, and we hope they will be introduced in praxis in the next future. We hope also that the broadband optical transmission combined with the mobile switching will offer a world-wide ubiquitous network which is completely service-provider independent. We would be happy if we will receive some up-to-date results from our reader. Our periodical is going to inform the telecom society about all the new recognition in this challenging period of telecommunications.

Dr. György Lajtha

Speech *F*0 estimation with enhanced voiced-unvoiced classification

TAMÁS BÁRDI

Department of Information Technology, Péter Pázmány Catholic University
bardi.tamas@itk.ppke.hu

Keywords: voicing determination algorithms, low-pass filter, basic parameter extraction

Pitch detectors for speech signal can only work correctly if the fundamental frequency estimation is linked with a reliable voiced-unvoiced decision. A pitch detection algorithm is presented with an enhanced voicing detection method, which gives less error rate than concurrent methods. This pitch detector is based on the well-known autocorrelation method with some modification. The robustness of the algorithm on voicing decision was evaluated over a database of speech recorded together with a laryngograph signal.

Although modern pitch perception models state that the subjective pitch of a sound is not always one to one relation with its fundamental frequency (*F*0), in speech signal processing *F*0 estimators are commonly known as *pitch detection algorithms* or PDA, pitch and *F*0 are treated often as synonyms. A reliably estimated pitch contour of a speech waveform can be useful for a wide range of application. Speech *F*0 variations plays important role in prosody analysis such as discriminating statements and questions. Automatic speech recognition in tonal languages such as mandarin Chinese or Vietnamese also needs a good pitch detector.

Many pitch determination methods have been proposed [10] in the literature and the most comprehensive review is that of Hess [7]. Most of them are moderate in performance but there are some outstanding. For example Bagshaw's eSRPD method [3,4] estimates *F*0 with less than 1% gross frequency error where voiced excitation exists in speech. But it detects the presence or absence of a voiced excitation with 3-4% error.

It is common in speech sciences that linguistically meaningful pitch can exist only where voicing exists. Hence the solution of voicing problem is a premise for the solution of the pitch determination problem. Voiced/unvoiced (V/UV) distinction is a must for speech recognition, since there are words differing from each other only in a voiced or unvoiced consonant, for example 'too' and 'do'.

Voicing determination algorithms (VDA) can be realized implicitly as a part of the PDA but also as a standing alone application. Several VDA has been proposed in the literature [7,12], some of them deserve attention unsparing theoretical invention and but mostly with not a persuasive performance. The rate of V/UV errors is usually higher than the *F*0 estimation error rates in PDAs. Atal and Rabiner presented a multi-parameter solution based on pattern recognition approach [5,6,7]. It gives 4% decision error but it solves a stronger task namely the voiced/unvoiced/silent (V/U/S) classification instead of voiced/unvoiced (V/UV) decision.

The present paper introduces an enhanced method for voicing detection built in a PDA. Our algorithm is based on the well-known Autocorrelation Function (ACF). Using our VDA the decision error falls nearly to 2%. Using Fast Fourier Transform to compute ACF our algorithms can be implemented with less than 2 megaflop per second computational cost assuming 8 kHz as sampling frequency.

Next sections of this paper follow the modular structure of the algorithm. Section 2 describes our unique preprocessor. It was designed above all to help V/UV distinction, and it plays an important role in achieving the error rate mentioned. After preprocessing typically 30-50 ms long windows of speech are sent to the basic extractor, which is described in section 3. This part computes the ACF and extracts parameters for V/UV decision and *F*0 estimation from it. We use there a special trick namely the "skeletonization" to reduce '*F*0 on the upper limit' type estimation errors.

Our very simple but efficient built-in VDA is in section 4. V/UV decision is based on two parameters, they are compared with thresholds. This two threshold method is essential attaching that good decision error rate. In the literature generally PDAs involve postprocessor which smoothes pitch contours. We do not apply postprocessor now, because this paper focuses only to the reliability of the voicing determination.

1. Preprocessing of speech signal

The usual realization of a PDA is subdivided into three main building blocks: 1) preprocessor, 2) basic extractor, and 3) postprocessor. The main task of the preprocessor is increasing the ease of pitch extraction or voicing determination.

The basic extractor normally works on 20-50 ms long windows of the speech signal. But distinguishing between the steps of preprocessing and basic extraction has just formal importance very often. When windowing

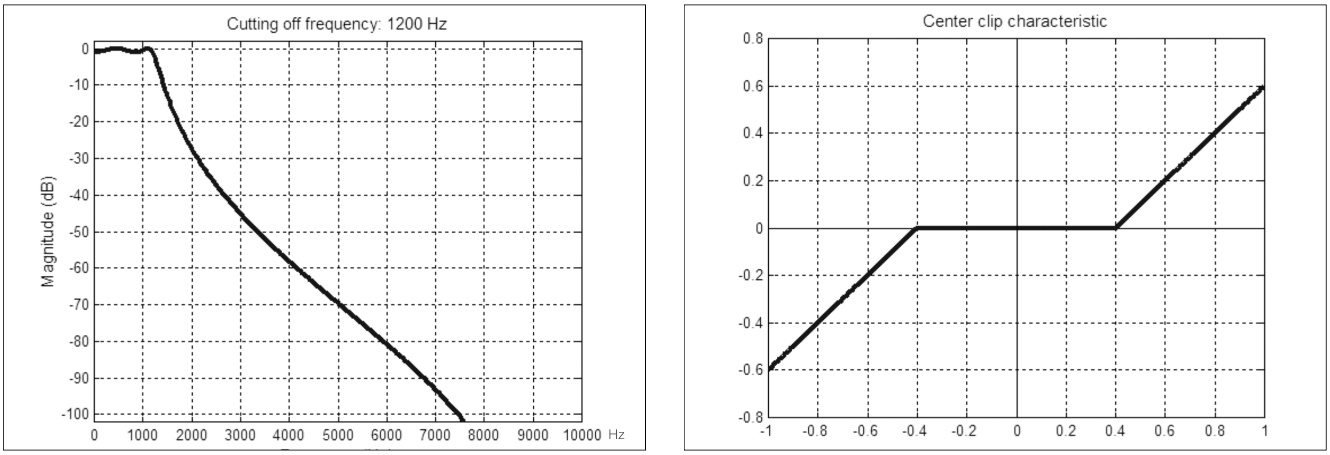


Fig. 1. Characteristics of low-pass filter and center clip applied in preprocessor

step precedes the preprocessor in the execution order of the algorithm we can not examine their work really separately and a lot of computations are duplicated if the windows are overlapped. Windowing before preprocessing makes impossible to listen by ear the output of the preprocessor connectedly. In contrast with that we suggest running the preprocessor on the complete speech signal, after then taking out windows form the output signal and sending them to the basic extractor. In this case we can make sensible an inner state of the algorithm. Creating sensual checkpoints inside a complicated speech processing system can help to optimize its parameters empirically. Our preprocessor is partly “optimized” by ear: fine tuning it we adjusted some parameters until we felt that the output sounds good.

In our preprocessor we use low-pass filter and center clipping. Those are both common in the literature of PDAs [6,9,11]. The characteristics of low-pass filter (Chebishev I type) and center clip used in our method are shown in *Fig. 1.*

The technique of adaptive center clipping applies time-varying clipping level which is adjusted according to the signal amplitude. Generally the varying clipping level is a fixed percentage of an envelope of the speech signal computed some way. The original innovation in our method that it combines the two step: the amplitude envelope is computed from the original speech signal and the low-pass filtered signal is center-clipped with 40% of the envelope. This method removes almost everywhere the speech segments with clearly stochastic excitation such as voiceless consonants. The output signal becomes zero where the low frequency component of the input signal represents the rate of total energy not high enough.

Fig. 2, 3 and *4* show the work of our preprocessor.

Fig. 2.
The original speech with its envelope and the low-pass filtered signal.

Fig. 3.
Low-pass filtered speech and the computed center clip level.

Fig. 4.
Speech signal before and after preprocessing.

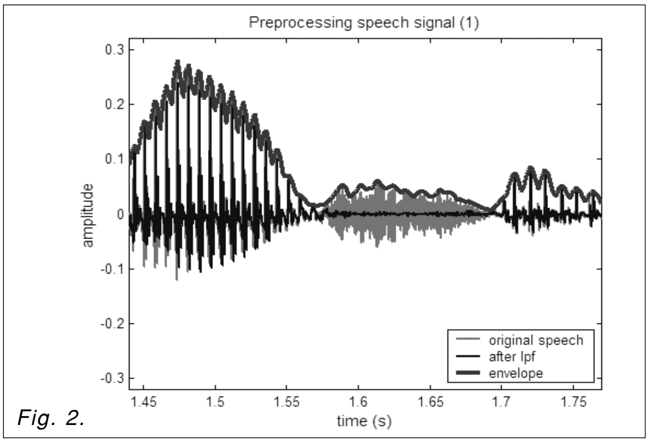


Fig. 2.

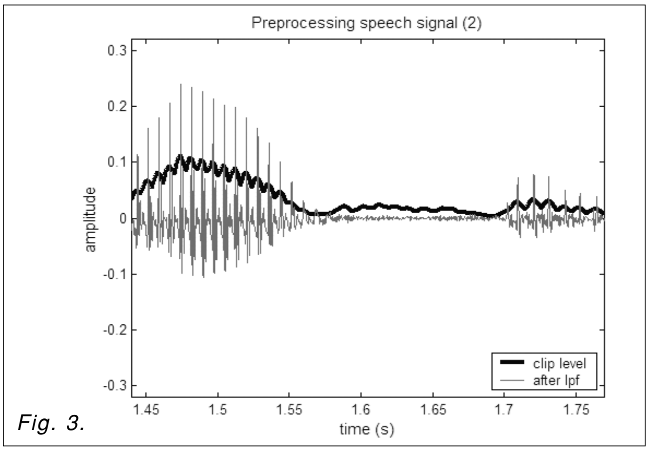


Fig. 3.

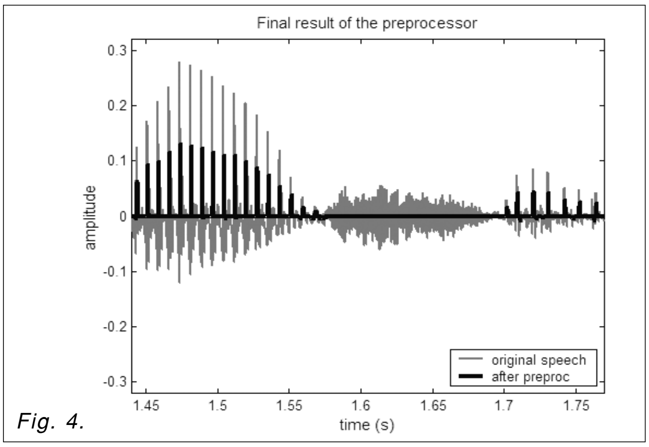


Fig. 4.

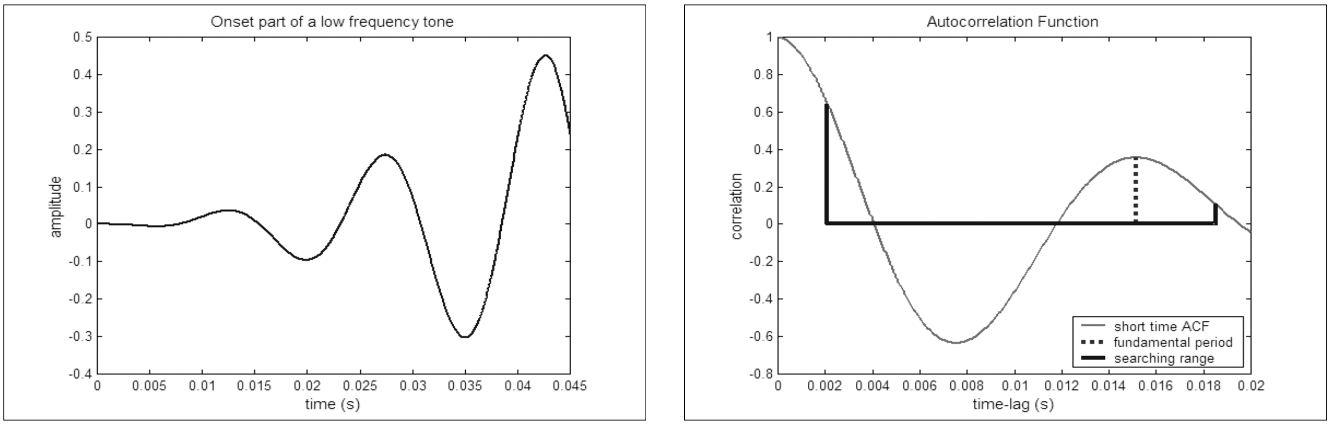


Fig. 5. Onset part of a low frequency tone (67 Hz) and its autocorrelation. The value of ACF at the fundamental period is lower than at the limit of the searching range.

2. Basic parameter extraction

This part of our PDA computes the Autocorrelation Function of the actual signal window and then the algorithm searches for the "best" local maximum of the ACF. The value of the selected peak serves as the main voicing decision parameter and its time lag is the estimation of the fundamental period. But how could we find the "best" peak? As you can see below, the "best" maximum is not so far definitely the global one.

First of all note that all in our formulas the time related symbols (τ, t, u, W) are meant in seconds and signals are meant continuous in time and amplitude hence we use integrals instead of sums. Signal amplitude is meant as the rate of maximal amplitude that can be processed in the system, so that $-1.0 \le x(t) \le 1.0$. These notations make our discussion independent of sampling frequency and bit-rate. Our integral type formulas can easily be converted to sums for concrete applications when sampling frequency and bit-rate are known.

Instead of the biased definition of ACF, which is common in signal processing, we use the unbiased definition, and we apply artificial biasing on it. (W denotes the window length; we set it to 32 ms for this investigation.)

$$r_t(\tau) = \frac{\int_{t-W/2}^{t+W/2} x(u)x(u-\tau)\,du}{\int_{t-W/2}^{t+W/2} x(u)^2\,du} \qquad (\tau,\ t,\ u,\ W \text{ in sec}) \quad (1)$$

and the artificial biasing (its degree can be tuned through the *gr* coefficient):

$$r_t^{biased}(\tau) = r_t(\tau) \cdot (1 - gr \cdot \tau) \qquad (2)$$

Computing ACF on the biased way it shows shrinking with increasing values of τ, which gives gain for the fundamental period against its multiples. Although this shrinking can be useful and attractive, its rate can only be tuned by adjusting W. De Cheveigné suggests [5] computing ACF on the unbiased way using fixed window length for all τ time lags and after then applying artificial bias on it. That enables us to adjust the rate of shrinking and window length independently.

For the onset part of a low frequency voice the maximum of the ACF frequently occur at the limit of the searching range. This phenomenon causes the "*F*0 on the upper limit" type errors. That can be seen in *Fig. 5.*

To avoid this sort of error we suggest using the skeleton function or "fishbone" method in other words. The skeleton of a function takes the values of the original function at its local maxima or minima and takes zero otherwise. For our purposes the most suitable definition of local extreme is an intermediate level one between the strict and non-strict version. *Fig. 6.* shows how we do mean local extreme.

Definition: $f: \mathbf{R} \to \mathbf{R}$ real function has local extreme at x, if f not stricly monoton and not plain at x.

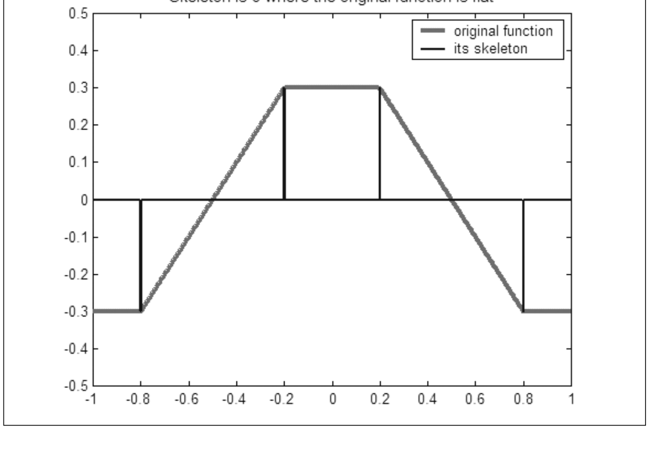
Definition: $g = skeleton(f)$ if and only if

$$g(x) = \begin{cases} f(x) & \text{if } f \text{ has local extreme at } x \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

Despite the artificial bias, for the release part of a clear voiced sound ACF tends to have higher peaks with increasing time lags as it can be seen in *Fig. 7.*

This symptom occurs only if the ACF is almost one or even greater than one at the fundamental period. We suggest applying a so called preference level to avoid this problem. Then our algorithm picks the first

Fig. 6. Skeleton function takes 0 where its original is plain.



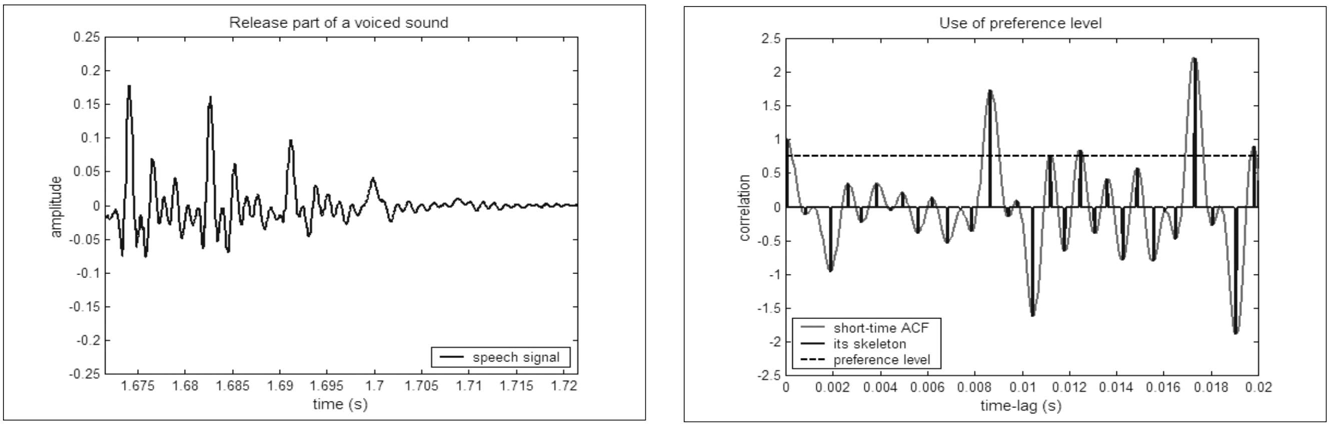


Fig. 7. Release part of a voiced sound and its autocorrelation.

peak that exceeds the preference level. If there is no peak exceeds the preference level the highest peak is chosen. We used 0.75 as preference level chosen it empirically.

Summarizing our basic parameter extractor now we list the algorithm's steps in the correct order:

Step 1: Compute unbiased ACF as in Eq (1).

Step 2: Skeletonization: $sr_t(\tau) = skeleton(r_t(\tau))$.

Step 3: Constrain the $F0$ searching range:

Let $[F0_{\min};F0_{\max}]$ the searching interval,

$$srl_t(\tau) = \begin{cases} -0.5 & \text{if } \tau < 1/F0_{\max} \\ sr_t(\tau) & \text{if } 1/F0_{\max} \le \tau \le 1/F0_{\min} \\ -0.5 & \text{if } \tau > 1/F0_{\min} \end{cases} \quad (4)$$

Step 4: Bias the skeleton:

$$srl_t^{biased}(\tau) = (1 - gr\cdot\tau)\cdot srl_t(\tau) \quad \text{with } gr = 1.75 \quad (5)$$

Step 5: $F0$ estimation:

Step 5/A: Applying the preference level:

$$\tau^* = \min\{\tau : srl_t^{biased}(\tau) \ge 0.75\} \quad (6)$$

Step 5/B:

If *5/A* did not succeed choose the highest peak:

$$\tau^* = \arg\max\{srl_t^{biased}(\tau)\} \quad (7)$$

and the estimated fundamental frequency:

$$F0^* = 1/\tau^*. \quad (8)$$

Step 6: Get voicing decision parameter:

$rm_t = srl_t(\tau^*)$ from the unbiased skeleton (9)

Fig. 8. shows an example for the algorithms work.

3. Voiced-unvoiced decision

Our VDA uses rm_t (9) as decision parameter and the logarithm of the signal energy on the analyzed window:

$$p_t = 10\cdot\log_{10}\left(\frac{1}{W}\int_{t-W/2}^{t+W/2} x(u)^2\,du\right) \quad \text{(dB)} \quad (10)$$

Consequently from the definition:

p_t=0 dB for a full-scaled square wave.

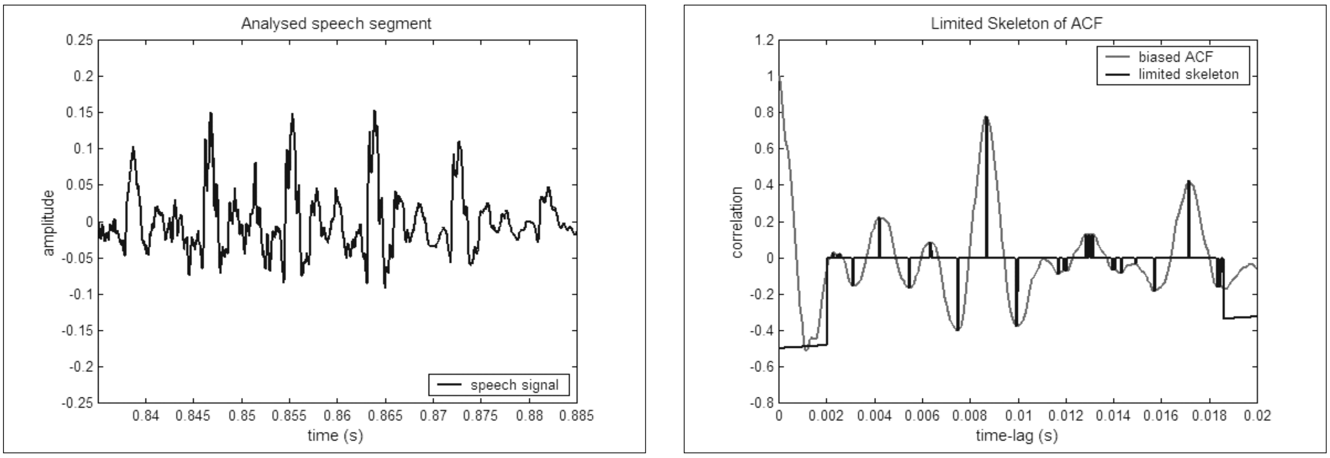
Parameters are compared with threshold so the voicing indicator function is:

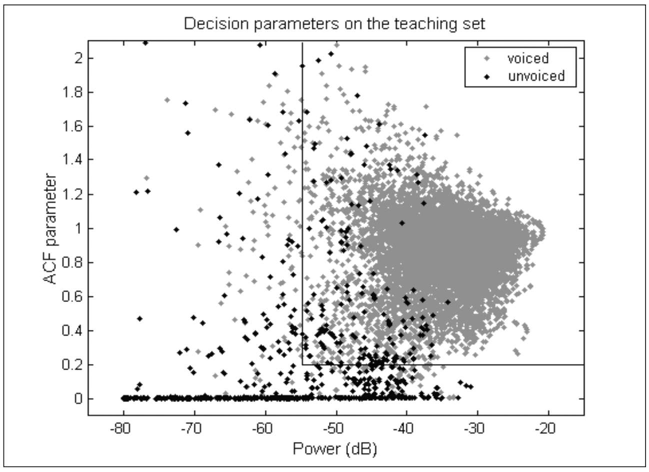
$$voicing(t) = \begin{cases} 1 & \text{if } (rm_t > rmth)\,\&\,(p_t > pth) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Where *pth* and *rmth* are the thresholds.

And now the only question is where to put these thresholds. Tuning procedure of thresholds is linked with the evaluation of voicing decision error rate. We divided the evaluation speech database into 2 parts: 1st half is the teaching set and 2nd half is the control set. The teaching set is for optimizing thresholds on it and the control set is for evaluating our VDA with the optimized thresholds. This evaluation method is correct only if the control set is disjoint from the teaching set. Partitioning both male and female speech into the teaching and control sets provides the maximum speaker independency of the optimization.

Fig. 8. The global maximum of srl(?) shows the fundamental period of the speech window.





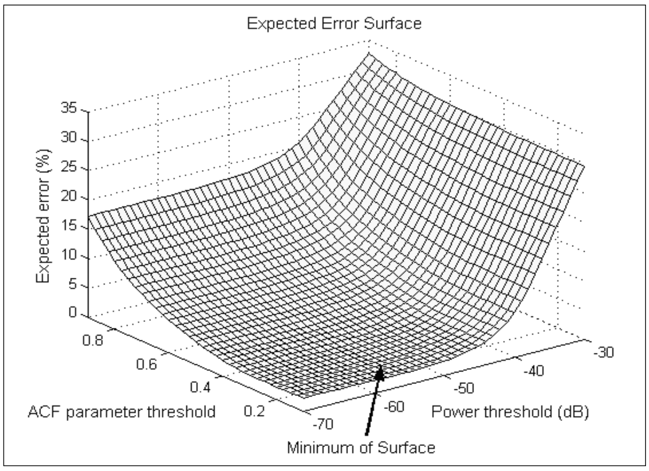


Fig. 9/a. Distribution of decision parameters. *Fig. 9/b. Expected error surface.*

Voicing decision parameters were extracted using *W*=32 ms window length and *F*0 searching range was between 55 and 480 Hz. *Fig. 9/a.* shows their distribution on the teaching set. Light points come from the voiced segments and dark points come from voiceless segments. The two perpendicular lines depict the two threshold classification method. As it can be seen they do not separate the voiced and voiceless sets perfectly.

Expected Error Surface can be derived from the distribution as a function of threshold pairs. The value of the surface at *(x,y)* represents the voicing decision error on the teaching set itself choosing *(x,y)* as thresholds. Minima of surfaces represent the optimal threshold. *Fig. 9/b.* shows the surface.

Optimized thresholds are: *pth* = –55.2 *dB* and *rmth* = 0.23. The value of error surface at that point is 1.95%. Applying these thresholds on the control set the error rate is *2.13%.* This error rate is the tested performance of our algorithm.

4. Summary

Surveying our algorithms we think that three original trick help us to achieve the 2.13% error rate. First is the combination of low-pass filter and center clip in preprocessor, the second is using skeleton in the basic extractor and the third is considering signal energy in voicing determination. The signal energy indicates voicing much more significantly after preprocessing than before. Precise formulation and correct execution order are also essential.

5. Evaluation database

Our algorithms were tested on the Fundamental Frequency Determination Algorithm (FDA) Evaluation Database recorded at University of Edinburgh, Centre for Speech Technology Research and authored by Paul Bagshaw.

This database is available via ftp from the URL: http://www.cstr.ed.ac.uk/˜pcb/fda-eval.tar.gz

It contains 0.12 h speech, 50 English sentences each spoken by one male and one female speakers. 37% out of the total time are voiced segments and 63% are voiceless (silent and unvoiced consonants together). Synchronously with speech signal laryngograph signal was also recorded, which was the basis of labeling voiced and unvoiced segments.

Acknowledgements

The author would like to thank to Dr. György Takács for his instruction and help, to the leaders of the Doctoral School of Department of Information Technology, Péter Pázmány Catholic University for their support and trust, and to Dr. György Lajtha for his help.

References

[1] B. S. Atal and L. R. Rabiner
"A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition"
IEEE Trans. Acoust., Speech, Signal Processing, - Vol. ASSP-24, pp.201–212, 1976.

[2] B. S. Atal and L. R. Rabiner
"Voiced-unvoice decision without pitch detection"
J Acoust. Soc. Am., Vol.58., 1975.

[3] P. C. Bagshaw Automatic prosodic analysis for computer aided pronunciation teaching PhD Thesis, Univ. Edinburgh, 1994.

[4] P. C. Bagshaw, S. M. Hiller and M. A. Jack
"Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching"
Proc. 3rd European Conf. on Speech Comm. and Technology, Vol.2., pp.1003–1006, Berlin, 1993.

[5] A. de Cheveigné and H. Kawahara
"YIN, a fundamental frequency estimator for speech and music"
J Acoust. Soc. Am., Vol.111., Apr 2002.

[6] J. R. Deller, J. H. L. Hansen and J. G. Proakis
Discrete-Time Processing of Speech Signals,
Macmillan, New York, 1993.

[7] W. A. Hess:
Pitch Determination of Speech Signals,
Berlin, Springer-Verlag, 1983.

[8] L. R. Rabiner:
"Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone quality speech"
Bell Syst. Tech. Journal, Vol.56, pp.455–482, 1977.

[9] L. R. Rabiner:
"On the Use of Autocorrelation Analysis for Pitch Detection" IEEE Trans. Acoust., Speech, Signal Processing,
Vol. ASSP-25, pp.24–33, 1977.

[10] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal:
"A Comparative Performance Study of Several Pitch Detection Algorithms"
IEEE Trans. Acoust., Speech, Signal Processing,
Vol. ASSP-24, pp.399–418, 1976.

[11] L. R. Rabiner and R. W. Schafer:
Digital Processing of Speech Signals,
Prentice Hall, Engelwood Cliffs NJ, 1978.

[12] L. S. Smith:
"A Neurally Motivated Technique for Voicing Decision and F0 Estimation for Speech"
Centre for Cognitive and Computational Neuroscience, Tech. Report, Vol. CCCN-22,
University Stirling, Scotland, 1996.

News

At Supercomm, **ECI Telecom** will display the latest enhancements to ist XDM platform that make it the first multi-service provisioning platform (MSPP) to fully integrate CWDM and DWDM. Fully integrating CWDM, LDWDM, Ethernet and SONET onto the same platform provides carriers and service providers with a solution that seamlessly connects and manages an optical network from the metro edge to the regional core with simplified operations, reduced costs and ent-to-end performance monitoring. The XDM converged platform with end-to-end multi-layer management allows for seamless traffic connectivity from subtending CWDM or SONET edge rings to metro core DWDM rings with performance monitoring for all services. Additionally, the XDM Build as you Grow architecture enables more freedom to choose the right technology for each service and application.

Veraz Networks will demonstrate on-the-fly services creation, customization, and services management for providers and their customers. The demonstration will highlight the ability to actually create, customize, deploy, and provision services II without the need for new software releases. This on-the-fly automation reduces the time traditionally required to take services from concept to revenue generation from months or years to hours. The demonstration will introduce Veraz´s built-in service management capabilities. It will demonstrate how providers can create services, group services into service bundles, provision and update service bundles. The service bundles can be made available hierarchically. With Veraz´s solution, service providers can create and customize new services for customers instantly to be able to meet and respond to individual needs faster than ever before.

Mr. Houlin Zhao, Director of ITU´s Telecommunication Standardization Bureau (TSB) commended the Chairman for his leadership, his ability to steer the work of the Assembly to a successful conclusion and for having achieved sound results consensually. „We agreed new tools, resolutions, decisions and guidelines that will make ITU-T more efficient and much stronger.“ – Zhao told delegates.
The main highlights of the Assembly include:

– A next-generation networks (NGN) focus spanning the work programme of all study groups
– The creation of a new Study Group on NGN
– The adoption of new resolutions on Internet-related issues (ENUM, spam, internationalized domain names, country code top level domain (ccTLD) names)
– The adoption of a resolution on cybersecurity
– The adoption of measures aimed at enhancing a greater involvement of developing countries in standardization activities
– A group to oversee the sector´s seminar and workshop programme and to monitor the market for new topic areas
– The inclusion of a gender perspective in the work of the ITU-T with the adoption of a resolution on gender mainstreaming

The setting up of 13 Study Groups with their areas of responsibility and the designation of their chairmen and vice-chairmen. WTSA also designated the chairman and vice-chairmen of the telecommunication standardization advisory group (TSAG). A request for a study on the economic effect of call-back and other similar calling practices in developing countries an how they impact on their ability to develop their telecommunication networks and services.

Short impulse-propagation in inhomogeneous plasma

ORSOLYA E. FERENCZ

Space Research Group, Eötvös University, Geophysical Department
spacerg@sas.elte.hu

Keywords: Maxwell's equations, Method of Inhomogeneous Basic Modes (MIBM), wave propagation

In this paper the problem of real impulse-propagation in arbitrarily inhomogeneous media will be presented on a fundamentally new, general, theoretical way. The general problem of wave-propagation of monochromatic signals in inhomogeneous media was enlightened in [1]. The former theoretical models for spatial inhomogeneities have some errors regarding the structure of the resultant signal originated from backward and forward propagating parts. The application of the Method of Inhomogeneous Basic Modes (MIBM) and the complete full-wave solution of arbitrarily shaped non-monochromatic plane-waves in plasmas made it possible to obtain a better description of the problem, on a fully analytical way, directly from Maxwell's equations. The model investigated in this paper is inhomogeneous of arbitrary order (while the wave-pattern can exist), anisotropic (magnetized), linear, cold plasma, in which the gradient of the one-dimensional spatial inhomogeneity is parallel to the direction of propagation.

The traditional theoretical descriptions of a monochromatic electromagnetic signal propagating in an inhomogeneous medium – e.g. eikonal-equation, W.K.B. method, generalized propagation-vector, application of Airy-functions in solving the Stokes equation etc. – have a common fundamental inaccurate assumption relating the physical concept of the structure of the signal. The model of the solution in these approaches is an additional sum of the different signal-parts, propagating forward or backward (scattered), which parts are supposed solutions of Maxwell's equations independently of each other. However, this way of description does not make it possible to recognize the influence of the real energetic coupling between these signal-parts, as for the real full-wave solution of Maxwell's equations always has to contain all the existing modes simultaneously.

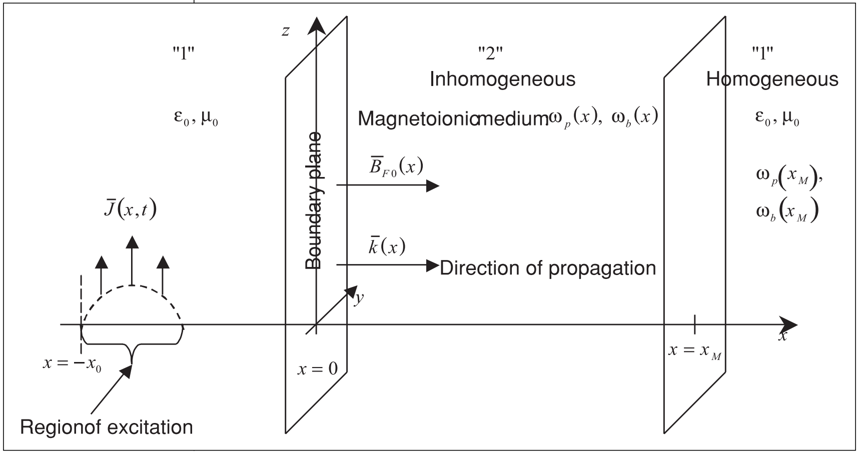
It is a well-known fact that the additional sum of solutions has to fulfill the original equations, if a linear differential equation-system has some solutions. Nevertheless, it does not mean that the parts of the solution fulfilling the equations – separated from each other by application of different theoretical points of view – would automatically be solutions of the full equation-system.

The resultant and existing signal is a solution of Maxwell's equations, but its parts (the backward and forward propagating signal-parts) are not.

Moreover, as it can be seen in the coupled W.K.B. philosophy, the influence of the reflection from the spatial inhomogeneity is neglected in the assumed signal-form in the traditional ways of thinking. The reflected signal-form is created on the same way – as if an independent mirror-source would exist –, but the coupling is supposed to be determined by a simple addition of these signals, although this influence was eliminated in their creation. A correct mathematical analysis of this situation will yield a contradiction for the resultant field, because the curl-equations become automatically over-determined neglecting the energetic coupling between the different signal parts.

A further theoretical problem, unanswered up to now, is originated from the fact, that a real physical signal excited by an impulse -– or switching on-off transient – is always non-monochromatic and its description is not enough accurate by superposition of monochromatic signals [2,3,4,5]. With other words, the problem of arbitrarily shaped signals does not make it possible to assume any exp(jωt)-type starting form of the solution at the beginning of the derivation of Maxwell's equations.

Fig. 1.
Model for longitudinal propagation.
Medium "1" is homogenous and "2" is inhomogeneous.



1. Derivation of the full-wave solution

The applied model and method

The example, by that the theoretical solution will be presented here, is a linear, time-invariant, anisotropic, cold plasma. The source is an arbitrarily shaped plane-wave (e.g. an "impulse plane"), the direction of the propagation and the gradient of the one-dimensional inhomogeneity are parallel to the superimposing magnetic field. The order of the magnitude of the inhomogeneity is not restricted, except the precondition that the wave front can be defined *(Fig. 1.)*

As it was briefly mentioned above, the main question is how to handle the continuous generation of the reflected (scattered) signal and the energetic coupling between the existing signal-parts during the propagation.

The Method of Inhomogeneous Basic Modes (MIBM) is well applicable for such problems. The philosophy of this method [6] considers the form of the final solution to be determined as a sum of the so-called basic modes, which are not the full solutions of Maxwell's equations independently in themselves.

$$\overline{G}(\overline{r},t)=\sum_i G_i(\overline{r},t) \tag{1}$$

where G = *E, D, H, B* and *i* is the number of the existing modes.

Substituting these basic modes into Maxwell's equations those can be disintegrated into two groups. One of them is valid even in a simple homogeneous medium; the other (the group of so-called coupling equations) characterizes the influence of the inhomogeneous medium. Boundary conditions remain as unknown variables in the full form of Maxwell's equations.

The final form of the solution can be determined by solving the coupling equations and describing these initial values.

Definition of the basic modes and the boundary conditions

For a suitable choice of the inhomogeneous basic modes it is necessary to take into consideration, that the solution must lead back to the known one valid in a homogeneous medium. (With other words, the solution for inhomogeneous case has to be a mathematical generalization of the homogeneous results.)

As a first step, the form of the solution excited by an arbitrarily shaped non-monochromatic plane-wave will be presented in homogeneous plasma.

The detailed mathematical derivation in the case of different plasma models can be found in [7,8]. If the signal to be investigated is non-harmonic, the exp(jωt) form or its superposition cannot be applied during the derivation of Maxwell's equations. The form to be determined remains open up to the final steps.

Apart from the detailed presentation (that would lead far beyond the scope of this work) the starting equations are

$$\begin{aligned}
&\overline{\nabla}' \; \overline{H}_2 = \overline{J}+\varepsilon_0\frac{\partial \overline{E}_2}{\partial t}, & & m\frac{\partial \overline{v}}{\partial t}=q\left(\overline{E}+\overline{v}\times\overline{B}_{F0}\right),\\
& & & \overline{J}=q\,N\,\overline{v}, \\
&\overline{\nabla}' \; \overline{E}_2 = -\,\mu_0\frac{\partial \overline{H}_2}{\partial t}, & & \overline{\nabla}\bullet\overline{J}+\frac{\partial\rho}{\partial t}=0,
\end{aligned} \tag{2}$$

where ε_0 and μ_0 the permittivity and permeability in vacuum, respectively. The electron density of the plasma is *N*, the superimposing magnetic field is $\overline{B}_{F0}$, *m* and *q* = –*e* are the mass and charge of an electron, $\overline{v}$ is the velocity of the electrons.

The plasma- and the gyro-frequencies are

$$\omega_b=\frac{eB_{F0}}{m} \qquad \omega_p^2=\frac{q^2N}{\varepsilon_0 m}\equiv\frac{e^2N}{\varepsilon_0 m}. \tag{3}$$

By some mathematical transformation of (2), the following differential equations can be obtained for the propagating field (4)

$$\frac{\partial^2E_y}{\partial x^2}=\frac{1}{c^2}\left\{\omega_p^2\int_0^t\frac{\partial E_y}{\partial\tau}\cos\omega_b(t-\tau)\cdot d\tau-\omega_b\omega_p^2\int_0^t E_z\cos\omega_b(t-\tau)\cdot d\tau+\frac{\partial^2E_y}{\partial t^2}\right\},$$

$$\frac{\partial^2E_z}{\partial x^2}=\frac{1}{c^2}\left\{\omega_p^2\int_0^t\frac{\partial E_z}{\partial\tau}\cos\omega_b(t-\tau)\cdot d\tau+\omega_b\omega_p^2\int_0^t E_y\cos\omega_b(t-\tau)\cdot d\tau+\frac{\partial^2E_y}{\partial t^2}\right\}.$$

As the form of the solution is completely unknown, (4) is not solvable in time-space domain. Therefore, it is necessary to apply the Laplace-transformation of the equations. This transformation takes into account the transient behavior of the signal.

Using the Laplace-transformation for (4) according to time and space, the unknown field components become separable. However, the application of this transformation makes it necessary to introduce initial conditions (boundary conditions) regarding the field, which will deliver the relation between the excitation and the signal propagating in the plasma.

By a suitable choice of the model structure, only two initial conditions or their Laplace-transformed forms will remain (*s* = jω)

$$\begin{aligned}
&E_z(x=0,t)\xrightarrow[s=j\omega]{L}e_{z0t}(\omega)=A_1(\omega)\\
&\left.\frac{\partial E_z(x,t)}{\partial x}\right|_{x=0}\longrightarrow e'_{z0t}(\omega)=B_1(\omega)
\end{aligned} \tag{5}$$

Solving the transformed forms of (4), the following formulas can be obtained for the propagation factor and the spectrum of the field (6)

$$\sum_{i=2}E_{yi}(\omega)=\frac{1}{4}\left\{-\left[\frac{B_1(\omega)}{k(\omega)}-jA_1(\omega)\right]e^{-jk(\omega)x}+\left[\frac{B_2(\omega)}{k(\omega)}+jA_2(\omega)\right]e^{jk(\omega)x}\right\},$$

$$\sum_{i=2}E_{zi}(\omega)=\frac{1}{4}\left\{j\left[\frac{B_1(\omega)}{k(\omega)}-jA_1(\omega)\right]e^{-jk(\omega)x}-j\left[\frac{B_2(\omega)}{k(\omega)}+jA_2(\omega)\right]e^{jk(\omega)x}\right\},$$

where *i* = 1 is the forward propagating and *i* = 2 is the reflected signal-part, and

$$k(\omega)=\frac{1}{c}\sqrt{\frac{\omega\omega_b\omega_p^2+\omega^2\left(\omega_p^2+\omega_b^2-\omega^2\right)}{\omega_b^2-\omega^2}} \tag{7}$$

The full-wave solution from (6) in homogeneous plasma can be found in [7,8]. This paper does not deal with these details further.

The amplitude functions of (6) contain the unknown boundary conditions that represent the connection to the excitation (or with other words, the previous state of the signal).

In the definition of the inhomogeneous basic modes, (6) is well applicable with some considerations, not forgetting the fact, that the basic modes are not solutions of the problem (of Maxwell's equations) in themselves but they make it possible to obtain a full, closed form description.

As the investigated medium is inhomogeneous, the constitutional parameters depend on space, which would mean $\bar{\varepsilon}(\bar{r})$ or at least $\bar{\varepsilon}(x)$ in the case of a monochromatic signal. However, in the case of arbitrarily shaped signals it is impossible to define the closed form of the constitutional parameters, because there is no supposed sinusoidal waveform. The only open way is to give a first assumption for a space-depending form of the "propagation factor", which is not the accurate propagation factor of the inhomogeneous solution, just a first step to yield information for the real phase pattern.

A trivial form of (7) for inhomogeneous case is

$$k(x,\omega)=\frac{1}{c}\sqrt{\frac{\omega\omega_b(x)\omega_p^2(x)+\omega^2\left[\omega_p^2(x)+\omega_b^2(x)-\omega^2\right]}{\omega_b^2(x)-\omega^2}} \tag{8}$$

where

$$\omega(x)_b=\frac{eB_{F0}(x)}{m} \text{ and } \omega_p^2(x)=\frac{e^2N(x)}{\varepsilon_0 m}. \tag{9}$$

A further consideration is requested for the definition of the inhomogeneous modes.

In (6) the amplitude functions of the transformed forms contain two unknown functions that are initial conditions characteristic for the previous state of the signal. In the homogeneous case, this investigated point can be found at the boundary surface of the plasma - the entering point of the signal into the plasma. But in a spatially inhomogeneous medium the wave pattern strongly depends on the continuously varying conditions, as for the forward propagating signal excites a forward and a backward directed (reflected) signal-part in every point of the medium.

This problem can be interpreted introducing an elementarily thin "sliding boundary surface" across the plasma, as if the "entering point" of the signal-part, at which the initial conditions are valid, traveled ahead together with the propagating signal-part.

With these considerations let the inhomogeneous basic modes be defined as (10)

$$\sum_{i=2}E_{yi}(x,\omega)=\frac{1}{4}\left\{-\left[\frac{B_1(x,\omega)}{k(x,\omega)}-jA_1(x,\omega)\right]e^{-j\int k(x,\omega)dx}+\left[\frac{B_2(x,\omega)}{k(x,\omega)}+jA_2(x,\omega)\right]e^{j\int k(x,\omega)dx}\right\},$$

$$\sum_{i=2}E_{zi}(x,\omega)=\frac{1}{4}\left\{j\left[\frac{B_1(x,\omega)}{k(x,\omega)}-jA_1(x,\omega)\right]e^{-j\int k(x,\omega)dx}-j\left[\frac{B_2(x,\omega)}{k(x,\omega)}+jA_2(x,\omega)\right]e^{j\int k(x,\omega)dx}\right\},$$

$$\sum_{i=2}H_{yi}(x,\omega)=\frac{1}{4Z_0}\left\{\left[-j\frac{B_1(x,\omega)}{k_0}-\frac{k(x,\omega)A_1(x,\omega)}{k_0}\right]e^{-j\int k(x,\omega)dx}-\right.$$
$$\left.-j\left[\frac{B_2(x,\omega)}{k_0}+j\frac{k(x,\omega)A_2(x,\omega)}{k_0}\right]e^{j\int k(x,\omega)dx}\right\}, \tag{11}$$
$$\sum_{i=2}H_{zi}(x,\omega)=\frac{1}{4Z_0}\left\{\left[-\frac{B_1(x,\omega)}{k_0}+j\frac{k(x,\omega)A_1(x,\omega)}{k_0}\right]e^{-j\int k(x,\omega)dx}-\right.$$
$$\left.-\left[\frac{B_2(x,\omega)}{k_0}+j\frac{k(x,\omega)A_2(x,\omega)}{k_0}\right]e^{j\int k(x,\omega)dx}\right\}.$$

where $Z_0=\sqrt{\mu_0/\varepsilon_0}$.

Further, it must be taken into account in the investigation that the inhomogeneity is extended from $x = 0$ to $x = x_M$ spatial point, but the medium is homogeneous beyond these points ($x < 0$ and $x > x_M$). The excitation exists in the $x < 0$ half-space. Further – as this half-space is homogeneous – there is no reflected signal from $x > x_M$.

In the presence of inhomogeneity, these yield the description of the energy-coupling between the signal parts – $i = 1$ and $i = 2$ – i.e. this case; after some rearrangement we get information regarding the coupling between the propagating and reflected modes, so these are the "coupling equations".

$$\frac{\partial}{\partial x}\left[-B_1(x,\omega)+jk(x,\omega)A_1(x,\omega)\right]e^{-j\int k(x,\omega)dx}$$
$$-\frac{\partial}{\partial x}\left[B_2(x,\omega)+jk(x,\omega)A_2(x,\omega)\right]e^{j\int k(x,\omega)dx}=0,$$

$$\frac{\partial}{\partial x}\left[-j\frac{B_1(x,\omega)}{k(x,\omega)}-A_1(x,\omega)\right]e^{-j\int k(x,\omega)dx} \tag{12}$$
$$+\frac{\partial}{\partial x}\left[j\frac{B_2(x,\omega)}{k(x,\omega)}-A_2(x,\omega)\right]e^{j\int k(x,\omega)dx}=0.$$

Let the following simplified notations be introduced in the amplitudes

$$jk(x,\omega)A_1(x,\omega)-B_1(x,\omega)\overset{\Delta}{=}C_1(x,\omega),$$
$$jk(x,\omega)A_2(x,\omega)-B_2(x,\omega)\overset{\Delta}{=}C_2(x,\omega). \tag{13}$$

After some mathematical rearranging (13) results in (14)

$$\frac{\partial C_1(x,\omega)}{\partial x}=\frac{1}{2k(x,\omega)}\frac{\partial k(x,\omega)}{\partial x}\left[C_1(x,\omega)+C_2(x,\omega)e^{j2\int k(x,\omega)dx}\right],$$
$$\frac{\partial C_2(x,\omega)}{\partial x}=\frac{1}{2k(x,\omega)}\frac{\partial k(x,\omega)}{\partial x}\left[C_2(x,\omega)+C_1(x,\omega)e^{-j2\int k(x,\omega)dx}\right]$$

The solution of (14) can be obtained by successive approximation. As the first step let $C_2 = 0$ be assumed. Then

$$\frac{\partial C_1(x,\omega)}{\partial x}=\frac{1}{2k(x,\omega)}C_1(x,\omega) \tag{15}$$

further

$$\ln C_1(x,\omega)=$$
$$=\frac{1}{2}\ln k(x,\omega)+c_0(\omega) \tag{16}$$

From (16)

$$C_1(x,\omega)=C_0(\omega)\sqrt{k(x,\omega)} \tag{17}$$

Using (17) the first approximation the propagating (i = 1) field is

$$E_{z1}(x,\omega)=-\frac{j}{4}\frac{C_0(\omega)}{\sqrt{k(x,\omega)}}e^{-j\int k(x,\omega)dx} \qquad (18)$$

As it can be seen in (18), the first step of the successive approximation process yields the known form of the non-coupled W.K.B. solution for weakly inhomogeneous medium. $C_0(\omega)$ delivers the connection of the solution with the excitation. It can be determined on the way presented in details in [8].

Writing back (18) into (14)

$$\frac{\partial C_2(x,\omega)}{\partial x}-\frac{1}{2k(x,\omega)}\frac{\partial k(x,\omega)}{\partial x}C_2(x,\omega)= \qquad (19)$$
$$=\frac{C_0(\omega)}{2}\frac{1}{\sqrt{k(x,\omega)}}\frac{\partial k(x,\omega)}{\partial x}e^{-2j\int k(x,\omega)dx}$$

(19) belongs to a known type of differential equations [9], as by the following notations

$$y(x,\omega)\longleftrightarrow C_2(x,\omega),$$
$$f(x,\omega)\longleftrightarrow -\frac{1}{2k(x,\omega)}\frac{\partial k(x,\omega)}{\partial x}, \qquad (20)$$
$$g(x,\omega)\longleftrightarrow \frac{C_0(\omega)}{2}\sqrt{\frac{1}{k(x,\omega)}}\frac{\partial k(x,\omega)}{\partial x}e^{-2j\int k(x,\omega)dx}. \qquad (21)$$

(20-21) can be written as

$$y'(x,\omega)+y(x,\omega)f(x,\omega)=g(x,\omega) \qquad (22)$$

If there is a known (ξ,η) point on the (x,y) plane – or with other words a single value of the field (the solution) is known at a given point of the medium – the solution of (22) is the following

$$y=e^{-F}\left(\eta+\int_{\xi}^{x}g(x)e^{F}dx\right)\text{, where }\quad F=\int_{\xi}^{x}f(u)du \qquad (23)$$

The value of the field (now C_2) is surely known at the $\xi = x_{max}$ point (at the end of the inhomogeneity), where $C_2\equiv 0$, as for no reflected signal-part arrives from the x > x_{max} homogeneous half-space. Furthermore

$$F=\int_{\xi}^{x}f(u)du=\int_{\xi}^{x}-\frac{1}{2k(u,\omega)}\frac{\partial k(u,\omega)}{\partial u}du=\frac{1}{2}\ln\frac{k(\xi,\omega)}{k(x,\omega)}$$
$$e^{-F}=\sqrt{\frac{k(x,\omega)}{k(\xi,\omega)}}\quad\text{and}\quad e^{F}=\sqrt{\frac{k(\xi,\omega)}{k(x,\omega)}} \qquad (24)$$

With (23) and (24) C_2 can be obtained as (25)

$$C_2(x,\omega)=C_0(\omega)\sqrt{\frac{k(x,\omega)}{k(\xi,\omega)}}\int_{\xi}^{x}\frac{\sqrt{k(\xi,\omega)}}{2k(u,\omega)}\frac{\partial k(u,\omega)}{\partial u}e^{-2j\int_0^u k(v,\omega)dv}du$$

From (10)

$$E_{z2}(x,\omega)=-\frac{j}{4}\left[\frac{C_2(x,\omega)}{k(x,\omega)}\right]e^{j\int k(x,\omega)dx} \qquad (26)$$

Substituting (25) into (26), the full-wave time-space function of the reflected field is obtained as (27)

$$E_{z2}(x,t)=-\frac{j}{8\pi}\int_{-\infty}^{\infty}\left[\frac{C_0(\omega)}{\sqrt{k(x,\omega)}}\int_{\xi}^{x}\frac{1}{2k(u,\omega)}\frac{\partial k(u,\omega)}{\partial u}e^{-2j\int_0^u k(v,\omega)dv}du\right]e^{\left(j\omega t-j\int_0^x k(u,\omega)du\right)}d\omega$$

(27) is the first approximation of the signal reflected from an arbitrarily strong inhomogeneity, during the propagation of the original signal.

As it is obvious, the more steps of the successive approximation are executed, the more accurate solutions can be obtained for the full-wave forms of the propagating and the reflected signals. The connection with the excitation is hidden in $C_0(\omega)$. On the way detailed in [7] the form of $C_0(\omega)$ coefficient originated from an arbitrarily shaped non-monochromatic signal is as follows

$$C_0(\omega)=I_{x=0}(\omega)\frac{k_0(\omega)\sqrt{k(x=0,\omega)}}{k_n(\omega)+k(x=0,\omega)} \qquad (28)$$

where the starting location of the inhomogeneity is x = 0, and the transformed form of the exciting signal at the boundary surface, $I_{x=0}(\omega)$, is

$$I_{x=0}(\omega)=\int_{-\infty}^{\infty}\left\{\int_{-x_0}^{0}J_0\left(l,t+\frac{l}{c}\right)dl\right\}e^{-j\omega t}dt \qquad (29)$$

The location of the excitation is in the x=[$-x_0$,0] spatial interval. J_0 is the exciting current density (see equation (1.67) in [8]).

It can be seen in (27), that the backward reflected signal-part at a given point contains the integrated influence of all the reflection generated in the medium from the end of the inhomogeneity back to the investigated point (see coordinate u). This term is determined by the complete forward propagating signal from the starting point of the inhomogeneity (see coordinate v). These integrals show well the complexity of the energy-relations of the signal at a given point of the inhomogeneous medium.

The $k(x,\omega)$=0 is a well-known mathematical singularity-problem in every lossless, ideal theoretical model. In the reality, this case never occurs, as the circumstances are never ideal (presence of loss, etc.). In the case of monochromatic signals, no propagation happens at this point, but the whole energy reflects. In the case of impulses, the problem is more complex, as the singularity-problem at a given spatial point will appear only for a frequency-segment (a given frequency) in the signal, but other parts of the signal will propagate.

As the behavior of the propagation-vector becomes rapidly varying in the surroundings of the cut-off and resonance points, the W.K.B. method cannot describe the problem (as for the W.K.B. is based on the elimination of the reflection and the assumption of constant Poynting-vector, etc.).

The new method based on MIBM takes into account the reflected energy in the signal-form (see eq. 27.) by using higher ordered steps of the successive approximation (the zero-ordered approximation, identical with the W.K.B. formula, does not contain the influence of the reflection).

Therefore the new formulas approach the singularities asymptotically for each frequency, and they remain numerically manageable (some samples around the

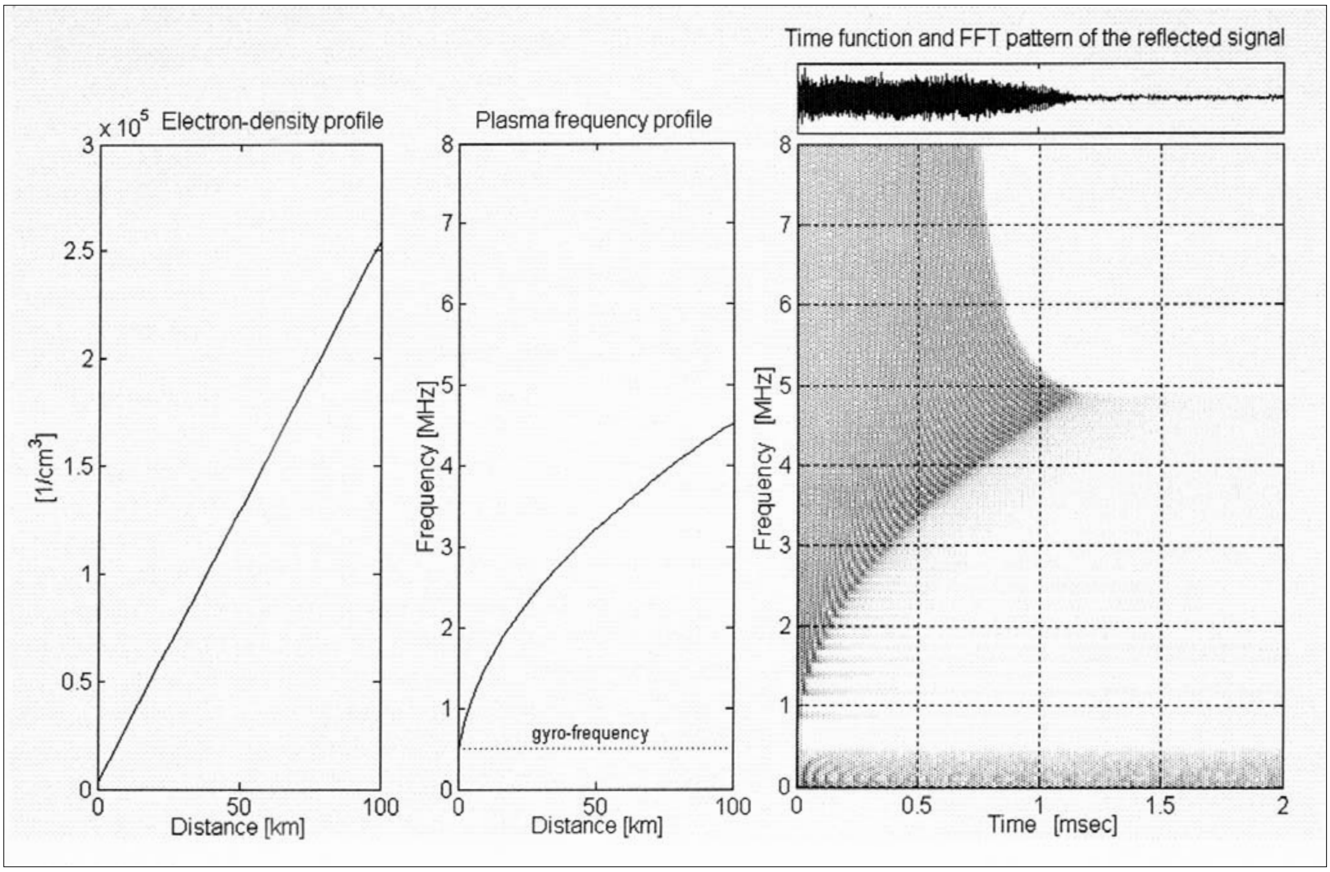


Figure. 2/a.

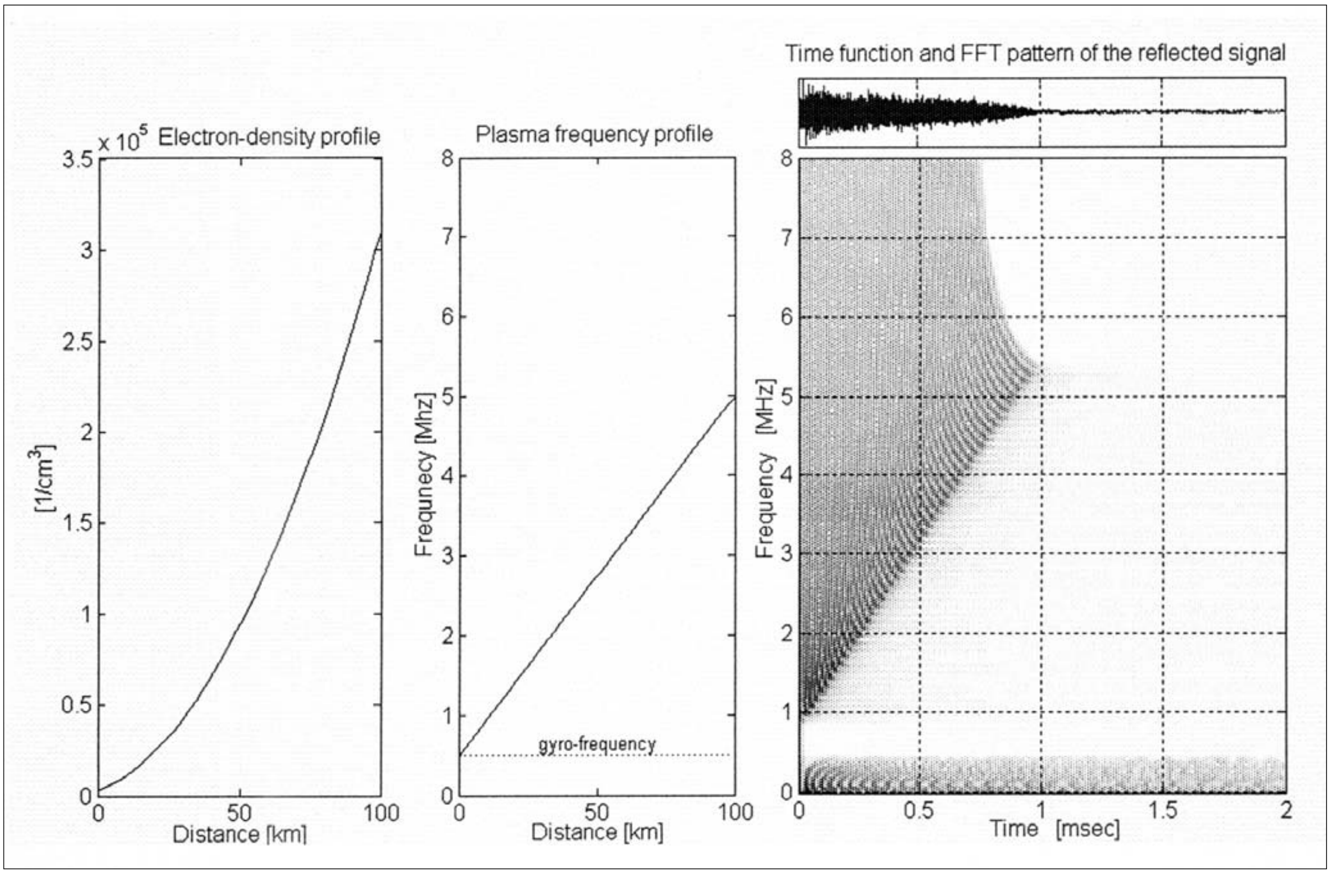


Figure. 2/b.
FFT-patterns and time-functions of calculated reflected signals for different electron-density and plasma frequency profiles.

singularities, at where the integrals become instable, are neglected from the model-calculation). This means physically that approaching this point, the reflection increases, finally up to the reflection of the forward propagating energy within a frequency-range (in which the singularity arises), so this signal-part never reaches actually this point.

Other parts of the whole signal will propagate beyond this point, up to their own singular points (if these exist somewhere).

(The comparison of the Airy-functions and the new method can be found in *Appendix A* – on the next page. The Airy-functions are good, asymptotic solutions of the Stokes-equation, but the Stokes-equation is not a good theoretical description of the reflection in inhomogeneous media.)

2. Results and conclusions

The closed-formed solution yielded from the new theoretical method opens the way of numerical model calculations. As a first computed result, the time function and the FFT-pattern of a calculated reflected signal can be seen in *Fig. 2/a.*, for a given assumed plasma frequency profile (when the electron-density is linear) and for Dirac-excitation, in the case of inhomogeneous, anisotropic electron-plasma.

The result shows that the measure and behavior of the reflection follows the given density profile like a snapshot of the inhomogeneous conditions of the medium. The signal in the lower and the higher frequency range is similar to the forward propagating signals, whistler-type appears in the lower range, while the well-known signal form can be seen in the higher range concordant with the detected TiPP-signals (Transionospheric Pulse Pairs, [8]).

But it is clearly seen that the reflection is continuous in the complete frequency-range, while the forward propagating signal crosses the inhomogeneity (the discrete lines in the FFT-pattern are caused by the finite resolution of the numerical description of the density-profile).

Another example is shown in *Fig.2/b.*, for linear plasma frequency profile and Dirac-excitation.

As it was shown above, the presented solving method of inhomogeneous problems does not lead to differential equation of Riccati-type, but it is possible to obtain the solution in closed form.

The first approximation results in the well-known (non-coupled) W.K.B. approximation for weakly inhomogeneous medium, but the full-wave result opened the way to obtain more and more accurate solutions for the propagating signal in the case of stronger inhomogeneities as well. Moreover, it became possible to determine the reflected signal in closed form. By applying these results, it is possible to compute some phenomena – e.g. the reflected part of ELF-VLF signal generated by lightning, traveling through the ionosphere up to the magnetosphere (whistler-precursors). The solving method is applicable in other media too, different from the presented plasma model (e.g. mine detection).

Acknowledgement

This work was carried out by the aid of the Bolyai János Research Grant of the Hungarian Academy of Sciences, the Research & Development funds no. OTKA T034831, T037611 and F037603 and by the assistance of the Hungarian Space Office.

References

[1] Ferencz, Cs.:
Real solution of monochromatic wave propagation in inhomogeneous media, PRAMANA Journal of Physics, Vol.62, No.4, 2004 April, pp.943–955.

[2] Osterberg, H.:
Propagation of plane electromagnetic waves in inhomogeneous media;
J. Opt. Soc. Am., Vol.48, pp.513–521, 1958.

[3] Simonyi, K.:
Foundation of Electrical Engineering,
Pergamon Press, New York, 1963.

[4] Budden, K.G.:
Radio waves in the ionosphere;
Cambridge University Press, London 1966.

[5] Marcuwitz, N.:
Interaction of Electromagnetic Fields,
(First Order Versus Reduced Formulation);
Proc. Fourth Coll. On Microwave Comm.,
Vol.III., ET-18/1-6. Academic Press, Budapest, 1970.

[6] Ferencz, Cs.:
Electromagnetic wave propagation in inhomogeneous media: Method of Inhomogeneous Basic Modes;
Acta Technica Ac.Sci.H.,
Vol.86(1-2), pp.79–92, 1978.

[7] Ferencz, O.E.:
Electromagnetic Wave Propagation in
Different Terrestrial Atmospheric Models;
Ph.D.Thesis, Budapest University of Technology and Economics, 1999.

[8] Ferencz, Cs., Ferencz, O.E.; Hamar, D.
and Lichtenberger J:
Whistler Phenomena, Short Impulse Propagation;
Kluwer Academic Publishers, Astrophysics and
Space Science Library, Dordrecht, 2001.

[9] Kamke, E.:
Differentialgleichungen, Lösungsmethoden und Lösungen, Band I. Gewöhnliche Differentialgleichungen; Akademische Verlagsgesellschaft, Geest & Portig K.-G., Leipzig, point 4.3., pp.16., 1951.

The Appendix:
“Comparison of the new model with the Airy integral” –
see on the next page.

Appendix

Comparison of the new model with the Airy integral

One of the most commonly known theoretical approximations of wave-propagation in inhomogeneous media is the solution of the Stokes-equation by Airy integral functions [4]. It is useful to investigate the differences between the new method (presented in this paper and in [1]) and the Airy-solution. This comparison will be demonstrated for (longitudinally propagating) monochromatic signals.

As it can be found e.g. in [4, Chapter 9 and 15], the Airy integral (Airy integral functions) is the mathematically correct solution of a type of differential equations (like Stokes-equation is):

$$\frac{{}^2E_y}{dz^2}+k_0^2q^2E_y=0 \qquad \text{(A.1)}$$

where $q^2 = n^2$ (longitudinal propagation), (A.2)
n is the refraction index, as usual.

The cornerstone of the new method based on MIBM is the realization of the physical fact, that only and exclusively the resultant sum of the forward propagating and reflected (scattered) signals can be an existing, real solution of Maxwell's equations.

As it is well seen in Budden's argumentation, the supposed starting form of the solution to be determined contains the resultant sum of forward and backward propagating signal-parts (see eq. (9.47) in [4]):

$$E_y=A\cdot \mathrm{e}^{-\mathrm{j}k_z z}+B\cdot \mathrm{e}^{\mathrm{j}k_z z} \qquad \text{(A.3)}$$

where $$k_z=k_0\cdot n=\frac{\omega}{c}\cdot n$$

Further, the detailed investigation of the derivation enlightens some important problems. Budden applies Maxwell's equations, and deduces the Stokes-equation from them (eq. (9.49)–(9.54) in [4]). As he states, this should refer to the signal-form (A.3), and a result of this deduction is the known Stokes-equation (eq. (9.58) in [4]). The Airy integral functions are valid for this differential equation type.

But it is important to recognize, that Budden supposes by the introduction of the signal form of (A.3), that the substitution of the forward and backward propagating signal parts separately into Maxwell'equations results formally identical equations, and he deduces the Stokes-equation for only the forward propagating part, independently. He does not apply the resultant sum of the signal-part in his computation in order to get the Stokes-equation. This form of the Stokes-equation cannot be obtained for the resultant sum from Maxwell' equations, only for the forward or backward propagating signal-parts independently (if we suppose that one of these signal-parts can exist alone). This assumption is not a special case of the new theoretical model presented in this paper, but means a fundamental contradiction between the former approximation and the new model.

In order to compare this solving method to the new (presented in this paper and in [1]), let us control Budden's derivation. If the whole solution-form shown in (A.3) is written back into the Stokes-equation and one derives the equations on a correct way, it will be obvious, that the result presented by e.g. Budden cannot be yielded.

Assuming (in concordance with Budden's work), that A and B are constant (it must be emphasized, that this precondition is a hard restriction of the validity limits in the case of spatial inhomogeneities) and substituting (A.3) into the Stokes-equation, the following will be obtained:

$$A=-B\cdot \mathrm{e}^{\mathrm{j}2k_z z} \qquad \text{(A.4)}$$

This is in obvious and fundamental contradiction with the starting precondition (A and B have to be constant). It well can be seen, that Budden's formulas are valid only for the forward and the backward propagating signal-parts separately, so this way of thinking implicitly considers this signal-parts as independent solutions of Maxwell'equations. (A.4 cannot be considered as some "reflection coefficient", because of the starting mathematical suppositions.)

If A and B are not constant, the result does also not lead back to the known formulas, from which the Airy functions are deducible, but gives a more complicated relation between A and B:

$$\begin{aligned}&\left[-2\mathrm{j}\frac{\mathrm{d}A}{\mathrm{d}z}(\mp k_0 q)-\mathrm{j}A\left(\mp k_0\frac{\mathrm{d}n}{\mathrm{d}z}\right)+\frac{\mathrm{d}^2A}{\mathrm{d}z^2}\right]\cdot \mathrm{e}^{-\mathrm{j}k_z z}+\\&+\left[2\mathrm{j}\frac{\mathrm{d}B}{\mathrm{d}z}(\mp k_0 q)+\mathrm{j}B\left(\mp k_0\frac{\mathrm{d}n}{\mathrm{d}z}\right)+\frac{\mathrm{d}^2B}{\mathrm{d}z^2}\right]\cdot \mathrm{e}^{+\mathrm{j}k_z z}=0\end{aligned} \qquad \text{(A.5)}$$

This relation, on the one hand, does not coincide with results published by Budden (and others), and, on the other hand, it results an underdetermined description of the problem (one equation containing two unknown variables), which is unsolvable.

This investigation obviously confirms, that Budden's theory (which finally leads to the Stokes-equation, as a description of wave-propagation in inhomogeneous media) at least implicitly contains the preconception, that the forward and backward propagating signals are independently existing solutions of Maxwell'equations.

Because of these fundamental theoretical differences, the new method never results the Stokes-equation, but uses a new and theoretically different way in order to solve the problem of spatially inhomogeneous media. This new method delivers mathematically correct answers for the problems presented above.

Furthermore, this fact is independent from the nature of the signal (monochromatic or impulse); this argumentation is equally valid for both cases. The direction of propagation (referring to the gradient of inhomogeneity; longitudinal, transversal or oblique) has also no influence on this theoretical difference.

Summarizing, the new method never leads back to Stokes-equation during the solving process, because of this it becomes possible to avoid the theoretical contradictions presented above and to obtain an exact solution.

Assessment of the errors in single point positioning

Bence Takács

Budapest University of Technology and Economics, Department of Geodesy and Surveying
bence@agt.bme.hu

Keywords: single point positioning, submeter accuracy, permanent stations, position errors

After turning off Selective Availability (SA) a new chapter began in the GPS-technique. The performance of GPS standard single point positioning technique was discussed in details [4]. It was stated, that in favourable conditions accuracy of several meters is achievable. Recently the number of GPS users has impressively increased; turning off SA has clearly played an important role in the propagation of GPS technique. Turning off SA is considered as a key point not only for the practice, but also for scientific researchers. It is well known, that compared to the artificial degradation of GPS accuracy, the effect of systematic and random errors on single point positioning is practically negligible. Some of the receivers do not take into account some systematic effects, because of the order of magnitude of SA error. Formerly errors on single point measurements could be invetigated only with limited efficiency. Turning off SA offers an opportunity to assess all the systematic and random effects in details; this paper will summarize the most important results of these investigations.

According to the GPS system operators the horizontal positioning error at 95% level of probability is around 13 m, the vertical error is around 22 m [2]. There are two possibilities to increase the achievable accuracy:

- Relative positioning instead of single point positioning, this is widely used in surveying and geoinformatics; or
- Using more effective models for taking into account the systematic errors.

The second method is also known as *precise single point positioning*. In a rigorous sense neither in this case one mention single point positioning, since modeling the systematic errors can be derived from special processing of permanent GPS stations' measurements. Major part of the users, because of convenience and practical aspects prefer to operate only one receiver. Hence this kind of relative technique is often represented as a single point positioning technique, since the user is not aware of using relative positioning.

Thanks primary to the activity of the International GPS Service (IGS), most of the systematic errors can be taken into account using precise models in post-processing. As a result of scientific processing of data from permanent GPS stations, satellite orbits are known to a few centimeter accuracy, the effect of satellite and receiver clock offsets (after converting them into distance) can be determined with the same accuracy. Finally, accurate maps of the Earth's ionosphere may contribute to the precise single point positioning.

The paper deals with the most important effects on single point positioning measurements. Algorithms and methods will be presented how to refine upon common models using products of permanent GPS stations, decreasing positioning errors to submeter level or even better.

Ionospheric effect

For the simplicity of computations, it is conventional to suppose that the electromagnetic waves on their whole paths have the same propagation velocity as in vacuum. Since GPS satellites orbit is around 20.000 km above the Earth's surface, signals travel in vacuum much part of their path, but before reaching the receiver antenna they have to cross the Earth's atmosphere, meanwhile their speed is modified significantly.

For the aspect of a decimeter radio signals, the atmosphere may be divided into two, totally different layers: the ionosphere and the troposphere. In the higher (between 40 km and 1000 km) ionospheric layer, particularly due to the ultraviolet radiation of the Sun, there are particles with electric charge. These particles modify the velocity of signals according to the signal frequency. So the ionosphere, from the point of a decimeter electromagnetic signals is a dispersive medium, its refractive index depends also on the signal frequency.

The effect of ionosphere could be taken into account using several methods. From the point of practice, two methods should be highlighted:

– with computations, using ionospheric models; or
– with elimination using dual frequency receivers, exploiting the properties of frequency dependency.

Now only the modeling method is discussed, since dual frequency receivers are used only in high precision scientific projects. In modeling it is supposed, that free electrons in the ionosphere are pressed into a single layer (also called a thin-shell model). The models can describe the Total Electron Content (TEC) of each point in the single layer.

Most frequently the effect of the ionosphere is taken into account using the Klobuchar-model with the parameters broadcast in the satellite navigation mes-

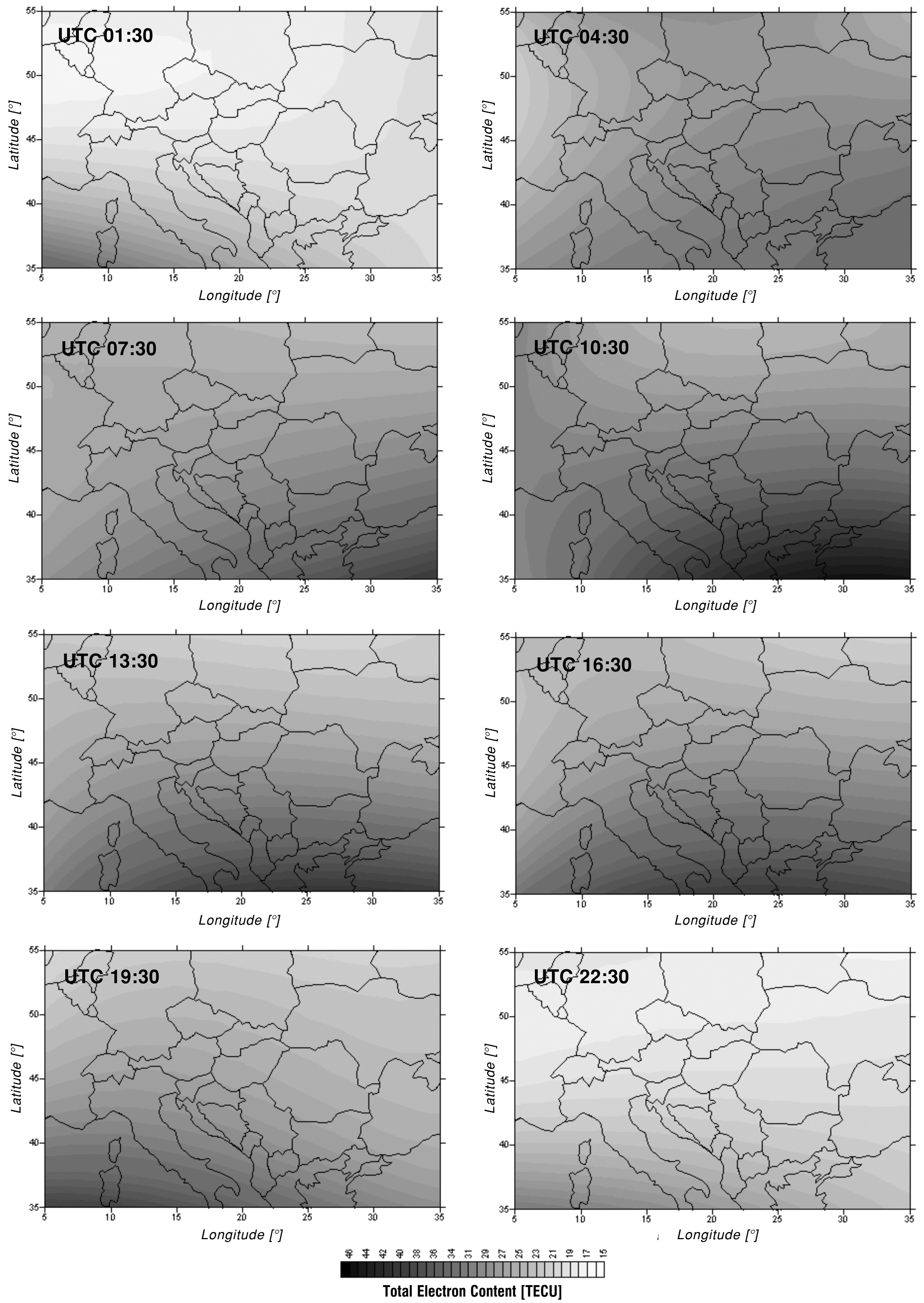


Fig. 1. Ionosphere maps from local models above Hungary and the neighbouring countries (16th June, 2002)

sage. This model is a simple cosine function, a more detailed description can be found for example in [3]. The most important advantage of the Klobuchar-model is that its parameters are transmitted in real-time by the GPS satellites themselves; hence for the positioning there is no need for external data. Its disadvantage is that the Klobuchar-model describes the ionospheric effect only with limited performance, according to the experiences with 50-60%.

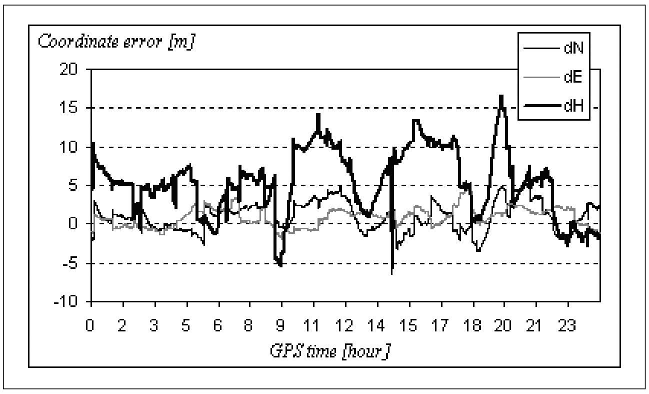
The ionosphere can be modeled more effectively, using so called local ionosphere models or global ionosphere maps. In the first case the single layer is described with low degree Taylor series, in the second one with harmonic spherical functions. The formulas can be found in the manual of the scientific post-processing software, BERNESE [1].

The main advantages of the local ionosphere models are that models, valid for a few thousands of square kilometers can be determined from measurements of a few permanent GPS stations, even in real-time using simple mathematical tools. At the same time global ionosphere maps are valid for the Earth, they are more precise but for the completeness of calculations the parameters could be determined only with some delay in time. For more details about global ionosphere maps, refer to the website of the IGS processing centre Berne (http://www.aiub.unibe.ch/ionosphere.html).

In a former paper [5] local ionosphere models valid for Hungary and the neighbouring countries were introduced. For the computations self-developed software was used. Without presenting the details, some maps demonstrating the total electron content for a given day are shown on the *Fig. 1.*

Henceforward the accuracy of single point positioning using different ionosphere models are presented. It is well known, that after turning off SA, the effect of the ionosphere became the largest impact [3]. For the analysis, 24-hour measurements from the permanent station of Budapest University of Technology and Economics were processed applying different ionosphere models and the position errors were compared. Self-developed software was used for the computations.

Fig. 2.
Errors of single point positioning at BUTE station (16th June, 2002, ionosphere: Klobuchar-model with broadcast parameters)



In the first case the Klobuchar-model with parameters from the navigation message was used. Most of the conventional GPS receivers use this method. According to *Fig. 2.* coordinate errors can reach 15 meters.

Then the ionosphere was taken into account using the local models shown on the *Fig. 1.* The improvement of the coordinate errors is clear on *Fig. 3.* especially in vertical sense. While the standard deviations of all the three components have been decreased equally with about 30%, the average systematic error in vertical sense has been decreased with 80%!

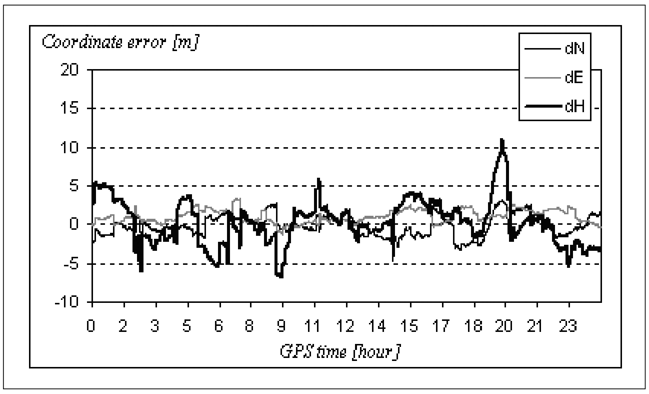
Orbits and satellite clocks

For GPS positioning it is essential to know the satellite orbits. The most conventional way to compute satellite positions is to use the broadcast ephemeris. The most important advantage of this technique is that the necessary information is broadcast by the GPS satellites, so external data sources are not required. Its most important disadvantage is that the broadcast satellite orbits are only a couple of meters accurate, so this technique is not precise enough for all kinds of applications. The accuracy may be improved using the so called precise satellite orbits based on the special processing of permanent GPS stations' data.

The different kinds of orbits can be compared with Bernese software. According to two data series the software can compute the differences of the satellite positions. In practice an orbital coordinate system consists of radial, tangential and out-of-the-plane components *(Fig. 4)*.

GPS is a passive positioning system, users do not transmit only receive signals from the satellites. Satellite-receiver distance determination is based on measuring the propagation time of signals from satellites to the user receiver. As a consequence it is necessary to have a clock both at the satellite and at the receiver side. These clocks have an individual offset from the so called GPS system time. The effect of satellite clock offset can be decreased significantly using a quadratic function with parameters taken from the navigation message.

Fig. 3.
Errors of single point positioning at BUTE station (16th June, 2002, local ionosphere models)



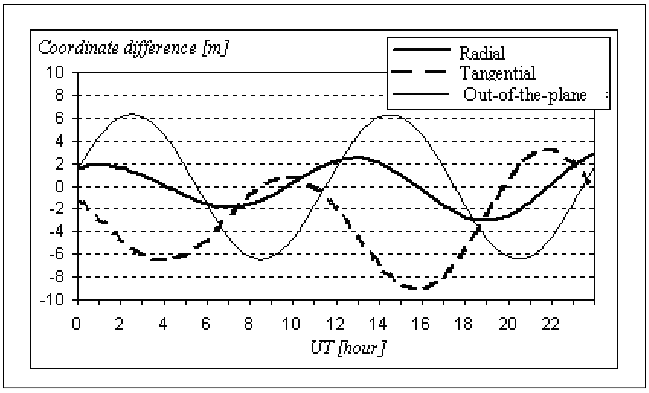


Fig. 4.
Differences of the satellite coordinates computed from the broadcast ephemeris and the IGS precise orbits (16th June, 2002, PRN: 08)

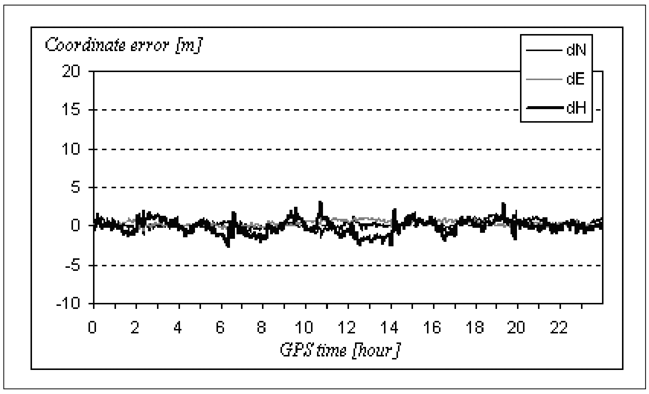
Similarly to the satellite orbits, International GPS Service determines the quasi correct satellite clock offset values based on processing permanent GPS stations, again. Using these values the accuracy of the parameters in the navigation message can be estimated and the differences can be applied as corrections *(Fig. 5)*.

It is possible to precise satellite orbits and clock offsets in the positioning. In this case the discrete values (usually given every 15 minutes) need to be interpolated using high ordered Lagrange polynomials.

From the computational point of view it is more elegant, if the differences of the precise orbits and the ones calculated from the broadcast ephemeris are taken into account as corrections. Its main advantage is that differences can be interpolated easily. The same algorithm can be used for the satellite clock offsets, as well.

On *Fig. 6.* the same measurements of the BUTE station are processed using the IGS precise orbits and satellite clock offsets. It can be clearly seen that the horizontal position errors are less than 2 m, the vertical error is less than 3 m. The average systematic error is practically the same in each component, less than 30 cm, the standard deviation is less than 1 m.

Fig. 6.
Errors of single point positioning at BUTE station (16th June, 2002, global ionosphere maps of Berne, IGS final orbits and satellite clocks)



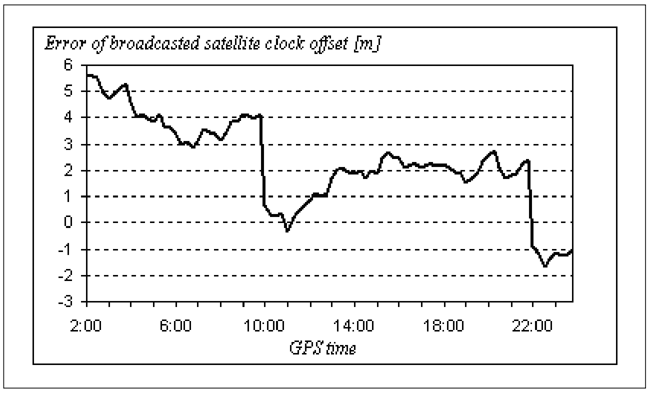


Fig. 5.
Errors of satellite clock offsets computed from the parameters of navigation message (16th June, 2002. PRN: 02)

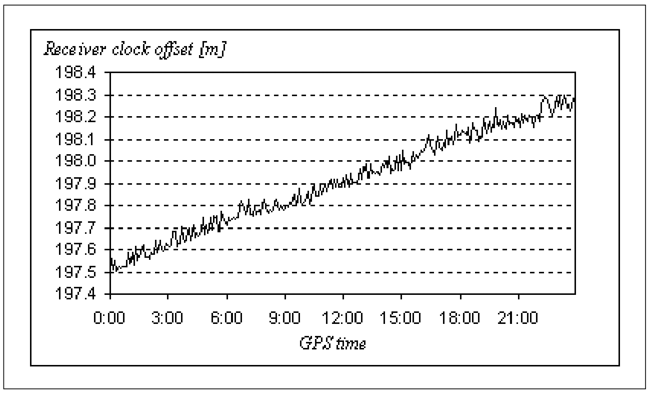
Receiver clock offset

The quartz frequency etalons in the receivers are much less accurate (with several orders of magnitude) than the atomic etalons on the satellites. In most of the cases the receiver clock offset is treated as unknown, its value is computed from the GPS measurements. That is why four satellites are required for the unambiguous three dimensional positioning, although from geometrical point of view three satellites were enough.

The International GPS Service beside the satellite clock offsets provides the receiver clock offsets of some permanent stations, again in discrete epochs (usually every 5 minutes). These products are available on the IGS home page (ftp://igscb.jpl.nasa.gov/pub/product/). The accuracy of these receiver clock offsets are estimated at a few centimeters level. (The receiver clock offset is treated as a time unit by nature, but for the better interpretation it is multiplied by the vacuum speed of light to get a metrical dimension.)

Unfortunately for our investigations the BUTE measurements are not usable in the followings, since BUTE is not part of the IGS network. Hence we selected another station: BRUS (Brussels, Belgium). *Fig. 7.* shows the receiver clock offsets of BRUS station according to the IGS product.

Fig. 7.
The receiver clock offset at BRUS station, according to the IGS product (16th June, 2002.)



As it can be seen on *Fig. 7.*, the clock offsets at BRUS station controlled by a hydrogen maser etalon can be modeled efficiently with e.g. linear regression. Now we present the accuracy of single point positioning if the receiver clock offset was not treated as an unknown, but as value taken from the above linear regression. Now the equation system contains only three unknowns in spite of the usual four.

On *Fig. 8.* we can see that the vertical position errors are not larger than the horizontal ones, the error of the three components is typically less than 1 m. It is worth mentioning that in “conventional” positioning the vertical position error is two times larger than the horizontal.

Two main disadvantages of the method based on the IGS receiver clock offsets are mentioned:

- it works only in the case of high precision frequency etalons, e.g. in a laboratory;
- the positioning can be carried out only with post-processing.

To encounter the second disadvantage we propose an algorithm based on Kalman-filtering. The single point positioning equation system is solved in two separate steps:

1) solving with the “conventional” conditions, using four unknowns;
2) smoothing the receiver clock offset with a Kalman-filter and solving the whole system again, but only with three unknowns.

Details are not presented because of the lack of space, but it is mentioned that this algorithm yields practically the same results as using IGS product, of course only in the case of stations with well-modeled receiver clock offsets.

Noise of code measurements

In this paper the effect of the most important systematic errors were presented. The random effects, especially the noise on code measurements have not been treated yet. It is well known that the noise on phase data is practically negligible in comparison with the code noise. In theory single point positioning can also be carried out using phase measurements only but the well known phase ambiguity problem is rather complicate in the practice. The optimal solution should be the use mixed phase and code data. The principle is to smooth the code distances with phase ones.

According to the most often used algorithm the phase ambiguity equals more or less to the difference of phase and code distance in each time epoch. Of course this value is affected by the code noise which can be decreased with simple mathematical tools, like running averaging. This algorithm is easy and efficient. One of its main disadvantages is however that using single frequency receivers the length of smoothing in time is limited, since the effect of the ionospheric delay has opposite sign on the code and phase measurements. Another disadvantage is that the quality of the smoothed data depends on every satellite, so once the smoothing has stopped at one satellite for whatever reason, it is no worth continuing the smoothing process for the other satellites either.

If we have a dual frequency receiver, the problem caused by the ionosphere can be almost totally neglected. A further advantage of the dual frequency receivers is the opportunity of cleaning the phase data from cycle slips using various types of linear combinations. This method is used in the Bernese software, too.

The effect of smoothing is presented in the case of BRUS station, using the above used data. *Fig. 9.* shows the positioning errors. The effect of systematic errors have been decreased using the most powerful algorithms presented before: ionosphere was taken into account using local ionosphere models, satellite orbits and clock offsets were modeled by the IGS final products, and receiver clock offsets were smoothed by a Kalman-filter approach.

On *Fig. 9.* it is clear that positioning error components were decreased significantly, however other systematic effects (e.g. troposphere and multipath) play still some role.

Fig. 8.
Errors of single point positioning at BRUS station (16th June, 2002, receiver clock offset modeled by linear regression of the IGS product)

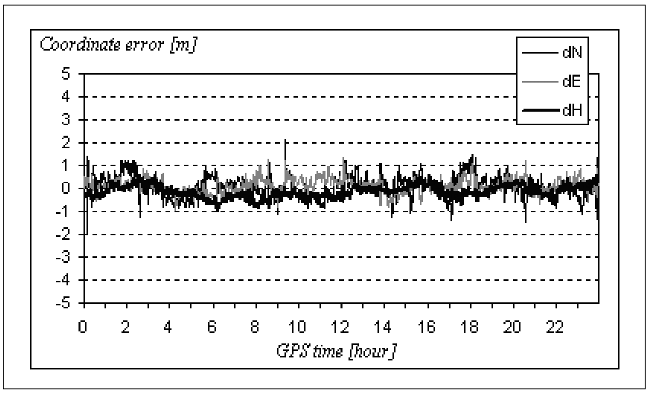
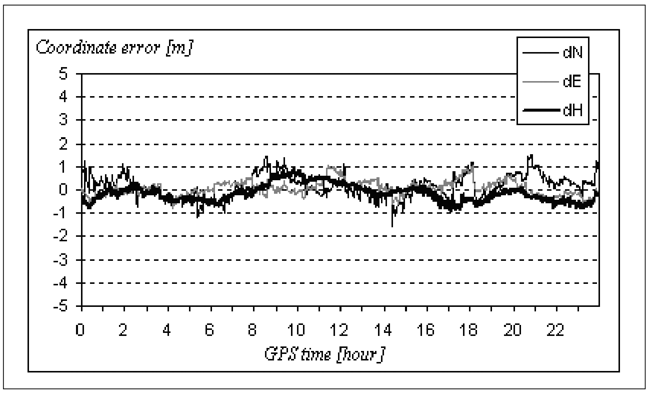


Fig. 9.
Errors of single point positioning at BRUS station (16th June, 2002, code distances are smoothed with phase data, carried out by Bernese)



Summary

The motivation of the investigations, presented in this paper was to develop the necessary algorithms for the one-meter-accurate, real-time GPS single point positioning. The method and results of the local ionosphere modeling were introduced. It was presented that with the precise IGS orbits and satellite clock offsets the standard deviation of the positions can be decreased below one meter, even in real-time. Further improvement in accuracy can be reached with receiver clock offset modeling and with smoothing code distances with phase data. Using dual frequency, high precision geodetic receivers controlled by atomic etalons the position error components are less than one meter.

It is worth mentioning that so called augmentation systems (EGNOS in Europe) have started their operation. Currently EGNOS is working in a test phase, but its results are well promising. The augmentation systems, beside so many other processes, use similar algorithms and models as presented in this paper. The corrections are transmitted from geostationary satellites and via Internet to the users.

In the newest generation of GPS infrastructure to get a homogenous corrections set for a region or a country, the corrections from the individual stations need to be co-ordinated. These algorithms are based again on similar methods, as presented in this paper.

References

[1] Beutler G., et al. (2001):
Bernese GPS Software Version 4.2: Documentation,
Astronomical Institute, University of Berne, Switzerland

[2] Global Positioning System Standard Positioning
Service Performance Standard, 2001 October 1,
http://www.navcen.uscg.gov/gps/geninfo/
2001SPSPerformanceStandardFINAL.pdf

[3] Parkinson, B.W., Spilker, J. J. (eds) (1996):
Global Positioning System: Theory & Applications I-II.
Vol. 164. Progress in Astronautics and Aeronautics,
AIAA Washington

[4] Takács B. (2001):
Műholdas helymeghatározás a korlátozott hozzáférés
(SA) felfüggesztése után,
Híradástechnika, Vol. LVI., No.6., pp.3–8.

[5] Takács B. (2003):
Lokális ionoszféra-modellek Magyarország területére,
Geodézia és Kartográfia, Vol. LV., No.6., pp.19–25.

News

ComArch takes telecommunications forward with OSS&BSS solutions supporting the technology of the future EDGE and UMTS and expands the possibilities of implementing new types of services for customers and gives new wide business opportunities for telecom operators and content providers. ComArch TYTAN Billing Platform and InsightNet Network Management Platform enable information and communication service providers to deploy, then effectively use 3G networks, and allow for the extension of their current functionalities within 2,5G networks.

The accounting of new type of UMTS services requires open and flexible billing systems to handle huge amount of data, new complex types of services and multiple service and content providers. The fully integrated TYTAN Billing Platform enables operators rapidly develop new revenue streams, handle new services and share the revenue with whole chain of service providers. Its real-time capability allows full prepaid and postpaid convergence. TYTAN Data Processing Server (DPS), as the kernel of the platform, is a very efficient rule-based rating engine, that has been benchmarked on major hardware platform to process the volume of data that UMTS will generate. TYTAN IneterPartner Billing component uses the DPS to respond to the new market requirements i.e. quick implementation and configuration of various revenue sharing scenarios.

ECI Telecom is presenting its Hi-FOCuS MiniRAM Outdoor Unit (ODU). The MiniRAM ODU is the first system with ADSL2+ capability and Gigabit Ethernet interface that is specifically designed to support triple play services. ECI´s MiniRAM ODU is a self-contained unit for installation in any outdoor location. This lowers service providers´ deployment costs and enables them to reach new triple play customers. In urban and suburban areas, the MiniRAM ODU enables Telcos deliver the higher bandwidth required for triple play. ECI´s MiniRAM ODU provides solution where existing remote network facilities, such as street cabinets, do not have sufficient space or conditions for broadband equipment.

The MiniRAM ODU can be equipped with various interfaces including T1 IMA, OC-3, and Gigabit Ethernet and provides 32 or 48 ADSL/2/2+ or 24 VDSL lines. A range of Hi-FOCuS products are available to fit a variety of applications, such as Remote Terminal and DLC cabinets, MTU/MDU and more.

Biometric authentication systems

DOMONKOS VARGA, ANDRÁS OLÁH

Budapest University of Technology and Economics, Department of Telecommunications
vargad@hit.bme.hu, olaha@hit.bme.hu

Keywords: personal identification, electronic scanning, security procedures

The security requirements of identification systems have increased considerably recently. This rapid change can partly be explained by the political trends that caused the people to be more and more concerned about their personal or proprietary safety. The conventional security solutions are no longer able to satisfy this demand. Therefore new authentication systems have to be introduced. Amongst these new systems the ones based on biometric authentication may play a decisive role.

Biometrics is one of the various technologies that utilize behavioral or physiological characteristics to determine or verify identity (such biometrics are fingerprint, face, hand geometry, speech, signature, typing dynamics, DNA, iris and retina). During biometric identification personal features are used rather than a conventional code or chip card. In the last decade several solutions were worked out. The motives that ensured the rapid development and spread of biometric systems are as follows [1]:

- the recent increase in the number of passwords causes the security to decrease.
- password management is of extreme cost.
- the importance of authentication required for accessing confidential information has been increasing gradually.
- there is a need for the integration of different security systems, and
- the most up-to-date security system has to be implemented.

Biometric security systems can be divided into two classes. The biometric verification (**BV**) systems make a one-to-one or one-to-few comparison. During verification the newly measured biometric samples are compared to the biometrics stored on disc or card.

The so-called biometric identification (**BI**) systems make a one-to-many comparison. The identity of the target person is determined from a great population. Usually these systems are quite slow as the identification is solely based on the incoming biometric information. Basically these incoming values are compared to the stored data and the most probable identity is chosen. The higher the population, the more complex the search, and the less reliable the result. The **BI** method is very comfortable since there is no need to use supplementary devices. Certain algorithms and special database structures can adequately reduce the time needed for proper identification. By increasing the reliability of the system, however, the false rejection rate may also increase, which in turn might irritate the user. An exact search time cannot be established since it depends on the user population.

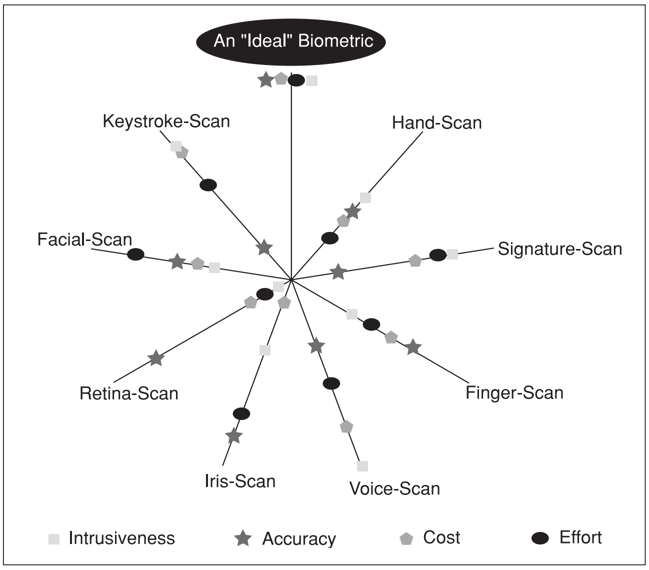
Amongst the advantages of **BV** systems as compared to BI systems are their better price, higher speed and accuracy, as well as their better error properties.

Nowadays the research focuses primarily on the identification systems, because there are several questions that have not been answered yet. One of the major problems is how an identification system can be implemented so that it would be fast and secure enough, and, in the same time, it would operate on large user population.

1. Main objectives

Vendors claim that there is no best biometric technology. A biometric identification technology can be defined as the most accurate, the easiest to use, the easiest to deploy or the cheapest solution, but no system meets all these criteria. For the proper comparison of the different systems an objective measurement tool has to be introduced with which the efficiency of the different implementations can be expressed. This objective tool is the so-called Zephyr chart *(Fig. 1).*

Fig. 1.



The Zephyr chart examines the different technologies from four main points of views, namely the ease-of-use, the cost, accuracy and the perceived intrusiveness. The further the parameter from the center, the better the property of that biometry [1].

1.1. Accuracy

There are two parameters for the measurement of accuracy:

- False Acceptance Rate (**FAR**) – that is the probability of the acceptance of unauthorized individuals.
- False Rejection Rate (**FRR**) – which is the probability that the system rejects an authorized individual.

From the security point of view the emphasis is on keeping as low false acceptance rate as possible. From the user point of view, however, it is just as important to establish fast and accurate recognition. (It is certainly not so fascinating to be able to get access just in the tenth trial.) The FAR for average biometric systems is between 1/100000 and 1/1000000.

1.2. Cost

At the beginning biometric systems were solely used in situations where providing extreme security was of great importance. Thanks to the decreasing costs biometrics is becoming more and more wide-spread. (Areas where biometrics has been gaining ground were computer system authentication, door control, work hour registration, alarm systems, etc.) A considerable part of the costs comes from the control and management software. The price of the installed peripheral devices may also vary between wide ranges. The overall cost of the systems, however, has been gradually decreasing.

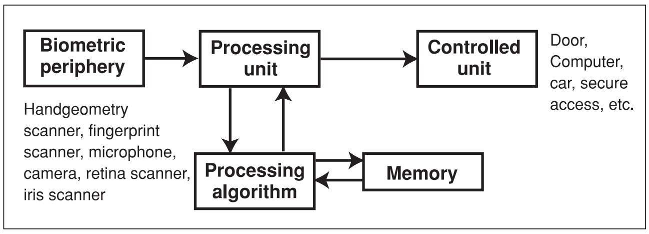
1.3. Effort

The usage of biometric systems needs to be user friendly as well as easy to learn.

1.4. Intrusiveness

There are certain factors other than technical ones that have to be taken into account during the design process. One of the most important non-technical feature is trust, i.e. whether the user can accept that the system reliably identifies them and only them. The other such feature is ergonomics, i.e. how comfortable it is to use the system in long term and how much effort is needed for the every day usage.

Fig. 2.


Biometric periphery
Processing unit
Controlled unit
Door, Computer, car, secure access, etc.
Handgeometry scanner, fingerprint scanner, microphone, camera, retina scanner, iris scanner
Processing algorithm
Memory

In the following section an overview of a general system is presented which helps to understand how the different biometric identification techniques work. (Of course there are other possible schemes as well.) The main functional blocks are depicted on *Fig. 2.*

- The role of the biometric peripheral device is to read biometric parameters and to convert them to digital signals so that they can be understood by the processing unit. The cost of this peripheral device largely depends on the sampled biometric parameter. (The price of a simple microphone or camera is just a tiny fraction of that of a fingerprint reader.)
- The processing unit controls the entire system, and also interacts with the user. It informs both parties whether the identification is successful and whether or not access can be granted for the examined person.
- The processing algorithm performs identification and recognition by comparing the original and the actual samples. The original samples are stored in memory. Although the memory is expandable, its maximum size is limited. This is mainly because the larger the memory the larger the population. High population means higher error probability and in case of identification systems the processing time also increases. If the population exceeds a certain limit it is more feasible to store the original samples in personal chip cards than in the memory of a central mainframe. This arrangement retains speed and security but has one major drawback: a card is needed for the identification. Further problems might arise if this card was lost.
- The controlled unit can be anything that requires security. In practice it is usually a computer system or a door lock system of a building.

In the next section the most common biometric identification systems are introduced.

2. The types of biometric identification systems

2.1 Face recognition

Face recognition systems are not that successful in practice. There are several features which the recognition can be based on. Usually the target person is identified by the contour of their face [1,2]. The geometric features of the face (the distance of the eyes from the edge of the face or from each other, nose length, mouth width, eye width), together with its profile or its thermogram are also used [1].

There are smaller systems for home use which can be utilized in smaller, family-size populations. Also, there exits larger systems that can cope with bigger population. The latter is for in hospital or in bank use and is for identifying patients or customers respectively.

The only way face recognition systems can be successful is when they either work on small population or they are combined with other solutions [1].

2.2. Identification based on iris patterns

The iris is the colored ring that surrounds the pupil. Iris biometric technologies analyze the complex pattern of the iris. The cornea, crypts, filaments, freckles, pits, radial furrows and striations make up such a complex system that cannot easily be duplicated and that possesses more than 400 different measurable parameters [3]. Even the irides of identical twins are different. Amongst all biometric systems this one is the most secure. The method can even distinguish between live and dead people's irides. The recognition process starts with the localization of the edge of the iris. It is followed by image capture. Then the artifacts caused by the shadowing effects of the eyelid and glimmering are compensated. The recognition itself is performed on this preprocessed image.

Performing the above steps results in a so called IrisCode record that serves as a reference for future recognitions. The system is so reliable that it can accurately recognize millions of people [1,3].

2.3. Identification by signature

The development of signature verification systems started a long time ago. Signature verification systems extract and measure a number of characteristics, such as the velocity and acceleration of the signature, the pressure exerted when holding the pen or the number of times the pen is lifted from the paper [4,5]. Two platforms were developed to measure these parameters. The first one is a special pen that is connected to the processing unit via a special cable [6]. The other one is a special plate [7]. The two platforms may also be combined.

Signature verification methods are very accurate with low false acceptance rate. Their use is optimal in situations where signature is a common and accepted way of identification.

It should be noted that making the signature recognition process adaptive is a key point since signature might drastically change by time. [4].

2.4. Identification by keystroke dynamics

Keystroke dynamics are also referred to as typing rhythms. This method analyzes the way a user types at a terminal by continuously monitoring the keyboard input. It has been shown that the typing can serve as biometrics. It is a security method only used in connection with computer systems.

The typing rhythm is checked a thousand times per second. Obviously, keystroke dynamics are behavioral and evolve over time as users learn to type and develop their own unique typing pattern. Although the security level of keystroke dynamic verification systems is not very high, it has got one main advantage: the target people are continuously monitored during their work [1].

2.5. Identification by retinal scan

The artificial duplication of the retina is impossible due to its unique properties. Retinal biometric technologies utilize low intensity infrared light together with an optical coupler to scan the blood vessel system pattern in the back of the eye ball known as fovea. The blood vessel system pattern changes with death, and it is impossible to remove it.

Although the retinal identification system is quite secure, it has a major disadvantage. Its use is uncomfortable, because the user has to look into a receptacle and focus on a given point, and their head must be held still for a couple of seconds [1].

2.6. Identification by voice recognition

Since the human voice has some unique characteristics it can also be used for identification. Voice recognition biometric technologies are based on sampling the spectrum (or rather cepstrum) of human voice and comparing it to the stored pattern.

Identification by voice is a plausible method. Some implementations use sensors attached to walls, while others are implemented as small devices that can be placed in conventional telephone receivers.

Voice verification biometrics is currently used in security control or as a door lock mechanism [1].

2.7. Identification by fingerprint verification

Two biometric property sets are used for identification by fingerprints:

Minutiae – these are the crossings, discontinuities, loops, marks, junctions, and bridges in fingerprints.

Pattern – patterns can be grouped into the following main classes: common arc, common loop, double loop, random line, sharp arc, and spiral.

By the appropriate analysis of the above characteristics the identification can be very accurate. About 100 such characteristics can be analyzed. (Even the fingerprints of identical twins differ to some extent.) A considerable part of biometric identification systems belongs to the so-called Automatic Fingerprint Identification Systems (ASIF). These are utilized by national, international, or federal institutions (FBI, INTERPOL, police etc.) in more than 30 countries.

Some implementations simply emulate the AFIS techniques (with the help of optical or capacitive sensors), others are based on pattern recognition or use ultrasonic scan to identify thresholds [1,8]. There are devices that can even determine whether the scanned fingerprint belongs to a live or a dead person. Presently, fingerprint biometrics is the most widely adopted biometric technologies in the industry.

It is certainly not surprising that due to their low cost, small weight and integrability fingerprint identification systems are used in computer jobs [1], in which the population is limited. Such systems are capable of replacing tradition passwords on computers and make it possible to use fingerprint information instead [8].

2.8. Identification based on hand geometry

Identification by hand geometry is the analysis of the hand geometry including the fingers. The scanned parameters make the hand geometry a useful and applicable identification system (more than 90 different parameter can be analyzed). The analysis is just as detailed as the micro-injuries of the hand would not affect the results of identification [9].

The hand geometry biometric system performs 3D scan to measure the physical characteristics of the hand and the fingers. The identification is based on the comparison of the geometric properties of the palm, or one or two fingers with the stored sample. (The three versions might also be combined.)

The hand geometry analysis is fast, accurate and easy to use. The method can be used well in case of large user population or if the users are with inexperienced [1]. The security provided by the system is not only high, but also allows for flexible acceptance level and easy configuration.

Hand geometry systems are widely used, especially in work hour registration [1,9]. They can easily be integrated with other systems through which a more secure system is formed.

2.9. Complex biometric identification systems

Each biometric identification system has its advantages and disadvantages. Although some of these systems are easy to use, they would not be able to guarantee adequate level of security in standalone mode. By combining three or four such biometric methods complex biometric systems are formed to provide for higher security and to utilize the advantages of the individual subsystems [10].

The overall uncertainty of the complex system is the product of the uncertainty of the subsystems. So if three systems were combined all of which had a false acceptance rate of 1/100, the combined system would have a FAR value of 1/1000000 [10]. And this means a significant increase in security.

3. Summary

Thanks to their ease-of-use and high reliability biometric systems now serve as a measure in the field of authentication applications. The overall cost of such systems varies between wide ranges. A part of the costs comes from the peripheral devices, from the processing unit, and form the memory, and a considerable part is needed to cover the expenses of the software.

The memory requirements also has an effect on the spread of the different biometric systems (the chart bellow depicts the typical memory sizes for the leading biometric technologies) [1].

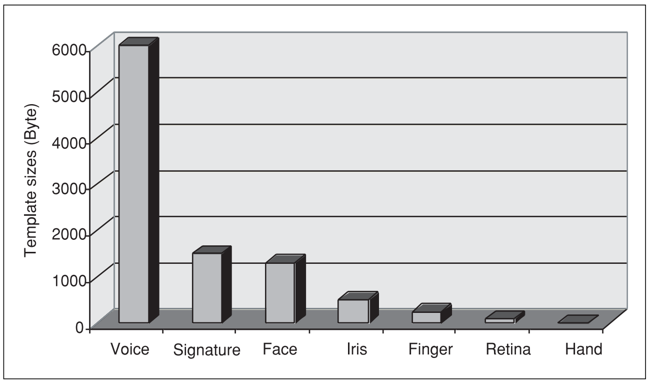


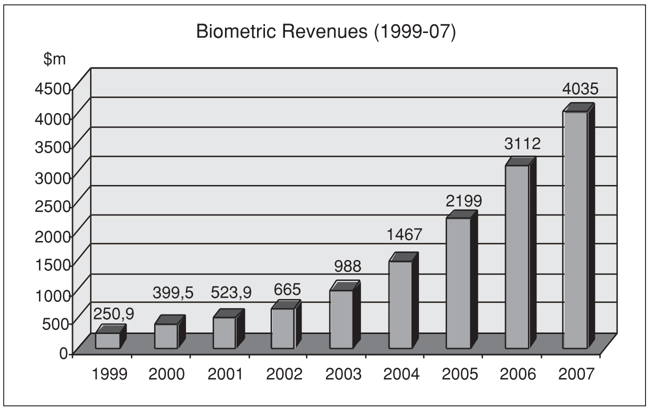
Fig. 3.

The presence of biometric systems in the civil sphere has been growing dynamically. This development is due to the great decrease in price and rapid advancements in technology. At the beginning the sophisticated (finger, iris and retina) solutions are expected to spread. This is followed by the spread of the easy-to-use techniques, because they are already part of our life. Such biometrics methods are voice and signature based identification systems.

Global 2004 industry revenues of 1467m USD are expected to reach 4.04b USD by 2007, driven by large-scale public sector biometric deployments, emergence of transactional revenue models, and the adoption of standardized biometric infrastructures and data formats.

The increase in biometric revenues is due to the rapid growth of PC/Network access and e-commerce. Although in the last decades biometric systems see-

Fig. 4.



med to be secure, the R&D corporations had to face new challenging difficulties. New techniques were developed for hacking biometric security systems. In conventional systems the security was assured by passwords or keys, which could be stolen. In case of biometrics, however, the key point is replicating the characteristics of a person, which obviously is a more difficult task.

Today the main problem is how to prevent the biometric security systems from being hacked. The fraud of a system is something completely different from the cracking of other security systems. (It is not known whether the system can be hacked as long as it is not deceived.) For example a vendor made an optical biometric identification system that identified only living finger. The system seemed secure until someone has breathed on the optical sensor. The system identified the breath together with the fingerprint still present on its sensor (the fingerprint of the previous person who gained access) as a living finger; and provided access. In another case a photocopy of a valid fingerprint was enough to fraud the system.

Until now the exact and fast identification was in the focus of the research, but as technology spreads, several new difficulties have to be taken into account. Until all the technological questions are answered biometric systems must be used very cautiously. For example in the case when one does not have count with the fake of security system because the identification is made under supervision (such situations can be when a security guard or camera is watching).

In the following years, biometric security systems are expected to spread further because identity verification has never been so important, as it has become recently. After September 11, 2001 fingerprint verification systems were set up in US airports. The EU also plans to install a global biometric verification system for identifying anyone entering the Union. Thanks to their decreasing price and increasing reliability, biometric systems have never been so popular, as in the last few years.

References

[1] International Biometric Group,
http://www.biometricgroup.com

[2] Xiaguang J.,
“Extending the feature set for automatic face recognition”,
thesis for the degree of doctor of philosophy, 1993.

[3] Iridian Technologies, http://www.sensar.com

[4] R. K. Abbas,
“A prototype system for off-line signature verification usig multilayered feedforward neural network”,
thesis, 1994.

[5] T. Wessels, C. W. Omlin,
“Hybrid system for Signature Verifivation”, 1999.

[6] H. S. M. Beigi,
“An overview of handwriting recognition”, 1994.

[7] CIC, http://www.cic.com

[8] ActiveCard, http://www.activcard.com/
/activ/products/other/biometrics/index.html

[9] A. Ross,
“A prototype hand geometry-based verification system”,
Biometrics Research, http://biometrics.cse.msu.edu

[10] BioID, http://www.bioid.com

News

NEC Corporation and Japan Science & Technology Agency have jointly succeeded in realizing 150-km-long single-photon transmission. This transmission enables secure network communication supported by the principles of quantum mechanical physics. Due to wide-area coverage (up to 150 km), this system can realize quantum cryptography transmissions in optical networks in metropolitan areas, and is expected to contribute to the realization of an optical fiber network system requiring advanced safety levels. The main features of this system are as follows:

- Stable one-way photon transmission which reduces the noise of backscattered photons.
- Suppression of the deterioration in photon-detection sensitivity that occurs due to the widening of the photon-pulse width.
- Tenfold increase in signal-to-noise ratio as compared with current systems.

The **Plenipotentiary Conference** creeated a Group of Specialists to review the management of the Union (the „GoS“). In October 2003, the ITU Council decided to mandate an external consultant to develop a plan for the implementation of the recommendation made by the Group of Specialists following this review. Based on the recommendations of the GoS, Dalberg staff worked from ITU premises to produce their report for consideration by the Council. The workshops and meetings were variously aimed at bringing together elected officials, representatives of the membership, ITU managers and staff to develop common approaches to the implementation of new processes. The consultants´ report includes various proposals with regard to information systems and management processes. The common view was to work objectives for the future, enhanced communication and information flows. What more suitable goal for an organization so deeply involved in helping the world to communicate?

Network architecture to provide secure anonymous communication

GERGELY TÓTH, ZOLTÁN HORNÁK

Budapest University of Technology and Economics, Dept. of Measurement and Information Systems
{tgm,hornak}@mit.bme.hu

Keywords: anonymity, network architecture, GPSAA, secure communication

Anonymity becomes more and more important in today's privacy-aware information society. Unfortunately the current network layer hierarchy does not support anonymous communication, thus new layers need to be introduced to provide anonymous, yet secure communication in a transparent and easy-to-use fashion. The introduced model of general-purpose secure anonymity architecture (GPSAA) aims to fulfill this purpose.

The rapid development in the area of computer science, software, hardware and communication made it possible to integrate information systems in a constantly increasing factor. This tendency with together with the spreading of Internet sets newer and newer challenges for the information science. For the first problems of bandwidth and reliable data transfer several general architectural solutions exist. In the last ten years new needs arose: besides the given features *secure* communication was also required. In the field of encryption, integrity-protection, remote authentication honored solutions are already known (such as SSL/TLS, SSH or IPSec).

The latest development brought the protection of personal data – *privacy* – into the spotlight. As more and more databases get connected and made – partially publicly – searchable, so will information gathering about persons become easier. As a kind of countermeasure that is why legal data-protection is needed. Anonymity can be understood as an extreme measure of such kind, where the identity of the subject needs to be hidden by technical means, this way eliminating (or at lease reducing to an acceptably small probability) the chance, so that an attacker may connect personal data to a person thus avoiding the creation of an unauthorized on-line profile [1].

For providing anonymity several different techniques exist, but we lack a uniform framework, where besides security techniques (e.g. encryption) arbitrary anonymity methods can be realized. The *general-purpose secure anonymity architecture* (GPSAA) aims to fill exactly this void – to combine security features with the following two kinds of anonymity techniques:

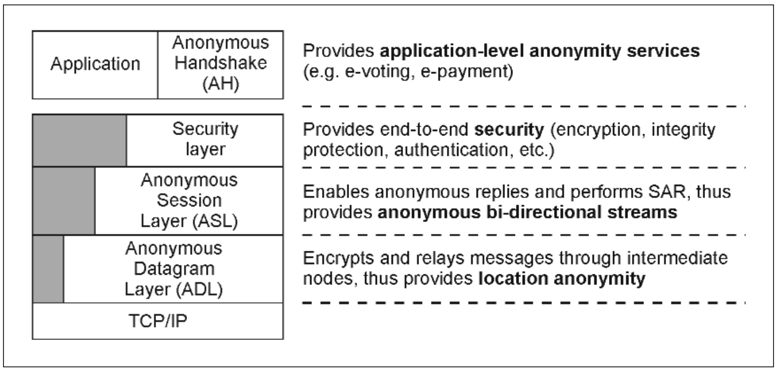
- *Anonymous message sending methods:* in the communication between two parties they make it hard to figure out (even with the help of eavesdropping on the communication channels) who sends messages to whom [2]. Typical scenarios include anonymous e-mail or anonymous web-surfing.
- *Anonymous authorization schemes:* with the help of an anonymity authority they make possible for a service provider to be certain that an anonymous subject is authorized to use a specified service. Typical application areas are: anonymous electronic payment (subject is the client, anonymity authority is the bank and the authorization is the payment) or anonymous electronic voting. In most cases in order to work correctly such schemes require anonymous message sending.

1. The Architecture

According to the introduction such a general framework is required, which enables to combine security features with techniques of the two above mentioned anonymity categories.

The anonymity problems during the electronic communication arise mainly because of the properties of the IP protocol family (i.e. each IP packet contains both the sender's and the recipient's IP address, which makes both of them back-traceable, not anonymous). However, since the popularity of the Internet changing these protocols is not an option, anonymity has to be provided in upper layers. The layer hierarchy of GPSAA was designed with these requirements *(Fig. 1)*.

Fig. 1. Layers of GPSAA


Application
Anonymous Handshake (AH)
Security layer
Anonymous Session Layer (ASL)
Anonymous Datagram Layer (ADL)
TCP/IP
Provides **application-level anonymity services** (e.g. e-voting, e-payment)
Provides end-to-end **security** (encryption, integrity protection, authentication, etc.)
Enables anonymous replies and performs SAR, thus provides **anonymous bi-directional streams**
Encrypts and relays messages through intermediate nodes, thus provides **location anonymity**

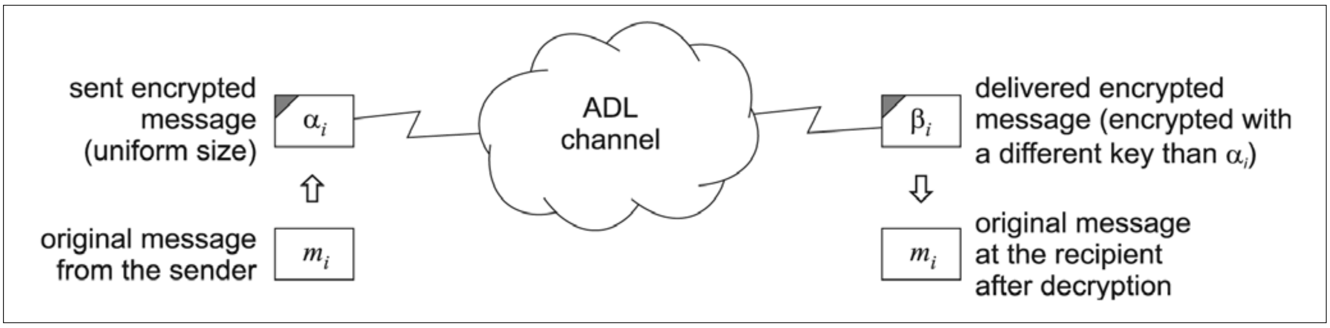

sent encrypted message (uniform size)
α_i
original message from the sender
m_i
ADL channel
β_i
delivered encrypted message (encrypted with a different key than α_i)
original message at the recipient after decryption
m_i

Fig. 2. Sending packets through ADL

The first layer above TCP/IP is the ADL *(Anonymous Datagram Layer)*, whose function is to deliver fixed size packets anonymously using relaying proxies that hide the source and destination of packages. The next layer, ASL *(Anonymous Session Layer)* builds on the services of ADL and is able to manage not just datagrams, but anonymous bi-directional streams. This layer has to solve also the problem of SAR (segmentation and reassembly). Based on anonymous streams, the conventional security layers can be applied. Finally, on the top resides AH *(Anonymous Handshake)*, which handles the application-level anonymous services, like authorization.

ADL – Anonymous Datagram Layer

Aim of ADL is to transport fixed size packets between two communicating parties anonymously *(Fig. 2)*. The packets will be encrypted at the sender, and will be delivered through an ADL channel. This channel is not necessary one physical unit, it can consist of several relaying proxies. Each relaying proxy encodes the packets further and reorders them as well so that by eavesdropping communication lines the routes of the packets cannot be followed.

How such encodings (encryption and decryption) work and how many relaying proxies are utilized is not part of the ADL specification. ADL aims to define an interface, which describes the requirements for possible implementations. By using such a general construction, one might use several different methods in an actual implementation.

ASL – Anonymous Session Layer

Building on ADL, the next step is to provide a bi-directional anonymous stream. Such bi-directional communication brings up some problems concerning anonymity, since with this scenario one party must not know the identity of the other. ASL handles anonymous addresses for this purpose. Besides that, no other anonymity functionality is defined for this layer, however its further task is to cut data received from upper layers into fixed size packets (as required by ADL) at the initiator side and to reorder (since in order to confuse attackers ADL reorders packages) and reassemble these at the responder side.

The security layer resides on top of ASL, which performs security functions on data streams. Although there is encryption in ADL as well, but it is used there only to provide protection against compromising anonymity so only provides link-by-link encryption. To provide end-to-end security, where even relaying proxies should not see the plaintext, the traditional security solutions should be applied (such as SSL, TLS or SSH).

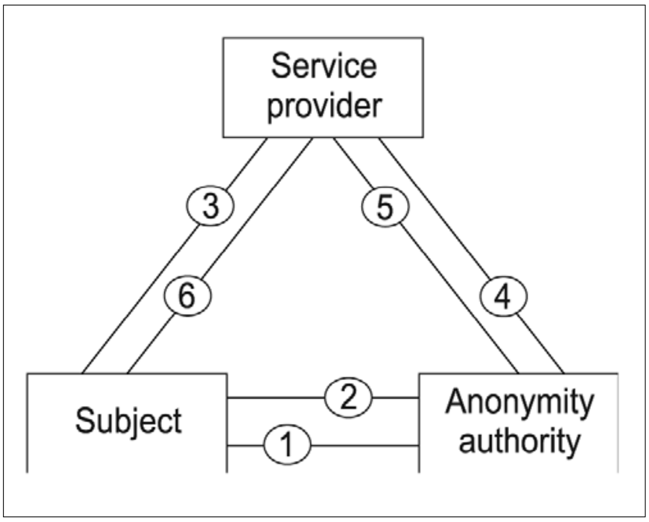
AH – Anonymous Handshake

Now we are ready to employ anonymous authorization above the secure bi-directional anonymous stream as part of AH *(Fig. 3)*.

Anonymous authorization consists of two phases. In the first phase the subject requests anonymity tokens from the anonymity authority (1) (2) during which the subject is not anonymous, on the contrary, he has to be authenticated. Actually accessing the service happens in the second phase, where the subject is already anonymous. He hands the tokens over to the service provider and requests the service (3). After this the service provider checks the tokens (4) and based on the answer from the anonymity authority (5) fulfills the request (6).

GPSAA only formulates requirements for the AH, no special algorithm is envisioned, the main emphasis is to be as general as possible, in order to allow interchangeable suites, like in other widely accepted solutions (e.g. SSL).

Fig. 3. General data flow of the anonymous authorization as part of AH



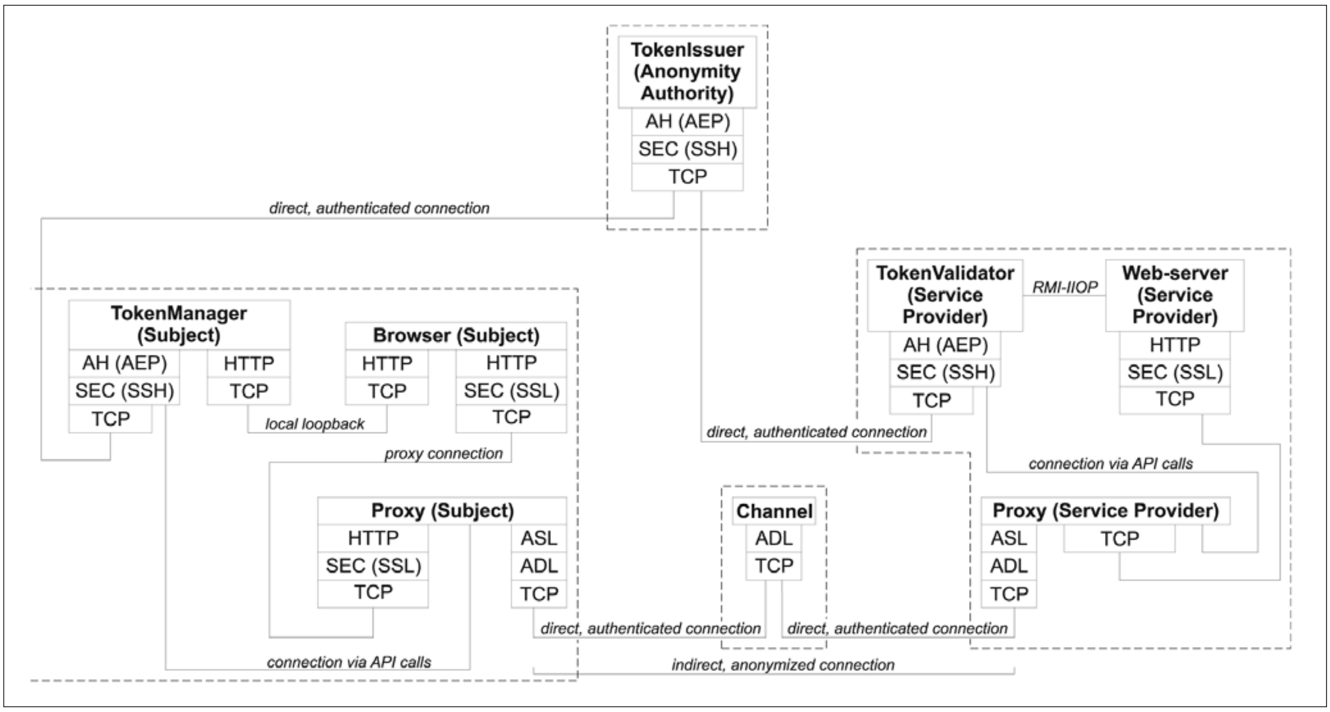


Fig. 4. Implementation scheme of GPSAA

2. GPSAA Implementation

Besides formulating interfaces and requirements as part of GPSAA, implementing the architecture is under way. This reference implementation follows the details of *Fig. 4.* For the first tests at ADL the PROB-channel [4], and at AH level the blind-signature scheme of Chaum [3] was implemented.

3. Conclusion

The general-purpose secure anonymity architecture is a general framework, which enables the combined usage of security functions and anonymity services. With the help of three newly introduced layers (ADL, ASL and AH) the current network hierarchy can efficiently extended to support secure anonymous data transfer – a key building block of today's communication.

References

[1] Froomkin, A. M.:
Flood Control on the Information Ocean:
Living with Anonymity,
Digital Cash and Distributed Databases.
1996.

[2] Reed. M., Syverson, P., Goldschlag, D.:
Anonymous Connections and Onion Routing.
IEEE Journal on Selected Areas in
Communication Special Issue
on Copyright and Privacy Protection,
1998., pp.482–494.

[3] Chaum, D.:
Blind Unanticipated Signature Systems.
US Patent #4 759 064,
1998.

[4] Tóth, G., Hornák, Z.:
Measuring Anonymity in a Non-adaptive,
Real-time System, Proceedings of the Privacy
Enhancing Technologies Workshop, Toronto,
2004.

LTRACK
A novel location management method

SÁNDOR IMRE, MÁTÉ SZALAY

Budapest University of Technology and Economics, Department of Telecommunications
imre@hit.bme.hu, szalaym@hit.bme.hu

Keywords: location management, mobility

In this paper we propose a new location management algorithm for mobility networks. Our algorithm is called LTRACK, it stands for “location tracking”. The signaling load that LTRACK puts on the network can be significally less than that of conventional location management algorithm.

Over the past few years there has been extreme growth in wireless communications. The mobile networks of today use cellular architecture. In a wireless network service access points are usually called base stations. Mobile nodes are connected to the networks via base stations, each of the base stations covers one cell. There are wireless links between the base stations and the mobile equipments. Base stations are interconnected with routers to form a network. This network usually uses fixed links. As the mobile node is moving around, it changes its connection point to the network from time to time. The event when the mobile equipment moves to a new service access point is called handover or handoff.

Any mobility protocol has to solve two separate problems: location management (sometimes called reachability) and session continuity. Location management means keeping track of the positions of the mobile nodes in the mobile network, session continuity means to make it possible for the mobile node to continue its sessions (e.g. phone calls) when the mobile node moves to another cell and changes its service access point. Several solutions exist to both problems [2,4,5,6]. This paper addresses the problem of location management.

Location management has to answer the following questions [1]:

- When should the mobile terminal update its location to the network?
- When a call arrives, how should the exact location of the called mobile equipment be determined?
- How should user location information be stored and disseminated throughout the network?

Of course these questions are not independent, and should be answered together.

Because of the growth of mobile communications and the limitations of resources (especially frequency), more and more efficient algorithms are needed for routing, call management and location management.

This paper is structured as follows: An overview of location management schemes of today’s mobile networks is given in Section 1. Then LTRACK is introduced in Section 2. After explaining the LTRACK network architecture and handover mechanisms, various qualities of LTRACK are examined in detail. In Section 3 LTRACK is compared to other location management schemes. In Section 4 we draw the conclusions.

1. Location management schemes

When an incoming call arrives to a mobile node, its exact location has to be determined. This requires location management. Today’s mobile networks (e.g. GSM) use location area (LA) based location management scheme[1,9]. It means that the cells are grouped into location areas. The network always knows which location area the mobile node is currently staying in, but does not have information about which cell it is in. At the time of the incoming call the network determines the exact location of the mobile equipment within the LA. This is called paging, see [1,9].

This introduces hierarchy into the network, which is an important property of modern mobile networks. For example IP micro mobility protocols use similar hierarchy in the IP based network [2,7,8].

1.1. GSM

In a GSM network (Global System for Mobile communications, the European cellular phone standard) the Home Location Register (HLR) stores the positions of the mobile nodes. If the mobile node does not have open sessions, and it is in idle mode, the HLR does not store the exact position just the Location Area Identifier that the mobile node is staying in. When a call arrives all the base stations within that specific location area broadcast a paging message through their broadcast channel. The mobile node must reply to the paging message, so the exact location can be determined. It is obvious that the mobile node has to update its location information, whenever it crosses a location area boundary. One drawback of this scheme is that when a mobile node moves back and forth between two neighbouring cells that belong to different location areas, a lot of location update messages have to be sent.

1.2. MobileIP

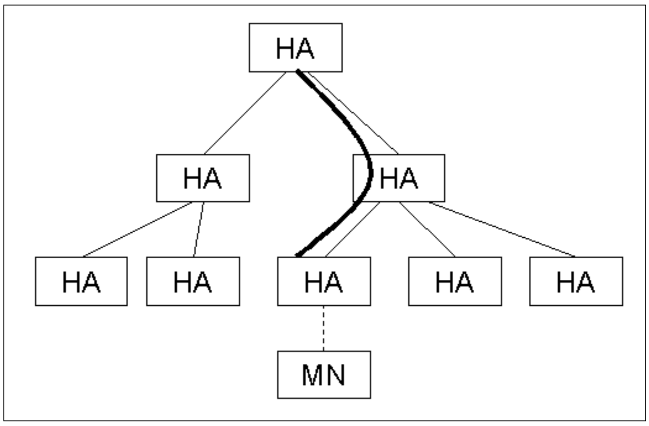
Not only the Internet is based on IP (Internet Protocol), various telecommunication networks can use an IP backbone. There is an increasing need for mobility within the IP world. MobileIP, the IP mobility solution is an extension of IPv4, the current IP version, but it is an integral part of IPv6.

In the MobileIP mobility scheme, a router called home agent (HA) keeps track of the actual position of the mobile node. All incoming calls arrive to the HA of the mobile node, so finding the mobile node is no problem, the HA has always up to date location information. The drawback is that the mobile node has to update its location every time it changes access points. It is a serious problem, because the goal of MobileIP is to provide global mobility within a large-scale IP network (e.g. the Internet). If the mobile node is far from its HA, MobileIP may generate a huge amount of signaling traffic on the network.

1.3. HMIP

HMIP (Hierarchical Mobile IP) uses the same approach as MobileIP, but instead of having one single HA for every mobile node, it uses a hierarchy of HAs [3,4]. Each HA of a given hierarchy level knows which HA of the next lower hierarchy level has location information about the mobile nodes in its subnetwork. One of the HAs at the lowest level knows the exact location, see *Fig. 1.* The mobile node still has to notify the HAs whenever it changes its access point, the advantage of HMIP over MobileIP is that it puts much less signaling load on the network.

Fig. 1.
HMIP Location Management Hierarchy

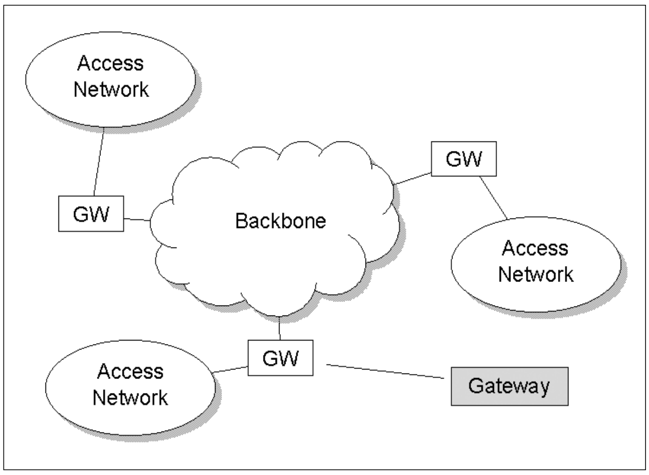


1.4. IP micro mobility

IP mobility protocols also introduce some kind of hierarchy into the mobile network, see [7,8] for examples. They usually interoperate with Mobie IP, Mobile IP is called macro mobility protocol in this environment. The network is partitioned into access networks (micro mobility networks), see *Fig. 2.*

The HA is notified only when the mobile node moves from one access network to another. This is different from the GSM scheme because the exact location of the mobile node within the access network is also stored in a (possibly shared) database. The gateway of the access network (that connects it to the core network) is responsible for finding the mobile node within the access network when an incoming packet arrives, the mobile node is not paged in this scheme, although some micro mobility protocols also allow paging areas (similar to location areas).

Fig. 2.
Micro Mobility architecture



2. LTRACK

2.1. Network architecture

LTRACK (Location Tracking) is a completely new approach for location management in mobile networks. An LTRACK network is built up from LTRACK nodes. A mobile node is connected to one of the LTRACK nodes in the network, and it can change its point of connection.

Every mobile node has an entry in a home LTRACK register (HLR). The basic idea behind LTRACK is to find a compromise between the Mobile IP scheme (where the HA has exact location information) and the GSM scheme (where only the LA and no further information is known). The HLR of LTRACK does not have exact location information, but when an incoming packet arrives, the exact location of the mobile node can be determined.

2.2. Locating the mobile node

In LTRACK each mobile node has a unique identifier similar to IP addresses or phone numbers. This unique identifier is connected to its home address. It is similar to the home address of the Mobile IP scheme.

For each of the mobile nodes, the HLR stores the last address where it received location update message from. It is a “next-hop” towards the node. The mobile node is either connected to that LTRACK node, or that LTRACK node knows a “next-hop” LTRACK node towards the mobile.

Once an incoming call arrives, there is a series of LTRACK nodes pointing from the HLR to the mobile node, see *Fig. 3.*

LTRACK nodes has to be able to find routes to each other. This can be easily solved by using LTRACK over an IP network, thus letting IP routing do the job.

2.3. Handover

When the mobile node moves from one LTRACK node to another, handover takes place. The LTRACK node that the mobile moves away from is called the old LTRACK node, the one it moves to is called the new LTRACK node.

There are two different kinds of handover in LTRACK: "normal handover" and "tracking handover". In a normal handover the mobile equipment updates its entry in the HLR. It sends the address of the new LTRACK node to the HLR. In case of a tracking handover the mobile sends the address of the new LTRACK node to the old LTRACK node.

Incoming calls always arrive to the home address of the mobile nodes, the HLR handles them. So the HLR has to locate the mobile node. It sends a request to the LTRACK node where it received the last normal handover message from. That LTRACK node either still has the mobile node connected to it, or knows a next hop LTRACK node towards the mobile node where it forwards the request. Thus, a normal handover can be followed by some tracking handovers before another normal handover takes place.

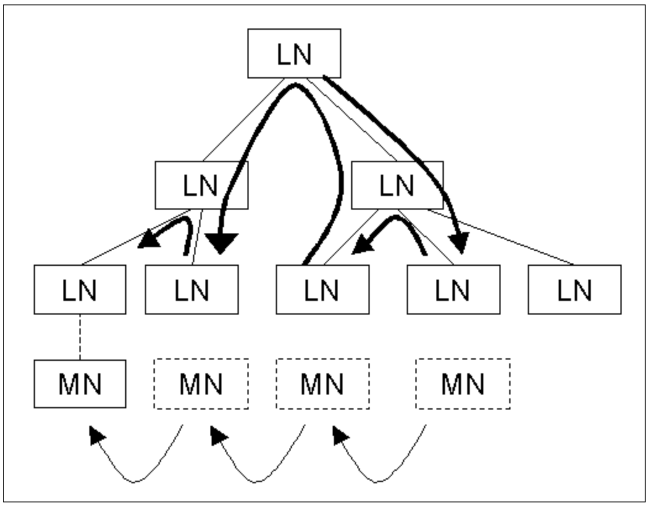
2.4. Advantages

If only normal handovers are used, the location management scheme becomes very similar to the Mobile IP scheme. The HLR always has exact location information about the mobile equipment.

The disadvantage of normal handovers is that they generate much more signaling traffic on the network than tracking handovers. The old and new LTRACK nodes are usually "close" to each other, the HLR can be further away, so this can be an important point.

Another advantage of tracking handovers is the following. Consider a series of tracking handovers between two normal handovers as the mobile node is wandering around in the LTRACK network. If it connects to the same LTRACK node two times on its path (i.e. it moves away from it and returns later) thus generating a loop, the locating request message coming form the HLR will not loop. The LTRACK node will directly forward the request to the LTRACK node towards which the mobile node left the last time it left. So if the mobile node moves back and forth between two LTRACK nodes, it would require a lot of signaling with normal handovers (Mobile IP scheme), but it is no problem with tracking handovers.

Fig. 3.
LTRACK locating of the MN



Who decides when should normal and tracking handovers be used? It may depend on our design goals. Either the mobile node can decide or the network can force a handover type. It is important that a normal handover can always be used, but there are some limitations on the use of tracking handovers. Handovers are usually initiated by the mobile equipment based on power or bit error rate measurements. A tracking handover can only be carried out successfully if communication between the mobile node and the old LTRACK node is also possible, not just between the mobile and the new LTRACK node. This means that tracking handover is a soft type handover. With a small workaround, a hard variant of tracking handovers can be defined that can be used even if the mobile can only communicate to the new LTRACK node. The old LTRACK node can be notified indirectly by sending the notification message to the new LTRACK node which forwards it the old one.

So by using tracking handovers we can minimize signaling traffic on the network. Why should normal handovers be used at all then? Obviously, if a lot of tracking handovers are used consecutively, the path from the HLR to the mobile node may get very long. It means that it will take several hops, and thus a long time for the HLR to locate the mobile node. As this should be avoided, normal handovers should also be used too. It is important to see, that this is not always the case. If the mobile visits only 5 LTRACK nodes on its path, but moves back and forth between them several times, locating the mobile node will no way need more than 5 hops.

What should decisions be based on?

It is possible to limit the number of tracking handovers allowed between two normal handovers. If no more than *n* tracking handovers are allowed between two normal handovers, locating the mobile node should not need more than *n* hops. It can take less as we have seen, but not more.

Another approach can be to limit the time allowed between normal handovers.

It is also possible to cluster the network to LTRACK areas (LTAs). A normal handover is required when the mobile node moves from one LTA to another one. While roaming around within the same LTA, tracking handovers are used. Locating the mobile node should not take more hops than the number of LTRACK nodes in the LTA. This scheme has the same drawback as the GSM scheme. If the mobile node moves back and forth between two neighboring cells that belong to different LAs, normal handovers will be used which results in generating a lot of signaling traffic.

These methods can also be combined. The network can be partitioned to LTAs, a normal handover is required when moving from one LTA to another, but the number of tracking handovers between normal handovers can also be limited within one LTA, so can the maximum time between normal handovers.

2.5. Functionality

LTRACK nodes can be routers in the real world. How are base stations connected the network of LTRACK nodes? There are three different approaches:

- Base stations are LTRACK nodes too.
- Base stations are connected to LTRACK nodes.
- Hierarchical approach.

The naive solution is to define base stations as LTRACK nodes too. When the mobile equipment moves from one base station to another, it moves from one LTRACK node to another. In this scheme the routing functionality and the base station functionality get mixed up, which is usually undesired.

An approach that uses a more structured network is to connect base stations to LTRACK nodes. In this scheme LTRACK nodes are similar to GSM Base Station Controllers. An LTRACK node can serve several base stations. When a mobile equipment moves from a base station to another one, and both base stations are served by the same LTRACK node, the old and new LTRACK nodes are the same. That is the only LTRACK node that has to be notified. When the old and new base stations are served by different LTRACK nodes, an LTRACK handover takes place. This solution decreases the number of required LTRACK handovers for the same number of handovers. Thus, a smaller number of hops will be needed when trying to find the mobile node.

The hierarchical approach is to define LTRACK nodes as small networks. The networks at the lower hierarchy level can be any kind of mobility networks, they can even be LTRACK networks. Thus a two or more level LTRACK network can be built.

3. Qualitative Analysis

Unlike the previously mentioned location management schemes, LTRACK allows different and dynamically controlled parameters for different users. This means that the actual LTAs do not have to be the same for all of the mobile nodes. Different mobile nodes can have different limits on time or on the number of consecutive tracking handovers. Moreover all these parameters can change in time. The system can be automatically fine tuned “on the fly” based on various measurements (e.g. traffic, delay or signaling load).

4. Quantitative Analysis

We have run some simulations using MATLAB. The purpose of the simulation was to compare the signaling load of different mobility schemes.

The simulated network consisted of 36 base stations arranged in a 6x6 grid and 14 routers interconnected to form a tree. Hierarchical Mobile IP is based upon a tree topology network, that is why we used this topology for comparisons.

We examined one mobile node making a random walk with the length of 100 handovers.

The network topology and the path of the mobile equipment were exactly the same in all cases.

Four protocols were examined:

- Mobile IP
- Hierarchical Mobile IP
- LTRACK ($t = 3$)
- LTRACK ($t = 10$)

Variable t denotes the maximum number of tracking handovers allowed between two normal handovers.

Signaling load was measured in hops. *Fig. 4.* shows the results of the simulations.

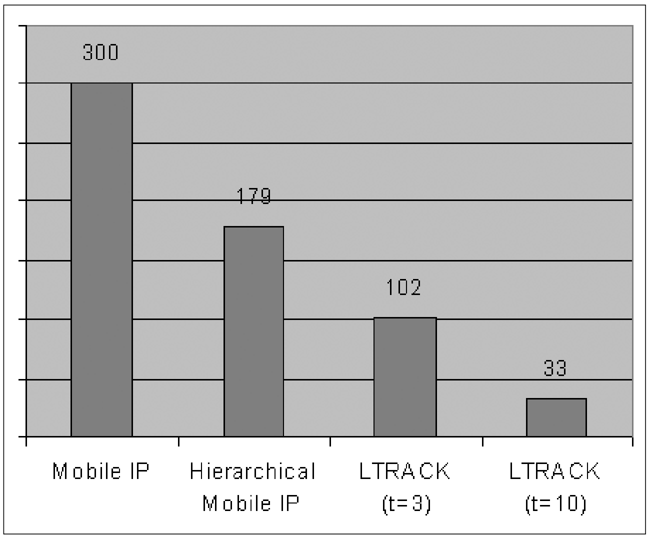


Fig. 4.
Signaling requirements of various protocols

Fig. 4. shows clearly that Hierarchical Mobile IP puts much less signaling load on the network than standard Mobile IP, but LTRACK does much better than them even with a small t value.

How can LTRACK be made efficient. LTRACK generates less traffic if:

- It is run over a more optimal topology than the tree.
- There are more than one normal handovers between two incoming calls.
- An LTRACK node serves more than one base stations.
- The mobile node visits some LTRACK nodes more than once on its path.

5. Conclusions

After defining location management and giving a brief overview of location management schemes we have introduced LTRACK, a new location management method. Its network structure and various handover mechanisms were explained in detail. After qualitative and quantitative considerations LTRACK was compared to MIP and HMIP location management.

Future works should include examination of various network topologies, how they suit LTRACK, and more simulations.

Acknowledgements

This research is supported by OTKA F042590 (Hungarian Scientific Research Found).

References

[1] Vincent W.-S. Wong, Victor C. M. Leung,
"Location Management for Next-Generation Personal Communications Networks",
IEEE Network,
September/October 2000, pp.18–24.

[2] Claude Castelluccia,
HMIPv6: A Hierarchical Mobile IPv6 Proposal.
ACM Mobile Computing and Communication Review (MC2R) – April 2000 issue

[3] Hierarchical MIPv6 mobility management (HMIPv6),
IETF draft, (draft-ietf-mobileip-hmipv6-04.txt)

[4] B. Gloss, C. Hauser,
"The IP Micro Mobility Approach",
EUNICE 2000, September 2000,
Eschende pp.195–202.

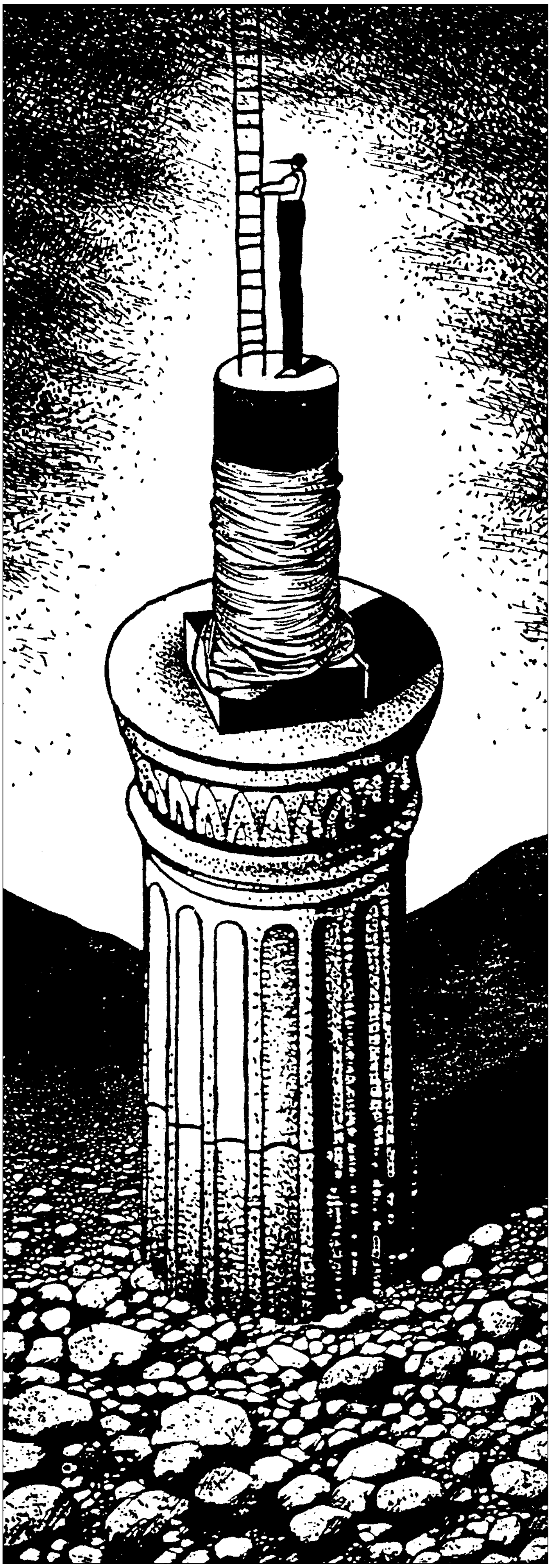
[5] H. Schulzrinne, J. Rosenberg,
"The Session Initiation Protocol:
Internet-Centric Signaling",
IEEE Communications Magazine,
October 2000, pp.134–141.

[6] Z. Turányi, Cs. Szabó, E. Kail, A. G. Valkó,
"Global Internet Roaming with ROAMIP,"
ACM SIGMOBILE Mobile Computer and Communication Review (MC2R),
Vol.4., No. 3., July 2000.

[7] A. T. Campbell, J. Gomez, C. Y. Wan,
S. Kim, Z. Turányi, A. Valkó,
"Cellular IP",
draft-ietf-mobileip-cellularip-00.txt,
IETF Internet Draft, 1999

[8] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan,
"HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks",
Seventh International Conference on Network Protocols,
Toronto, Canada, 1999

[9] Yi-Bing Lin, Imrich Chlamtac,
"Wireless and Mobile Network Architectures",
John Wiley and Sons, 2001.



On-board autonomy of lander units for comet nucleus exploration

ATTILA BAKSA

KFKI RMKI

baksa@rmki.kfki.hu

Keywords: space research, spacecraft, lander, computer, software, autonomy, high reliability

Keeping lander units functional in the hostile, energy-lacking environment in the outskirts of our Solar system is a great challenge. An autonomous real-time control system of a lander is expected to response on board to both nominal and non-nominal events without any external intervention. Recent developments in microelectronics make it possible to use such space-qualified microprocessors that allow the development of highly autonomous on-board software systems. But the increased computing power itself is not all - equally advanced software methods are also needed to provide real autonomy. Considering the complexity of a number of mutually interacting tasks, it is necessary to model them by well-described abstract logical modules. Our focus was on managing the static and dynamic behaviour of the system separately and eventually we developed the Mission Sequencing Object Model Language for describing the long-term autonomous mission control mechanism. This model was implemented on the Philae Lander for the Rosetta mission of the European Space Agency, which was successfully launched on 2 March 2004.

Many space missions are nowadays under preparation, which use the most advanced space technology to explore the unknown depths of our Solar system. The most recently developed microelectronic devices, such as low power consumption high-speed microprocessors, FPGAs, high efficiency solar cell modules and high storage density batteries open the way to keep functional even in hostile, energy lacking environment in the outskirts of our Solar system.

Problems

Operating so far from the Earth poses not only the problem of the rocket engines capable to get there but also causes a long dead time in the remote controlled operations issued from the Earth. While in the distance of Mars the radio systems signal propagation delay is just 20 minutes then, for example in the distance of Jupiter the control loop delay can take several hours. In the cold of the outer solar system there comes an additional problem: it is necessary to heat all the electrical equipment to keep them functional, but we have very low energy budget. We can, however, use radioactive energy source, which is not recommended for ecological reasons.

Fig.1. Rosetta lander (Philae)

It can easily happen that we have to acquire solar energy for days but this energy will be enough just for a few ours of scientific activity. It is clear that a spacecraft far from the sun shouldn't waste its energy and time by waiting for control signals from the Mission Control Centre on the Earth. Direct control is potentially dangerous because the energy balance of the whole spacecraft may collapse in case of an unexpected problem, due to a time-consuming command exchange.

The solution

The only solution for these situations is to increase the rate of on-board autonomy. We have to rely on a built-in intelligent, adaptive control system, which provides the following functions:

- Managing the scientific operations continuously without any interactions with the Control Centre
- Adaptation to the non-predictable timing requirements during the scientific operations
- Giving real-time autonomous reaction to the nominal and non-nominal external events
- Handling emergency situations
- Taking care of the energy balance
- Capability to store the measured scientific data, even in case of energy loss

Having a control system without these capabilities can easily lead to a failed mission. The earlier surface modules had insufficient computing power for long term autonomous missions. In the past in most cases it was unfeasible to plan missions not requiring the Earth intervention for more than a few days. Recent developments in the field of microelectronics make it possible to use such space-qualified microprocessors, which allows creation of on-board software systems having high rate of autonomy. Although these processors still do not provide enough computing power to employ real Artificial Intelligence but a good design model makes it possible to create a real-time autonomous system which can handle all the tasks of a lander unit even in long-term missions.

Considerations

The extent of fault-tolerance of the software system is especially important when forming autonomous strategy of the on-board software. In the design of the software model we have to take the following policies:

- For the sake of the safe operation the measured environment values must be verified according to limits of values, trends, etc. It is equally important to verify the software variables before they are used as actuator signals.
- The internal control model has to be sensible also for the anomalies of its environment.
- There has to be state transitions for every events. This condition was hard to achieve In the case of high number of potential events or bad predictability of the complete event set in the traditional models. What is innovative about our conceptual model is that it provides solution to reconfigure the state transition definitions even during live operation.
- We have to apply timeouts in management of every states.
- In order to keep the reaction time low we have to minimise the execution time in critical or non-interruptible states.
- To minimise the danger of crash or malfunction of the system the received telecommands have to be fully decoded, checked and verified.
- The model can not have logical path causing system deadlock

The task

We've studied in details the requirements for a probe or rover, which should operate on the surface of a planet, asteroid or a comet. We divided all the requirements into the following topics:

- Controlling the Lander unit during approach, descent, landing and surface operations
- Keeping the Power and Thermal balance of the lander unit
- Management and execution of the scientific program
- Collecting and storing the experimental data acquired by the payload and service subsystems
- Management of telecommunication functions (Radio link management, telecommand reception and telemetry transmission)
- Providing fault tolerance by handling the built-in hardware redundancy

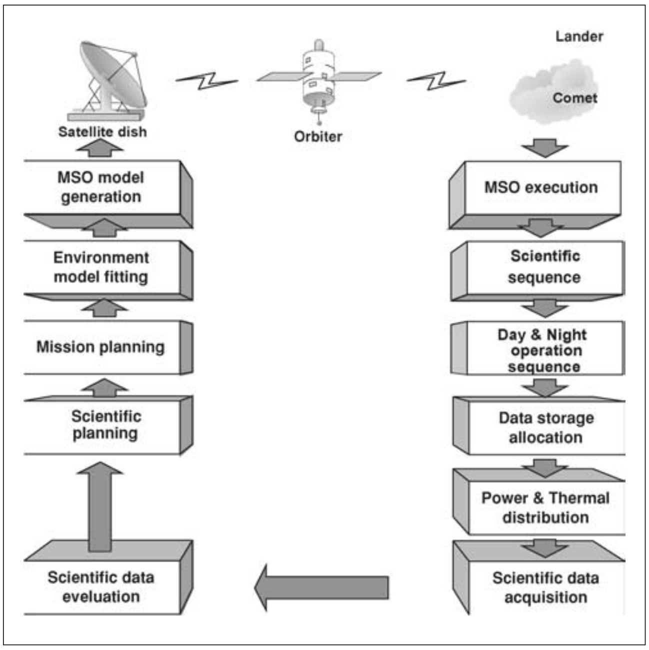
Examining each aforementioned items we came to the conclusion that all of them interact to the others. So an appropriate central logic should provide interaction between them. In implementation, however, it can lead to a very complex control algorithm that is very hard to implement software. To fulfil all the mentioned requirements in a manageable way it is necessary to construct an abstract architectural model which is rather flexible, but as simple as possible.

The model in general

Our central idea is to separate the static and the dynamic behaviour of the system. This modelling approach has many advantages opposite to the concept of the software design of earlier surface modules. This method minimises the required data traffic between the lander and the Mission Control Centre because the various combination of the static and dynamic algorithm reduces the required telecommands for the control. It is an essential point because the upload speed of the telecommands through the communication link is 10-200 bits/second at most from the Earth and the communication session length is usually limited to 10-20 minutes because of the high signal-to-noise ratio. We continuously kept in mind to provide the possibility to easily reconfigure the whole central logic during any mission phase in the lifetime of a lander. So we broke down both our models into a set of individual basic parts. A single part is called Mission Sequencing Object (MSO). The size of an MSO is varying to fit into the telecommand packet. The link between the MSO items make it possible to design them independently which gives us an understandable, easy-to-use man-machine interface for the Mission Control Team. Using this modelling language makes it possible to translate the Mission Control Information to a data format, which is uplinkable to the on-board Central Control Computer via Telecommands.

It is also possible to attach MSOs to the Mission Control Information Database of the Knowledge Management System. Additionally we designed an advanced storage and retrieval algorithm for the on-board control software for storing MSOs to and retrieving them from the on-board memory. This algorithm has dependable but space saving data storage capability and a very short seek-time for accessing further MSO items.

Fig.2. Mission sequencing objects generation and usage on Rosetta lander (Philae)



The static model

The main task of the static model is to generate the actual operational state of the system. The basic MSO of the static model is the SMSO (Static Mission Sequencing Object). An SMSO is responsible for the following system attributes:

- Parameters of the current operation mode
  - Operation speed
  - Rate of failure tolerance
  - Rate of energy saving
- Current set-up of the scientific payload instruments
- Parameters of the data collection for scientific instruments
- Data transfer quotas for the optimised distribution of data storage capability
- Protection for the critical operation phases
- Set up priority order for the currently operating experiments from the following point of views:
  - Energy distribution
  - Data collection and storage
  - Service speed

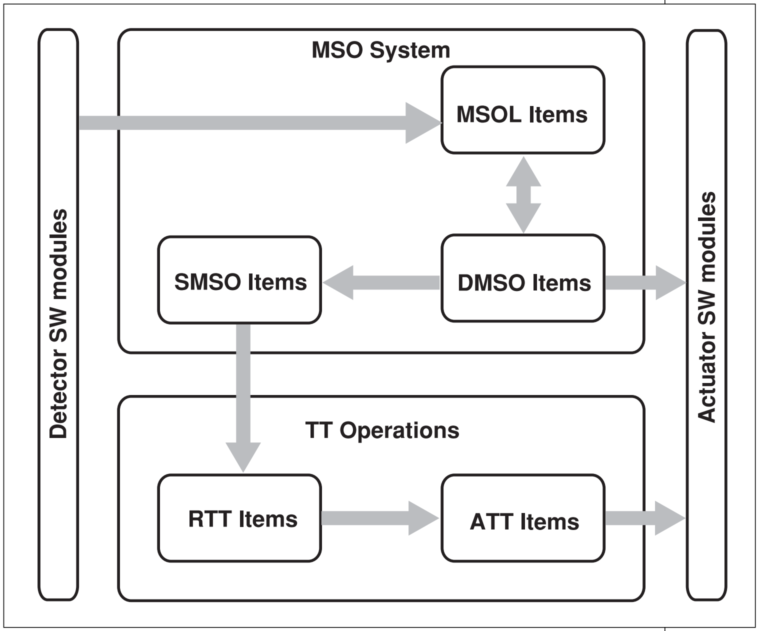


Fig. 3. Structure of the Mission Sequencing Object Model

The dynamic model

The dynamic model describes the required reactions to nominal and non-nominal events and defines the state changes in the static model. The basic MSO of the dynamic model is the DMSO (Dynamic Mission Sequencing Object). A DMSO is responsible for the following system attributes:

- Reference to the current SMSO item
- Nominal and non-nominal events definition
- Reactions to nominal and non-nominal events, which are categorised as follows
  - control-, failure prevention-, failure handler-, recovery- and safe mode-algorithm
- Time tagging and time-out mechanisms
- Link and connection definitions to other DMSO items. Possible connection types are the following:
  - chain like, call like and jump like

The implementation

This model is implemented for the ESA (European Space Agency) cometary mission called Rosetta. The on-board central computer (Command and Data Management Subsystem) of the Philae Lander in the Rosetta mission is equipped with this technique.

The scientific mission of the Lander has not defined yet because there are a lot of uncertainties concerning the attributes of the target object. Using this controlling technique made that possible to finish the on-board software development without detailed information about the scientific program of the mission. The final scientific program will be translated into MSO items and will be uplinked to the Lander via telecommands, just before starting the descent to the comet surface in 2014. Rosetta was successfully launched on 2 March 2004 and its is now on its decade long way to comet 67P/Churyumov-Gerasimenko.

We hope our model will help Philae to accomplish its landing and operating on an ice mountain bouncing around its three axis. In case of success this event may open a new chapter in the history of the Solar system exploration.

The software environment

The on-board software of the central computer of the Philae Lander consists of a real-time operating system and 8 application tasks. All these software modules are specially developed by our team for the Harris RTX2010RH microprocessor. The co-ordination of the scientific program and the overall control of the algorithms used by the application tasks are done by the MSO modelling language.

References

[1] Ron S. Kenett, Emanuel R. Baker:
Software Process Quality ,
1999 New-York

[2] Savio Chau, Abhijit Sengupta,
Tuan Tran, Ali Backhshi:
Ultra Long-life Spacecraft for
Long Duration Space Exploration Missions
Space Technology Vol. 23, 2003

[3] David P. Youll:
Making Software Development Visible,
1990, Chichester

Land ahoy! Land?
A cautious market snapshot of the infocom-industry

GYÖRGY BÖGEL

Strategic consultant at KFKI Computer Systems Corp., Adjunct Professor at CEU Business School, and Associate Professor at the University of Debrecen

gybogel@kfki.com

Keywords: e-business, Carlota Perez's model, mass commodity

The ragged sails of the ship weave along the still-fierce waves of the sea. Sailors are busy cobbling together a new rudder, patching the sails, and repairing the damage caused by the storm. An exhausted crewman sits in the crow's nest, wearing his striped uniform and with his eyes on the horizon. At first he only sees a faint blur; he doesn't dare give the signal, perhaps it's only a cloud or a patch of fog in the distance. But as the ship sails on, he gets a clearer image - yes, those are cliffs and that yellow must be sand in the foreground. Finally he raises the cry: "Land ahoy! Land!"

"Land?" – the captain asks himself in his cabin. "That's all well and good, but what kind of land is it? There are no maps of this part of the world. That patch in the distance is as likely to be a tiny island as a new continent, or it might even a dangerous reef. What can we expect, for what should we be prepared? Will we find a quiet place to set anchor? Will there be food, water? Can the land be settled, is it fertile? Will there be room for everyone, or are we in for an extended struggle with hostile natives?"

After the storm

The info-communications industry has seen some difficult years. After the long boom of the 90s, demand for info-communications products and services fell back. Stocks began to plummet, investor confidence waned, and the captains of the industry were no longer in the headlines. The storm broke quickly, and in its wake were leaky companies with tattered sails, as well as people wracked with uncertainty.

The older industrial sectors weren't as surprised by the *recession*; they had been through this sort of thing numerous times. But information technology is a relatively young industry, and for many this was the first serious trial of their lives. Fortunately, youth also implies a readiness to learn. Most of the companies that survived the storm did the same as the old sea-dogs: they cut costs, reduced their capacities to be in line with shrinking demand, restructured or closed divisions that were losing money, streamlined their profiles, reduced their debts, and focused more on efficiency and productivity. And the richer ones looked around at the battered market with an eye to what can be easily and cheaply bought up, in order to build a bigger and stronger ship.

The fleet has been reorganised and the weak have sunk. A few illusions have been lost, some big lies have been exposed, and everyone has become a bit older and more experienced. These are ancient problems and solutions.

It appears that the storm has now passed. Land can be seen in the distance, and the sails that were torn down during the storm are hoisted anew. A fresh, cheerful breeze fills the canvas sheets.

The info-communications market is once again expanding, and many market players feel that they have recovered. The stock exchange is optimistic; prices are on the rise. People are showing a keen interest in products from the digital industries: they take their laptops with them in search of WI-FI cafés, they upgrade their computers, buy digital cameras, trade their mobile phones in for later models, sample 3G services, use broadband to surf the Internet, and install cable or wireless networks in their homes so that every family member can enjoy the bounty. Thanks to lower prices and products tailored to their needs, smaller businesses are pleased to find the world of e-business waiting with open arms. Larger companies are still somewhat wary of investing in major projects, but their desire to get the most out of their already-built systems means that they still give service providers a great deal of systems integration work. E-commerce indicators are improving, sometimes by leaps that catch even serious analysts by surprise.

Our imaginations and fantasies are once again being stimulated by technical novelties and innovations, such as "grids", utility computing, software on demand, web services, radio technologies that sometimes clash with and sometimes complement each other, and "smart dust", dust-like "grains" complete with sensors and transmitters which may someday replace the barcode. China, a key driver of the global economic boom, is showing massive demand towards the technology sector. "E-Biz Strikes Again!" trumpets the headline from an issue of Business Week this past May, and even the conservative and restrained periodical, The Economist, has made optimistic pronouncements about the future of digital industries. Following a three-year freeze, four new Internet companies have appeared on the American stock market in 2004, and a further 24 have submitted the required documentation and are awaiting approval. Even more surprisingly, 20 out of these 28 companies are showing a profit, something that was true only for 4% of companies that made their IPO between 1998 and 2000. In 2003, these same 28 companies showed a 56% growth in income and a 490% growth in net profits compared to the previous year.

"Land ahoy! Land!" comes the cry from the crow's nest, and sure enough mountains are emerging, and the sandy shores are visible. "Land?" The captain mutters to himself. "What kind of land? Where is this ship going?" According to market indicators, the info-communications industry is once again on the rise, which is good news indeed. We can perhaps refer to the years of recession as a minor accident, a temporary slump after which things have once again returned to the right track. Market economies work in cycles: upswings are followed by declines, then by newer upswings. It appears simple enough - we need to weather the times of recession, get rid of the ballast and trim the lines, and after the storm has passed, hoist those sails and continue onwards.

Unfortunately, the alarming fact is that the actual situation differs from this. In all likelihood, the start of the new millennium has signalled the end of an era for this industrial sector. The new times ahead will require new strategies. We can continue sailing, but in a different manner than in the "golden age" of the 90s.

Riding the waves

There are a number of *models* available to help us understand the reshuffling, the cyclical movements, the repeating patterns and lasting trends. The models that are of use to us here are those that have something to say about the connection between the info-communications industry and consumer/users, about the relationships between supply and demand, in terms of indicators for both quantity and content/ quality. Some of these models are macroeconomic, while others are best applied in a more concrete way to help us understand the market movements of individual products or types of service.

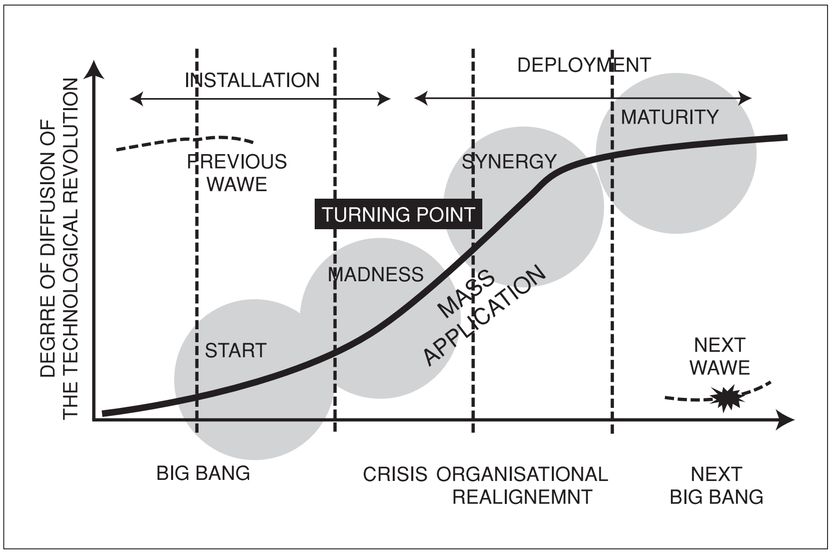
There are some models that, by analysing the economic and social effects of the major *technological innovations,* (the steam engine, railroad, electricity, etc.), seek to find regularity and repeated patterns. According to Carlota Perez [8], an oft-cited Venezuelan researcher, waves of innovation can be broken into two major periods: the installation period and the deployment period *(Fig. 1).*

As its name indicates, the *installation* period is the time during which the infrastructure for the new technology is established. Some concrete examples include the building of a network of railways; the equipping of factories with electronic motors; the appearance of automotive plants along with petrol and service stations springing up like mushrooms beside new motorways. The installation period itself can be broken up into two phases. During the first phase, the new technology incubates, seeking its place while its potential is not fully known yet. The second phase is characterised by a "big bang" that awakens general interest: entrepreneurs see the technology's "grand opportunities", investors open their wallets, a great deal of excitement is generated, and demand often exceeds supply. This is the period during which illusions and false hopes can appear. After all, the bubble must burst in time. Because of this, the installation period generally ends in a crisis.

Fortunately, the crisis is only temporary, and it doesn't signify the end of the wave of innovation. The second phase of the installation period can see irrational behaviour: an excess of wonderment at the new technology, inflated and foolhardy investment, blind gambling on the stock market. This "madness" actually serves and important function: it aids the quick development of the infrastructure built on the new technology. Tracks are laid at a lightning pace, machinery is replaced, roads are built, cabling is completed, and service providers are created. In a sense, the following crisis creates order. The weak are weeded out, overblown capacities are scaled back, the stock market calms down, and a more sober reality prevails.

During the installation period, businesses that build/ spread the new technology typically envision the wide open spaces of new hunting grounds before them, and they consequently focus on growth and raising capital, pouncing on scant resources. They want to sell, and care little for what actual use the thing they sell is put

Fig. 1.
Carlota Perez's model of the phases of technical innovation



to, they care only for the next eager customer that pops up. Since interest is high it is easy to raise funds, and new businesses quickly multiply.

During the *deployment* period, the new infrastructure is largely in place. People, businesses, and various organisations are increasingly routine in their use of the technological innovations. After a time, the existence of the new infrastructure is regarded as part of the natural order of things: we think nothing of travelling by car, calling grandma in another city, or the fact that there is electricity in the wall socket. It's all ordinary, and cause for alarm only on the rare occasions when the system fails: when there is a blackout, air traffic is backed up, or the phone line is dead. It is only at this point that the wave of innovation truly reorganises economic and social structures, in a manner that is quieter yet deeper than during the previous period. Inexorably it becomes part of everything: factories, offices, homes, culture, the state, and politics.

Businesses are established that base their competitiveness on their ability to apply the technological innovation intelligently. A consumer lifestyle develops that utilizes the new infrastructure. New procedures and habits evolve. The issue of the day is no longer, for example, whether to quickly build a new railway line, but rather how to create a simple, standardised, and unified system of railroad use, or how production and sales can take advantage of the existence of the railway, or establishing where we should place the mines and factories and how large a geographical region we can select our workforce from. The question is not "how can we lay more cable underground and in the ocean", but "how can we encourage people to use their phones more often", and "how can the new technology be applied to education, business and governance?"

This deployment period is a longer process for the technology, its infrastructure, the economy and society. It is not as loud as the colourful and spirited world of the previous period, but the effects are longer-lasting and of greater consequence. This continues until the given technological innovation's wave dies out and something new takes its place.

The typical business during this period operates in a more consolidated market, since the crisis at the end of the installation period has decimated and reorganised the field. Growth slows, and the start-up fever abates. Consumers/users soon realise that the ball is now in their court. They are cautious and suspicious; flashy ad campaigns are less effective now. In this phase, consumers/users are developing their use of the technology, becoming increasingly creative with it, but they are also more deliberate, with an eye to balancing the usefulness of the technology against the costs. What interests them is not so much owning the technology, but *using* the technology. Whoever wants to sell a product or service needs to focus on its application, and guarantee its profitability. Efficiency and productivity are the watchwords both on the buyer-side and the seller-side. Buyers of the technology want to be more efficient and competitive, and sellers can only turn a profit in this consolidating and maturing market if they rein in their expenses. Relationships are crucial, since at this point if a customer is lost, it is very difficult to find another to take his place.

If we observe the events of the recent past, we should come to the following logical conclusion: the *info-communications cycle of innovation* experienced an installation period in the 90s, has gone through its crisis at the end of this period, and the ship is now in the deployment period, making its way towards the less turbulent waters of adoption. One of the most important "products" of deployment and adoption, the "integrated, real-time, extended electronic business" [1], is developed during this second period. Its walls are built upon the foundation of the already-developed infrastructure. Info-communications technology fills the space step-by-step. First, individual tasks are automated via this technology, then entire functions and processes. After this, these isolated systems are integrated, followed by supply chains that connect multiple businesses [6]. The process seems unstoppable, and its long-term effects are unpredictable.

The deployment period demands different strategies and conduct than the installation period. *IT services* provide a good example of the shifts in emphasis and the transformation of strategies. This is an industry with an annual turnover of 520 billion dollars, yet which, in spite of the explosive growth experienced in the 90s, could only show a growth of 3% for each of the last two years. Sensing the limits of the market, small and large businesses in this sector have become rather ingenious in finding ways to increase efficiency and reduce costs, at the same time demonstrating what the technology is capable of. They seek cheaper labour, and move some of their activities to countries such as India (which has quickly and rather cleverly leapt at the opportunity); remote real-time control is now technically possible, spurring a wave of outsourcing. Inflow Inc., which operates as a data centre serving hundreds of clients can be found in a 2000 square metre building packed with humming equipment; one or two employees hover about, but everything is essentially automated. Packages are compiled for companies in various industries using Accenture software and services, after which the given packages can easily and quickly be customised.

Wipro Technologies, one of the crown jewels of the IT industry in India, has automated software-development processes and boasts of a program that can translate from six European languages into English with 99% accuracy. Its sister company, Infosys Technologies, built upon web-service technologies to create a standard library out of reusable software modules. When developers at the company are given a new task, they take these modules off the shelf and combine them to suit the particular demands at hand. Getronics, a Dutch company, has automated the process of diagnosis and support for desktop computers, thanks to which they were able to cut their necessary workforce in half. Thanks to automation, certain server software installation tasks that previously took 5-10 days at IBM, can now be performed in just a few hours.

Though the technology is new, from an economic perspective the methods applied in the above examples are old. Even today, the businesses that manage to increase employee productivity are those that are able to exploit the advantages of mass production, standardise operations, use finished parts, find cheaper suppliers, carry over any advantages from one activity to the next, learn quickly, etc. These "basic methods" are employed by Chinese companies specialising in mass production as well as by flagship American and European companies, though in different ways and with different content.

IT services must take care not to cause problems for their customers when reducing costs and automating their processes. Fortunately, there is a convergence of interests here: efficient and cheaper service providers in turn allow their customers to be more efficient and cheaper and ultimately more competitive. To this end understanding the technology is not sufficient, since using the technology effectively requires human and organisational changes. In education as in application development, the question today is no longer "technology or business"; the people most in demand are those who are at home in both areas. It is no accident that IT and telecommunications companies are eager to climb up the value-chain beyond production and basic services to include high-level business consulting.

The *terms of a contract* can influence goals, attitudes, and expectations. For a long time, the trend was to bill for business IT consulting based on the hours worked. It was in the service provider's interest for ever more people to work on a given job. Nowadays the situation has changed. According to well-known market analyst company IDC, only 20% of today's consulting contracts are based on a traditional hourly rate, as compared to 85% just four years ago. Currently, fees tend to be performance-based. Service providers receive their fees if they manage to increase efficiency, if turnover increases according to plans, or if costs are reduced, and the number of consultants working on a specific job is irrelevant to the client.

According to Carlota Perez's model, madness, crisis, and sobering up predictably follow each other. The Gartner Group's well-known "hype cycle" implies something similar, but on the level of individual products or product ranges rather than on the macroeconomic level. Technological innovations generate great interest, which is kindled by manufacturers, marketing professionals, newspapers, consultants and conference organisers alike. After all this hype comes the inevitable disillusionment ("this isn't the panacea that will cure all my ills"). Realism follows the disillusionment ("well, it may not be the wonder drug, but it can be useful for treating this and this particular illness"). And finally, the innovation finds its place in the world. Carlota Perez's model, introduced earlier, demonstrates what happens when a technological revolution causes an entire industry to enter the hype cycle.

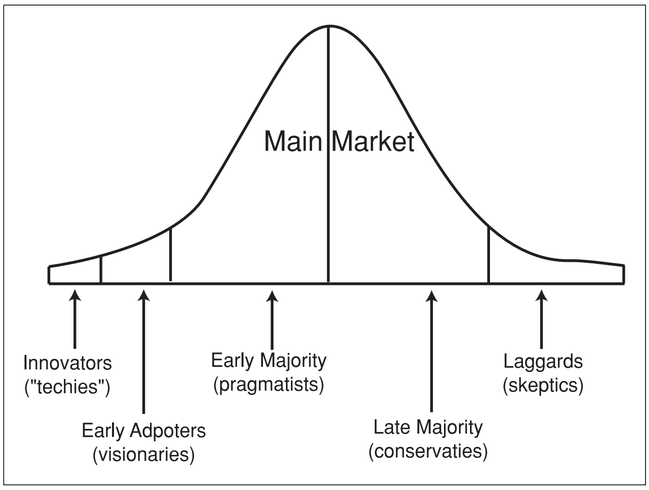
Changing customers

Geoffrey Moore's model [5] is also instructive, and its application can help explain a number of features of the info-communications market. According to the founder and president of the Chasm Group, the market adopts new technologies step by step *(Fig. 2)*. The individual groups adopting the technology differ not only in size, but in needs, expectations, and habits. They have different interests and can be inspired by different things. You may conquer one group, only to find that the same methodology is a complete failure with the next group. Those who do not take this into account and fail to change in time are trapped by their own success.

The first group to notice an innovation has but a few members. They are people who are interested in technology for its own sake, rather than in what can be done with technology. We are speaking about *technocrats,* enthusiastic and curious, whose tables and pockets are filled with all sorts of gizmos, but who are rarely decision-makers. The innovation only interests them as long as they are unfamiliar with it. Afterwards, they turn elsewhere and wait for the next innovation. Next in the process of adopting a technology comes those who see great *strategic opportunities* in the innovation, those who say "here's something that can put me ahead of the pack!" They think in terms of business rather than technology. They are daring and willing to take risks, but unfortunately they too are few in number.

Next comes a more populous group, the pragmatists of the *early majority.* They are not revolutionaries and are averse to taking risks. They wait for the technology to prove itself and for getting positive recom-

Fig. 2.
Geoffrey Moore's technology adoption model


Main Market
Innovators ("techies")
Early Adpoters (visionaries)
Early Majority (pragmatists)
Late Majority (conservaties)
Laggards (skeptics)

mendations. They believe their own eyes rather than “the hype”. They are willing to learn and invest, but they do not want to be first at any cost. According to them, “the prairies are filled with pioneers with arrows in their back.” They do not expect using the technology to result in radical changes or great leaps, and they prefer smaller and safer steps. They plan for expected profits, they are careful with expenses, and choose their suppliers carefully. They are numerous, and the first serious recommendations are likely to come from this group.

These recommendations are important, because the early majority is followed by the *late majority*. The typical member of this group favours inexpensive, tried-and-tested solutions. This group is won over by a technology’s obvious benefits and ease-of-use. They wait patiently for the new technology to become a mass commodity, and then go shopping. They are wary of technology, perhaps even a bit afraid of it. They are reluctant to understand it, and would rather that the technology understand them. If they are frustrated, they quickly retreat and can lose their taste for the whole affair. They prefer simple, easily-understood solutions which they will stick to if things pan out. They don’t want to build a generator in the basement, they want the electricity to come out of the socket: simple, cheap, and reliable.

If you still want to win more people over, the late majority is followed by a *laggards* group that you can set your sights on. It won’t be easy. Members of this group question everything, and they will gleefully refer to failures (easy enough to find among IT projects). They will bring your attention to the often vast difference between the promises and realities of a technology. They frequently exclaim “the emperor has no clothes!” They are a tiresome lot, but there is much to learn from their observations, misgivings, and questions.

The classic marketing *lifecycle model* implicitly fits beside Geoffrey Moore’s abovementioned model. It presents a simple, often-experienced pattern whereby the life of a product or product range sees a standard progression of phases one after the other: introduction, growth, maturity, and decline. It is easy to see the parallels between the two models. The freshly-introduced product at first only interests the technology-obsessed. Growth occurs as the visionary risk takers catch on, and expands with the early majority. The mature and proven market belongs to the late majority, and finally a few reluctant laggards may be won over.

And where is the info-communications industry as a whole in this adoption lifecycle model? Many indicators show that it is conquering the late majority, in the mature phase. The era of “garage-assembled” and difficult-to-handle machines is over. We were also witness to major strategic leaps: some of these visionaries sank in the storm, others (e.g. eBay, Dell, Amazon.com) are truly on the edge of modern technology. The early majority has already built their internal infrastructures, purchased and installed their systems, and, as mentioned, they are striving to increase their efficiency. Now is the time to win over the late majority, with a strategy and battle plan suited to a mature market.

The mass commodity rebellion

Mature markets are characterised by mass commodities, which is exactly what the late majority craves. Mass commodities are the basis and the engine of the upswing that followed the info-communications industry crisis, though this is not necessarily good news for everyone. However, market drivers such as consumer and small business products and services would never have developed without this.

A product that is to become a mass commodity must be standardised, inexpensive, easily replaced, easily learned, as well as compatible and connectable to most everything. And these happen to be the catch-phrases and developmental direction of the info-communications industry.

The leading product of the industry, the desktop computer, provides a good example of the process of becoming a mass commodity. In a relatively short time, the PC has become a standardised, easily installed, and simple-to-use product. Most run on the same operating system, the same microprocessors, and the same software. They can be connected to anything, especially one another, which was an early fundamental condition of the Internet. The prices have shrunk, and can no longer be considered a serious obstacle. They can be found in stores and plazas everywhere, just put them in your shopping cart.

The same thing is happening today to servers, work stations, as well as network and storage tools: the most popular of these are inexpensive, easily installed and upgraded. Google, the company behind the popular Internet search engine, bought its hardware off-the-shelf, based its system on older microprocessors, and used inexpensive or free open-source software. The newspapers are filled with the news of the huge sums saved by General Electric and Amazon.com with the purchase of inexpensive mass commodity IT equipment. Dell is positioning itself as a provider of computers for the masses and refocuses its cheaper product lines, while spending less on research and development that, for example, Sun.

One typical feature of the mass commodity orientation of the info-communications industry is “overdevelopment”: products that are capable of far more that the average consumer expects from them. This explains why such leading companies as Google, GE, and Amazon.com can forego a constant push towards innovation and be satisfied with earlier generations of computers. Without the presence of “overdevelopment”, Dell would be less successful as well, seeing as its cost-saving ploy is based on earlier technological innovation.

The concepts of "utility computing" [7] and "software on demand" are good symbols of a possible direction for mass commodification. In this way, with the participation of some significant players, the same thing may happen to IT as happened to water and electricity provision. Nowadays, nobody runs their own generator or waterworks when they can access a tap or electricity from a socket. According to the vision, future users won't purchase and install applications and systems for themselves, but rather rent a service from a "public utility" whenever they require something (for example, a customer relations management application). When the user begins using the product, the taxi meter starts running, and when the user turns it off the meter stops. The user need not worry about maintenance and development – just leave it to the service provider. This is an entirely new economic model than that of purchases and installation based on large investments and fixed budgets. There is a growing number of examples for this, including stock exchange hopeful Salesforce. com, Taleo and Right-Now Technologies who offer software over the Internet at roughly 65 USD per user per month.

The move towards mass commodities comes with some unpleasantness for representatives of the infocommunications industry. Differentiation becomes more difficult, competition becomes more fierce, profit margins shrink and companies have to work much harder to achieve similar results. (Just look at the PC market: because of the simultaneous drop in prices, the boom in quantity meant only a small increase in turnover for manufacturers and retailers.) In spite of this, the process is self-perpetuating and unstoppable. The logic that we need to standardise, work with small units, make things compatible, avoid monopolies, distribute and spread everything throughout a wide sphere, and the ideas that we must be organised, that standard things must be handled in a standard fashion, etc. – all this is burned deep into the behaviour of the info-communications market, on the buyer-side and seller-side alike [11].

A typical response to cost-based competition and the move towards mass commodification is the urge to commence *factory-type operations,* which can be seen in many manufacturers and service providers. Take software development as an example. At the end of the 50s there were barely 20,000 software experts in the entire world. Today their number is estimated to be nearly nine million. Back when one had to work in machine language, writing software was a complicated and difficult process. Today, countless tools are at the disposal of software developers, making their work easier. As the software demands of companies become standardised and as software becomes increasingly modular, its development (at least a large part of its development) becomes increasingly like routine manufacturing. In this regard, software development migrates towards those parts of the world where such manufacturing activities are cheap and well-run. The global service model of Indian companies is predicated on this logic (e.g. Infosys, Wipro, Tata, and Satyam). Over the course of a project the client must conduct the specific situational analysis, while the "manufacturing" is built up from modules by the inexpensive, well-organised, hinterland outfit...

"Land ahoy! Land!" – comes the cry once again from the crow's nest. It would be good to have a more precise map and to see more clearly, thinks the captain to himself. Then he goes to the bridge and gives out his orders. We'll see what happens. Come what may, we must sail...

References

[1] Bőgel György, Forgács András (2004): Informatikai beruházás – üzleti megtérülés. Műszaki Könyvkiadó, 2004.

[2] Carr, N. (2004): Does IT Matter? Harvard Business School Press, Boston

[3] Gates, B. (1995): The Road Ahead. Viking, New York

[4] Kocsis Éva, Szabó Katalin (2000): A posztmodern vállalat. Oktatási Minisztérium, Budapest

[5] Moore, G. (2002): Crossing the Chasm. Harper Business, New York

[6] Murphy, T. (2002): Achieving Business Value from Technology. John Wiley & Sons, Hoboken, New Jersey

[7] Ördög Péter: Utility computing – Informatikai közművek. Diplomadolgozat, Debreceni Egyetem, Közgazdaságtudományi Kar, 2004.

[8] Perez, C. (2002): Technological Revolutions and Financial Capital. Edward Elgar, Cheltemham, U.K.

[9] Porter, M. (2001): Strategy and the Internet. Harvard Business Review, march-april

[10] Salamonné Huszty Anna (2000): Jövőkép- és stratégiaalkotás. Kossuth Könyvkiadó, Budapest

[11] Shapiro, C., Varian, H. (1999): Information Rules. Harvard Business School Press, Boston

Automated test suite generation from formal protocol specification

GÁBOR VINCZE

Budapest University of Technology and Economics, Dept. of Telecommunications and Telematics
vincze@alpha.ttt.bme.hu

Keywords: conformance testing, test generation, mutation analysis, evolutionary algorithms, bacterial algorithm

In this paper, we present a method for automatic test generation from the formal SDL specification of a protocol. Protocol testing is an important step in the development process, but the creation of test suites is a time consuming task. Automating this phase reduces the time necessary for implementation, and cuts an important error source. We show how Mutation Analysis can be used to match test criteria and test cases obtained with a graph exploration algorithm applied on the SDL description of the system. We then use evolutionary algorithms to select an optimal subset from this initial set of test cases. Using these methods, we build a complete process for the automated generation of a test suite from the formal specification of a protocol.

As telecommunication companies had to offer more of services every day, while trying to integrate their networks, telecommunication protocols became increasingly complex. At the same time, the reliability of these networks had to meet ever-higher standards as well.

With this increase in complexity, the effort needed for the specification of protocols became a serious burden, and the need for reliability and interoperability between manufacturers called for more extensive testing. These problems gave birth to formal specification methods, and formal testing methods to verify if implementations behaved according to the specifications.

The most widely used formal languages in the world of telecommunications are the Specification and Description Language (SDL, [1]) for system specification, which models a system as parallel Communicating Finite State Machines (CEFSM), and Tree and Tabular Combined Notation (TTCN, [2]) for black-box test description. Today highly integrated and widely used development tools [3] exist to aid designers in the specification and testing process. However, the creation of a formal test suite still requires considerable effort, and the human factor remains the most expensive and error-prone component in the process. As these tests often have to be run several hundreds or thousands of times, execution time and hardware requirements are also a crucial factor.

In this paper, we present a method for automatic test generation from the SDL description of a system. The test generation process has four main steps:

1) formal specification in SDL
2) creation of a set of test cases by a state-space exploration algorithm
3) mutation analysis
4) selection of an optimal subset of test cases

We will first explain the mutation analysis method in detail; then, we will show how we use evolutionary algorithms to select an optimal subset of test cases from the resulting set, and finally, we will illustrate the whole test generation process by an example on the INRES protocol.

1. Mutation analysis

1.1. Overview

Mutation analysis is a white-box test case development method, which means we possess knowledge on the internal working of the system. Traditional mutation analysis has been developed to find errors in program code, but we use it here on formal protocol specifications instead to select the appropriate black box test cases.

In a mutation analysis system, we need to define a set of mutation operators [4], where each operator represents an atomic syntactical modification. The use of these operators is convenient for two reasons: They allow the formal description of error types, and they allow the automatic generation of mutants. By applying systematically the operators on the specification, we can generate a set of mutants.

A mutation analysis system is made up of three basic components:

– The original system;
– The mutant system, which contains a small syntactical modification compared to the original system. Mutants are obtained by applying the mutation operators, each operator representing a small syntactical modification;
– An oracle – a human or, in most cases, a program, which differentiates the original system from the mutant by observing its interactions with the environment.

We assume that the original CEFSM specification is close to the requirements, and thus test cases detecting syntactical modifications of the specification are useful.

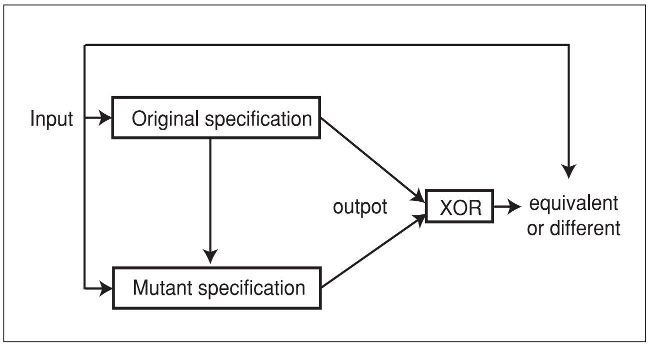


Fig. 1. Mutation analysis

We only produce first-order faults – we apply one mutation at a time – because test cases detecting simple modifications will also detect complex modifications created as a sequence of simple modifications [5]

Test cases distinguish a mutant from the original if it gives a different output. However, part of the mutants generated by the operators might be semantically equivalent to the original system: they give exactly the same output on all possible inputs. We call these mutants *equivalents*. We call *pseudo-equivalents* mutants that are semantically different from the original system, but give exactly the same output on all possible inputs. We should ignore all equivalents during testing, but should consider all non-equivalents during test case selection. This creates a serious problem in mutation analysis, since it is generally not possible to automatically identify equivalents, and the distinction between equivalents and non-equivalents needs human interaction.

1.2. Mutation operators

It is a very important consideration that mutation operators do not create pseudo-equivalents, and minimize the number of equivalents. The basic principles of mutation operator definition are:

- Operators should model atomic faults;
- They should only create first order mutants;
- We should only generate syntactically correct mutants;
- To allow test case generation, we should only create semantically correct mutants;
- Operators should generate a finite, and the lowest possible number of mutants.

Five classes of operators have been defined [4] for CEFSMs, depending on the part of the CEFSM they modify:

- state, input, output, action, and predicate modifying operators.

For each class, we can give three types of operators depending on the type of fault they represent:

- augmenting, reducing and exchanging operators.

1.3. Test case – test criteria matching

The following algorithm allows us to assign a set of test criteria to each test case of a finite sized, unstructured, and highly redundant test suite, which we can obtain for example by a state-space exploring algorithm exploring the system specification. If we apply the mutation operators to observe inopportune inputs, this initial test suite must also contain inopportune test cases.

Let C be a two dimensional matrix of Boolean values.

```
0) Generate a set of test cases;
1) Apply a mutation operator on the CEFSM to
   create the $i^{th}$ mutant;
2) Run all the test cases on the mutant specification,
   and observe inconsistencies: if the test case
   gives a different result from the original specifi-
   cation, the test case detects the given mutant;
3) Create column vector $Ci$ ($i^{th}$ column of the C matrix)
      – Let $Ci[j] = 0$ if the $j^{th}$ test case
        cannot detect the $i^{th}$ mutant;
      – Let $Ci[j] = 1$ if the $j^{th}$ test case
        detects the $i^{th}$ mutant;
4) Repeat steps 2-4. where i goes from 1 to N,
   where N is the number of all the possible
   mutants;
5) Acquire the C matrix of criteria,
   where rows represent test cases of the original
   set, and columns represent the mutants.
```

2. Test selection with evolutionary algorithms

The aim of the selection process is to obtain an optimal subset of test cases from an already existing unstructured, highly redundant set. To achieve this goal, we applied three different soft algorithms: the Genetic Algorithm (GA), the Pseudo-Bacterial Genetic Algorithm (PBGA) and the Bacterial Evolutionary Algorithm (BEA).

We chose to use evolutionary algorithms for test selection because they provide high quality solutions in acceptable time, can handle very complex cases, and can be easily integrated into the test generation process [6].

2.1. General considerations

Individuals: An *individual* is a possible solution of the problem, an optimized test suite in our case. We had two different ways of representing test suites: either a fixed length string of N bits, where N is the number of all the test cases in our original set, each bit's value being 1 if the test suite includes the corresponding test case (which we called *bit-string* individuals), or a variable size set of values between 1 and N, where each number represents the number of a test case in the original set (which we called *pointer-set* individuals).

In the latter case, it is of course possible to have test suites that incorporate the same test cases twice, but these will have an increased execution cost without any added value and will be eliminated during the selection process. We used one or both representation methods, depending on the algorithm.

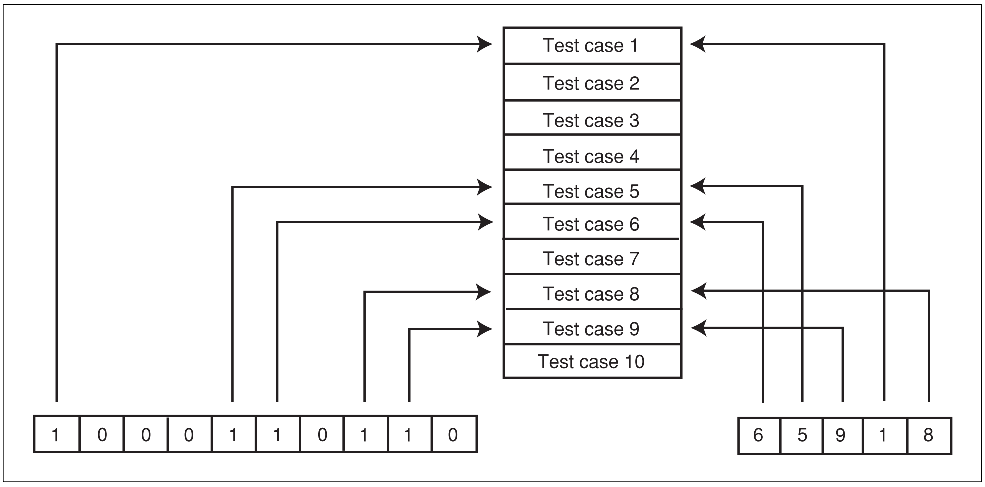


Fig. 2. Bit-string and pointer-set individuals

Test suite cost: test suite cost represents the execution cost of a test suite, which can mean execution time as well as hardware requirements.

Let T = {t1,t2,...,tn} be the test suite containing test cases t1,t2,...,tn , and R = {r1,r2,...,rk} the test requirements covered by this test suite. We assign the c : T → R positive function to each set of test cases. The execution cost of any given T set of test cases is then defined as

$$c(T) = \sum_{t \in T} c(t) \qquad (1)$$

The cost of the individual test cases can be arbitrarily assigned, or measured during the mutation analysis phase. We will suppose here that the checking of each test requirement needs a certain amount of resources and execution time, and the initialization of each test case requires a certain amount of resources as well.

Thus the cost of a test case will be given by

$$c(t) = c1 + c2*L \qquad (2)$$

where *c1* is the initialization cost, *c2* the cost required to check each test requirement, and *L* the number of covered test requirements.

Objective function: the function that evaluates the quality of each individual, and which the algorithm tries to minimize. To obtain the desired test suites, the objective function should take into account the following:

– Execution cost of the test suite should be minimized, by minimizing the redundancy in the test criteria covered by the test cases;
– The test suite should cover all test criteria.

Our objective function is the sum of the execution cost of all the test cases in the test suite, and a penalty for each untested requirement

$$O = c3*C + c4*M \qquad (3)$$

where *C* is the cost of the individual, *M* is the number untested requirements, and *c3* and *c4* are weighting constants, which must be chosen so that it isn't economical to omit test cases.

2.2. Genetic Algorithm

The Genetic Algorithm is an optimization method trying to model the process of natural selection [7].

The canonical GA, which we employed here, works as follows:

```
Initialization
Create initial population
Evaluate of initial population
generation := 0

Generational loop
{
    Calculate fitness values
    Selection
    Recombination
    Mutation
    Evaluate of new individuals
    Insert new individuals into population

    generation := generation + 1
    } while generation < max. generation
```

The individuals are bit-string individuals, as the implementation of the crossover step was much more intuitive in this way.

Let's examine each algorithm step in detail:

Fitness: Fitness of individuals is evaluated by the linear rank-based method, where the *Fi* fitness of the i^{th} individual is given by

$$F_i = 2 - sp + 2*(sp-1)*\frac{pos(fi)-1}{N_{ind}-1} \qquad (4)$$

where *sp* is the selection pressure (here *sp*=*2*), *pos(fi)* is the position of the i^{th} individual based on the value of the objective function, and N_{ind} is the population size.

Selection: Individuals are selected for breeding by the Stochastic Universal Sampling method: individuals are mapped on an axis where each individual has a length equal to its fitness.

We then generate a random number in the [1..nb_parents] interval, where *nb_parents* is the number of individuals we want to select for breeding. We then add i*(sum of all fitnesses)/(nb_parents) to this value, where i ∊ [0 .. nb_parents – 1], and select each time the individual to which this value points on the axis.

Recombination: We use the uniform recombination method: we generate a random bit pattern; bits of both parents are then inverted where the value of this mask is 1, giving the offspring.

Mutation: All offspring are mutated with a small probability to allow drastic changes. Beginning at a random position, we invert each bit of a predefined length section with *Pm* probability.

2.3. Pseudo-Bacterial Genetic Algorithm

Bacterial algorithms, developed in the second half of the '90s, model evolutional processes of bacteria. The simplest bacterial algorithm is the Pseudo-Bacterial Genetic Algorithm [8].

At the beginning of the algorithm, we create a random individual, on which we apply the bacterial mutation. We make *n – 1* copies (clones) of the original individual. Then we randomly select a part of the chromosome, which we mutate in each clone, but leave unchanged in the original individual. After the mutation, we evaluate each individual, and transfer the mutated part of the best individual to the other clones.

We repeat this mutation-evaluation-selection-reinsertion cycle until we have mutated all parts of the chromosome. We then select the best individual, and annihilate the others. We can repeat the cycle until we have a satisfactory solution, or we have reached a predefined generation number.

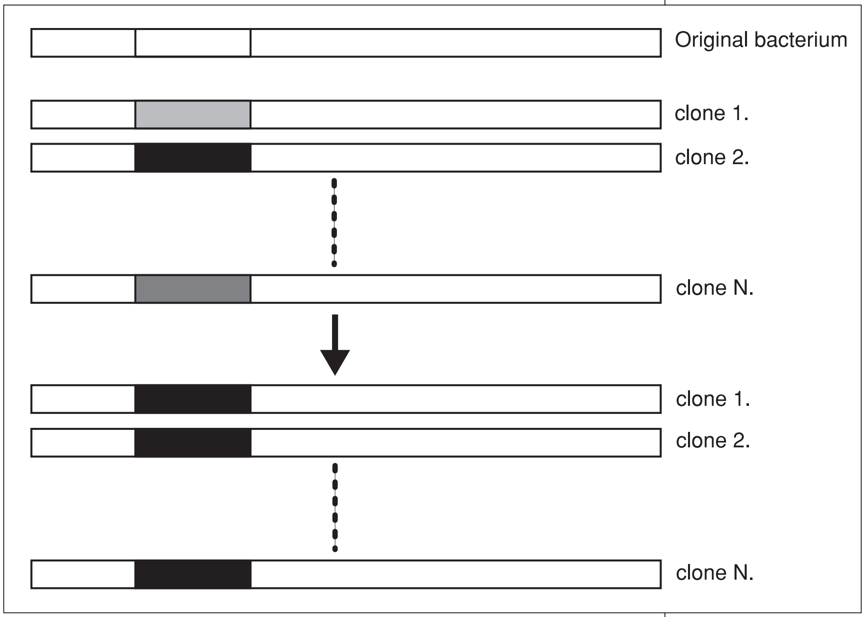


Fig. 3. The Pseudo-Bacterial Genetic Algorithm

We have implemented this algorithm with both types of individuals. In the case of bit-string individuals, the mutation step is the same as in the case of the GA. In the case of pointer-set individuals, a mutation has to allow changes in the length of the mutated individual (since individuals have no predefined length, and we have no a priori knowledge of the optimal length).

Thus, mutation can induce three kinds of modification:

- The substitution of a test case by another test case;
- The deletion of a test case;
- The addition of a test case.

2.4. Bacterial Evolutionary Algorithm

The Bacterial Evolutionary Algorithm is an improved version of the PBGA, which works on many individuals in parallel. It was inspired by the gene transfer ability of bacterial populations [9].

The algorithm works in the following way:

1. We create a random population of *n* individuals.
2. We apply the bacterial mutation (as shown in 2.3) on all individuals.
3. We apply the gene transfer operation *Ninf* times, where *Ninf* is the number of infections: during this step, we divide the population into an upper half (better individuals), and a lower half (worse individuals), and transfer genes from the upper half into the lower half.
4. We repeat steps 2-4. until we get a satisfactory solution, or reach the maximum number of generations.

In the case of this algorithm, we had to modify individuals so that they contain distinct genes, since the gene transfer operation requires a metric measuring how "good" each gene constituting the individual is. We took pointer-set individuals, and divided them into a predefined number of genes, which are groups of a variable number of test cases. We have implemented two different versions of gene fitness:

First version

In this first implementation, the goodness of a gene is given by the average cost at which a gene covers the requirements:

the fitness of a gene is given by

$$F=\frac{\sum_{i\in I} C_i}{R} \qquad (5)$$

where *F* is the fitness of the gene, *Ci* the cost of the test cases, *I* the set of test cases of the gene, and *R* the number of requirements covered by the gene.

During the gene transfer operation, we take one of the superior half of the bacteria, and insert its best gene to replace the worst gene of one of the bacteria of the lower half:

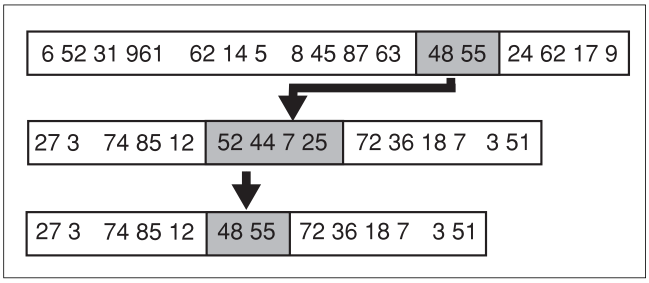


Fig. 4. Gene transfer 1

Second version

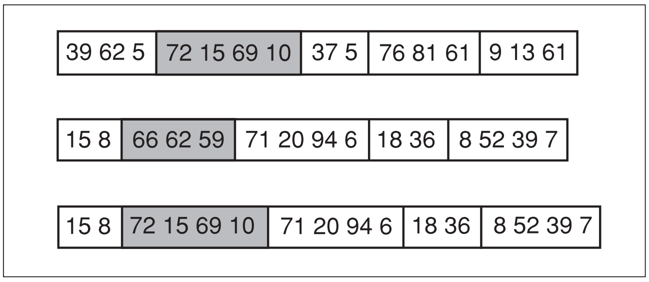
In this approach, we divide the test requirements in as many parts as there are genes in the bacteria. The goal is that each gene covers a specific part of the requirements. The goodness of a gene is defined in the same manner as the objective function of individuals in the previous cases, but the missed requirements are only taken into account in the interval covered by the gene. Thus, the fitness of a gene is given by

$$F = c1 * C + c2 * M_i \qquad (6)$$

where F is the fitness of the gene, C is the cost of the gene, M_i is the number of missed requirements on the set to be covered by the gene, and *c1, c2* are weighting constants.

During gene transfer, we take a bacterium from the upper half, and another from the lower half. We take a random gene of the source bacterium, and if it is better than the corresponding gene of the destination bacterium, we replace the corresponding gene of the destination bacterium with it:

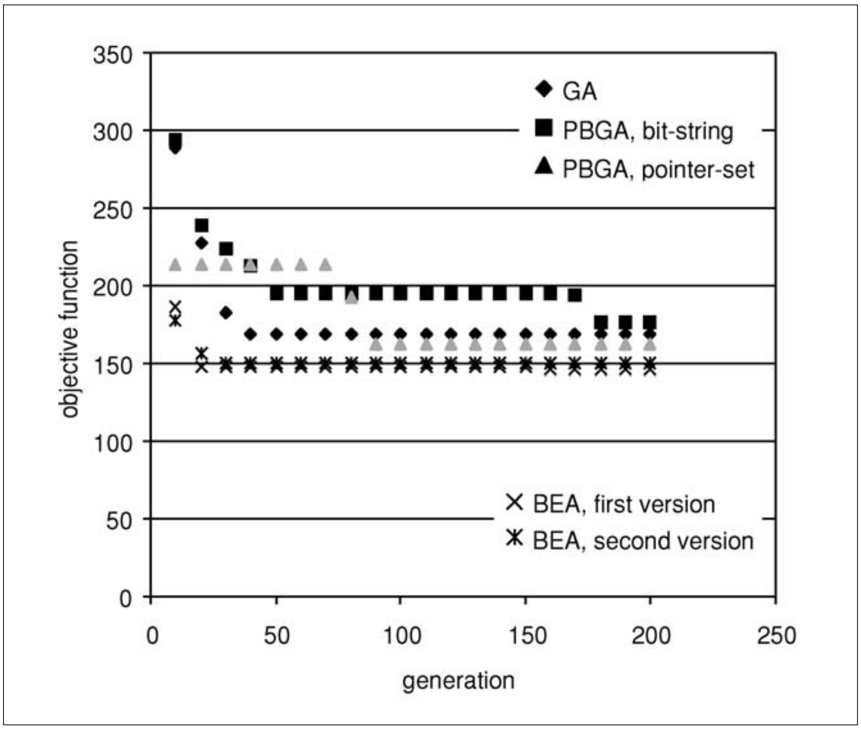
Fig. 5. Gene transfer 2



2.5. Algorithm Comparison

To compare the effectiveness of these algorithms in test case selection, we ran them on a fictive set of 100 test cases (as it will be shown later, the initial set of test cases for the INRES protocol only contains 41 test cases, which is too few to show differences in the convergence of these algorithms). The convergence of the different algorithms can be seen on *Fig. 6.:*

Fig. 6. Algorithm convergence



3. Automated test suite generation

We will now show the whole test generation process. We will illustrate this process by an example on the well-known INRES sample protocol.

The automated test generation process:

0) We create a formal specification of the protocol in SDL. Highly developed tools exist for this purpose ([3]). *Fig. 7.* shows the system overview of the INRES protocol SDL specification.
1) We apply a state-space exploration algorithm on the SDL specification, which gives us a highly redundant, unstructured set of MSC test cases.
2) With mutation analysis, we determine the matrix of test criteria for this set of test cases. *Fig. 8.* shows the full test suite resulting from the exploration of the SDL specification of the INRES system, containing 41 test cases, with the cost of individual test cases and the whole test suite, where the cost of test cases was calculated according to (2), with c1=20 and c2=5
3) We select an optimal subset of test cases from this set with one of the evolutionary algorithms presented above. This gives us a test suite that covers all test criteria, with minimal redundancy and execution cost. *Fig. 9.* shows the reduced set of test cases for the INRES protocol.

(Note: In this case, the selection of test cases is quite simple, and although it is not necessarily the case with very large test suites, all evolutionary algorithms found the same reduced test suite in a few generations.)

4. Conclusion

Conformance testing has become a crucial part in the development process of telecommunication protocols. Since the creation of test suites is a very time consuming process, automated test generation plays an increasingly important role in the development process.

We have shown here a complete method for the automatic generation of test suites from the SDL description of a system. We have only shown a simple example for illustration purposes. However, Mutation Analysis has been shown to work well on real-life cases ([4]) the motivational force behind the development of Evolutionary Algorithms was also the handling of extremely complex problems.

This test suite generation process is easily implementable, and should provide a working solution for the automated test generation of real-world telecommunication protocols.

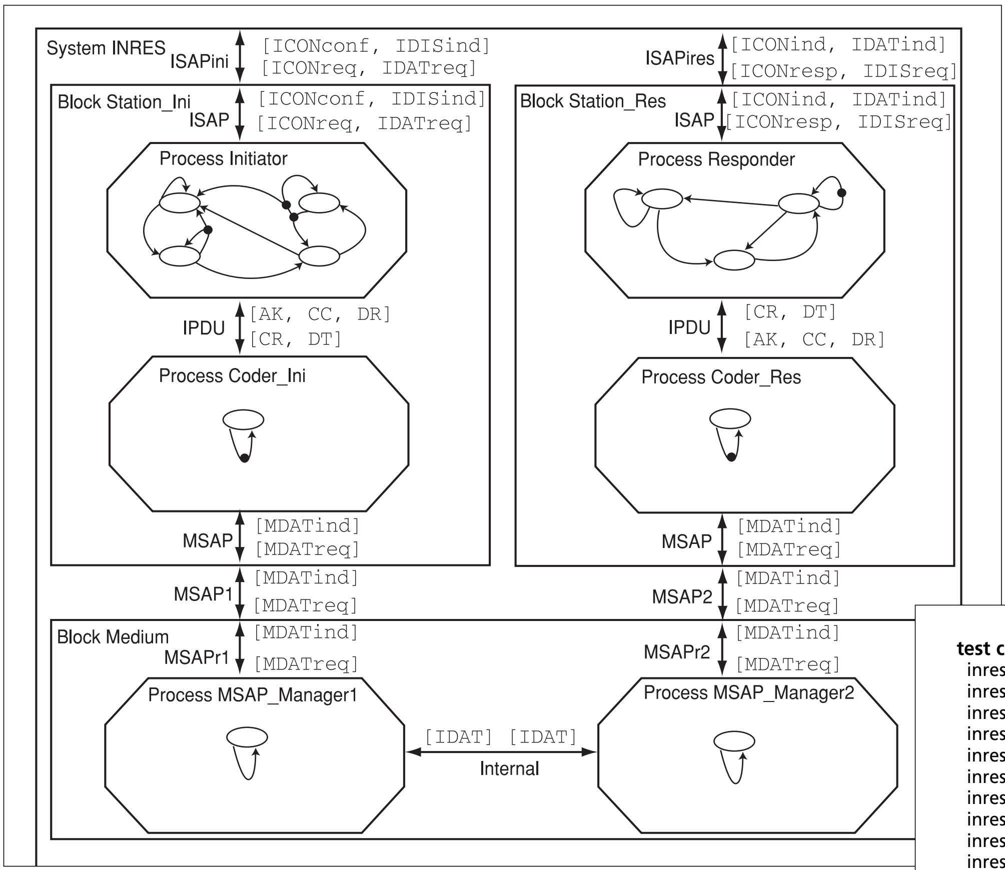


Fig. 7.
Overview of INRES SDL specification

Fig. 8.
Initial set of test cases

test case	cheked test criteria	test case cost
inres01	48	260
inres02	19	115
inres03	36	200
inres04	21	125
inres05	44	240
inres06	34	190
inres07	46	250
inres08	21	125
inres09	27	155
inres10	60	320
inres11	11	75
inres12	46	250
inres13	89	465
inres14	59	315
inres15	58	310
inres16	49	265
inres17	17	105
inres18	46	250
inres19	47	255
inres20	66	350
inres21	21	125
inres22	65	345
inres23	24	140
inres24	82	430
inres25	25	145
inres26	26	150
inres27	78	410
inres28	29	165
inres29	71	375
inres30	30	170
inres31	36	200
inres32	34	190
inres33	66	350
inres34	62	330
inres35	35	195
inres36	88	460
inres37	37	205
inres38	39	215
inres39	84	440
inres40	41	225
inres41	48	260
total test suite cost:		**10145**

Fig. 9.
Reduced set of test cases

test case	cheked test criteria	test case cost
inres10	60	320
inres13	89	465
inres14	59	315
inres23	24	140
inres27	78	410
inres28	29	165
total test suite cost:		**1815**

References

[1] ITU-T. Recommendation Z.100:
Specification and Description Language (SDL), 1992.

[2] CCITT. Recommendation X.292:
The Tree and Tabular Combined Notation (TTCN), 1992.

[3] Telelogic Tau. http://www.telelocig.com

[4] Black P.E., Okun V., Yesha Y.:
Mutation Operators for Specifications.
In The Fifteenth IEEE International Conference on Automated Software Engineering 2000,
Proceedings ASE 2000, pp.81–88.

[5] Gábor Kovács, Zoltán Pap, Gyula Csopaki:
Automatic Test Selection based on CEFSM, 2002.
Acta Cybernetica 15, pp.583–599.

[6] B. Kotnyek, T. Csöndes:
Heuristic methods for conformance test selection.

[7] J.H.Holland:
Adaptation in Nature and Artificial Systems:
An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence,
MIT Press, Cambridge, 1992.

[8] M.Salmeri, M.Re, E. Petrongari, G.C.Cardarilli:
A Novel Bacterial Algorithm to Extract the Rule Base from a Training Set,
Dept. of Electronic Engineering, University of Rome, 1999.

[9] N.E.Nawa, T.Furuhashi:
Fuzzy System Parameters Discovery by Bacterial Evolutionary Algorithm,
IEEE Tr. Fuzzy Systems 7 (1999), pp.608–616.

Linear and nonlinear analysis of a testing effort model

Gábor Stikkel, Gábor Szederkényi

stiko@compalg.inf.elte.hu, szeder@sztaki.hu

Keywords: software maintenance and testing, state observer, parameter estimation

The maintenance and testing effort is modelled as a predator-prey model in the well-known Lotka-Volterra form in order to facilitate tracking and estimating maintenance and testing resources. Our aim is to find a solution for a crucial software development decision problem, namely for determining the end of testing and releasing the product. The method is based on tools of system theory and is applied on real project data.

Software maintenance and testing activities consume most of software project resources. This fact motivates research in the field of planning, estimating and tracking maintenance and testing resources.

One of the most promising approach for modeling maintenance and testing effort was suggested by Calzolari et.al. [2]. This model considers as prey the software faults which cause environmental needs and corrective actions. Predators are the testers or developers observing and removing the prey. The dynamical change of the number of faults in the testing process or after release shows similarities to predator-prey competition. The only difference is that faults can not reproduce new faults.

Similar models was introduced in the literature previously. Lehman et.al. [5,6] used dynamic models to describe the evolution of relevant software engineering metrics. Those models were successful in describing the changing of the size of software systems among releases.

Another approach which is close to the one presented here is in [1] and [8]. The authors gave a comprehensive system dynamics model of the software development process. The equations they suggest The outcome of their simulation can help in predictions and making decisions. On the other hand the construction of these models and the estimation of the parameters is a hard, human intensive task. As far as predator-prey like model is concerned, the parameter estimation can be automated.

1. Modeling testing effort by differential equations

The classical predator-prey model was proposed by V. Volterra and A.J. Lotka. In this model a system of two differential equations model the variation of two populations. This model was adapted to maintenance and testing activities by Calzolari et.al. [2] in the following way: corrective interventions are considered to be predating software faults and the associated effort is fed by the discovery of faults.

The result of the adaptation is two new models a linear and a nonlinear one. The dynamics generated by these models represents the effort evolution within a given release, hence when a new release is delivered the dynamics starts again with another initial values.

1.1. The linear model

The linear model is defined by the following differential equations [2]:

$$\dot{x}_1 = \qquad -ax_2$$

$$\dot{x}_2 = bx_1 - cx_2.$$

The first variable denotes the residual faults in the underlying software system while the second one is the testing or maintenance effort. Parameters a, b and c are positive. The first equation describes the decrease in the number of residual faults as a function of actual value of testing effort. The latter quantity can increas with a rate proportional to the available fault number. The decrease term in the second equation represents the intrinsic *mortality* of this *population*.

The modification of the classical Volterra-Lotka model was needed because the faults can not reproduce themselves. A possible system evolution is depicted in *Fig. 1.*

1.2. The nonlinear model

The nonlinear model is described by the following differential system:

$$\dot{x}_1 = -h(x_1)x_2 \tag{1}$$

$$\dot{x}_2 = eh(x_1)x_2 - mx_2. \tag{2}$$

The negative term in the first equation contains $h(x1)$ and $x2$. The first term represents the *functional response of the predators(testers)*. It describes that how many faults will be found by each tester depending on the number of residual faults.

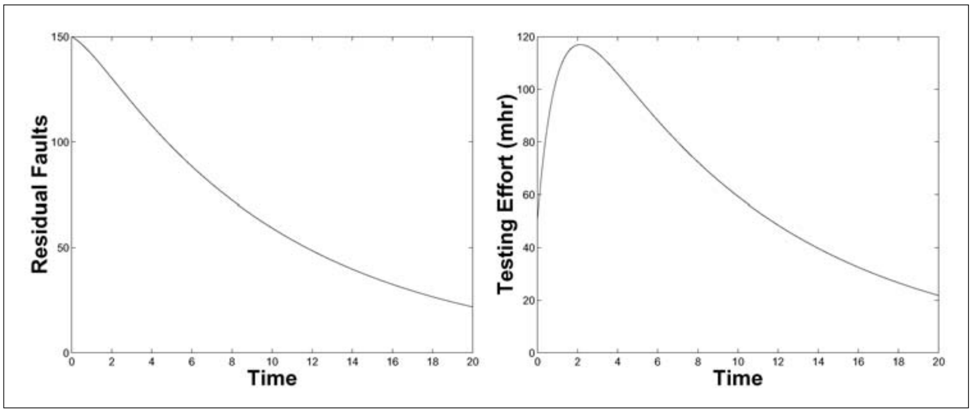


Fig. 1. A possible evolution of the linear model with parameters a=0.1, b=1 and c=1.1

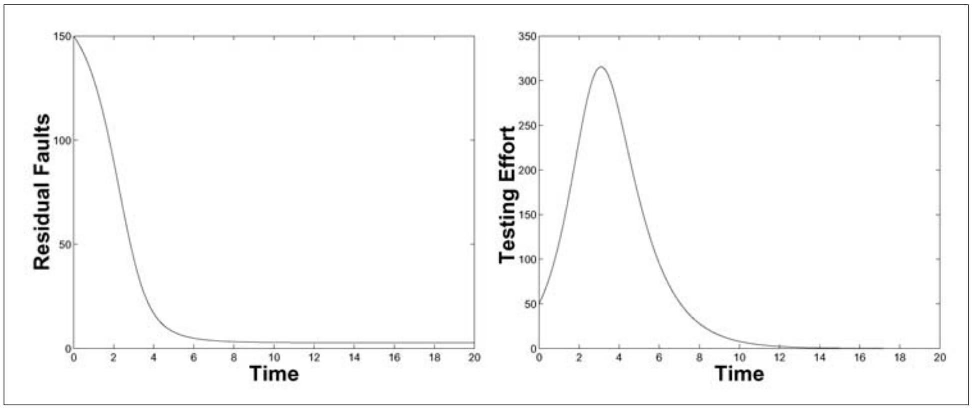


Fig. 2. A possible trajectory of the nonlinear model with parameters a=0.5, b=130, e=6 and m=0.7

Here the Holling function of type 2 is used to model the functional response:

$$h(x_1) = \frac{ax_1}{b + x_1}.$$

Parameter *a* is the asymptotic fault fix rate, while *b* is the fault level for which the fault fix rate is halved. The second equation could be decomposed into two parts. Current fault fix rate is converted into new corrective effort through the *effciency* factor *e*. The second term in the second equation shows the intrinsic decrease of corrective effort over time, with *mortality* factor *m*. Possible trajectories of the two variables can be seen in *Fig. 2.*

2. Observer design for the linear model

2.1. Observer theory for linear time invariant systems

The theory of linear time invariant systems is dealing with the following system of linear differential equations:

$$\dot{x} = Ax + Bu \tag{3}$$

$$y = Cx \tag{4}$$

where *x* represents the state of the system, *u* is the input while *y* is the output.

In the proposed linear model [2] the matrices are

$$A = \begin{bmatrix} 0 & -a \\ b & -c \end{bmatrix}, \qquad B = 0, \qquad C = [0 \;\; 1].$$

A method for estimating the parameters was presented in [2] for both linear and nonlinear models. In the nonlinear case we will suggest another method but for the linear case parameters are supposed to be known in the remaining.

From system theoretic point of view it is an interesting question whether the state *x* can be reproduced from the input *u* and the output *y*. The answer is affirmative if the system is observable [7], i.e. *y*(*t*) ≡ 0 implies that *x*(0) = 0. An equivalent characterization of observability of linear systems is that the kernel of the matrix

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

contains only the zero vector.

The observability can be carried out by a state observer [7] which is another dynamical system of the form

$$\dot{\hat{x}} = A\hat{x} + Gu + Hy. \tag{5}$$

The system (5) is called state observer for system (3) if for all initial states x_0, $\hat{x}_0$ and for all input *u*

$$\lim_{t\to\infty} \hat{x}(t) - x(t) = 0.$$

2.2. Observer design

The computation of the matrices *G,H* can be found in [7]. It can be shown that in our special case *G* = *H* = 0, hence

$$\dot{\hat{x}} = A\hat{x}$$

is the state observer for system (3).

The effectiveness of the observer in applications depends on the initial condition $\hat{x}(0)$. If a project manager can guess the exact value of $x_1(0)$ then by knowing $x_2(0)$ the observer initial state can be set to $\hat{x}(0) = [x_1(0)\ x_2(0)]^T$, and the estimation error $(\hat{x}-x)$, will be identically zero. It means that the manager is able to track the number of the residual faults. It is a difficult task to give an accurate estimation of $x_1(0)$. The next section two methods are suggested how to handle this problem.

2.3. Estimation methods for the initial value of the observer

2.3.1. Expert judgements

The first proposed method is rather heuristic. It is hard to guess the exact value of the initial residual faults. Instead, one might use data from testing i.e. number of faults found until a certain time instant of the project. This number will be a lower bound on $\hat{x}(0)$ and as the project progress the manager can give more and more accurate estimates of the number of residual faults.

Suppose that initially there were 150 faults in a system. With this initial condition the model tells us that there will be approximately 21 faults remaining after the 20. day of testing. The project manager guess an initial fault containment of 100 which gives an estimation of 14 faults on the 20. day. However, after the 10th day of testing it has turned out that already 68 faults has been found. It makes the manager improve his guess to 170 which results an estimation of 24 faults after the 20. day.

2.3.2. Estimation based on software reliability growth models

The second method is based on software reliability models presented in [9] and [10]. These models are dealing with the cumulative number of faults found during testing as a function of time. Logistic [9] and Gompertz [10] differential equations and their discrete variants are fit on empirical data. The input of this procedure is a few data points i.e. number of faults found during the first days of testing. It is necessary to have at least as many data points as the fifth of the whole in order to have fairly good parameter estimation of the logistic curve [9].

By knowing the parameters we can predict the total number of faults found. These predictions can be a basis of the estimation of the initial value of the observer. The method is depicted in *Fig. 3.* Other methods for predicting residual faults and failures can be found in [3].

3. Stability analysis of the nonlinear model

A very brief stability analysis can be found in [2]. The stability of the linear model is no matter of discuss: if the parameters are positive then both eigenvalues of the system matrix have negative real parts implying the asymptotic stability of the system. The latter means that the states of the system tends to zero whatever the initial condition is. Another qualitative property of an equilibrium point is its attractiveness. Such a point is attractive if there exist a neighborhood of it so that if the system is initialized from that neighborhood than the solutions tend to the equilibrium point. Necessary attractive equilibriums should be isolated. But, as we will see the equilibriums of the nonlinear model constitute a half line, i.e. they are not attractive as was stated in [2]. This fact motivates the investigations presented in this section.

Both state in the proposed nonlinear model are supposed to have non-negative values. It can be seen that in the case when $ea - m \leq 0$ the second variable will be decreasing which is somewhat contradictory to the process observed in real life situations (testing effort is used to increase for a certain amount of time). Hence we can continue the analysis with the reasonable assumption that $ea - m > 0$.

The equilibrium points of (1) can be determined by solving the equations $\dot{x}_1^*= 0$ and $\dot{x}_2^*= 0$. The solutions are $(x_1^*, 0)$, where x_1^* is a non-negative number. These points constitute a half line which immediately implies that only local stability results make sense in this context. In order to restrict our attention to the stability analysis of the zero state, the equilibrium point is shifted to get the following equations:

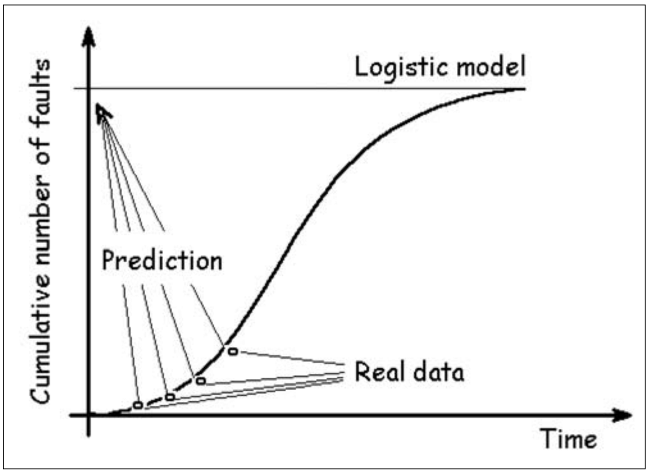
$$\dot{x}_1 = -\frac{a(x_1 + x_1^*)x_2}{b + x_1 + x_1^*} \tag{6}$$

$$\dot{x}_2 = \frac{ea(x_1 + x_1^*)x_2}{b + x_1 + x_1^*} - mx_2. \tag{7}$$

To investigate local stability of the system the Jacobian matrix of its map will be used:

$$F = \begin{bmatrix} 0 & -h(x_1^*) \\ 0 & -m + h(x_1^*). \end{bmatrix}$$

Fig. 3.
Estimating initial value of the observer based on logistic reliability groowth model



Eigenvalues of the Jacobian are 0 and $-m+h(x_1^*)$. It can be seen that the condition for the second eigenvalue to be negative is $x_1^* < mb/(ea-m)$. Hence equilibrium state $(x_1^*, 0)$ is unstable if $x_1^* \geq mb/(ea-m)$.

The case of system whose Jacobian has some eigenvalue with zero real part is commonly referred to as a *critical case of the asymptotic analysis*. Center manifold theory is of great help in analyzing critical cases.

The argument presented in [4] starts with considering instead of 6, the following form:

$$\dot{y} = Ay + Pz + g(y,z) \tag{8}$$

$$\dot{z} = Bz + f(y,z), \tag{9}$$

where A is a matrix having all eigenvalues with negative real part, B is a matrix having all eigenvalues with zero real part. Then we have to solve the following partial differential equation:

$$\frac{\partial \pi}{\partial z}(Bz + f(\pi(z),z)) = A\pi(z) + Pz + g(\pi(z),z).$$

Theorem 1

The (asymptotic) stability of the zero state of

$$\zeta = B\zeta + f(\pi(\zeta),\zeta) \tag{10}$$

implies the (asymptotic) stability of (y, z) = (0, 0) *of (8).*

In our case

$A = -m + eh(x_1^*)$, $P = 0$, $g(y,z) = \frac{ea(z+x_1^*)y}{b+x_1^*+z} \quad eh(x_1^*)y$,

$B = 0$, $g(f,z) = -\frac{a(z+x_1^*)y}{b+x_1^*+z}$

and the underlying differential equation is ordinary having the trivial solution namely $\pi(z) \equiv 0$. It results that the reduced equation (10) takes the simple form of

$$\dot{\zeta} = 0$$

implying that all the equilibrium points which satisfies $0 \leq x_1^* < mb/(ea-m)$ are stable.

4. Parameter estimation

4.1. Parameter estimation of the linear model

Consider again the linear model of dynamic variation of the testing effort and residual faults. This model has three positive parameters. The problem of parameter estimation is that we have to determine the values of *a, b* and *c* from the data on the second variable.

Suppose we have data on x_2 in discrete time instances denoted by $x_2(0), \hat{x}_2(1), \ldots, \hat{x}_2(N)$. We would like to choose parameter values such that the squared error

$$\sum_{i=0}^{N}(x_2(i) - \hat{x}_2(i))^2$$

is minimal. It was carried out by creating a simulation model in MatLab and running a built-in optimization tool which uses simplex method. In order to have a fairly good estimation we need so many data points such that the peak of the testing effort curve is reached.

4.2. Nonlinear parameter estimation by algebraic elimination

As it was visible in section 2, the physical nature of the testing effort model is basically nonlinear. Moreover, linear modeling often cannot provide models of satisfactory quality for certain purposes (e.g. control). Therefore the parameter estimation of the nonlinear model structure given in (1)-(2) is carried out in this section.

4.2.1 Calculation of a nonlinear input-output model

Considering the assumption that the only measurable state variable of the nonlinear model (1)-(2) is x_2, the purpose of this section is to formulate a nonlinear inputoutput model in the following form:

$$F(y,\dot{y},\ddot{y},u,\dot{u}) = 0 \tag{11}$$

where *y* denotes the measurable output (i.e. $y = x_2$) and *u* is a fictitious input which is assumed to be known. An obvious and computationally simple selection for *u* is *m* in (2).

The aim is now to eliminate the non-measurable state variable x_1 from the state equations and thus to obtain a model of the form (11). For this, let us rewrite the state equations and the fictitious output equation as

$$\dot{x}_1 + \frac{ax_1x_2}{b+x_1} = 0 \tag{12}$$

$$\dot{x}_2 - \frac{eax_1x_2}{b+x_1} + ux_2 = 0 \tag{13}$$

$$y - x_2 = 0 \tag{14}$$

From (14) and (7) we get

$$\dot{y} - \frac{eax_1x_2}{b+x_1} + ux_2 = 0,$$

from which x_1 can be expressed as

$$x_1 = \frac{b(\dot{y}+ux_2)}{-\dot{y}-ux_2+eax_2} =: \frac{n(t)}{d(t)} \tag{15}$$

using the notations *n* and *d* for the numerator and denominator of (15) respectively. Taking the time derivative of (15) gives

$$\dot{x}_1 = \frac{\dot{n}(t)}{d(t)} - \frac{n(t)\dot{d}(t)}{d^2(t)}. \tag{16}$$

Using eqs (16), (12) and the notation in (15) we can eliminate $\dot{x}_1$ from the state equations i.e.

$$-\frac{a\frac{n(t)}{d(t)}x_2}{b+\frac{n(t)}{d(t)}} = \frac{\dot{n}(t)}{d(t)} - \frac{n(t)\dot{d}(t)}{d^2(t)}, \tag{17}$$

which can be rewritten as

$$-\frac{an(t)x_2}{d(t)b+n(t)} = \dot{n}(t) - \frac{n(t)\dot{d}(t)}{d(t)}. \tag{18}$$

Computing the time derivative of *n* and *d* gives the required input-output model (that can be easily rearranged to the form (11))

$$-\frac{\dot{y}+uy}{ey} = b(\ddot{y}+u\dot{y}) - \frac{b(\dot{y}+uy)(-\ddot{y}-u\dot{y}-\dot{u}y+ea\dot{y})}{-\dot{y}-uy+eay}. \tag{19}$$

Using the fact that in our case the parameter $u = m$ is constant and thus $\dot{u} = 0$, a simpler form of the model is obtained

$$-\frac{\dot{y}+uy}{ey} = \frac{bea(y\ddot{y}-\dot{y}^2)}{-\dot{y}-uy+eay}. \tag{20}$$

Introducing the transformed parameters

$$c_1 = be^2a, \qquad c_2 = ea \tag{21}$$

gives the following model

$$(\dot{y}+uy)^2 = c_1(y\ddot{y}-\dot{y}^2) + c_2y(\dot{y}+uy), \tag{22}$$

which is *linear in the new parameters*. Now, c_1 and c_2 can be estimated by any of the standard techniques, measuring only x_2.

4.2.2. Parameter estimation

For the practical implementation of the parameter estimation it's useful to calculate a discrete-time model from (22). First, let us introduce the δ operator for the notation of the numerical derivatives (calculated by using a simple Euler-approximation) as follows

$$\delta z(k) = \frac{z(k+1)-z(k)}{t_s}, \tag{23}$$

where z denotes a discrete time sequence and t_s is the sampling time.

Using (23) the input-output model (22) in discrete time is written as (24)

$$(\delta y(k)+uy(k))^2 = c_1(y(k)\cdot\delta^2y(k)-(\delta y(k))^2)+c_2y(k)(\delta y(k)+uy(k))$$

which is in a standard regression form

$$w(k) = \phi^T(k)\theta \tag{25}$$

with (26-28)

$$\begin{aligned} w(k) &= (\delta y(k)+uy(k))^2 \\ \phi^T(k) &= [y(k)\cdot\delta^2y(k)-(\delta y(k))^2 \quad y(k)(\delta y(k)+uy(k))] \\ \theta &= [c_1 \quad c_2]^T. \end{aligned}$$

Therefore c_1 and c_2 can be estimated using the well-known least squares method which minimizes the quadratic criterion

$$V(n,\theta) = \frac{1}{N}\sum_{i=1}^{N}(w(i)-\phi^T(i)\theta)^2 \tag{29}$$

with respect to θ.

Note that in this case we have to assume that one parameter from *a, b* and *e* is known, and then we can solve (21) for the remaining two unknown parameters. This is a strict limitation of the usefulness of the parameter estimation method, however it can be used in order to refine or to check the goodness of previously estimated parameters.

5. Application of the linear observer to real data

Our study project was an open, standards based, modular and distributed application. The source has a length of approximately 250,000 lines of code (LOC – without comments) and was written in C++ with an effort of approximately 75,000 man-hours. We have effort data on this project (x_2 variable) from which parameters can be estimated, see *Fig. 4.*

We also have data on faults found during testing from which we could estimate the initial value of the observer. The goodness of the observer can be tested because the number of residual faults (i.e. the observed x_1 variable) should be approximately equal to the difference of our estimated initial value ($x_1(0)$) and the number of faults found during testing.

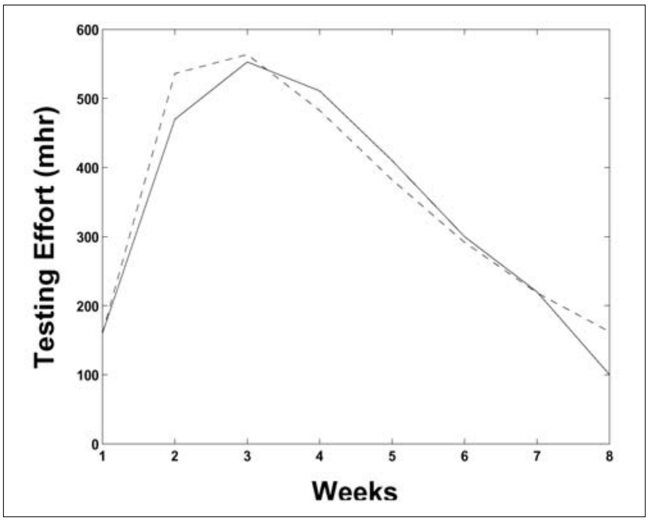
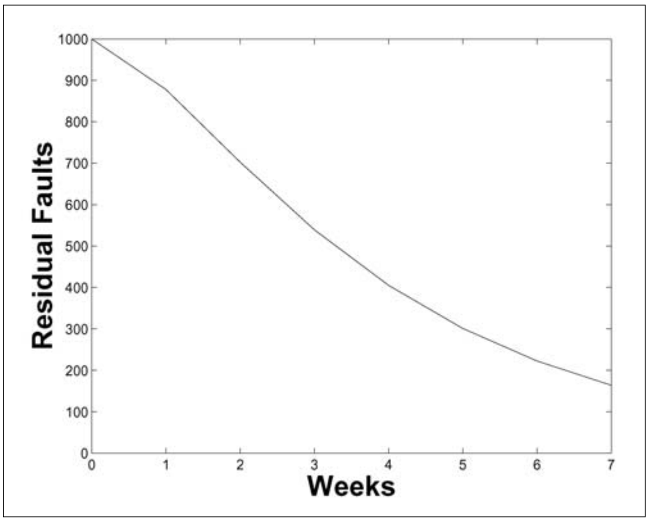


Fig. 4. Effort data on the project (solid) and the testing effort given by the estimated parameters (dashed; a = 0.31, b = 0.9 and c = 1.22)

The number of faults found during testing was 848. It makes us to estimate the initial value of the residual faults as 1,000. Hence the model can be accepted if the simulation of the first variable results that there are ≈152 faults in the system after test.

Fig. 5. shows the simulation results which tells us that there are 163 residual faults in the system, hence the error of the estimation is below 10 percent.

Fig. 5. Simulation results (residual faults)



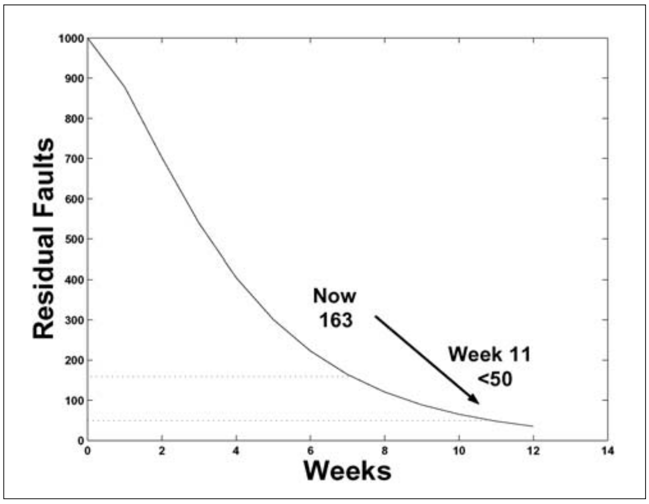


Fig. 6. Simulation result of the first variable

Accepting this model we arrived to answer the question posed in the title of the paper. Suppose that the manager would like to continue testing until the estimated fault content of the software system is below 50. How long shall we test? Running again the simulations with the identified parameters we can depict *Fig. 6.*

It suggests that the testing process should continue for four more weeks to reach the specified quality. The model can also be used to answer what-if questions regarding the trade off between testing effort and software quality.

6. Conclusion

Proposed maintenance and testing effort based on linear and nonlinear Lotka-Volterra systems was revisited and was investigated from system theoretic point of view. We have found that state observer for the linear model can be used to predict residual number of faults in a software system. It can also give estimation for the manager how long the testing phase should be continued in order to reach a specified software quality. Future work will be focused on model extension and observer design for the nonlinear model.

Acknowledgements

The authors would like to thank Domonkos Asztalos (Ericsson Telecommunications Hungary) and Gusztáv Bártfai (previously Ericsson Telecommunications Hungary) for project data and for valuable comments.

References

[1] T. K. Abdel-Hamid:
The dynamics of software project staffing:
a system dynamics based simulation approach.
IEEE Transactions on Software Engineering,
15(2). 1989, pp.109–119.

[2] F. Calzolari, P. Tonella, G. Antoniol:
Maintenance and testing effort modeled by
linear and nonlinear dynamic systems.
Inform. and Software Techn., 43(2001), pp.477–486.

[3] M. Grottke, K. Dussa-Zieger:
Prediction of Software Failures Based on
Systematic Testing. Electronic Proc. 9th European
Conference on Software Testing Analysis and Review,
(EuroSTAR), Stockholm, 2001.

[4] A. Isidori:
Nonlinear Control Systems. Springer-Verlag, 1995.

[5] M. M. Lehman, D. E. Perry, J. F. Ramil:
Implication of evolution metrics on
software maintenance. Proceedings of the
International Conference on Software Maintenance,
Bethesda, MD, 1998, pp.208–217.

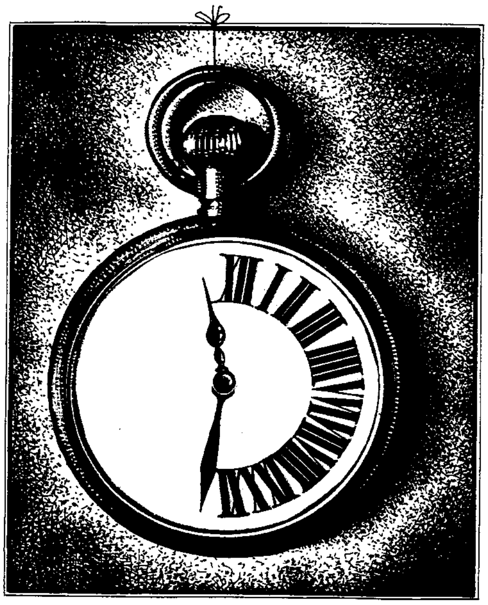
[6] M. M. Lehman, D. E. Perry, J. F. Ramil:
On evidence supporting the feat hypothesis and
the laws of of software evoution. Proceedings of
the 5th International Symposium on Software metrics,
Bethesda, MD, 1998.

[7] J. M. Maciejowski:
Multivariable Feedback Design. Addison-Wesley,
Wokingham, U.K, 1989.

[8] R. Madachy:
System dynamics modelling of an inspection-based
process, Proceedings of the International
Conference on Software Engineering,
Berlin, 1996. pp.376–386.

[9] D. Satoh:
A Discrete Gompertz Equation and
a Software Reliability Growth Model.
IEICE Transactions on Information and Systems,
E83(2000), No.7, pp.1508–1513.

[10] D. Satoh, S. Yamada:
Parameter Estimation of Discrete Logistic Curve
Models for Software Reliability Assessment,
Japan Journal of Industrial and Applied Mathematics,
19(2002), No.1, pp.39–53.



Optical filter type influence on transparent WDM network's size

ÁKOS SZÖDÉNYI

Budapest University of Technology and Economics, Dept. of Telecommunications and Media Informatics
szodenyi@tmit.bme.hu

Keywords: wavelength multiplexing, transparent optical multiplexer, optical filters, bit error-rate

The number of nodes (size) of an optical transparent network-island is limited according to Bit Error Rate (BER) estimation of the optical signals that cross transparently the optical nodes. Three optical add-drop multiplexer – based on different filter technology and therefore different architecture – is cascaded and compared by BER-degradation estimation in a special network architecture environment.

Technologies and new concepts for optical networking are advancing rapidly as a result of notable progresses in all-optical technologies and emerging bandwidth greedy applications. Telecom operators are forced, in consequence, to adapt in the near future, their deployed optical fiber communication systems so as to cope with these challenging advances. Deploying "islands" wherein the optical signals benefit from the advantages of transparency may be more feasible than replacing totally the current conventional digital systems by all-optical technologies.

In this paper the size of a metropolitan "transparent island" (the "island" is in the non transparent network "ocean") is assessed by computer simulations depending on the architecture of the all-optical add/drop multiplexer (OADM) used. In effect, three architectures of OADM were on focus to compare between their performances after cascading several optical nodes. Optical signal quality represented by BER estimation is used as the metric that determines the size of a transparent island.

To the best of my knowledge, this is the first time an estimation of the size (hop number) of transparent islands is given depending on applied optical devices used and the target BER.

Three optical filter types

Multiplexing and demultiplexing functions both employ narrowband filters, cascaded and combined in other ways to achieve the desired result. Particular techniques that have been used to perform such filtering include thin film filters, fiber Bragg or bulk gratings and integrated optics (AWG).

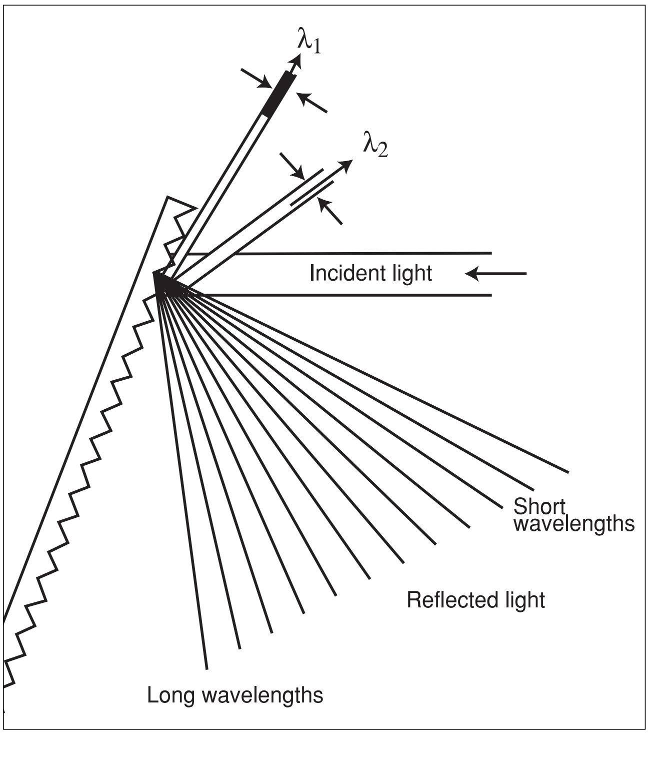
Diffraction grating (mux)

A bulk-optic diffraction grating [1] reflects light at an angle proportional to wavelength and the underlying physical principle is constructive and destructive interference.

For each wavelength of incident light, there is an angle for which light waves reflecting from individual grating lines will differ in phase by exactly one wavelength-spacing. At this angle, the intensity contribution from each line will add constructively, so this will be the angle of maximum transmission for that specific incident wavelength.

Designing a mux or demux using a diffraction grating is a matter of positioning the input and output optics to select the desired wavelength. Although they are difficult to manufacture and expensive, devices based on diffraction gratings have an insertion loss that is essentially independent of the number of channels, rendering this technology one of the more promising for high channel count systems. However, polarization control requires critical attenuation.

Fig. 1. Bulk diffraction grating



Arrayed Waveguide Grating

AWG [2,3] is also known as phased-array gratings (PHASARS), or waveguide grating routers (WGR). In the contrary of bulk grating filter, AWGs are wavelength insensitive (in a given range of optical frequencies of course) and therefore periodical. The performance of AWG is similar to multiplexer/demultiplexers, as it can separate and also multiplex different wavelength which are propagating in a SM fiber. Just the ways – how they do it – differ.

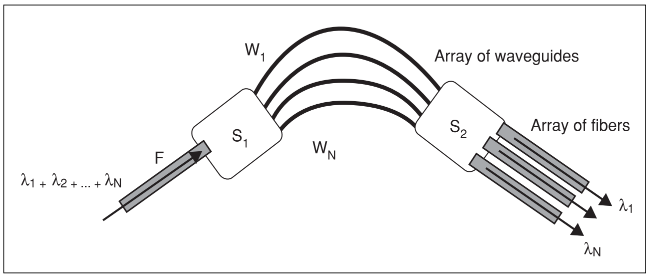


Fig. 2. Arrayed Waveguide Grating

AWG is based on interferometry. The architecture of AWG is depicted in *Fig. 2.* The incoming beam-carrier-fiber guides several different wavelengths to the first cavity (S$_1$), which is coupled to an array of waveguides. As many wavelength there are in the fiber, as many waveguides are set up in S$_1$. The lengths of these waveguide-fibers are different. Because of the optical lengths are not the same, wavelength-dependent phase shifts can be achieved in the second cavity (S$_2$), where an array of fibers is coupled. The phase difference of each wavelength interferes in such a manner that each wavelength's intensity contributes maximally at *one* of the output fibers. Considering that all points on the emerging wavefront must have the same phase (modulo 2π), two adjacent optical path form the incident wavefront to the emerging wave front must have optical path length difference. This difference is equal to an integer multiplied with the wavelength.

AWGs can be used as all optical routers too when not only one input port is manufactured. Considering a 2x2 AWG with two inputs and two output port both fibers carries λ_1 and λ_2. Thus the input on A port is λ_{1a} and λ_{2a} at input B λ_{1b} and λ_{2b}. AWG routers are able to interchange same wavelength without mixing the containing modulation. Therefore upper output port can transmit λ_{1a} and λ_{2b} while second output λ_{1b} and λ_{2a}. These type of routers are very promising candidates in future transparent networks, moreover a special network architecture is under patent request in the US for the Technical University of Berlin, where these type of routers play the most important role in network architecture (Ringostar).

Fiber Bragg Grating

These components are filters, which let all wavelengths through with low attenuation, expect one, which it is designed for and will be reflected. FBG can be tunable or fixed. In contrast of its name, it is not a grate. It is called so after all because light behaves like it would have met with a grating.

The reflection of a specific band of wavelength can be reached by periodical refractive index changes in the core of the fiber. This is what light feels like a grating. There are at least two technologies to create it. One is more popular and is called: UV technology. The Germanium doped core is exposed by an ultraviolet pattern, which causes interference, and also refractive index periodical variation in the core. This pattern is in strong relationship with the selected wavelength. Different patterns are for different wavelength to reflect. The longer the FBG is manufactured, the narrower the reflected wavelength-band is. On the other hand the longer an FBG is manufactured, the higher its insertion loss is.

Different FBGs can be cascaded to reflect more than one wavelength. To combine an FBG with a circulator it is easy to drop wavelength from the WDM fiber. Another application is to use FBG's for chromatic dispersion compensation. These types are known as 'chirped FBGs' and have the grating linearly variable "chirped".

Architecture of the optical nodes

The test bed features an ASON/GMPLS network formed by a transport plane of three reconfigurable optical add/drop multiplexers (OADM), a control plane and a management plane to allow for dynamic and intelligent optical channel provisioning.

Three different implementation of an OADM architectures are considered. The first one uses a multiplexer and a demultiplexer (based on bulk gratings) in addition to a set of 2x2 optical switches, and the second one uses an AWG and a set of 2x2 optical switches, the third one 4 FBGs *(Fig 3)*. Eight ITU channels with 100 GHz (0.78nm) spacing can be allocated [from 193.0 THz (=1553.33nm) to 193.7 THz (=1547.715nm)]; however, up to four channels can be added/dropped locally within each optical node but in the simulation no channel has been dropped out at the nodes.

The investigation aim also to explore to which extent the all-optical test bed metro network can evolve to future extension and/or transparent interconnection with other test-beds.

Results and discussions

It is considered in the simulations [4], an optical signal traveling along several optical nodes in a metro optical network composed by one of the three described OADM architectures (based on the use of Mux/Demux or AWG or FBG architectures) with an EDFA pre-amplifier at each node and a fiber length between adjacent nodes of 35 km. All the components data introduced in the simulations are in accordance with the worst-case experimental data of the optical component's datasheets.

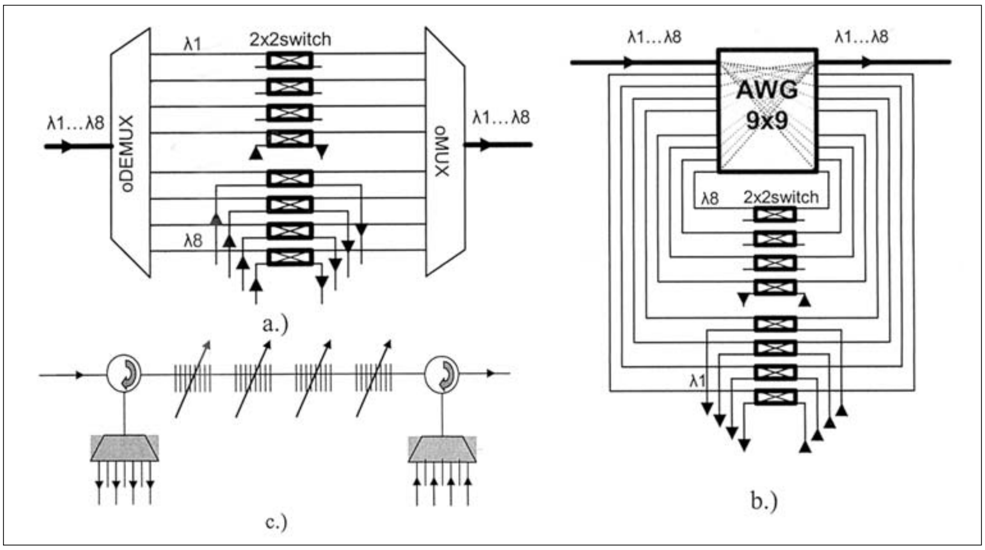


Fig. 3. OADM's architectures:
a. Diffraction grating (mux/demux) based; b. AWG based; c. FBG based

Mostly important for comparing between the three architectures, in our case, were the insertion losses of the FBG (2.8 dB) AWG (10 dB) and of the mux/demux components (4 dB) as the other data were similar in both cases. The optical signals, which are not to be dropped at a given node, undergo a double amount of attenuation due to the fact that they should be demultiplexed, first, and then multiplexed again at the passing through nodes.

The VPI program estimates the BER by the following equation:

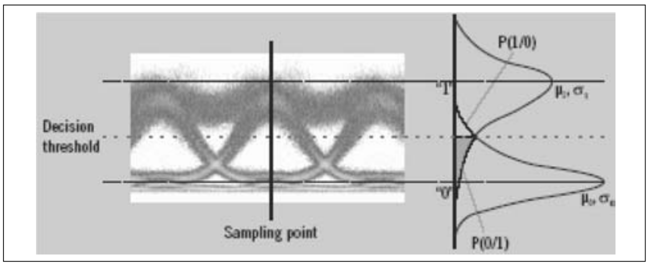
$BER = \frac{1}{2} erfc\left(\frac{Q}{\sqrt{2}}\right)$ where the $erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{+\infty} e^{-t^2}\, dt$

and Q is the quality factor [5].

The 'Q' factor is a measure of the digital signal eye aperture; it adopts the concept of S/N ratio in a digital signal and is an evaluation method that assumes a normal noise distribution [6] and can be calculated from the following formula: $Q = \frac{|\mu_1 - \mu_0|}{\sigma_1 - \sigma_0}$

where in μ_1 is the average value of the logical '1' level and μ_0 of logical '0' level. The σ_1 and σ_0 are the standard deviation values of the noise distribution on the '1' and '0' rails, respectively as it is depicted on *Fig. 4*. By the variation of the decision threshold of the receiver diode, the sensitivity of the system can change. The estimation of the Q factor is based on this sensitivity change evaluation.

Fig. 4. An eye aperture is shown to illustrate the calculation of Q factor



This is the "Variable Threshold Method" which is discussed in more details in ITU-T G.976 (1997) recommendation, and in the North American patent [7].

Fig. 5. shows the BER estimated values when the signals pass through a given number of optical nodes (we consider a limit of 10 nodes spaced by 35 km for our metropolitan optical network). Fig. 5. gives the extent of transparently crossed optical nodes versus the BER estimated value for each of the three OADMs architectures.

After passing transparently 10 optical nodes, the optical signal quality decreases to 10^{-4} in case of the mux/demux (bulk gratings) based architecture, decreases to 10^{-6} in case of AWG based architecture, and decreases 10^{-9} in case of FBG based architecture of the OADMs.

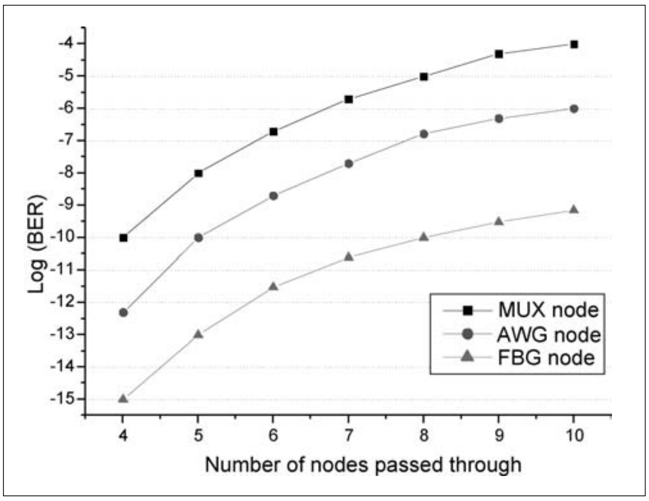
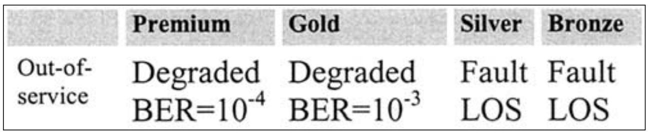


Fig. 5.
Number of crossed optical nodes vs. BER foreseen:
a. mux/demux based, b. AWG based, c. FBG based

If optical services were classified according to four categories, *Table 1,* in the limit of 10 optical nodes of a transparent island and according to Fig. 5., provisioning all services, even Premium services, which require the most stringent requirements of quality, could be expected for all OADM architectures (according to BER

Table 1.
Out-of-service criterion for different classes of services [8]



	Premium	Gold	Silver	Bronze
Out-of-service	Degraded BER=10^{-4}	Degraded BER=10^{-3}	Fault LOS	Fault LOS

estimated values). However, other criteria such as connection set-up times and recovery times should also be taken into account to actually ensure the provisioning of optical services.

On the other hand if BER were fixed to a higher quality level such as 10^{-9}, the situation is quite different for the three OADM architecture of the testbed, a transparent island based on mux/demux and AWG architecture would have a maximum reach of only four and five optical nodes, whereas for the FBG architecture, the transparent island would reach 10 optical nodes.

In case of 50 GHz channel spacing the BER estimation will give a bit worse result, especially at FBG node. This is due to that FBG node does not filter the incoming spectrum, while mux node and AWG node do it producing more or less the same result as in case of 100 GHz spacing.

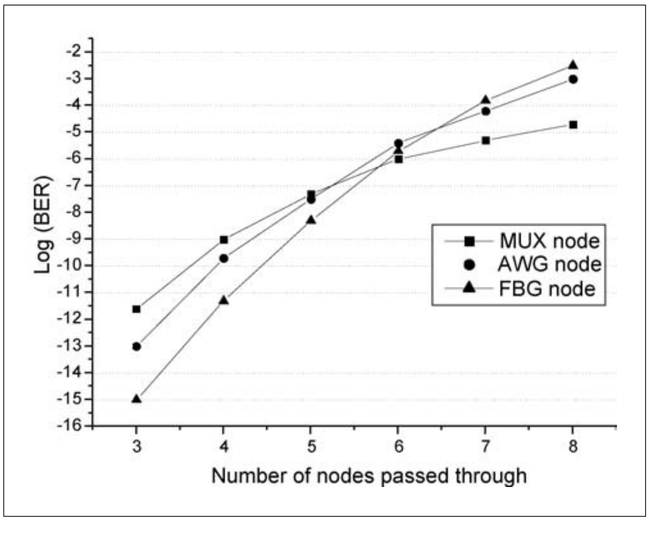
As the channels are closer to each other, the crosstalk and noise density is higher if there is no filtering when the light crosses an FBG node. Results based on this case are depicted on *Fig. 6.*

Conclusions

In the expectation of improving the cascading performances of all-optical components and systems, a migration path could be envisaged where in hybrid all-optical and opto-electronic technologies coexist. In this scenario, transparent "optical islands" would be conceived and could be bridged by 2R/3R regeneration systems. From an ASON/GMPLS network perspective [9], these "optical transparent islands" could be seen as different "domains" at the control and management planes to help distributing and/or partitioning the control and management tasks.

Correct sizing of these islands would be of prime importance for ensuring the desired SLS (service level specification) requirements (including quality measurement of BER).

Fig. 6.
Number of crossed optical nodes vs. BER foreseen when the channel spacing is 50GHz:
(a. mux/demux based, b. AWG based, c. FBG based)



In this paper, it has been demonstrated that the choice of the OADM architecture is a determinant factor in the size of an optical transparent island. In addition, for the considered cases, OADM architecture based on the multiplexers/demultiplexers and 2x2 optical switches is more advantageous than the AWG based architecture for enlarging the transparent island size.

It is worth noting that other reasons, besides technical aspects, may also contribute to limit the transparent islands' sizes, such as regulation constraints, interoperability, cost and network management/control issues.

Acknowledgement

The author is grateful to BUTE for funding, and to the Optical Lab of the Telecommunication Technical Center of Catalonia in Barcelona where the work was done.

References

[1] Ghislain Lévesque, Vance Oliver:
"Guide to WDM Technology and Testing"
2000 EXFO Electro-Optical Engineering Inc.,
Quebec City, Canada, ISBN 1-55342-001-2

[2] K.A. McGreer:
"Arrayed Waveguide Gratings for Wavelength Routing",
IEEE Comm. Magazine, Dec. 1998, pp.62–68.

[3] S.V. Kartalopoulos:
"Introduction to DWDM Technology"
USA, IEEE Press, IEEE, ISBN 0-7803-5399-4

[4] VPI photonicsTM

[5] ANRITSU CORPORATION, Application Notes:
Q Factor Measurement/Eye Diagram Measurement,
SDH/SONET Pattern Editing
http://www.eu.anritsu.com/files/
MP1632_1763_1764_EF1100.pdf

[6] MSZ EN 61280-2-8:
Fibre optic communication subsystem test procedures – Digital systems, Part 2-8:
Determination of low BER using Q-factor measurements (IEC 61280-2-8:2003),
Hungarian National Patent

[7] TIA/EIA-526-9: OFSTP-9:
Accelerated Measurement Procedure for
Determining BER and Q-factor in
Optical Transmission Systems Using the
Variable Threshold Method

[8] W. Fawaz et al.:
"Service level agreement and provisioning in optical networks", IEEE Communications Magazine,
Vol.42 (Jan 2004), pp.36–43.

[9] G.8080/Y.1304(11/2001)

VLAN tag-based cross-connection function in video network architecture

PÉTER SZEGEDI

MATÁV PKI Telecommunications Development Institute
szegedi.peter3@ln.matav.hu

Keywords: digital studio, Ethernet, transport network, label switching

Growing the magnitude of digital television techniques in studios as well as at the end users the solution of video communication over heterogeneous networks become more relevant. Broadcasters want to transport and exchange their digital streaming video and audio traffic between the studios in a packet-based manner. Innovative video communication services where new technologies enable additional opportunities are introduced in this paper. On the following pages the motivations and advantages of the integrated VLAN tag-based cross-connection function in Third-Generation SDH equipments are illustrated via hypothetical Hungarian video transport network architecture.

1. Introduction and motivations

Nowadays, the most of the modern television studios are equipped by digital studio techniques. Digital camcorders, continuity desks, audio-mixers, content storage devices and many different hardware equipment are installed enabling to use special softwares and applications such as virtual 3D studios, special video editing and so on.

In studios there are Local Area Networks (LANs) to realize communication and inter-working between different applications running on these high-performance computers and professional digital devices. Since the LANs are using Ethernet technology it is obvious to handle video and audio stream in Ethernet format. However, Gigabit Ethernet is more than just transport, it is the basis of the Next-Generation Digital Video Network.

Representing the merging of synchronous audio/video (A/V) systems and asynchronous data networks in studios the leading players in the broadcast industry developed an open file format called MXF to transfer video and audio streams as files over Ethernet networks. The Material eXchange Format (MXF) is an open file format, targeted at the interchange of audiovisual materials with associated data and metadata. It has been designed and implemented with the aim of improving file-based interoperability between various applications used in the television production chain. The transportation of these different files is independent of contents (e.g. not compression scheme specific) and the applications of manufacturer specific equipment are not required [1].

Parallel with this trend the main brands (like SONY) improve their product line and announce new camcorders support Ethernet or wireless LAN interfaces and new professional decks support up to five times faster-than-real-time transfer of full-resolution video over Gigabit Ethernet interfaces, in addition to MXF file transfer over a 100-Base-T network connection. Testing of live streaming video systems the *Level 3 Connections'* engineers successfully transferred 50 and 30 Mbps broadcast-quality digital video segments across a local network [2].

The traditional and widespread MAN or WAN core network transport technology is the SDH with the underlying optical cable infrastructure or WDM systems. As the video transport solutions in studios are shifting to Ethernet and Gigabit Ethernet technology it is obvious to use Ethernet Private Line or Ethernet Virtual Private Line services [3] for video transport between studios over the MAN and WAN. Therefore the service providers and network operators have to offer Ethernet connectivity over the existing SDH and WDM technology (migrating to Next-Generation and Third-Generation SDH) to meet the broadcasters' new requirements detailed in the next section. The relevant network functions and the tag-based cross-connection solution are described in the third section. An application example and the proper network and node architecture are proposed in the fourth and fifth sections to model the next-generation network functions. Finally, in the sixth section the technical and economical advantages of the VLAN tag-based cross-connection are illustrated by a real case study.

2. New client requirements

The audio and video stream transport over the asynchronous data networks requires strict packet loss, latency and jitter constraints. The broadcasters want to transport their compressed or uncompressed video streams in Ethernet frames through the MAN or WAN network according to different SLAs. The main requirements are:

- Reliable video stream transport
- Guaranteed bandwidth service

- SLA-guaranteed interconnection.
- MXF data transport.
- Just in Time service provisioning.

The modern television applications (like regional news, nationwide interactive games, region-dependent advertising, and so on) require flexibility of the video connections. Mainly Just in Time services are required in a customer-driven connection-provisioning manner between studios.

The different SLAs allow the broadcasters to scale the transport services to their demands. For example stricter, and more expensive as well, protected Ethernet connections for the live streaming video transport are needed from the local studios to the main studio or the distribution points and non-protected Ethernet connections for the stored content exchange are needed between the studios.

3. Network functions and architecture

From the transport service provider point of view the native switched Ethernet core networks have well-known restrictions in fields of guaranteed bandwidth, QoS and fast (<50ms) protection/restoration mechanisms. But on the other hand the Ethernet technology's advantages (e.g. plug-and-play installation, good scalability and granularity, VLAN security, simplicity of operation, low cost, etc.) provide good economics of scale for the service providers. To eliminate the disadvantages of the native Ethernet and bring Ethernet economics with SLA guarantees to the existing SDH infrastructure Next-Generation SDH (NG-SDH) functions for the Ethernet services are implemented in the service providers' networks.

The Ethernet Private Line and Ethernet Virtual Private Line services [3] over the NG-SDH network provide such performance as traditional private line services. With the GFP, VCAT and LCAS techniques [6] the Ethernet over NG-SDH network architecture has great performance to meet clients' specific requirements.

Traditionally the NG-SDH equipment provides a time-slot cross-connect that allows time-slots from one physical interface to be cross-connected to a different physical interface. As the amount of the Ethernet services in the SDH network is growing, it is worth to integrate frame-based cross-connect into the SDH equipment [4].

Each Ethernet service is identified and segregated by tagging the Ethernet frames, using VLAN tags. Benefits of this integrated frame cross-connect include:

- Multiple Ethernet services can be presented on a single physical interface *(Fig. 1.)*, thus the client's Ethernet equipment can be connected to the VLAN tag-based cross-connect cloud in more economical way.
- Multiple Ethernet services can share the given SDH bandwidth, namely point-point VLAN connections can be multiplexed into virtually concatenated SDH payloads.
- Statistical gain can be achieved to overbook the SDH bandwidth, i.e. in some cases the individual peak bit rates could be higher than average.

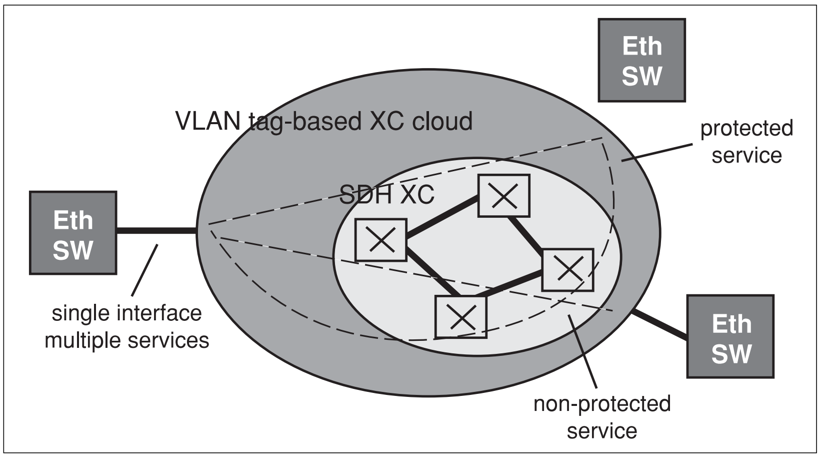
The physical integration of the frame cross-connect into the SDH equipment reduces the number of interfaces by an order of magnitude or more. In addition to reducing the cost of physical interfaces, provisioning of multiple services on a single Ethernet interface increases the traffic density and reduces operation costs. Based on the ability to provide multiple, SLA-guaranteed services per Ethernet interface, carriers can now grow the revenue per customer without incremental investment in access infrastructure.

The service provider prefers to provide guaranteed bandwidth point-point VLAN connection over the NG-SDH network. The automatic provisioning of VLAN connections gives flexibility to the network. The management of the VLAN tags in the Ethernet layer is the main issue form the feasible network operation point of view. The connection set up and the valid VLAN tag administration is supported by the GVRP (Generic VLAN Registration Protocol). GVRP remove the burden of manually installing and managing VLANs from the network administrator's hands, provides a mechanism for dynamic maintenance of VLAN Active Filtering Database and for propagating the information they contain to other VLAN-aware switches. GVRP was namely never designed to set up point-point VLANs (the VLAN topologies would rather resemble a sub-tree of the Spanning Tree topology). The required GVRP modifications [7] can be done correctly, but these are out of this paper's scope.

4. Application example

To illustrate the advantages of the integrated VLAN tag-based cross-connects in the networks, typical Hungarian video transport architecture among the main re-

Fig. 1. Network architecture



gional studios (located in the capital and in largest country towns) is proposed. The network architecture contains an Ethernet and a NG-SDH layer. The regional studios are connected to the redundant Ethernet switches. Between the switches the dual star logical topology provides redundant paths against the failures in the Ethernet layer. The logical loops are avoided by the STP algorithm.

Fig. 2. Ethernet logical topology

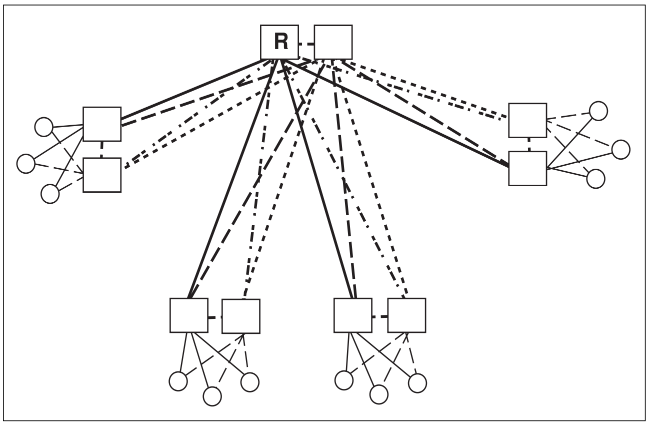
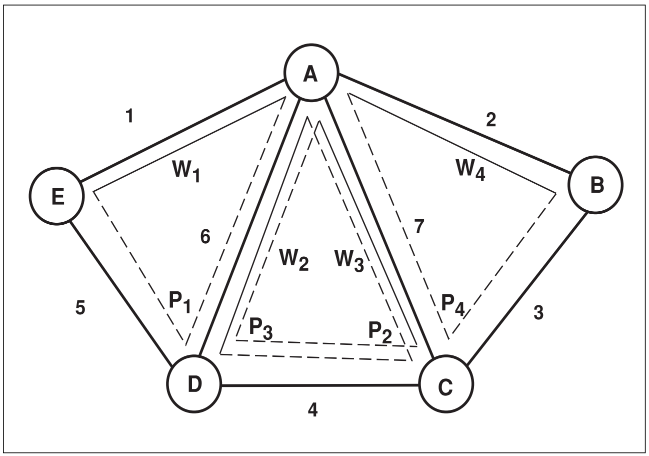


Fig. 3. SDH physical topology



The underlying SDH layer has more connected physical topology, this provides to establish link-independent physical connectivity for the Ethernet services. The Ethernet layer provides protection only against the failure of the Ethernet layer. The physical link failures are protected by the SDH layer's 1+1 path protection because the convergence time of the STP algorithm is not feasible for the clients' requirements. The physical link failures are hided from the Ethernet by the SDH protection.

In the regional studios the streaming video connection demands are separated into different VLANs by the customer Ethernet equipment's GVRP protocol, than the VLANs are mapped into the right sized virtually concatenated VC-4 payloads of the SDH to the desired direction. According to the different SLAs there are protected VLAN demands and non-protected VLAN demands on the customer side. Because of the optimal SDH bandwidth utilization the service provider wants to protect only the protected VLAN demands in the transport SDH layer. So the VLANs can be identified and grouped by the service provider equipment. In case of no VLAN tag-based cross-connecting function, the different VLANs should be identified by tributary ports, but if there is frame cross-connection in the SDH layer the different VLANs are identified by the VLAN tags on the single tributary port.

5. Network and node models

To illustrate the benefits of the application of integrated frame cross-connects detailed node models are introduced in this paper. The main functions are both identified in case of Next-Generation SDH and so called Third-Generation SDH [5] equipments with VLAN tag-based cross-connecting function.

Only one direction of one connection is shown on the next figures *(Fig. 4.)*

Fig. 4.
Ethernet – NG-SDH node models (a, b)

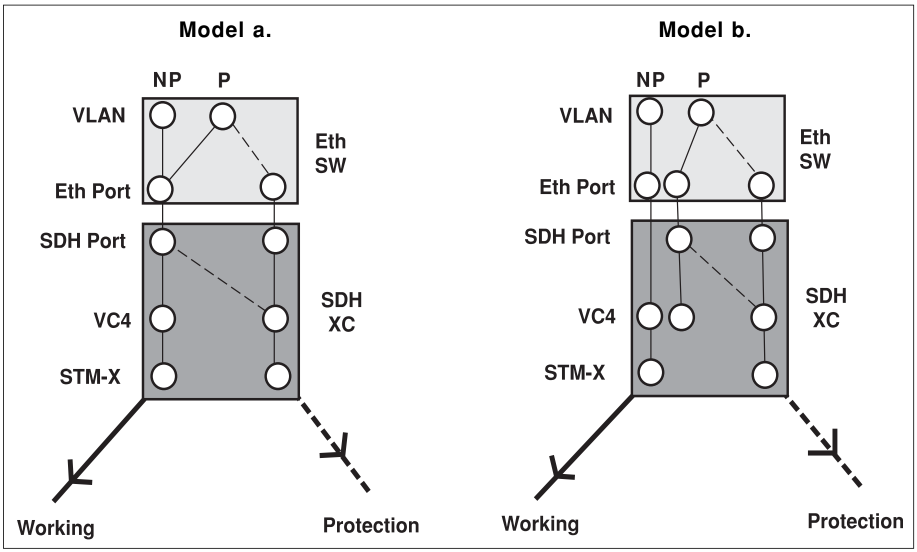
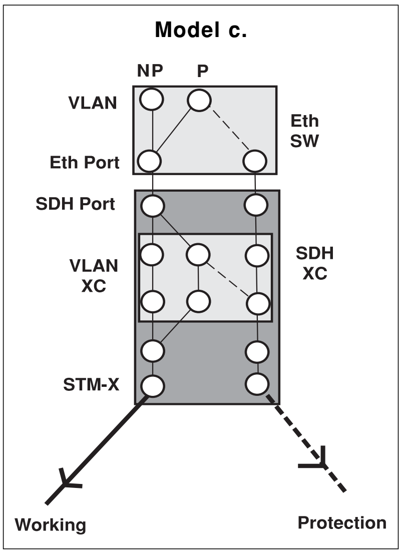


Fig. 5.
Ethernet – TG-SDH node model (c)



In case of NG-SDH transport architecture without selective protection all type of services are presented on a single Ethernet-SDH interface *(Fig. 4/a).*

The Ethernet switch (Eth SW) provides protection only against the interface failure and the SDH cross-connect (SDH XC) provides protection against the link failure. In case of selective protection the protected (P) and the non-protected (NP) services had to be separated into different interfaces *(Fig. 4/b.)*, thus more ports are required. But based on the SDH tributary ports the SDH cross-connect can provide protection only for the protected services, so fewer SDH capacity is required.

If the bandwidth of the protected and non-protected VLANs (VLAN), the Ethernet port capacities (GbE) and the SDH concatenation unit size (VC4) are known the following formulas describe the number of required ports (#port) and SDH working and protection transport capacities (#W_VC4, #P_VC4).

Model a – No selective protection

$$\# port_{1a} = \left\lceil \frac{\sum VLAN_P + \sum VLAN_{NP}}{GbE} \right\rceil + \left\lceil \frac{\sum VLAN_P}{GbE} \right\rceil$$

$$\# \ddot{U}\_VC4_{1a} = \left\lceil \frac{\sum VLAN_P + \sum VLAN_{NP}}{VC4} \right\rceil$$

$$\# V\_VC4_{1a} = \left\lceil \frac{\sum VLAN_P + \sum VLAN_{NP}}{VC4} \right\rceil$$

Model b – Port-based selective protection

$$\# port_{1b} = 2 \times \left\lceil \frac{\sum VLAN_P}{GbE} \right\rceil + \left\lceil \frac{\sum VLAN_{NP}}{GbE} \right\rceil$$

$$\# \ddot{U}\_VC4_{1b} = \left\lceil \frac{\sum VLAN_P}{VC4} \right\rceil + \left\lceil \frac{\sum VLAN_{NP}}{VC4} \right\rceil$$

$$\# V\_VC4_{1b} = \left\lceil \frac{\sum VLAN_P}{VC4} \right\rceil$$

In case of Third-Generation SDH with selective protection based on the VLAN tag-based cross-connection function the all type of services are presented on a single Ethernet-SDH interface *(Fig. 5* – Model c).

Thanks to the frame cross-connection capability in SDH (VLAN XC) the protected and non-protected services can be identified by VLAN tags on a single interface as well. Thus, the SDH can provide selective protection for the protected VLANs against the link failures. This solution requires fewer ports and fewer SDH capacities than the NG-SDH selective protection solution (Model b).

The following formulas describe the needed port numbers (#port) and SDH transport capacities (#VC4).

Model c – VLAN tag-based selective protection

$$\# port_2 = \left\lceil \frac{\sum VLAN_P + \sum VLAN_{NP}}{GbE} \right\rceil + \left\lceil \frac{\sum VLAN_P}{GbE} \right\rceil$$

$$\# W\_VC4_2 = \left\lceil \frac{\sum VLAN_P + \sum VLAN_{NP}}{VC4} \right\rceil$$

$$\# P\_VC4_2 = \left\lceil \frac{\sum VLAN_P}{VC4} \right\rceil$$

6. Case studies

Above the network architecture and topology described in Section 4., assuming a given traffic matrix witch contains the number of point-point VLAN demands between regional studios. The bandwidth of this broadcast-quality, uncompressed connection demands is 165 Mbps (IEC-601). The higher-order SDH virtual concatenation unit is one VC-4 (139,264 Mbps), because the maximum number of virtually concatenated lower-order containers (e.g. VC-12) is 64, so this is not enough for the VLAN's bandwidth [6].

In normal cases one VLAN requires one VC-4-2v payload and two different VLANs require a VC-4-4v payload (2xVC-4-2v). In case of VLAN tag-based cross-connection one VLAN requires one VC-4-2v payload as well but two VLANs requires only a VC-4-3v payload because the connections are identified by the VLAN tags.

The detailed network and node models enable to analyze many of technical and economical case studies. Based on the simple example above it is easy to understand that the VLAN tag-based solution requires fewer SDH protection capacities and lower number of interfaces to satisfy the client's requirements.

On the network level, instead of the total number of required network resources in the service provider point of view the most interesting question is the number of unused resources and the possible points of the capacity upgrades. Based on the proposed models in this paper, assuming given link capacities the following figures show the unused resources and the capacity upgrade points in function of the relative traffic load *(Fig. 6).*

As it is shown, it is obvious the port-based selective protection (Model b) enables to extend the capacity upgrade point compared to the no selective protection case (Model a) because of the fewer protection capacity needs. The proposed architecture with VLAN tag-based cross-connection (Model c) enables to provide more VC-4s and more VLANs *(Fig. 7.)* in the same network, so enables to more extend the capacity upgrade point as the relative traffic load is increasing.

In connection with this, the required interface card numbers by the VLAN tag-based cross-connection case

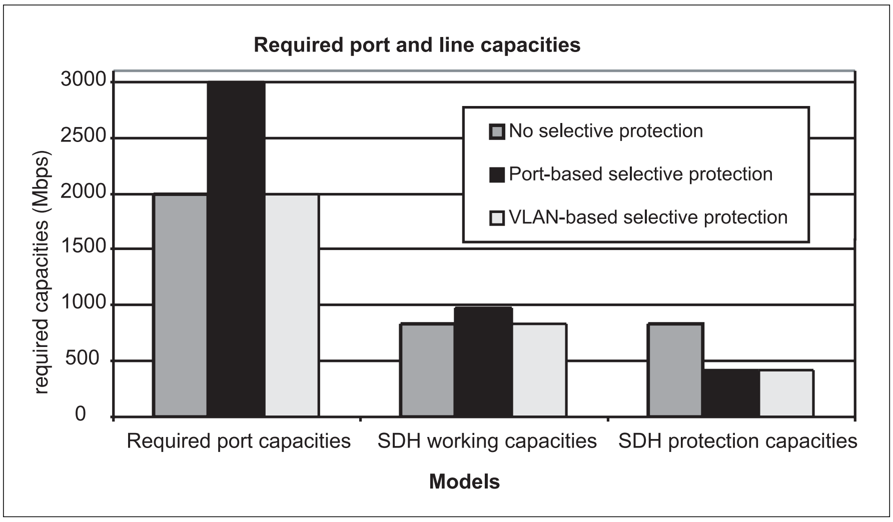


Fig. 6. Capacity requirements

(Model c) are the same as the case of no selective protection (Model a), so beside better network performance the total network cost can be lower than in port-based selective protection case.

7. Conclusions

Broadcasters want to transport and exchange their digital streaming video and audio traffic between the studios in a packet-based manner. The Next-Generation SDH equipments and the integrated VLAN tag-based cross-connecting function (e.g. Third-Generation SDH) enable to flexibly establish guaranteed bandwidth VLAN connections over the transport network.

Based on the proposed models the advantages of the integrated frame cross-connection function are manifested in the network resources, the network utilization, the interface cost and the total network cost as well.

References

[1] EBU Techn. Review – B. Devlin: MXF – the Material eXchange Format, 2002.

[2] Sony Press Release (2003): Sony and Level 3 transfer broadcast video segments across Ethernet network directly from tape playback, http://news.sel.sony.com/pressrelease/4035

[3] Appian Communications – Carrier-class Ethernet: A service definition, 2001. http://www.appiancom.com/solutions.htm

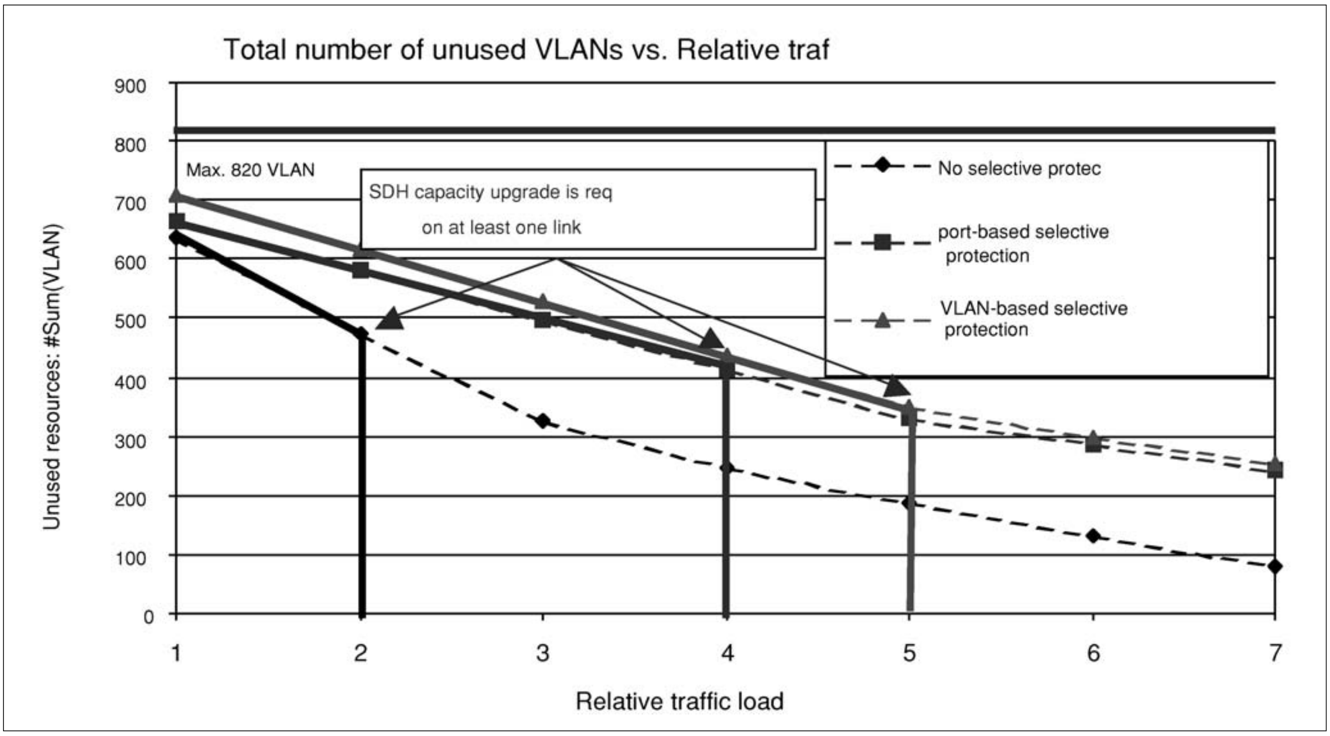
[4] Marconi – G. W. Rees (2002): Physical integration of SDH switching and Ethernet switching – Analyzing the opportunities and constraints.

[5] Heavy Reading – The future of SONET/SDH, Vol.1, No. 6, November 14., 2003.

[6] ITU-T G.7041/Y.1303 GFP; ITU-T G.707 VCAT and ITU-T G.7042/Y.1305 LCAS

[7] F. V. Quickenborne, F. De Greve, P. V. Heuven, F. De Turck, B. Vermeulen, S. V. den Berghe, I. Moerman, P. Demeester: Tunel set-up mechanisms in Ethernet networks for fast moving users, 2004.

Fig. 7. Capacity upgrade points



GUI design and structure of a teleeducational multimedia courseware

JÁN TURÁN[1], ĽUBOS OVSENÍK[1], JÁN TURÁN JR.[2], KÁLMÁN FAZEKAS[3]

[1] University of Technology Košice, Slovakia
[2] 3D People GmbH, Reutlingen, Germany
[3] Budapest University of Technology and Economics, Hungary

Keywords: rapid transforms, graphical user interface design, multimedia and educational telecommunication courseware

In the paper a systematic approach to multimedia communication graphical user interface design is evaluated. The proposed application is in the field of radioengineering teleeducation. The design starts with a standard task analysis using task knowledge structures as the basis of the task model. Then description of the media resources available to the system and their access is evaluated. Finally, the task information is elaborated by attaching dialogue acts to specify the desired communicative effects for each task step. This method is a frame for a software mean developing of an interactive multimedia course: Rapid Transform and Its Application.

Multimedia graphical user interfaces are currently created by intuition. They are usually designed and developed without exact analysis of multimedia information presentation. The most nowadays multimedia applications present possibilities of the technical means but they have not respected a user-centred approach. The multimedia interfaces developed by this way cannot achieve the maximum effect [1–3].

In the paper a systematic approach to multimedia communication graphical user interface design is evaluated. The proposed application is in the field of radioengineering teleeducation. The design starts with a standard task analysis using task knowledge structures as the basis of the task model. Then description of the media resources available to the system and their access is evaluated. Finally, the task information is elaborated by attaching dialogue acts to specify the desired communicative effects for each task step. This method is a frame for a software mean developing of an interactive multimedia course: *Rapid Transform and Its Application*.

1. Multimedia interface design

The exact definitions of human – computer – interface terminology and user-centred -design are very important for understanding of the multimedia design problems. We must view multimedia interfaces from an appropriate perspective. Multimedia is one of the most innovative ways of using a telecommunications network to achieve effective communications between people and for access to information. In [3–7] was pointed out that multimedia approach could be viewed either from a technological perspective or from a user-centred perspective.

The technological perspective is defined through lists of technical characteristics of system claiming to be multimedia systems such as multidimensional presentation techniques, multimodal interaction or hypermedia techniques. User-centred perspective focuses on the possibilities offered by the technology. A user centred definition characterise multimedia systems as systems enabling the usage of multiple sensory modalities and multiple channels of the same or different modality enabling the user to perform several tasks at the same tasks at different times [3,9,10]. The understanding of these two definitions is important for view the key question of multimedia interface design when to use which media and in what combination to achieve the maximum effect.

There are several alternatives for classifying broadband multimedia communications according to information types, communication types (dialogue, messaging, retrieval, distribution), organisation (tree-structured media, hypermedia), functions (e.g., interactive, link navigation, co-operative teleworking), and other criteria. One suitable alternative classifies the utilisation of information types with respect to user or terminal. This is particularly advantageous for the definition of telecom services to effectively support multimedia communication.

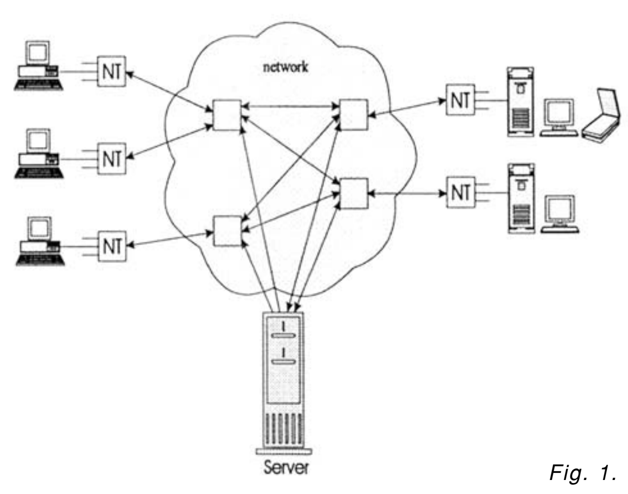


Fig. 1.

The target of future multimedia communication is: to efficiently support manifold applications; to provide worldvide communication capabilities of user as large as possible; unproblematic and cost-effective operation and utilisation on the basis of standardised components with high production volumes; and simple interworking [1,4,5,7].

A modern teleeducations must be thought of in terms of networked organisation *(Fig.1).* The objective of co-operative teleworking among students and teachers (with simultaneously possible using of databases on a learner side) is the provision of some degree of "telepresence" for geographically distributed persons and teaching materials in a quality comparable to that of a real-world lecture (conference). Co-operative teleworking enables a group of distant participants to jointly view, discuss, and edit multimedia documents while at the same time using communication and computing resource. This can be considered as an extension of conventional audio/video conferencing access, and collaborative work assistance. A desktop multimedia workstation allows the student to create, retrieve, and manipulate and activate a "hotline" to a teacher (central specialist). Co-operative teleworking represents a case of complex and dynamic communication, which encompasses a number of participants, connections, information types, systems, and functions [8–12].

Emerging access techniques (such as XDSL over copper network, HFC on PON) and FTTx have result into a number of demonstrators and field trials in the are of educational telecommunication. The users are connected to the SDH backbone network by means of a copper or fiber access network that transports ATM [7–12].

2. Systematic design for task related multimedia interfaces

Multimedia GUI is usually designed by intuition. This way is not suitable to use all the available resources and utilising the multimedia effect in maximally way. So it is important to develop a systematic approach to multimedia GUI design [7,8]. The agenda of issues, which a method must address, were first, the creation of a task model incorporating specification of information requirements and presentational effects, accompanied by a resources model describing the information media available to the designer.

The GUI design method should advise on selecting appropriate media for the information needs and scripting a coherent presentation for a task context. The design must with directing the user's attention to extracting required information from a given presentation and focus on the correct level of details. In addition, the design method should guide the designer to the cognitive issues underlying a multimedia presentation such as selective attention, persistence of information, concurrency and limited cognitive resources such as working memory. *Fig. 2.* gives an overview of a systematic method for teleeducation graphical user interface design based on the methods of task components. This is based on the following components: model, information flow, process, source, destination [8–11].

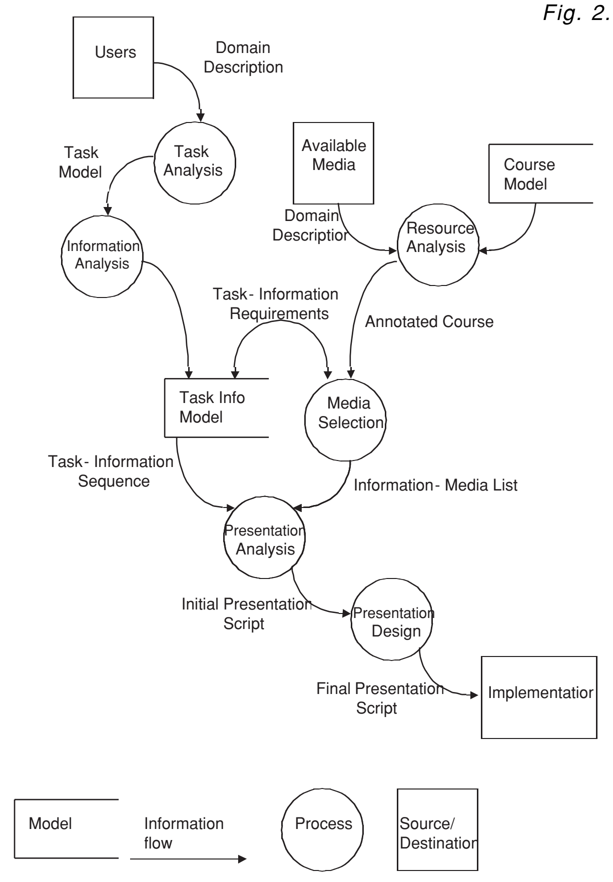


Fig. 2.

3. Multimedia rapid transform courseware description

This course is an interactive multimedia course based on use of multimedia document and visual simulations programme package for teleeducation purposes. As a particular example the rapid transform and its application (RT&IA) was chooses [13]. The course structure and some of its interactive features are noticed on *Fig. 3.*

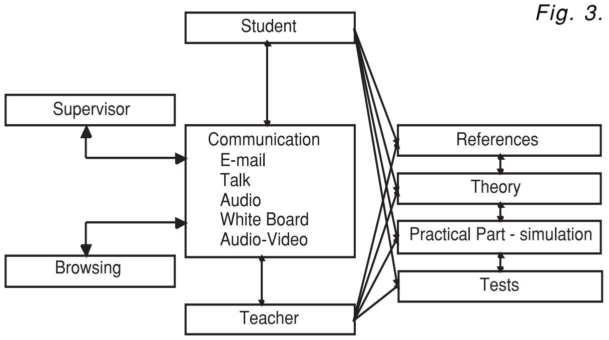
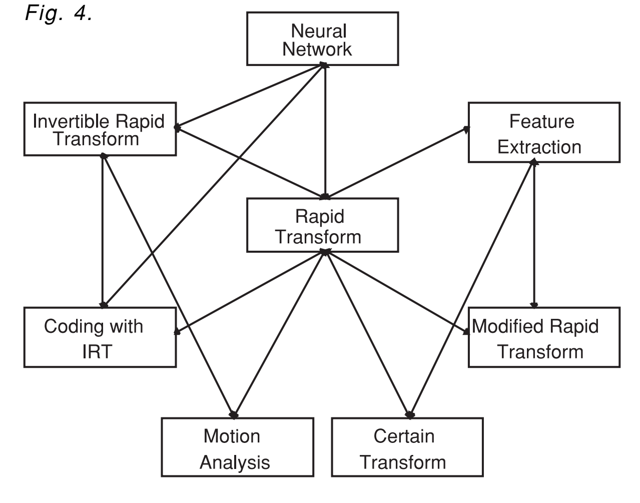


Fig. 3.

Student and teacher have access to an interactive multimedia document stored in a server. Teacher as a master has the possibility of changing this document if necessary. There are possibilities of interactive multimedia communications between student and teacher using various tools (E-mail, talk, audio, White Board, Audio-Video).

Teacher has the possibility to supervise of students work and able to monitor his/her progress and interactively change-tailor the course content.

Fig. 4.



The basic organisation of the course consist from four parts:

Theoretical part – this is an interactive multimedia document about the theory of rapid transform and related class of fast translation invariant transforms (CT). The block in the figure represents of the CAD or CAE oriented programme routines of programme packages CT-CAD or IRT-MA able to solve problems in the area of application of rapid transform *(Fig. 4)*.

Practical part – this is an interactive multimedia based simulation programme package able to solve CAD and CAE problems in the area of applications of rapid transform and other transform from the class CT in the area of DSP, pattern recognition, image processing (image coding, motion vector estimation) for various applications *(Fig. 5)*. As shown the basis of this part are programme packages CT-CAD and IRT-MA with applications to:
- Pattern recognition
- Image or generally signal coding
- Motion estimation
- Polar plot of motion vectors

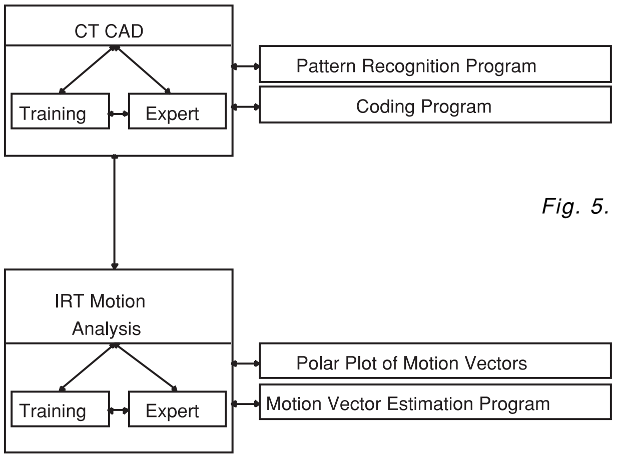

CT CAD
Training
Expert
Pattern Recognition Program
Coding Program
IRT Motion Analysis
Training
Expert
Polar Plot of Motion Vectors
Motion Vector Estimation Program

Fig. 5.

The practical part has two modifications:

A. Training Simulations

This is by teacher preselected and automatically sequenced (with student interactivity) solved examples of various applications of rapid transform. The sequence of multimedia documents obtains control points for student and teacher and may be interactively changed by student and teacher. There is possible to establish a hot-line consultation with teacher.

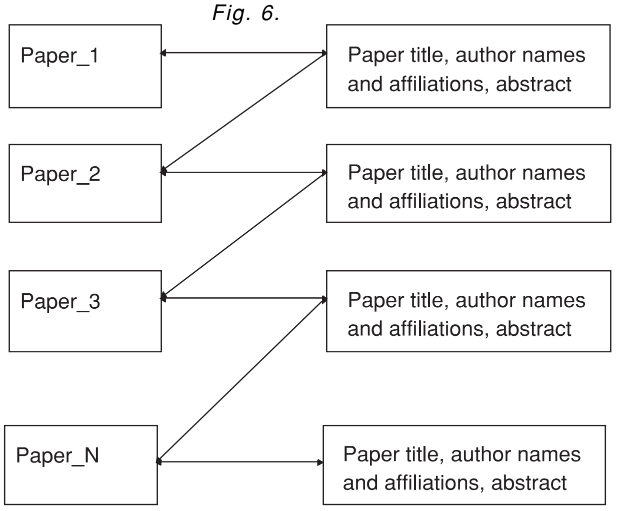
B. Expert Simulations

This is a possibility to creation of new problems and solves them with the use of CAD and CAE capabilities of in course embedded programme packages (CT CAD and IRT MA). Students have entered this possibility after there properly learn the part 1 and part 2A. There is possibility to interact remote teacher if there occurs any problem.

Part References – this is a multimedia document about every published documents related to rapid transform or its applications *(Fig. 6)*.

Part tests – the tests embedded to the courseware are entitled to evaluated the knowledge, routines and working shills obtained by students trough the learning process *(Fig. 7)*.

Fig. 6.


Paper_1
Paper title, author names and affiliations, abstract
Paper_2
Paper title, author names and affiliations, abstract
Paper_3
Paper title, author names and affiliations, abstract
Paper_N
Paper title, author names and affiliations, abstract

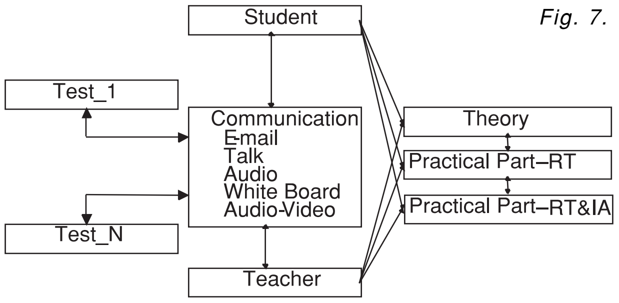


Fig. 7.

4. Feedback in the rapid transform courseware

The architecture of feedback used in the courseware is depicted on the *Fig. 8.*

It consists of five feedback loops, which are realised on the both level of the course. The simplest way of feedback is the study and practicing solved examples embedded to the courseware. The quality of the courseware and the student progress in the course may be evaluated using predefined Questionnaire and the course statistics available to the teacher (course supervisor). Course statistics deals with registration and multimedia document utilization (users data, data and time using, working on course, results of evaluation etc.). Course Questionnaire deals with questions about course structure, optimal material selection, multimedia document quality etc. Tests embedded to the courseware are entitled to evaluate the knowledge, routines and working skills obtained by students trough the learning process. The test is structured trough the course content and may consist from the questions, unsolved examples and simulation problems. If there is any problem with the student progress in the course the student is able activate a hot line to the teacher, but only in consultation hours. At the present level of development of the course and available technology it may be only an E-mail contact with the remote teacher. Outputs from the feedback is structured, saved and statistically processed to be used for improving the course quality in next development step.

Multimedia Graphical User Interface designed with use of proposed systematic design approach can achieve an increasing effect in the practical use of human computer interface for teleeducation purposes. The method was practically tested in the described application in the field of DSP teleeducation. The implementations of the RT&IA multimedia courseware demonstrate the key features of the systematic GUI design method.

Acknowledgements

The authors are thanking for the financial support from the COST 276 and COST 292 grant and VEGA grant No.1/0381/03.

References

[1] Altly, J. L., Bergan, M.: The design of multimedia interfaces for process control. Proc. 5th IFIP, IFAC, IFORS, IEA Conf. on Man-Machine Systems, 1992, pp.249–255.

[2] Altly, J. L., McCartney, C. D.: Design of a Multimedia Presentation System for a Process Control Environment. In.: Pro. Eurographics Workshop, Stockholm, 1991.

[3] Wright, D. J.: Broadband – Business Services, Technologies and Strategic Impact. Boston/London, Artech House, 1993.

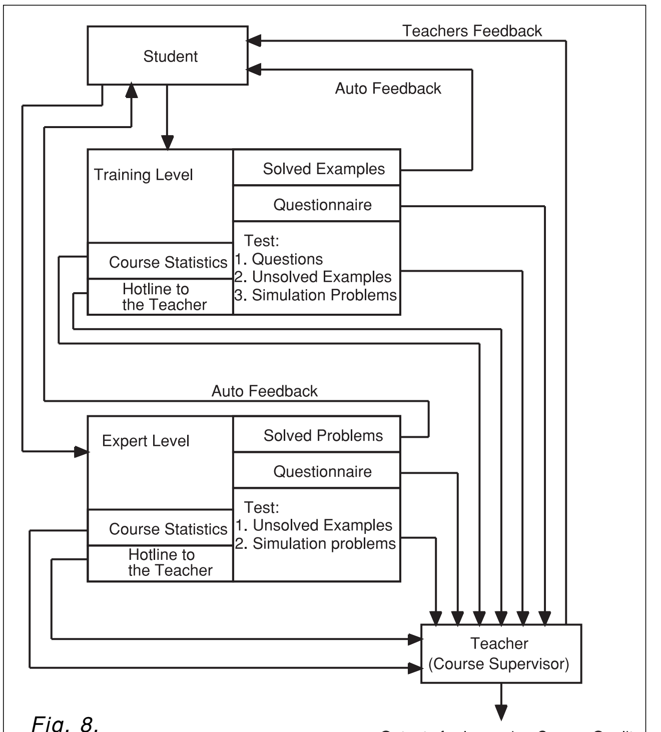


Fig. 8.

[4] Barbosa, L. O., Georganas, N. D.: Multimedia Services and Applications. Europ. Trans. Telecom., Vol.2, No.1, 1991, pp.5–19.

[5] Armbrüster, H., Wimmer, K.: Broadband Multimedia Application Using ATM Networks: High-Performance Computing, High-Capacity Storage and High-Speed Communication. IEEE JSAC, Vol.10, No.9, 1992, pp.1382–1396.

[6] Rosenberg, J. et al.: Multimedia Communications for Users. IEEE Commun. Mag., May 1992, pp.20–36.

[7] Thiriet, J. M. et al.: Towards a Pan-European Virtual University in Electrical and Information Engineering. IEEE Trans., Vol. ED V-45, No.2, 2002, pp.152–160.

[8] Kövesi, M., Turán, J., Ovseník, L., Kövesi, L.: GUI Design and Structure of a Multimedia Teleeducation Course. DSP'97, Herlany, Slovak Rep., Sept. 3-4, 1997, pp.96–100.

[9] Ovseník, L.,Turán, J., Kövesi, M., Kövesi,L.: Graphical User Interface for Teleeducation. DSP'97, Herlany, Slovak Rep., Sept. 3-4, 1997, pp.101–105.

[10] Turán, J., Fazekas, K., Ovseník, L., Kövesi, M., Kövesi, L.: Tools for Rapid Transform Multimedia User Terminal. COST 254, Budapest, Hungary, Feb. 1997, pp.1–9.

[11] Turán, J., Ovseník, L., Turán, J. jr.: Web-based Multimedia Courseware: Applied Photonics. Proc. EC-VIP-MC 2003: 4th Eurasip Conference, Zagreb, Croatia, 2003, pp.741–746.

[12] Tran, S. M., Lajos, K., Balázs, E., Fazekas, K., Csaba, Sz.: A Survey on the Interactivity Feature of MPEG-4. Invited paper: ELMAR 2004, Zadar, Croatia, 2004, 30-38.

[13] Turán,J.: Fast Translation Invariant Transforms and Their Applications. Elfa Press, Kosice, 1999.

THEORETICAL STUDIES

SPEECH F0 ESTIMATION WITH ENHANCED VOICED-UNVOICED CLASSIFICATION

Keywords: voicing determination algorithms, basic parameter extraction

Pitch detectors for speech signal can only work correctly if the fundamental frequency estimation is linked with a reliable voiced-unvoiced decision. A pitch detection algorithm is presented with an enhanced voicing detection method, which gives less error rate than concurrent methods. This pitch detector is based on the well-known autocorrelation method with some modification. The robustness of the algorithm on voicing decision was evaluated over a database of speech recorded together with a laryngograph signal.

SHORT IMPULSE-PROPAGATION IN INHOMOGENEOUS PLASMA

Keywords: Maxwell's equations, wave propagation, MIBM

In this paper the problem of real impulse-propagation in arbitrarily inhomogeneous media will be presented on a fundamentally new, general, theoretical way. The general problem of wave-propagation of monochromatic signals in inhomogeneous media was enlightened in [1]. The former theoretical models for spatial inhomogeneities have some errors regarding the structure of the resultant signal originated from backward and forward propagating parts. The application of the Method of Inhomogeneous Basic Modes (MIBM) and the complete full-wave solution of arbitrarily shaped non-monochromatic plane-waves in plasmas made it possible to obtain a better description of the problem, on a fully analytical way, directly from Maxwell's equations. The model investigated in this paper is inhomogeneous of arbitrary order (while the wave-pattern can exist), anisotropic (magnetized), linear, cold plasma, in which the gradient of the one-dimensional spatial inhomogeneity is parallel to the direction of propagation.

ASSESSMENT OF THE ERRORS IN SINGLE POINT POSITIONING

Keywords: single point positioning, submeter accuracy, permanent stations, position errors

After turning off Selective Availability (SA) a new chapter began in the GPS-technique. It was stated, that in favourable conditions accuracy of several meters is achievable. Recently the number of GPS users has impressively increased; turning off SA has clearly played an important role in the propagation of GPS technique. Turning off SA is considered as a key point not only for the practice, but also for scientific researchers. It is well known, that compared to the artificial degradation of GPS accuracy, the effect of systematic and random errors on single point positioning is practically negligible. Some of the receivers do not take into account some systematic effects, because of the order of magnitude of SA error. Formerly errors on single point measurements could be invetigated only with limited efficiency. Turning off SA offers an opportunity to assess all the systematic and random effects in details; this paper will summarize the most important results of these investigations.

SECURITY AND AUTHENTICATION

BIOMETRIC AUTHENTICATION SYSTEMS

Keywords: personal identification, electronic scanning, security

The security requirements of identification systems have increased considerably recently. This rapid change can partly be explained by the political trends that caused the people to be more and more concerned about their personal or proprietary safety. The conventional security solutions are no longer able to satisfy this demand. Therefore new authentication systems have to be introduced. Amongst these new systems the ones based on biometric authentication may play a decisive role.

NETWORK ARCHITECTURE TO PROVIDE SECURE ANONYMOUS COMMUNICATION

Keywords: anonymity, architecture, GPSAA, secure communication

Anonymity becomes more and more important in today's privacy-aware information society. Unfortunately the current network layer hierarchy does not support anonymous communication, thus new layers need to be introduced to provide anonymous, yet secure communication in a transparent and easy-to-use fashion. The introduced model of general-purpose secure anonymity architecture (GPSAA) aims to fulfill this purpose.

MANAGEMENT

LTRACK – A NOVEL LOCATION MANAGEMENT METHOD

Keywords: location management, mobility

In this paper we propose a new location management algorithm for mobility networks. Our algorithm is called LTRACK, it stands for "location tracking". The signaling load that LTRACK puts on the network can be significally less than that of conventional location management algorithm.

ON-BOARD AUTONOMY OF LANDER UNITS FOR COMET NUCLEUS EXPLORATION

Keywords: space research, spacecraft, autonomy, high reliability

Keeping lander units functional in the hostile, energy-lacking environment in the outskirts of our Solar system is a great challenge. An autonomous real-time control system of a lander is expected to response on board to both nominal and non-nominal events without any external intervention. Recent developments in microelectronics make it possible to use such space-qualified microprocessors that allow the development of highly autonomous on-board software systems. But the increased computing power itself is not all – equally advanced software methods are also needed to provide real autonomy. Considering the complexity of a number of mutually interacting tasks, it is necessary to model them by well-described abstract logical modules. Our focus was on managing the static and dynamic behaviour of the system separately and eventually we developed the Mission Sequencing Object Model Language for describing the long-term autonomous mission control mechanism. This model was implemented on the Philae Lander for the Rosetta mission of the European Space Agency.