International Journal on Advances in Systems and Measurements

2009 vol. 2 nr. 1

The *International Journal On Advances in Systems and Measurements* is Published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal On Advances in Systems and Measurements
Volume 2, Number 1, 2009

- Igor Podebrad, Commerzbank, Germany
- Leon Reznik, Rochester Institute of Technology, USA
- Chi Zhang, Juniper Networks, USA

Sensor Networks

- Steven Corroy, University of Aachen, Germany
- Mario Freire, University of Beira Interior, Portugal / IEEE Computer Society - Portugal Chapter
- Jianlin Guo, Mitsubishi Electric Research Laboratories America, USA
- Zhen Liu, Nokia Research – Palo Alto, USA
- Winston KG Seah, Institute for Infocomm Research (Member of A*STAR), Singapore
- Radosveta Sokkulu, Ege University - Izmir, Turkey
- Athanasios Vasilakos, University of Western Macedonia, Greece

Electronics

- Kenneth Blair Kent, University of New Brunswick, Canada
- Josu Etxaniz Maranon, Euskal Herriko Unibertsitatea/Universidad del Pais Vasco, Spain
- Mark Brian Josephs, London South Bank University, UK
- Michael Hubner, Universitaet Karlsruhe (TH), Germany
- Nor K. Noordin, Universiti Putra Malaysia, Malaysia
- Arnaldo Oliveira, Universidade de Aveiro, Portugal
- Candid Reig, University of Valencia, Spain
- Sofiene Tahar, Concordia University, Canada
- Felix Toran, European Space Agency/Centre Spatial de Toulouse, France
- Yousaf Zafar, Gwangju Institute of Science and Technology (GIST), Republic of Korea
- David Zammit-Mangion, University of Malta-Msida, Malta

Testing and Validation

- Cecilia Metra, DEIS-ARCES-University of Bologna, Italy
- Krzysztof Rogoz, Motorola, Poland
- Rajarajan Senguttuvan, Texas Instruments, USA
- Sergio Soares, Federal University of Pernambuco, Brazil
- Alin Stefanescu, SAP Research, Germany
- Massimo Tivoli, Universita degli Studi dell'Aquila, Italy

Simulations

- Tejas R. Gandhi, Virtua Health-Marlton, USA
- Ken Hawick, Massey University - Albany, New Zealand
- Robert de Souza, The Logistics Institute - Asia Pacific, Singapore
- Michael J. North, Argonne National Laboratory, USA

Additional reviews by:

- Carlos Alexandre Barros Mello, Universidade de Pernambuco, Brazil

International Journal On Advances in Systems and Measurements
Volume 2, Number 1, 2009

Foreword

Following the proposed targets announced in the end of 2008, we are ready now with the 2009, Vol. 2, No. 1 of the *International Journal On Advances in Systems and Measurements* published by IARIA. This issue is completed due the valuable support of the reviewers, together with the enthusiastic contribution of the editorial team.

The submission to this journal is based on invitation only; the awarded and outstanding papers presented at IARIA conferences are the reliable candidates. Nevertheless, the enhanced versions of these articles have to follow a standard review process in order to be published. This assures that high quality scientific papers are finally selected to appear in this journal.

Among the received submission for this issue, eleven papers were qualified for publication. They cover several research topics in the field of Systems and Measurements. In the first article, Subhendu Das et al. make an interesting extension of the sampling theory, showing potential applications for digital communication systems. Jyrki T.J. Penttinen, the author of the second paper, presents a method to collect and analyze the key performance indicators of the DVB-H radio interface, using a mobile device as a measurement and data collection unit. In the third article, Nicolas Repp et al. develop an integrated approach towards distributed service level agreements (SLA) monitoring and enforcement. The fourth paper by Ramón José Aliaga et al. proposes a mixed approach to artificial neural networks training, based on a system-on-chip architecture on a reconfigurable device. The authors of the fifth article propose a least-squares adaptive algorithm based on the QR-decomposition, which is suitable for implementation on fixed-point DSP platforms. Gerardo Arturo López et al. develop in the sixth paper several methods for the acquisition of bidimensional thermal images in acoustic fields produced by ultrasonic transducers used for different therapy. In the seventh article, Luis Rojas Cardenas et al. introduce several functionalities into access point equipments in order to improve the handover performances in the context of mobile communications systems. Spyros Veronikis et al. present in the eighth paper the design methodology of creating hybrid spaces in an academic library, also providing the evaluation method for such a system. In the ninth article, Rangarao Muralishankar and H. N. Shankar investigate the noise robustness of three techniques, i.e, the warped discrete Fourier transform cepstrum (WDFTC), perceptual minimum variance distortionless response (PMVDR), and Mel-frequency cepstral coefficients (MFCC), providing several insightful features in the framework of speech recognition systems. Marc Gilg et al. follow with an evaluation of fairness in Single Star and Double Star network topologies. Last but not least, Jon G. Hall and Lucia Rapanotti present the case of an assurance driven design as opposed to traditional approaches where assurance is considered post development.

We hope that the content of this journal issue will be appealing for the readers, but also a motivation for the researchers in the field to consider the IARIA conferences and journals for publication of their works.

Constantin Paleologu, Editor-in-Chief

International Journal On Advances in Systems and Measurements
Volume 2, Number 1, 2009

CONTENTS

The Sampling Theorem for Finite Duration Signals

Subhendu Das, *CCSI, West Hills, California, subhendu.das@ccsi-ca.com*
Nirode Mohanty, Fellow-IEEE, *CCSI, West Hills, California, nirode.mohanty@ccsi-ca.com*
Avtar Singh, *San Jose State University, San Jose, California, avtar.singh@sjsu.edu*

Abstract

The Shannon's sampling theorem was derived using the assumption that the signals must exist over infinite time interval. But all of our applications are based on finite time intervals. The objective of this research is to correct this inconsistency. In this paper we show where and how this infinite time assumption was used in the derivation of the original sampling theorem and then we extend the results to finite time case. Our research shows that higher sample rate is necessary to recover finite duration signals. This paper validates, with detailed theory, the common industrial practice of higher sample rate. We use the infinite dimensionality property of function space as the basis of our theories. A graphical example illustrates the problem and the solution.

Keywords: Sampling methods, Communication, Linear system, Wavelet transform, Modulation.

1. Objective

This paper is an extended version of [1]. It provides more details of the theories and presents many related ideas including the re-sampling process. The objective of this paper is to extend the original sampling theorem [2] to finite duration signals. It is shown here that the proof of the Shannon's sampling theorem assumed that the signal must exist for infinite time. This assumption came because the proof used Fourier transform theory which in turn uses infinite time. We give a new proof that does not require infinite time assumption and as a result of elimination of this assumption we get a new theory.

Our research shows that more you sample more information you get about the signal when your signal measurement window is finite. We provide some theoretical analysis to justify our results. A very fundamental and well known concept in mathematics, infinite dimensionality of function space, is used as a basis of our research. Thus the main focus of the paper is on sampling theorem and on the number of samples. Since the result establishes a new view in signal processing, we apply the result to few other areas like signal reconstruction and up-down sampling.

In engineering practice most of the applications use two to four times the Nyquist sample rate. In audio engineering much higher rate is used [3]. So the results of this paper are not new ideas in the practical world. However, this engineering practice also points out that there is something wrong somewhere in our theory. There is also this (mis)conception that higher sample rate provides redundant information. Therefore we examine the core issues and assumptions behind the original theory of [2], make some changes, and provide a theoretical proof of the high sample rate concept. It should be noted that the theory in [2] is not wrong, we are only changing one of the assumptions that is more meaningful in the present technology.

Besides sampling theorem, another objective is to highlight the infinite time assumption behind the existing theories. This infinite time assumption is not practical in engineering. Thus we emphasize the infeasibility of the approaches based on transfer function and Fourier transform. All of them use infinite time assumption. In the past many engineers have rejected these approaches because they are useful for only Linear Time Invariant (LTI) systems. Now we have another reason – the infinite time assumption. Interestingly enough, we show that LTI systems do not exist in engineering.

This research leads us to realize that the concept of finite time duration of signals is the backbone of all our engineering systems. Therefore we need to do something about it, i.e., we should start a research in reducing these inconsistencies between the theory and the practice. Eventually, if we can successfully provide a new direction, then our technology will be more predictable and reliable. We may get significant product quality improvements. It may also be possible to reduce waste and thus help to create a greener technology [4].

In this paper our objective is not really to make a big jump in this new research on finite time direction

but occasionally we have touched upon the various related topics, problems, and solutions. We believe that this is an important area of investigation even in mathematics. It should be noted though that all time domain approaches are closer to finite time reality. However unless we create or change some basic engineering definitions all our theories will remain somewhat inconsistent and unsatisfactory.

During the publication process of this research many colleagues and reviewers have made many comments and questions on the subject of this paper. We have tried to include our answers to many of them. As a result, the paper got little bit defocused from its original goal and the contents got diluted. We hope the integration of all these subjects still maintains some coherency and novelty.

The contents of this paper can be described using the following high level summary. We first show, in Section 2, that infinite time assumption is not really needed in engineering. Then we present a new modulation method, in Section 3, and show how we encountered this infinite time issue in a practical engineering problem. To solve the problem over finite time and to provide its theoretical foundation we discuss in details the concept of infinite dimensionality of function space in Section 4. Using this infinite dimensionality concept, in Section 5, we show that finite rate sample representations actually converge to the original function as rate increases to infinity. In Section 6, we provide new proofs of the original sampling theorem and provide a numerical example in Section 7. We also discuss briefly using a numerical example, in Section 8, how approaches based on analytical expressions rather than samples can help to resample a finite duration signal. Finally, in Sections 9 and 10, we discuss the nonlinear nature of engineering systems and explain why time domain approach with high sample rates is more meaningful.

2. Infinite Time

In this section we show that the assumption of infinite time duration for signals is not practical and is not necessary for our theories. In real life and in all our engineering systems we use signals of finite time durations only. Intuitively this finite duration concept may not be quite obvious though. Ordinarily we know that all our engineering systems run continuously for days, months, and years. Traffic light signaling systems, GPS satellite transmitters, long distance air flights etc. are some common examples of systems of infinite time durations. Then why do we talk about finite duration signals? The confusions will be cleared when we think little bit and examine the internal design principles, the architecture of our technology, and the theory behind our algorithms. Originally we never thought that this question will be asked, but it was, and therefore we look here, at the implementations, for an explanation.

The computer based embedded engineering applications run under basically two kinds of operating systems (OS). One of these OS uses periodic approaches. In these systems the OS has only one interrupt that is produced at a fixed rate by a timer counter. Here the same application runs periodically, at the rate of this interrupt, and executes a fixed algorithm over and over again on input signals of fixed and finite time duration. As an example, in digital communication engineering, these signals are usually the symbols of same fixed duration representing the digital data and the algorithm is the bit recovery process. Every time a symbol comes, the algorithm recovers the bits from the symbol and then goes back to process the next arriving symbol.

Many core devices of an airplane, carrying passengers, are called flight critical systems. Similarly there are life critical systems, like pacemaker implanted inside human body. It is a very strict requirement that all flight critical and life critical systems have only one interrupt. This requirement is mainly used to keep the software simple and very deterministic. They all, as explained before, repeat the same periodic process of finite duration, but run practically for infinite time.

The other kind of applications is based on the real time multi-tasking operating systems (RTOS). This OS is required for systems with more than one interrupts which normally appear at asynchronous and non-periodic rate. When you have more than one interrupts, you need to decide which one to process first. This leads to the concept of priority or assignment of some kind of importance to each interrupt and an algorithm to select them. The software that does this work is nothing but the RTOS. Thus RTOS is essentially an efficient interrupt handling algorithm.

These RTOS based embedded applications are designed as a finite state machine. We are not going to present a theory of RTOS here. So to avoid confusions we do not try to distinguish among threads, tasks, processes, and states etc. We refer to all of these concepts as tasks, that is, we ignore all details below the level of tasks, in this paper. These tasks are executed according to the arrival of interrupts and the design of the application software. The total application algorithm is still fixed and finite but the work load is distributed among these finite numbers of tasks. The execution time of each task is finite also. These tasks process the incoming signals of finite time and produce the required output of finite size.

An example will illustrate it better. A digital communication receiver can be designed to have many tasks – signal processing task, bit recovery task, error correcting task etc. They can be interconnected by data buffers, operating system calls, and application functions. All these tasks together, implement a finite state machine, execute a finite duration algorithm, and process a finite size data buffer. These data buffers are originated from the samples of the finite duration signals representing the symbols.

We should point out that there are systems which are combinations or variants of these two basic concepts. Most commercial RTOS provide many or all of these capabilities. Thus although all of the engineering systems run continuously for all time, all of them are run under the above two basic OS environment. Or in other words for all practical engineering designs the signal availability windows, the measurement windows, and the processing windows are all of finite time. For more details of real time embedded system design principles see many standard text books, for example [5, pp73-88].

The signals may exist theoretically or mathematically for infinite time but in this paper none of our theories, derivations, and assumptions will use that infinite time interval assumption.

In the next section we describe the concept of a new digital communication scheme [6][7] to demonstrate the need for high sample rate. This scheme will also give the details of how finite time analysis can be used in our engineering systems.

3. Motivation

Almost all existing communication systems use sinusoidal functions as symbols for carrying digital data. But a sine function has only three parameters, amplitude, frequency, and phase. Therefore you can only transmit at most three parameters per symbol interval. That is a very inefficient use of symbol time. If instead we use general purpose functions then we can carry very large amount of information, thus significantly increasing the information content per symbol time. However, as we show below, these general purpose functions will require a large number of samples over its symbol time, and hence a high sample rate, to represent them precisely. We present a new digital communication system, called function modulation (fm) [6], to introduce the application of non-sinusoidal functions and the need for a new sampling theorem.

Figure 1 shows an fm transmitter. The left hand side (LHS) vertical box shows four bits, as example, that will be transmitted using one symbol, s(t), shown in the right hand side (RHS) graph. Each bit location in the LHS box is represented by a graph or a general function. These functions, called bit functions, are combined by an algorithm to produce the RHS graph or function. A very simple example of the algorithm may be to add all the bit functions for which the bit values are ones and ignore whose bit values are zeroes. We call this algorithm a 0-1 addition algorithm. Since the bits in the LHS vertical box are continuously changing after every symbol time, the symbol s(t), $t \in [0,T]$, is also continuously changing.

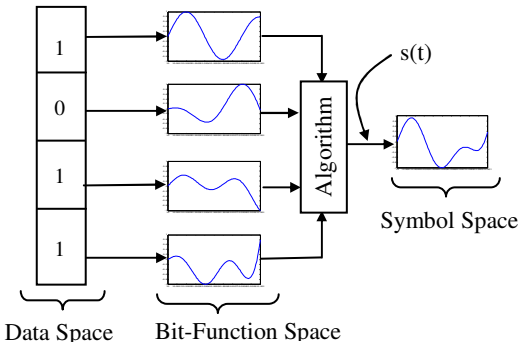


Figure 1. fm Transmitter

For this 0-1 addition algorithm we can write:

$$s(t) = d_1 g_1(t) + d_2 g_2(t) + ..+d_M g_M(t) \quad (1)$$

where $d_i \in \{0,1\}$ are the bit values. If we select $\{g_i(t), t\epsilon[0,T], i=1\ldots M\}$ as a set of independent bit functions then we will be able to recover the bits if we know s(t). Here M is the number of bits to be transmitted using one symbol. This process of recovery of $\{d_i\}$ from s(t) will require very precise knowledge of s(t). That can be achieved only by providing large number of samples for s(t) and for each member of $\{g_i(t)\}$. Note that in (1) s(t), $\{g_i(t)\}$, and $\{d_i\}$are all known quantities. In a later section we highlight the similarity of expression (1) with Fourier series and its consequences.

The functions used in fm are not defined over infinite time interval; they are defined only over the symbol time, which are usually very small, of the order of microseconds or milliseconds, and should not be considered as infinite time intervals. The Nyquist rate will provide very few samples on these small intervals and will not enable us to reconstruct them correctly. We use these general classes of functions to represent digital data, because they have higher capacity to represent information compared to simple sine wave functions. Modern Digital Signal Processors (DSP) are ideally suited to handle them also. The DSP technology, high speed and high resolution Analog to Digital Converters (ADC), along with the analytical functions are quite capable of handling powerful

design methods, which cannot be implemented using hardware based concepts like voltage controlled oscillators or phase lock loops etc.

The fm receivers are more complex than the fm transmitters. Mainly because the objective of the fm receiver is to decompose the received functions into the component bit functions that were used to create the symbol at the transmitter. The decomposition process is usually more complex than the composition process. However, if the bit functions are orthogonal then the decomposition process is very trivial [7]. Figure 2 shows the block diagram of a fm receiver based on orthogonal bit functions. Note however that the transmitter design is the same whether we use orthogonal or non-orthogonal bit functions.

In Figure 2, the functions $\{g_i(t)\}$are assumed to be orthogonal bit functions. They are M in number, the

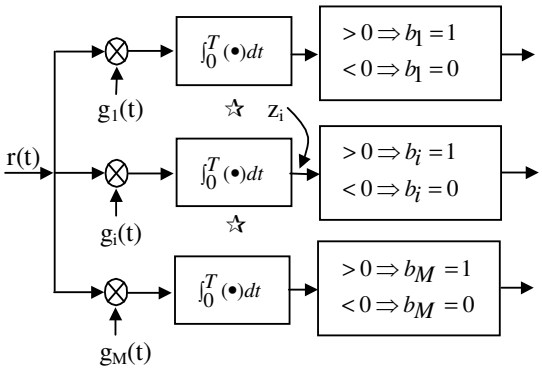


Figure 2. fm Receiver for orthogonal functions

number of bits to be transmitted using one symbol. The received symbol r(t) is generated by the corresponding fm transmitter using 0-1 addition algorithm as shown in (1). We can write the symbol r(t) using the following relation:

$$r(t) = g_1(t)x_1 + g_2(t)x_2 + .. + g_M(t)x_M + w(t) \quad (2)$$

Where the set $\{x_i\}$ in (2) represents the unknown bit values and w(t) is the additive white Gaussian noise term. If the functions in $\{g_i(t)\}$ are orthogonal then we can get the estimate for $\{x_i\}$ using the following integration (3). Here w_i is the projection of the noise on the i-th orthogonal function.

$$x_i = \int_0^T r(t)g_i(t)dt + w_i \quad (3)$$

The integration process (3) is shown in the Figure 2, along with the thresholds for detecting the bits.

Unlike similar figures in communication text books, here all parallel paths produce bit data. Therefore all integrations in all paths must be very precisely performed. This process will also require large volume of samples as well as powerful numerical integration methods, preferably based on analytical approaches. The receiver for non-orthogonal functions [6] is more complex and also demands large sample rate. Later we point out that the fm method is essentially an implementation of the concept behind the finite term Fourier series.

The fm design provides a method for using general purpose functions for digital data communication. General functions can carry more information than sinusoidal functions. We highlight in many different ways the well known fact that any general continuous function defined over any finite time interval has infinite dimension and therefore can carry infinite amount of information. Intuitively this concept and its consequences in communication engineering may not be very clear, so we describe it in many details beginning with the following section. We also show that to extract this infinite information content we have to sample the functions, theoretically, at infinite rate. Thus the motivation for this research is to establish the theoretical foundation for the function modulation method. The engineering foundation of the fm method over real time voice band telephone line has already been demonstrated and presented [6].

4. Infinite Dimensionality

We will use the following basic notations and definitions in our paper. Consider the space $L_2[a,b]$ of all real valued measurable functions defined over the finite, closed, and real interval [a,b]. We assume that the following Lebesgue integral is bounded:

$$\int_a^b |f(t)|^2\, dt < \infty\,, \quad \forall f \in L_2[a,b] \quad (4)$$

And then we define the norm:

$$\|f\| = \left[\int_a^b |f(t)|^2\, dt\right]^{1/2}, \quad \forall f \in L_2[a,b] \quad (5)$$

Measurable functions form an equivalence class, in the sense that each function in this class has the same integral value. Two such functions in the same equivalent class differ on some countable discrete set whose measure is zero thus without affecting the integral value. We can always find a continuous function that can represent this equivalent class [8, pp418-427] in the sense of $L_2[a,b]$ norm. Thus for all engineering purposes we can think about continuous functions only [9, pp27-28].

The space $L_2[a,b]$ is a complete space. This completeness property ensures that every convergent sequence $\{f_n\}$ converges to a function f that belongs to $L_2[a,b]$ space. That is, $L_2[a,b]$ includes all the limit

points. The norm (5) is used to prove the convergence because it embeds the concept of distance between the elements of a sequence.

We also define the inner product for L_2[a,b] as:

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt \quad \forall f, g \in L_2[a,b] \tag{6}$$

For L_2[a,b] the integral in (6) exists and therefore the inner product is well defined. In a finite dimensional vector space the inner product of two vectors u = $\{u_i\}$ and v = $\{v_i\}$ is defined by

$$u'v = u_1v_1 + u_2v_2 + \cdots + u_nv_n$$

Which is similar to (6) when you take the limit of the approximation given by the following:

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt$$
$$\approx \Delta t[f(t_1)g(t_1) + f(t_2)g(t_2) + \cdots + f(t_n)g(t_n)]$$

Thus a function space is very intimately linked with the concept of finite dimensional linear vector space when we look at it as nothing but a collection of infinite samples.

Under the above conditions the function space, L_2[a,b], is a Hilbert space. Hilbert space is defined as a complete inner product space. The inner product comes from the definition (6) and the completeness from the norm (5). The inner product helps to introduce the concept of orthogonality in the function space. We also define the distance between two functions in L_2[a,b] space by:

$$d(f, g) = \|f - g\| = \left[\int_a^b |f(t) - g(t)|^2\, dt\right]^{1/2} \tag{7}$$

The metric d in (7) defines the mean square distance between any two functions in L_2[a,b].

One very important property of the Hilbert space [9, pp31-32] related to the communication theory, is that it contains a countable set of orthonormal basis functions. Let $\{\varphi_n(t), n = 1.2..., t \in [a,b]\}$ be such a set of basis functions. Then the following is true:

$$\langle \varphi_m, \varphi_n\rangle = \delta_{mn} = \begin{cases} 0 & if\ m \neq n \\ 1 & if\ m = n \end{cases} \tag{8}$$

Also for any f ϵ L_2[a,b] we have the Fourier series

$$f(t) = \sum_{n=1}^{\infty} a_n\varphi_n(t), \quad \forall t \in [a,b] \tag{9}$$

The above expression (9) really means that for any given ε > 0 there exists an N such that

$$\left\|f(t) - \sum_{n=1}^{k} a_n\varphi_n(t)\right\| < \varepsilon, \quad \forall k > N \tag{10}$$

In this context we should also mention that the coefficients in (9) can be obtained using the following expression:

$$a_n = \int_a^b f(t)\varphi_n(t)dt\,, \quad n = 1,2, \ldots \tag{11}$$

In this paper we will consider only the continuous functions and their Riemann integrability. Riemann integration is the normal integration process we use in our basic calculus. We note that the continuous functions are measurable functions and the Riemann integrable functions are also Lebesgue integrable. For continuous functions the values for these two integrals are also same. Thus the Hilbert space theory (4-11) and the associated concepts will still remain applicable to our problems. We should point out though that the space of continuous functions is not complete for the L_2[a,b] norm defined by (5). That means, there exists a sequence of continuous functions that does not converge to a continuous function under the L_2[a,b] norm. However it will converge to a measurable function under L_2 norm, that is, in the mean.

Equality (9) happens only for infinite number of terms. Otherwise, the Fourier representation in (10) is only approximate for any finite number of terms. In this paper ε in (10) will be called as the measure of approximation or accuracy estimate in representing a continuous function. The Hilbert space theory ensures the existence of N in (10) for a given ε. The existence of such a countably infinite number of orthonormal basis functions (8) proves that the function space is an infinite dimensional vector space. This dimensionality does not depend on the length of the interval [a,b]. Even for a very small interval, like symbol time, or an infinite interval, a function is always an infinite dimensional vector. The components of this vector are the coefficients of (9).

Hilbert space theory shows that a function can be represented by equation (9). The coefficients in (9) carry the information about a function. Since there are infinite numbers of coefficients, a function carries an infinite amount of information. Our digital communication theory will be significantly richer if we can use even a very small portion of this infinite information content of a function. The function modulation approach provides a frame work for such a system. The fm scheme essentially implements equation (9), for fm transmitter, where the coefficients used are zero or one instead of any real number. Similarly Figure 2, the fm receiver for orthogonal functions, implements expression (11). For an fm receiver (11) will produce zero or one as the values for $\{a_n\}$. It is clear that if we can find a band limited set of orthogonal functions then equations (9) and (11) will

allow us to create a fm system with almost unlimited capacity [4].

It is not necessary to have orthonormal basis functions for demonstrating that the function space is infinite dimensional. The collection of all polynomial functions $\{t^n, n=1,2,..\}$is linearly independent over the interval [a,b] and their number is also countable infinity. These polynomials can be used to represent any analytic function, i.e., a function that has all derivatives. Using Taylor's series we can express such a f(t) at t as:

$$f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!} (t-c)^n \quad (12)$$

around the neighborhood of any point c. Thus the above polynomial set is also a basis set for the function space. Therefore using the infinite Taylor series expression (12), we prove that a function is an infinite dimensional vector over a finite interval. Here the information content of the function is defined by the derivative coefficients and the polynomial functions. Expression (12) also shows that the information content of a general function is infinity.

The above two theories prove that the dimension of the function space is infinity. The number of such functions in this function space is also infinity, actually uncountable infinity. This is illustrated using the following logic. Consider any coefficient in the right hand side of (9). You will get a new function every time you change that coefficient. Since that coefficient can be adjusted to any value in the interval [0,1] you get a continuum of functions. Thus the cardinality of the function space is uncountable infinity whereas the dimensionality is countable infinity.

We say that to represent a function accurately over any interval we need two sets of information: (A) An infinite set of basis functions, not necessarily orthogonal, like the ones given by (8) and (B) An infinite set of coefficients in the infinite series expression for the function, similar to (9). That is, these two sets completely define the information content in a mathematical function. In most cases the basis set described in (A) will remain fixed. We will distinguish functions only by their coefficients described in (B). Each function will have different coefficients in its expression for (9).

We normally represent vectors as rows or columns with components as real numbers. As an example, a three dimensional vector has three real components. Similarly an n dimensional vector has n real components. Along that line an infinite dimensional vector will have infinite number components. We can represent a function by an infinite dimensional vector by selecting the coefficients of (9) as components of this vector. In this sense every function is an infinite dimensional vector. We will show later that the samples of a function can also be used to represent these components of an infinite dimensional vector. Thus these samples will bring this mathematics to engineering, because the ADCs can produce these samples.

We now show that a band limited function is also infinite dimensional and therefore carries infinite amount of information. Consider a band limited function f(t), with bandwidth [-W,+W]. Then f(t) is given by the following inverse Fourier Transform [2]:

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(w)e^{iwt}dw$$

$$= \frac{1}{2\pi}\int_{-2\pi W}^{+2\pi W} F(w)e^{iwt}dw \quad (13)$$

In (13) t is defined for all time in (–∞,+∞). But the frequency w is defined only over [–W,+W], and it can take any value: integer, rational, or irrational frequencies, within that range.

The second line in expression (13) shows that the band limited function f(t) has uncountably infinite number of frequencies. That is, f(t) is created using infinite number of frequencies and therefore is an infinite dimensional vector. This is true even when we consider a small interval of time for the function f(t). In that small interval the function still has all the infinite frequency components corresponding to the points in [–W,+W]. This is another way of showing that a band limited function is an infinite dimensional vector over a finite measurement window.

We have been talking about countable and uncountable infinities. To refresh our memory, countable infinity is the number of elements in the set of integers {1,2,…} and the uncountable infinity is the number of points in the interval [0,1]. The set of rational numbers is also countable. Clearly uncountable infinity is larger than the countable infinity. However, one interesting fact is that any real number can be represented as a limit of a sequence of rational numbers. This fact is mathematically stated as the set of rational numbers is a dense set in the set of real numbers [8, pp43-45]. Therefore when we talk about uncountable infinity we can in many cases think in terms of countable infinity also. The relationship between measurable functions and continuous functions are similar as mentioned before.

We point out here that a constant function f(t) = C, as an element of function space, is also an infinite dimensional vector. The only difference is that all sample values are same. In terms of Taylor series the coefficients for a constant function are {C,0,0,….}, which is an infinite dimensional vector.

The infinite dimensionality idea of a function can be understood in another very interesting way.

Consider the real line interval [0,1]. We know that it has uncountably infinite number of points. If we stretch this line to [0,2] we will still have all these uncountable number of points inside it. Now if we bend it and twist it all the points will still be there but the line will now become a function, a graph, in the two dimensional plane. Thus a function has uncountably infinite number of points. Every sample you take will have different coordinates and therefore different information. Therefore we can prove that a function can be exactly represented by this infinite number of samples and that more samples you take over this finite interval better will be the representation of the function.

The definition of dimension should be clearly pointed out. The dimension of a vector space is the number of basis vectors of the space. For function space, both Fourier series (9) and the Taylor series (12) show, that the number of basis vectors is countable infinity. Therefore the dimension of the function space is countable infinity. Any element of this vector space will also have the same number of components in its representation as a vector. Therefore the total number of components in a vector is also called the dimension of the vector.

We now show that samples can also be used to represent this infinite dimensionality. We prove that it is theoretically necessary to sample a function that is defined over finite time interval, infinite number of times, to extract all the information from the function.

5. Sample Convergence

Let f(t) be a continuous function defined over the real time interval [a,b]. Assume that we divide this finite time interval [a,b] into n > 1 equal parts using equally spaced points $\{t_1, t_2, .. t_n, t_{n+1}\}$. Where $t_1 = a$ and $t_{n+1} = b$. Use the following notations to represent the t-subintervals

$$\Delta t_i = \begin{cases} [t_i, t_{i+1}), & i=1,2,..,n-1 \\ [t_n, t_{n+1}], & i=n \end{cases}$$

Define the characteristic functions:

$$X_i(t) = \begin{cases} 1, & t \in \Delta t_i \\ 0, & t \notin \Delta t_i \end{cases} \quad i = 1,2,..,n \tag{14}$$

In this case the characteristic functions, $X_i(t)$ are orthonormal over the interval [a,b] with respect to the inner product on L_2 [a,b], because

$$X_i(t)X_j(t) = 0, \; i \neq j, \; \forall t \in [a,b] \tag{15}$$

Also define the simple functions as:

$$f_n(t) = \sum_{i=1}^{n} f(t_i)X_i(t) \quad \forall t \in [a,b] \tag{16}$$

Here $f(t_i)$ is the sampled value of the function f(t) at time $t = t_i$ that is, at the beginning of each sample interval Δt. It is easy to visualize that $f_n(t)$ is a sequence of discrete step functions over n. Expression (16) is an approximate Fourier series representation of f(t) over [a,b]. This representation uses the samples of the function f(t) at equal intervals, $f_n(t)$ uses n number of samples.

We show that this approximate representation (16) improves and approaches f(t) as we increase the value of n. Which will essentially prove that more you sample more information you get about the function.

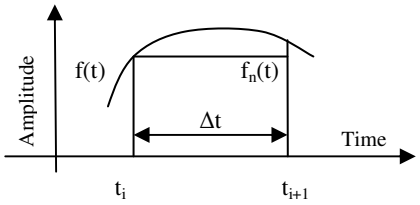


Figure 3. Simple function approximation

Thus the higher sample rate is meaningful and does not produce any redundant information. The following theorem is quite intuitive; its proof is also very simple. However, its consequence is very profound in the field of digital signal processing and in communication engineering.

Theorem 1: $f_n \rightarrow f$ in $L_2[a,b]$ as $n \rightarrow \infty$.

First consider Figure 3, where we show the simple function $f_n(t)$ and the original continuous function f(t), between two consecutive sample points on the time line. It is geometrically obvious that the maximum error between the two functions reduces as the interval Δt reduces. Mathematically, consider the error

$$\Delta y_n = \max_t |f(t) - f_n(t)|, \; \forall t \in [a,b] \tag{17}$$

It is then clear that $\{\Delta y_n\}$ is a monotonically decreasing sequence of n since the function f(t) is continuous over the closed interval [a,b]. Therefore, given any ε>0 we can find an N such that $\Delta y_n \leq \varepsilon/\sqrt{(b-a)}$ for all n > N. We can also verify that

$$\|f - f_n\| = \left[\int_a^b |f(t) - f_n(t)|^2 dt\right]^{1/2}$$

$$= \left[\int_a^b |f(t) - \sum_{i=1}^{n} f(t_i)X_i(t)|^2 \, dt\right]^{1/2} \tag{18}$$

Since $X_i(t) = 1$ for all t we can rewrite the above expression without affecting the integral as

$$= \left[\int_a^b |\sum_{i=1}^{n} f(t)X_i(t) - \sum_{i=1}^{n} f(t_i)X_i(t)|^2 \, dt\right]^{1/2}$$

Rearranging the terms we can write

$$= \left[\int_a^b \left[\sum_{i=1}^n \big(f(t) - f(t_i)\big) X_i(t)\right]^2 dt\right]^{1/2}$$

Now performing the squaring operation, noting that (15) holds, we can write the above as:

$$= \left[\int_a^b [\sum_{i=1}^n [f(t) - f(t_i)]^2 X_i^2(t)] dt\right]^{1/2}$$

$$\leq \left[\int_a^b [\sum_{i=1}^n [\Delta y_n]^2 X_i^2(t)] dt\right]^{1/2}$$

$$= \left[\Delta y_n^2 \int_a^b (\sum_{i=1}^n X_i^2(t)) dt\right]^{1/2}$$

$$= [\Delta y_n^2 (b-a)]^{1/2} = \sqrt{(b-a)}\Delta y_n \leq \varepsilon$$

Thus from (18) we see that $\forall n \geq N$

$$\|f(t) - \sum_{i=1}^n f(t_i) X_i(t)\| \leq \epsilon, \forall t \in [a,b] \qquad (19)$$

Which means:

$$f(t) = \sum_{i=1}^{\infty} f(t_i) X_i(t)\ ,\ \forall t \in [a,b] \qquad (20)$$

This concludes the proof of Theorem 1.

The expression (20) says that a function is an infinite dimensional vector and can be correctly represented by all the infinite samples, while the expression (19) can be used for approximate representation with accuracy given by ε. Essentially Theorem 1 proves that infinite sample rate is necessary to represent a continuous function correctly over a finite time interval. Another important interpretation of Theorem 1 is that the information content of a function is available in the samples of the function. Thus the amount of information these general purpose functions can carry is actually infinity. A communication system can be designed to extract a large amount of information from this infinite content. The fm system uses such a general class of function and can be used to carry more information than conventional designs.

It is clear that the Theorem 1 does not depend on the bandwidth of the function f(t). However, for any given ε>0 the number N will depend on the bandwidth.

Theorem 1 is similar to the one described for measurable functions in [8, pp389-391]. But the coefficients are not the sampled values in that theorem. For measurable functions, samples are usually taken on the y-axis. Another proof can be found in [10, pp247-257] where the Bernstein polynomial has been used instead of the characteristic function. Although Bernstein polynomial functions are not orthogonal, like the characteristic functions used in Theorem 1, but they are defined over finite and closed interval, occasionally know as functions with compact support. We will see that it has an important consequence when we reconstruct the function using the samples.

Theorem 1 shows that the approximating functions (16) converge in the mean, because we have used the L_2 norm. In engineering we normally like pointwise, that is point by point convergence. We say that a sequence of functions, like $\{f_n\}$ in (16), converges uniformly to a function f on the closed interval [a,b] if for every ε > 0 there exists an N > 0, depending only on ε and not on t in the interval [a,b], such that $|f_n(t) - f(t)| < \varepsilon$ for all n >N, and for all t in [a,b]. It should be pointed out that uniform convergence implies pointwise convergence.

Since the function f(t) is continuous over a closed interval, the sequence (16) is bounded, because the sample values are bounded. Therefore if we show that the supremum or the maximum of (16) converges then obviously the sequence will converge uniformly also [8, pp308-311]. That is we have to show that

$$\sup_{a \leq t \leq b} |f(t) - f_n(t)| < \varepsilon\ ,\ \forall n > N \qquad (21)$$

Incidentally, the supremum and the maximum are the same thing for continuous functions over closed interval and also the difference really does not matter for engineering problems. Over every small interval Δt_i we can write

$$|f(t) - f_n(t)| = |f(t)X_i(t) - f(t_i)X_i(t)|$$

$$= |f(t) - f(t_i)| X_i(t)$$

According to the mean value theorem, there exists a c_i within every interval Δt_i such that

$$f(t) = f(t_i) + (t - t_i) f'(c_i)$$

Therefore using (16) we can write

$$|f(t) - f_n(t)| = \sum_{i=1}^n |f(t) - f(t_i)| X_i(t) \qquad (22)$$

$$= \sum_{i=1}^n |(t - t_i) f'(c_i)| X_i(t)$$

$$= \sum_{i=1}^n |t - t_i| |f'(c_i)|\, X_i(t)$$

$$\leq \Delta t\ M\ \sum_{i=1}^n X_i(t) = \frac{b-a}{n} M \qquad (23)$$

Where, M is the upper bound of the derivative of f(t) on the entire interval [a,b]. Since f(t) is a continuous function over a closed interval, M always exists. So in the above proof we have assumed that the function is differentiable. Thus the right hand side of (23) is independent of t and therefore is an uniform bound for all t for the left hand side of (22). Thus we can see that the difference expressed in the left side of (22) goes to

zero as n goes to infinity. This shows that (20) is true for all t, thus proving the uniform convergence. The above derivation also proves the intuitive assertion made in (17).

We have shown that the approximations generated by sample values of the original function f(t) converges to the original function. As it converges the number of samples increases, because that is the way we constructed the approximating function $f_n(t)$. Since the approximations improve, the $f_n(t)$ improves, confirming that more samples are better and does not generate redundant information. These infinite samples collectively define the function. A complete description of the function can only be obtained by these infinite numbers of samples, which can be considered as components of an infinite dimensional vector. The above basic tools now can help us to give theoretical proofs of sampling theorem.

6. Sampling Theorem

Consider the simple sinusoidal function

$$s(t) = A \sin (2 \pi f t + \theta) \quad (24)$$

and assume that it is defined only for one period for simplicity, although not necessary. We can think of this sinusoidal function as the highest frequency component of a band limited signal. So if we can recover this function by sampling it then we will be able to recover the entire original signal, because the other components of the band limited signal are changing slowly at lower frequencies. We try now to determine how many samples we need to recover the sine function.

We can see from the above expression (24) that a sinusoidal function can be completely specified by three parameters A, f, and θ. That is we can express a sine function as a three dimensional vector:

$$s = [A, f, \theta] \quad (25)$$

However (25) is very misleading. There is a major hidden assumption; that the parameters of (25) are related by the sine function. Therefore more precise representation of (25) should be:

$$s = [A, f, \theta, \text{“sine”}] \quad (26)$$

The word sine in (26) means the Taylor's series, which has an infinite number of coefficients. Therefore when we say (25) we really mean (26) and that the sine function, as usual, is really an infinite dimensional vector.

We can use the following three equations to solve for the three unknown parameters, A, f, and θ of a sinusoidal function:

$$\begin{aligned} s_1 &= A \sin (2 \pi f t_1 + \theta) \\ s_2 &= A \sin (2 \pi f t_2 + \theta) \\ s_3 &= A \sin (2 \pi f t_3 + \theta) \end{aligned} \quad (27)$$

where t_1 , t_2 , t_3 are sample times and s_1 , s_2 , s_3 are corresponding sample values. Again a correct representation in terms of samples would be

$$s = [(s_1,t_1), (s_2,t_2), (s_3,t_3), \text{“sine”}]$$

Hence with sinusoidal assumption, a sine function can be completely specified by three samples. The above analysis gives a simple proof of the sampling theorem. We can now state the well known result:

Theorem 2: A sinusoidal function, with the assumption of sinusoidality, can be completely specified by three non-zero samples of the function taken at any three points in its period.

From (27) we see that if we assume sinusoidality then more than three samples, or higher than Nyquist rate, will give redundant information. However without sinsoidality assumptions more samples we take more information we get, as is done in common engineering practice. It should be pointed out that Shannon's sampling theorem assumes sinusoidality. Because it is derived using the concept of bandwidth, which is defined using Fourier series or transform, and which in turn uses sinusoidal functions.

Theorem 2 says that the sampling theorem should be stated as $f_s > 2f_m$ instead of $f_s \geq 2f_m$ that is, the equality should be replaced by strict inequality. Here, f_m is the signal bandwidth, and f_s is the sampling frequency. There are some engineering books [11, p63] that mention strict inequality.

Shannon states his sampling theorem [2, p448] in the following way: “If a function f(t) contains no frequencies higher than W cps, it is completely determined by giving its ordinates at a series of points spaced 1/2 W seconds apart”. The proof in [2] is very simple and runs along the following lines. See also [12, p271]. A band limited function f(t) can be written as in (13). Substituting t = n/(2W) in (13) we get the following expression:

$$f\left(\frac{n}{2W}\right) = \frac{1}{2\pi}\int_{-2\pi W}^{+2\pi W} F(w) e^{iw\frac{n}{2W}}\, dw \quad (28)$$

Then the paper [2] makes the following comments: “On the left are the values of f(t) at the sampling points. The integral on the right will be recognized as essentially the nth coefficient in a Fourier-series expansion of the function F(w), taking the interval –W to +W as a fundamental period. This means that the values of the samples f(n/2W) determine the Fourier coefficients in the series expansion of F(W).” It then

continues "Thus they determine F(w), since F(w) is zero for frequencies greater than W, and for lower frequencies F(w) is determined if its Fourier coefficients are determined".

Thus the idea behind Shannon's proof is that from the samples of f(t) we reconstruct the unknown Fourier transform F(w) using (28). Then from this known F(w) we can find f(t) using (13) for all time t. One important feature of the above proof is that it requires that the function needs to exist for infinite time, because only then you get all the infinite samples from (28). We show that his proof can be extended to reconstruct functions over any finite interval with any degree of accuracy by increasing the sample rate. The idea behind the proof is similar, we construct F(w) from the samples of f(t).

In this proof we use the principles behind the numerical inversion of Laplace transform method as described in [13, p359]. Let F(w) be the unknown band limited Fourier transform, defined over [-W,+W]. Let the measurement window for the function f(t) be [0,T], where T is finite and not necessarily a large number. Divide the frequency interval 2W into K smaller equal sub-intervals of width Δw with equally spaced points $\{w_j\}$ and assume that the set of samples $\{F(w_j)\}$ is constant but unknown over that j-th interval. Then we can express the integration in (13) approximately as:

$$f(t) \approx \frac{1}{2\pi}(\Delta w)\sum_{j=1}^{K} e^{itw_j} F(w_j) \qquad (29)$$

The right hand side of (29) is a linear equation in $\{F(w_j)\}$, which is unknown. Now we can also divide the interval [0,T] into K equal parts with equally spaced points $\{t_j\}$ and let the corresponding known sample values be $\{f(t_j)\}$. Then if we repeat the expression (29) for each sample point t_j we get K simultaneous equations in the K unknown variables $\{F(w_j)\}$as shown below:

$$\begin{bmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_K) \end{bmatrix} = \frac{\Delta w}{2\pi} \begin{bmatrix} e^{it_1w_1} & e^{it_1w_2} \dots e^{it_1w_K} \\ e^{it_2w_1} & e^{it_2w_2} \dots e^{it_2w_K} \\ & \vdots \\ e^{it_Kw_1} & e^{it_Kw_2} \dots e^{it_Kw_K} \end{bmatrix} \begin{bmatrix} F(w_1) \\ F(w_2) \\ \vdots \\ F(w_K) \end{bmatrix} \qquad (30)$$

These equations are independent because the exponential functions in (29) are independent.

We recall that a set of functions $G(t) = \{g_i(t), i = 1 \dots M, t \in [0,T]\}$ is called dependent over the interval if there exists constants c_i, not all zero, such that

$$g_1(t)c_1 + g_2(t)c_2 + \cdots + g_M(t)c_M = 0$$

for all t in [0,T]. If not, then it is independent [14, pp177-181]. The above expression is a linear combination of functions. Here the coefficients $\{c_i,\ i = 1 \dots M\}$ are all real numbers. It essentially says that one function cannot be constructed using other functions.

Therefore we can solve (30) for $\{F(w_j)\}$. Theorem 1 ensures that the sets $\{F(w_j)\}$ and $\{f(t_j)\}$ can be selected to achieve any level of accuracy requirements in (13) for either f(t) or F(w). For convenience we assume that the number of terms K in (29) is equal to $Tkf_s = 2kWT$. Here f_s is the Nyquist sample rate and $k > 1$. We state the following new sampling theorem.

Theorem 3: Let f(t) be a band limited function with bandwidth restricted to [-W,+W] and be available over the finite measurement window [0,T]. Then given any accuracy estimate $\varepsilon > 0$, there exists a constant $k>1$ such that 2kWT equally spaced samples of f(t) over [0,T] will completely specify the Fourier transform F(w) of f(t) with the given accuracy ε. This F(w) can then be used to find f(t) for all time t.

Note that we did not say that the function does not exist over the entire real line. We only said that our measurement window is finite. What happens to the function beyond the finite interval is not needed for our analysis. The main point of our paper is that we do not need to be concerned with the existence of our signals over the entire real line.

We have given, as in (30), a very general numerical method of solving an integral equation. The method can be applied also to the case when f(t) in the left hand side of the equation (13) is unknown. The equation (13) itself can be generalized too. A well known [15] generalization is given below:

$$f(t) = \int_a^b K(t,w)F(w)dw \qquad (31)$$

Here we have replaced the sinusoidal function by the kernel function K(t,w). Expression (31) represents a relationship between frequency components with the time functions for finite duration signals. In that sense it gives the bandwidth information of any given function f(t). It may even be possible to solve the equation (31) analytically over finite time and frequency ranges for some specific kernel functions. More we research in this very practical finite duration engineering problem better will be our definitions and theories and they will be closer to reality. We point out here that there is a need for extending the definition of bandwidth of a function from infinite time to finite time.

In a sense Shannon's sampling theorem gives a sufficient condition. That is, if we sample at twice the bandwidth rate and collect all the infinite number of samples then we can recover the function. We point out that this is not a necessary condition. That is, his theorem does not say that if T is finite then we cannot recover the function accurately by sampling it. We have confirmed this idea in the above proof of Theorem 3. That is if T is finite we have to sample at infinite rate to get all the infinite number of samples. Or in other words more we sample more information we get. This is because a function is an infinite dimensional vector and therefore it can be correctly specified only if we get all the infinite number of samples.

Shannon proves his sampling theorem in another way [2]. Any continuous function can be expressed using the Hilbert space based Fourier expression (9). He has used the expression (9) for a band limited function f(t), defined over infinite time interval. He has shown that if we use

$$\varphi_n(t) = \frac{sin\{\pi f_s[t-(n/f_s)]\}}{\pi f_s[t-(n/f_s)]} \quad (32)$$

Then the coefficients of (9) can be written as

$$a_n = f(n/f_s) \quad (33)$$

Thus the function f(t) can be expressed as:

$$f(t) = \sum_{n=-\infty}^{\infty} f(n/f_s) \frac{sin\{\pi f_s[t-(n/f_s)]\}}{\pi f_s[t-(n/f_s)]} \quad (34)$$

Here $f_s = 2W$, where W is the bandwidth of the function f(t). The set $\{\varphi_n\}$ in (32) is orthogonal only over $(-\infty,+\infty)$.

Observe that the above is very similar to our proof of theorem 1. Shannon used sinc functions as the orthogonal basis functions, whereas in our theorem 1 we used rectangular pulses as the orthogonal basis functions. We know that the sinc function is the Fourier transform of the rectangular pulse. Only difference is that the sinc functions require infinite time interval.

We make the following observations about (34):

1. The representation (34) is exact only when infinite time interval and infinite terms are considered.
2. If we truncate to finite time interval then the functions φ_n in (32) will no longer be orthogonal, and therefore will not form a basis set, and consequently will not be able to represent the function f(t) correctly.
3. If in addition we consider only finite number of terms of the series in (34) then more errors will be created because we are not considering all the basis functions. We will only be considering a subspace of the entire function space.

We prove again, that by increasing the sample rate we can get any desired approximation of f(t), over any finite time interval [0,T], using the same sinc functions of (32). From calculus we know that the following limit holds:

$$\lim_{x\to\infty} \frac{\sin x}{x} = 0 \quad (35)$$

Assume that f_s is the Nyquist sampling frequency, i.e., $f_s = 2W$. Let us sample the signal at k times the Nyquist rate. Here k>1 is any real number. Then using (35), we can show that given any T and a small $\delta > 0$, there exists an N such that

$$\left|\frac{sin(\pi k f_s t)}{\pi k f_s t}\right| < \delta, \forall k > N, \forall t \geq T \quad (36)$$

Thus these orthogonal functions (32) substantially go to zero outside any finite interval [0,T] for large enough sampling rate and still maintain their orthogonality property, substantially, over [0,T]. Therefore, for a given band limited function f(t), with signal capture time limited to the finite window [0,T], we can always find a high enough sample rate, kf_s so that given any $\varepsilon>0$ the following will be true:

$$\left\| f(t) - \sum_{n=0}^{K} f(\frac{n}{kf_s}) \frac{sin\{\pi k f_s[t-(n/kf_s)]\}}{\pi k f_s[t-(n/kf_s)]} \right\| < \varepsilon \quad (37)$$

$$\forall k > N, \forall t \in [0,T]$$

The number of functions in the above series (37) is now K, which is equal to the number of samples over the period [0,T]. Thus $K= kf_sT = 2kWT$. As k increases the number of sinc functions increases and the distance between the consecutive sinc functions reduces thus giving higher sample rate. See the numerical example given below. The original proof [2] for (32-34) still remains valid as we increase the sample rate. That is, the sinc functions in (32) still remain orthogonal. It can be shown using the original method that the coefficients in (33) remain valid and represent the sample values. Of course, the original proof requires the infinite time interval assumption. Thus the system still satisfies the Hilbert Space theory expressed by (4-11) making expression (37) justified. Now we can state the following new sampling theorem.

Theorem 4: Let f(t) be a band limited function with bandwidth restricted to [-W,+W] and available over the finite measurement window [0,T]. Then given any accuracy estimate ε there exists k>1 such that 2kWT equally spaced samples of f(t) over [0,T] along

with their sinc functions, will completely specify the function f(t) for all t in [0,T] at the given accuracy.

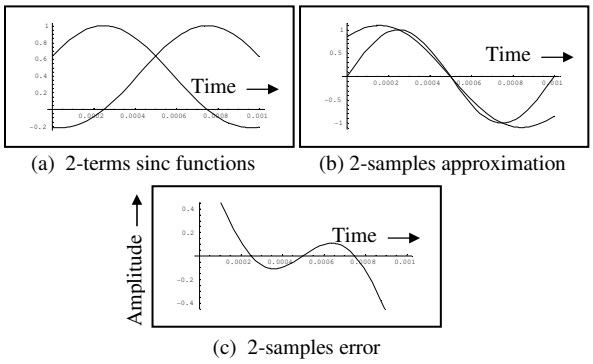
We should point out, like in Theorem 2, that if we assume infinite time interval then faster than the Nyquist rate will also not give redundant information. This concept is also easily seen from the Fourier series expression (9). To solve for the coefficients of (9) we need infinite number of samples to form a set of simultaneous equations similar to (30). As we increase the sample rate the solution of (30) will only become better, that is, the resolution of the coefficients will increase and the unknown function will also get better approximations.

For finite time assumption higher sampling rate is necessary to achieve the desired accuracy. The reason is same; the concept of infinite dimensionality must be maintained over finite time interval. That can be achieved only by higher sample rate. We also repeat, if you know the analytical expression then the number of samples must be equal to the number of unknown parameters of the analytical expression. This case does not depend on the time interval.

A lot of research work has been performed on the Shannon's sampling theorem paper [2]. Somehow the attention got focused on the WT factor, now well known as the dimensionality theorem. It appears that people have [16][17] assumed that T is constant and finite, which is not true. Shannon said in his paper [2] many times that T will go to infinite value in the limit. No one, it seems, have ever thought about the finite duration issue. This is probably because of the presence of infinite time in the Fourier transform theory. The paper [15] gives a good summary of the developments around sampling theorem during the first thirty years after the publication of [2]. Interestingly [15] talks briefly about finite duration time functions, but the sampling theorem is presented for the frequency samples, that is, over Fourier domain which is of infinite duration on the frequency axis. Now we give a numerical example to show how higher rate samples actually improves the function reconstruction.

7. A Numerical Example

We illustrate the effect of sample rate on the reconstruction of functions. Since every function can be considered as a Fourier series of sinusoidal harmonics, we take one sine wave and analyze it. This sine function may be considered as the highest frequency component of the original band limited signal. The Nyquist rate would be twice the bandwidth, that is, in this case twice the frequency of the sine wave. We are considering only one period, and therefore the Nyquist rate will give only two samples of the signal during the finite interval of its



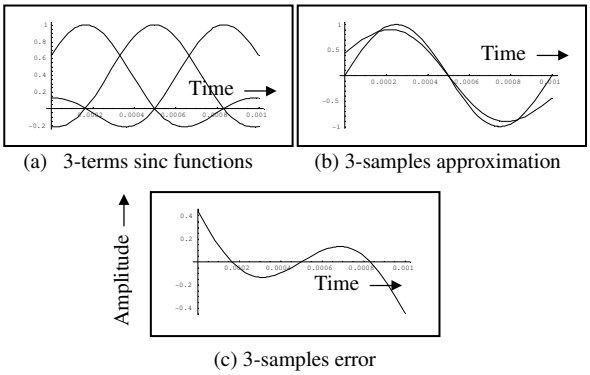
(a) 2-terms sinc functions (b) 2-samples approximation

(c) 2-samples error

Figure 4. Reconstruction using two samples

period. We are also assuming that we do not know the analytical expression of the sine wave that we are trying to reconstruct.

In this example we use the sinc functions of Shannon's theory, equation (37), to reconstruct the signal from the samples. In Figures 4-6 x-axis represents the angles of the sine function in terms of degrees or samples or time. The y-axis represents the amplitude of the sine function. Figure 4 shows the reconstruction process using two samples of the signal. In part (a) we show the sample locations and the corresponding sinc functions over the interval of one period. In part (b) we show how the construction formula (37) reproduces the function. Part (c) shows the error between the reconstructed sine function and the original sine function. We can see that the reconstruction really did not work well with two samples. Thus for finite time interval signals this process of recovering the function using expression (37) and the Nyquist rate provides very poor results.



(a) 3-terms sinc functions (b) 3-samples approximation

(c) 3-samples error

Figure 5. Reconstruction using three samples

Therefore in all engineering application we cannot use the Nyquist rate.

In Figure 5 we repeat the process described for Figure 4 with three samples. In part (b) of Figure 5 we still have significant errors. The error is prominent at the two edges because the sinc functions do not stop at the end, they continue up to infinity. This is where Bernstein polynomials, with compact support, will fit better [10]. Usually Bernstein polynomial converges very slowly and thus will require large number of samples. We mention about the Bernstein polynomial because it has some good analytical qualities [10, pp247-258]. However when we have large number of samples we may have many other better options for reconstructing the functions as described in the next section.

Figure 6 shows how the reconstruction process improves when we substantially increase the sample rate to six samples per period or three times the Nyquist rate. In this figure we still use the sinc function approach, i.e., expression (37). Notice that the errors at the edges are also reduced. This is because, as we increase the sample rate the sinc functions become narrower as predicted by the expression (36) and major part of the functions remain inside the signal interval.

The graphs show that the error decreases as we increase the sample rate as predicted by the new sampling theory and infinite dimensionality of the function space. It is clear from the examples that, for the same number of samples, a different recovery function, instead of sinc function, will give different result. In the next section we discuss this different reconstruction approach.

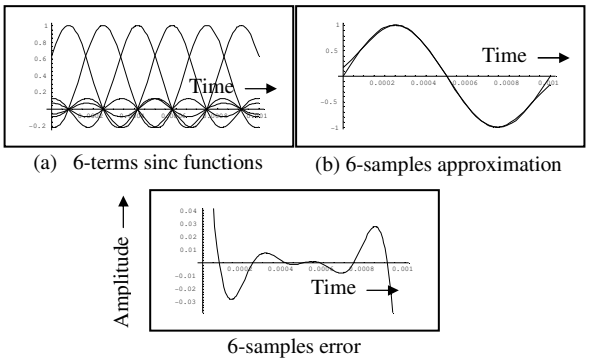
8. Re-sampling and Reconstruction

In many applications in engineering we may require different sampling rates mainly to control the computation time of the processor. The theory presented in this paper essentially says that, sample as fast as you can. That will give you maximum amount of information about the signal you want to process. However, we may not be able to select the desired Analog to Digital Converters (ADC) to sample the signal at the rate we want, because of many reasons. The two most important of them are the cost and the power required by the ADC chips.

After we have all the samples, collected at the highest feasible sample rate, then to reduce the sample rate we can simply drop few samples. This approach will maintain the quality of the signal representation. This is can be justified from the expression (30). Note that the Fourier series expression (9) can also be converted into a form given by (30). Normal decimation process changes the bandwidth thus losing the accuracy.

However, if we want to increase the sample rate, after collecting all the samples from the ADC, then we have to interpolate the samples and resample the analytical expression, thus obtained, using the new higher rate.

We emphasize the idea of using the analytical expression for the received signal in our algorithms. Instead of focusing on the samples we should focus on the mathematical expression and on the design of the algorithms around that mathematical expression. If we can achieve that then the number of samples will play a very minimal role. We will still have the complete expression of the signal even when we use very few samples.



(a) 6-terms sinc functions (b) 6-samples approximation

6-samples error

Figure 6. Reconstruction using six samples

Since in this paper we are dealing with finite time intervals, we do not yet have the proper definition of bandwidth. We also cannot use the conventional filters because they use transfer function, which uses Laplace transform, and which in turn is defined only over infinite time interval. Classical linear system theory is inappropriate for finite time interval problems. Note that the normal decimation and interpolation methods used in digital signal processing techniques [18, pp303-316] cannot be used also, because they use bandwidth related and transfer function based filters. Thus all of our analysis must rely on the time domain approaches.

We have described several analytical approaches for signal definition from the samples. Any method based on Fourier series (9), similar to (37), can be used as an analytical expression. Once you have the expression you can generate any number of samples from that expression. Approximation theory [10], a time domain approach, is also very rich in the area of interpolation using analytical expressions and can be used for re-sampling. This analytical expression approach requires that we use the entire batch of data. This batch allows us to see the entire life history of the system. This history can be more effective in signal processing than a sample by sample approach.

We want to add another layer of information to our signal processing approach. All of the above methods assume that we do not have any total system level information about the origin of the signal we are trying to reconstruct. More specifically, in digital communication receiver, for example, we know how the received signal was constructed at the transmitter. We can use that information to reconstruct an analytic expression from the samples at the receiver and then go for re-sampling, if necessary. This will produce more realistic results than straight forward application of approximation theories to the samples.

We give a numerical example to illustrate this global or system level concept of re-sampling and reconstruction of signal analysis. In function modulation [6] for example, at the transmitter, we used linear combination of a set of sinusoidal frequencies. We also used some constraints on the coefficients of this linear combination. The purpose of these constraints was to control the bandwidth (defined using the conventional sense) of the stream of concatenated symbols existing over infinite time. At the receiver we may not be able to use these constraints, but definitely we can interpolate using these specific frequency signals and achieve a higher level of accuracy in the reconstruction process.

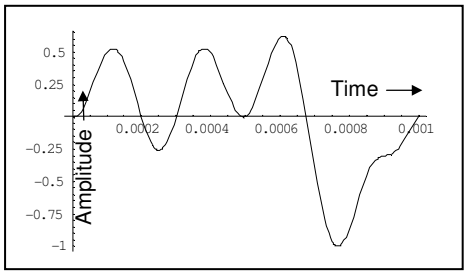


Figure 7. Transmitted symbol for fm

In our experiment, we transmitted the signal shown in Figure 7. We used the voice band telephone line to transmit the signal. The signal sample rate both at the transmitter and at the receiver was 16 kHz at 16 bits resolution. The telephone companies sample the voice signals at 8 kHz rate and at 8 bits resolution. This sample rate difference or some other unknown reasons distorted the received signal very significantly as shown in the Figure 8. As we can see from the figures the received signal has two positive peaks as opposed to three positive peaks in the transmitted signal. As if the second trough of the transmitted signal got folded up in the received signal.

It is clear that the conventional signal recovery methods, that use local concepts, no matter how many samples we take, cannot bring the received signal back to the transmitted form. However, a global approach or a systems approach, where we use the knowledge of the entire system can definitely help. We used the same sine wave frequencies of the transmitter, to interpolate the samples at the receiver. Here, of course,

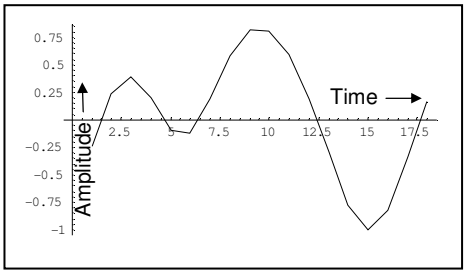


Figure 8. Received symbol for fm

the high sample rate played an important role in the least square interpolation method. The details of the signal processing, is quite involved, and is not given here. The large sample rate and the systems approach helped us to bring the received signal back to a shape that is very close to the transmitted signal, as shown in Figure 9, which allowed us to detect the bits correctly. Thus we can see that a total system level or global approach in signal processing can perform miracles.

At this end, we repeat again that all of the existing theories that are based on infinite time assumptions

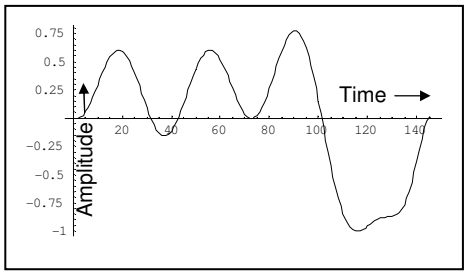


Figure 9. Best fit of the received symbol

should be carefully reviewed, redefined, and recreated for their finite time applications. We do not want any infinite time assumptions behind any of our finite time applications. That is not theoretically correct. More research will be required to generate analytical results for finite time systems. We should design our theories based on the engineering constraints. Our technology has advanced significantly. We can now use many mathematical theories that we could not use before.

Besides infinite time assumption most of our theories assume linearity also. We point out here in the next section how linearity concepts are deeply embedded in all our theories thus ignoring some basic engineering constraints. It is to be noted that the original sampling theorem used linearity assumption also, because it was based on Fourier theory.

9. Linearity Assumptions

All of our engineering systems are nonlinear because of two very important reasons. We briefly discuss them here to point out in another section that the transform methods that are based on linearity assumptions cannot be effectively used in engineering problems. Also we want to raise the importance of this nonlinearity to create a concern for the validity of our theories for engineering.

The most important reason for this nonlinearity is very well known though, but we probably never think about them. Hardly any text book [19, pp196-199] talks or provides any theory [20, pp167-179] for solving them. We call it saturation nonlinearity. Every engineering variable, like voltage, current, pressure, flow, etc. has some upper and lower bounds. They cannot go beyond that range. In terms of mathematical equation this situation can be described as

$$m_x \leq x \leq M_x,$$

where x is any physical engineering variable, m_x is the lower bound and M_x is the upper bound. Graphically the above equation can be represented by Figure 10. The figure shows that whenever the engineering variable x is within the range of [a,b] the output is linear and is equal to x. If x goes outside the boundary it gets clipped by M_x on the higher side and by m_x on the lower side. Note that m_x can be zero or negative also.

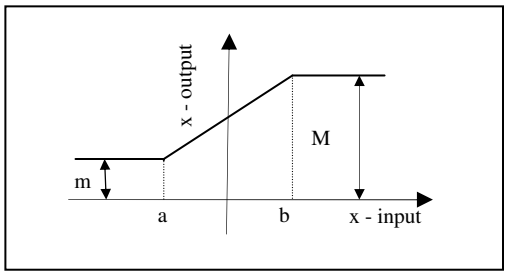


Figure. 10. Saturation Non-Linearity

Clearly Figure 10 shows severe nonlinearity and there is no escape from this in any engineering problems. We should not ignore its presence in engineering. These constraints are a kind of natural laws for our technology. Therefore we cannot treat our engineering problems using simple Linear Time Invariant (LTI) systems theory. Simply because there are no real LTI systems in engineering. If we do use such a theory then the performance of the system will be compromised.

It should be pointed out that it is almost impossible to keep the variables in the linear region. In most practical engineering problems there are hundreds of variables and hundreds of equations. It is not feasible to ensure that all variables will be within the linear range all the time. Their relationships are very complex. In addition there are many transients in all systems that can significantly alter the domain size of the variables. There is yet another important case where we have to demonstrate that the systems we build must behave normally when it goes to these limits. In addition, many systems have strict requirements that variables must go to this limit and stay there, like actuators in airplane wings. The wing flaps must reach their maximum angular positions and stay there for certain period of time.

In all engineering software, any conscientious programmer will always include the above nonlinearity test, usually called anti-windup, in their source code. And this code will automatically make the software, and hence our algorithm, nonlinear. A software engineer, barring few exceptions, does not know how mathematics works but knows how to make systems work. If you see any source code you will find many such patch work or kludge in the source code that are necessary to make the systems work. This necessity originates not only because of the lack of theoretical foundations of our algorithms but it will also be due to the RTOS and the interrupts of the background software which interferes with our theory.

All electronic hardware automatically includes such saturation nonlinearity in their systems. An automatic gain control mechanism, for example, actually is nothing but a nonlinear method of keeping the variables inside their linear regions. Because of this non-linearity any application of the LTI theory in engineering will violate the mathematical assumptions of the LTI theory. All transfer functions based approaches, like Laplace, or Fourier, are inappropriate for all engineering methods. They cannot work, because they violate the basic engineering assumptions. The transform approaches not only assume infinite time, they also assume linearity. If we use correct theory consistent with engineering models then we will definitely get much better results from our systems.

There is another very natural reason for using the non-linear theory in the design of our systems. Every engineer knows this one also but we want to mention it to strengthen our argument against the application of the LTI system theory. Most of the engineering requirements for today's technology are very stringent and we have experienced that our technology in most cases can support them to some extent. Because of these highly advanced and sophisticated requirements a simple linear model of engineering systems cannot achieve the objectives. We must make use of advanced

nonlinear and dynamic or time varying adaptive system models.

A very well know and well established example that engineers have developed during the last thirty years is the Inertial Navigation System (INS). Today our requirement says that the first missile will make a hole in the building and the second missile must go through that hole. That is a very sophisticated and precise demand. The INS development shows that simple Newton's law of force equals to mass times acceleration cannot work. The equation has been extended to fill books of hundreds of pages. They derive starting with simple linear Newton's equation [21, p4] a complete and very highly nonlinear set of equations [22, pp73-77] to describe the motion of an object. These equations include, among many other things, Coriolis forces, earth's geodetic ellipsoidal models etc. Even after including all the nonliniearities that we know of we still had to integrate the INS with a Global Positioning System (GPS) to satisfy many of our requirements.

The simple Linear Time Invariant (LTI) system equation like the one given below:

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t) \qquad (38)$$

cannot solve most of our engineering problems, simply because the laws of nature are too complex and the demands from our technology are too precise. In engineering many nonlinear problems have been attempted using successive linearization methods with the application of theories of (38). These approaches do not work also, mainly because there are no theories with very general assumptions that establish their convergence, stability, and optimality. The above reason of nonlinearity is very well known to all of us.

We have added this section to highlight the need for the signal processing approach presented in this paper. We are proposing a software radio approach with finite time batch data processing, very high sampling rate, time domain theories, and global system level simultaneous interactions. We believe that this direction, has some theoretical foundations, and can be augmented to many nonlinear dynamic approaches.

The wavelets have become a very popular signal processing tools. It also has been used to extend the sampling theorem applications [23]. So we briefly talk about it in the next section.

10. Wavelet theory

The wavelet theory has many relations with the theories discussed in this paper. The Haar wavelet systems starts [24, p431] with the characteristic function of the [0,1] interval similar to the one defined by (14). These are also orthogonal functions as described in (15). The Shannon's scaling functions [24, p422] are the sinc functions $\sin \pi t/\pi t$ similar to (32). Wavelets are very useful signal processing tools for many image and voice related problems. However it is still at its developmental phase. It has not been demonstrated yet that it can be integrated, similar to Fourier or Laplace methods, into dynamic systems governed by differential equations.

The wavelet theory also gives a down sampling process [25, pp31-35]. Like Fourier theory the wavelet down sampling provides a lower resolution, as a consequence of the multi-resolution analysis, reconstruction of the original function. This is in contrast to the method presented in previous section, where the down sampled version still maintains the original resolution quality. The down sampling process is required for reducing the throughput requirements of the processor. Down sampling should not therefore reduce the quality of the signal.

Continuous wavelet transform [24, p366] of a function f(t) has been defined by the following integral:

$$W_\psi[f](a,b) = \int_{-\infty}^{\infty} f(t)\overline{\psi_{a,b}(t)}\,dt$$

Where the wavelet $\psi_{a,b}(t) \in L^2(R)$ is defined by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), \qquad a,b \in R, a \neq 0$$

It is clear from the above definitions that the wavelet theory uses the infinite time assumption and therefore is not appropriate for signals of finite duration. It has the same application problem that the Fourier transform has. As mentioned before all practical problems are based on finite time assumptions. Since the scaling functions and the wavelets are used with translations on time axis, only very few and finite number of translations can be used over a finite duration interval.

It is also well known that the wavelet transform uses the linearity assumptions [24, p378]. Therefore it has the same problems discussed in Section 9 above and should not be used for most engineering applications.

Because many wavelets are orthogonal functions they will be very helpful in implementation of the function modulation systems described in Section 3. However it is not really known at this time how many band limited orthogonal wavelets can be constructed over a finite duration symbol time. Although wavelets are band limited but their bandwidth appears to be very high.

11. Conclusion and Future Work

We have given various proofs to show that k times, k>1, the Nyquist sample rate is necessary to improve the accuracy of recovering a function that is available only over finite time measurement window. We have shown that this k can be selected based on the required accuracy estimate ε.

The foundation of our derivations used the infinite dimensionality property of the function space. The concept essentially means that an infinite number of samples are necessary to precisely extract all the information from a function.

We have pointed out that many of our existing definitions and theories depend on the infinite time assumptions. We should systematically approach to eliminate this requirement from all our theories to make them realistic for our engineering problems.

12. Acknowledgement

The first author wishes to express his sincere thanks to our friend and colleague Hari Sankar Basak for his extensive and thorough review of our original manuscript and for his many valuable questions and comments.

13. References

[1] S.Das, N.Mohanty, and A. Singh, "Is the Nyquist rate enough?", ICDT08, Bucharest, Romania, available at IEEE Xplore, 2008

[2] C. E .Shannon, "Communication in the presence of noise", Proc. IEEE, Vol 86, No. 2, pp447-457, 1998

[3] Xilinx, "Audio sample rate converter. Reference design for Xilinx FPGAs", PN-2080-2, San Jose, California, 2008

[4] S.Das, N.Mohanty, and A. Singh, "Capacity theorem for finite duration symbols", ICN09, Guadalupe, France, available at IEEE Xplore, 2009

[5] P.A.Laplante, *Real-time systems design and analysis,* Third Edition, IEEE Press, New Jersey, 2004

[6] S.Das, N.Mohanty, and A.Singh, "A function modulation method for digital communications", WTS07, Pomona, California, available at IEEE Xplore, 2007

[7] S.Das and N.Mohanty, "A narrow band OFDM", VTC Fall 04, Los Angeles, California, available at IEEE Xplore, 2004

[8] M.A.Al-Gwaiz and S.A.Elsanousi, *Elements of real analysis*, Chapman & Hall, Florida, 2007

[9] Y.Eideman, V.Milman, and A.Tsolomitis, *Functional Analysis, An Introduction*, Amer. Math. Soc, Rhode Island, 2004

[10] G.M.Phillips, *Interpolation and Approximation by Polynomials*, Can.Math.Soc., Springer, New York, 2003

[11] V.K.Ingle and J.G.Proakis, *Digital Signal Processing using Matlab*, Brooks/Cole, California, 2000

[12] T. M. Cover and J.A.Thomas, *Elements of Information Theory*, Second Edition, John Wiley, New Jersey, 2006

[13] R. Bellman, *Introduction to Matrix Analysis,* McGraw-Hill, New York, 1970

[14] J. Farlow, J. E. Hall, J. M. McDill, and B. H. West, *Differential equations linear algebra*, Prentice Hall, New Jersey, 2002.

[15] A. J. Jerri, "The Shannon sampling theorem, its various extensions and applications – a tutorial review", Proc. IEEE, Vol 65, No. 11, 1977

[16] D.Slepian, "Some comments on Fourier analysis, uncertainty and modeling", SIAM review, 1983

[17] D. Slepian, "On bandwidth", Proc. IEEE, Vol. 64, No. 3, March 1976, pp379-393

[18] R.G.Lyons, *Understanding Digital Signal Processing*, Addison Wesley, Massachusetts, 1997

[19]G.F.Franklin, J.D.Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Third Edition, AddisionWesley, Massachusetts, 1994

[20]S.E.Lyshevski, *Control Systems Theory with Engineering Applications*, Birkhauser, Boston, 2001

[21] A.B.Chatfield, *Fundamentals of High Accuracy Inertial Navigation*, AIAA, Vol 174, Massachusetts, 1997

[22] R.M.Rogers, *Applied Mathematics in Integrated Navigation Systems*, Second Edition, AIAA, Virginia, 2003

[23] C.Cattani, "Shannon wavelets theory", Mathematical Problems in Engineering, Vol. 2008, ID 164808, Hindawai Publishing Corporation, 2008.

[24] L.Debnath, *Wavelet Transforms & Their Applications*, Birkhauser, Boston, 2002

[25] C.S.Burrus, R.A.Gopinath, and H.Guo, *Introduction to Wavelets and Wavelet Transforms - A primer*, Prentice Hall, New Jersey, 1998

DVB-H Field Measurement and Data Analysis Method

Jyrki T.J. Penttinen
Member, IEEE
jyrki.penttinen@nsn.com

Abstract

The field measurement equipment that provides reliable results is essential in the quality verification of DVB-H (Digital Video Broadcasting, Hand held) networks. In addition, sufficiently in-depth analysis of the post-processed data is important. This paper presents a method to collect and analyze key performance indicators of the DVB-H radio interface by using a mobile device as a measurement and data collection unit.

Index Terms—*DVB-H, mobile broadcast, radio network planning, network performance evaluation.*

1. Introduction

The verification of the DVB-H quality of service level can be done by carrying out field measurements within the coverage area. Correct way to obtain the most relevant measurement data, as well as the right interpretation of it, are fundamental for the detailed network planning and optimization as described in the paper [1] (ICDT 2008).

During the normal operation of the DVB-H network, there are only few possibilities to carry out long-lasting and in-depth measurements. A simple and fast field measurement method based on mobile DVB-H receiver provides thus added value for the operator. The mobile equipment is easy to carry both in outdoor and indoor environment, and it stores sufficiently detailed performance data for the post-processing.

The measurements are required for the network performance revisions and for the indication of potential problems. As an example, the transmitter site antenna element might move due to the loose mounting, which results outages in the designed coverage area. The antenna feeder might still remain connected correctly, keeping the reflected power in acceptable level. As there are no alarms triggered in this type of instance, and as the basic DVB-H is a broadcast system without uplink and its related monitoring / alarming system, the most efficient way to verify this kind of fails is to carry out field tests.

This paper is an extension to the publication [1], presenting a method to post-process the field test data collected with DVB-H mobile terminal. The method can be considered as an addition to the usual network performance testing carried out by the operator and is suitable for the fast revisions of the radio quality levels and possible network faults.

The results presented in this paper are meant as examples and for clarifying the analysis methodology. The more detailed guidelines and examples of the related results can be found e.g. in [2, 3, 4, 5, 6]. The data was collected with a commercial DVB-H hand-held terminal capable of measuring and storing the radio link related data. In this specific case, a Nokia N-92 terminal was used with a field test program. The program has been developed by Nokia for displaying and storing the most relevant DVB-H radio performance indicators.

2. Coding of DVB-H

In order to carry the DVB-H IP datagrams of the MPEG-2 Transport Stream (TS), a Multi Protocol Encapsulator (MPE) is defined for DVB. Each IP datagram is encapsulated into single MPE section. Elementary Stream (ES) takes care of the transporting of these MPE sections. Elementary Stream is thus a stream of packets belonging to the MPEG-2 Transport Stream and with a respective program identifier (PID). The MPE section consists of 12 byte header, 4 byte CRC-32 (Cyclic Redundancy Check), as well as a tail and payload length as described in [2, 7].

The main idea of MPE-FEC is to protect the IP datagrams of the time sliced burst with an additional link layer Reed-Solomon (RS) parity data. The RS data is encapsulated into the same MPE-FEC sections of the burst with the actual data. The RS part of the burst belongs into the same elementary stream (MPE-FEC section), but they have different table identifications.

The benefit of this solution is that the receiver can distinguish between these sections, and if the terminal does not have the capability to use the DVB-H specific MPE-FEC, it can anyway decode the bursts although with lower quality when it experiences difficult radio conditions, according to [7, 8, 9, 10].

The part of the MPE-FEC frame that includes the IP datagrams is called application data table (ADT). ADT has a total of 191 columns. In case the IP datagrams do not fill completely the ADT field, the remaining part is padded with zeros. The division between ADT and RS table is shown in Figure 1.

In DVB-H system, the number of the RS rows can be selected from the values of 256, 512, 768 and 1024. The amount of the rows is indicated in the signaling via the Service Information (SI). RS data has a total of 64 columns.

For each row, the 191 IP datagram bytes are used for calculating 64 parity bytes of RS rows. Also in this case, if the row is not filled completely, padding is used. The result is a relative deep interleaving as the application data is distributed for the whole bust.

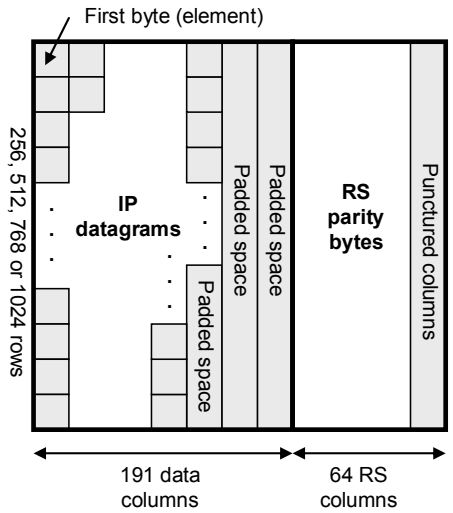


Figure 1. The MPE-FEC frame consists of application data table for IP datagrams and Reed-Solomon data table for RS parity bytes.

The error correction of DVB-H is carried out with the widely used Reed-Solomon coding. The Reed-Solomon code is based on the polynomial correction method. The polynomial is encoded for the transmission over the air interface. If the data is corrupted during the transmission, the receiving end can calculate the expected values of the data within the certain limits that depends on the settings.

The RS data of DVB-H is sent in encoded blocks with the total number of m-bit symbols in the encoded block is $n=2^m-1$. With 8-bit symbols the amount of symbols per block is $n= 2^8-1 = 255$. This is thus the total size of the DVB-H frame.

The actual user data inside of the frame is defined as a parameter with a value k, which is the number of data symbols per block. Normal value of k is 223 and the parity symbol is 32 (with 8 bits per symbol). The universal format of presenting these values is $(n,k) = (255,223)$. In this case, the code is capable of correcting up to 16 symbol errors per block.

RS can correct errors depending on the redundancy of the block. For the erroneous symbols which location is not known in advance, the RS code is capable of correcting up to $(n-k)/2$ symbols that contains errors. This means that RS can correct half as many errors as the amount of redundancy symbols is added in the block.

If the location of errors is known (i.e. in case of erasures) then RS can correct twice as many erasures as errors. If N_{err} is the number of errors and N_{ers} the number of erasures in the block, the combination of error correction capability is obtained according to the formula $2N_{err} + N_{ers} < n$.

The characteristic of RS error correction is thus well suited to the environment with high probability of errors occurring in bursts, like in DVB-H radio interface. This is because it does not matter how many bits in the symbol are erroneous – if multiple errors occur in byte, it is considered as a single error.

It is possible to use also other block sizes. The shortening can be done by padding the remaining (empty) part of the block (bytes). These padded bytes are not transmitted, but the receiving end fills in automatically the empty space.

FEC (Forward Error Correction) is widely used in telecommunication systems in order to control the errors. In this method, the transmitting party adds redundant data to the message. On the other side, the receiving end can detect and correct the errors accordingly without the need to acknowledge the data correction. The method is thus suitable for especially uni-directional broadcast networks.

FEC consists of block coding and convolutional coding. RS is an example of the block coding, where blocks or packets of bits (symbols) are of fixed size whereas convolutional coding is based on bit or symbol lengths.

In practice, the block and convolutional codes are combined in concatenated coding schemes; the convolutional coding has the major role whereas the block code like RS cleans the errors after the convolutional coding has taken place. In mobile communica-

tions, the convolutional codes are mostly decoded with the Viterbi algorithm. It is an error-correction scheme especially suitable for noisy digital communication links.

The MPE-FEC has been designed taking into account the backwards compatibility. The DVB-T demodulation procedure contains error correction so both DVB-T and DVB-H utilizes it for the basic coding with Viterbi and Reed-Solomon decoding, whereas DVB-H can use also a combination of Viterbi, Reed-Solomon and additional MPE-FEC to improve the *C*/*N* and Doppler performance [7]. The detection of the presence of MPE-FEC is done based on the single demodulated TS packet, as its header contains the error flag. MPE-FEC adds the performance in moving environment as it uses so called virtual interleaving over several basic FEC sections.

In this study, the bit rates before and after the Viterbi as well as the frame error rates of DVB-H MPE-FEC and DVB-T specific FEC were observed by collecting and analyzing data in radio interface. The measurement principles of WingTV guidelines [2] were taken into consideration in the selection (limiting) of the parameter values.

The guideline defines the MFER (MPE Frame Error Ratio) as a ratio of the number of residual erroneous frames that can not be recovered to the total number of the received frames:

$$MFER(\%) = 100\frac{residual\_erroneous\_frames}{received\_frames}$$

There are possibilities to obtain the MFER value by observing the frames during certain time (e.g. 20 seconds), or as proposed by DVB, at least 100 frames should be collected in order to calculate the MFER with sufficient statistical reliability.

The measurement principles of WingTV were used as a basis in the study by collecting 25 minutes of field data during a drive test with varying speeds, and the data was later post-processed in order to obtain the more specific relation between the FER, MFER and received power level categories. Although WingTV does not recommend the use of QEF as a criterion, also the bit errors were analyzed by utilizing the same principles for the comparison purposes.

3. Test setup

A test setup with a functional DVB-H transmitter site was utilized in order to investigate the presented field test methodology with respective post-processing and analysis. The site antenna safety distance zone was estimated by applying [11]. The investigated area represents relatively open sub-urban environment as shown in Figure 2. There are mainly relatively small trees, residential areas and highways in the investigated area with nearly-LOS in major part of the investigated area.

Figure 2. The environment in main lobe of the transmitter antenna.

The methodology was verified by carrying out various field tests mostly in vehicle. There was also static and dynamic pedestrian type of measurements included in the test cases in order to verify the usability of the equipment and methodology for the analysis.

The DVB-H test network consisted of a single 200 watt DVB-H transmitter and a basic DVB-H core network. The source data was delivered to the radio interface by capturing real-time television program. The program was converted to DVB-H IP data stream with standard DVB-H encoders. There was a set of 3 DVB-H channels defined in the same radio frequency, with audio / video bit rates of 128, 256 and 384 kb/s. The number of FEC rows was selected as 256 for MPE-FEC rate of 1/2, and 512 for MPE-FEC rate of 2/3. The audio part of the channel was coded with AAC using a total of 64 kb/s for stereophonic sound. The bandwidth was 6 MHz in 701 MHz frequency.

The antenna system consisted of directional antenna panel array with 2 elements as shown in Figure 3. Each element produces 65 degrees of horizontal beam width and provides a gain of +13.1 dBi.

Figure 3. The antenna system setup.

The vertical beam width of the single antenna element was 27 degrees, which was narrowed by locating two antennas on top of each others via a power splitter. Taking into account the loss of cabling, jumpers, connectors, power splitter and transmitter filter, the radiating power was estimated to be +62.0 dBm (EIRP) in the main lobe.

The transmitter antenna system was installed on a rooftop with 30 meters of height from the ground. The environment consisted of sub-urban and residential types with LOS (line-of-sight) or nearly LOS in major part of the test route, excluding the back lobe direction of the site which was non-LOS due to the shadowing of the site building. Each test route consisted of two rounds in the main lobe of the antenna with a minimum received power level of about -90 dBm. The maximum distance between the antenna system and terminals was about 6.4 km during the drive tests.

The terminals were kept in the same position inside the vehicle without external antenna, and the results of each test case were saved in separate text files. The terminal setup is shown in Figure 4. The external antenna was not used because the aim was to revise the quality that the end-user experiences in normal conditions inside the moving vehicle. On the other hand, the test cases were not designed for certain coverage area, but the aim was to classify the performance indicators in function of the received power levels. It does not matter thus if the received power level is interpreted via external or internal antenna.

Figure 4. The terminal measurement setup.

If the relevant data can be measured from the radio interface and stored in text format, the method presented in this paper is independent of the terminal type. It is important to notice, though, that the characteristics of the terminal affects on the analysis, i.e. the terminal noise factor and the antenna gain (which is normally negative in case of small DVB-H terminals) should be taken into account accordingly. On the other hand, unlike with the advanced field measurement equipment, the method gives a good idea about the quality that the DVB-H users observe in real life as the terminal type with its limitations is the same as used in commercial networks.

There were a total of 3 terminals used in each test case for capturing the radio signal simultaneously. Multiple receptions provide respectively more data to be collected at the same time, which increases the statistical reliability of the measurements. It also makes possible the comparison of the differences between the terminal performances.

4. Terminal measurement principles

The DVB-H parameter set was adjusted according to each test case. The cases included the variation of the code rate (CR) with the values of 1/2 and 2/3, MPE-FEC rate with the values of 1/2 and 2/3 and interleaving size FFT with the values of 2k, 4k, 8k, in accordance with the Wing TV principles described in [3], [4], [5] and [6]. The guard interval (GI) was fixed to 1/4 in each case. The parameter set was tuned for each case, and the audio / video stream was received with all the terminals by driving the test route two consecutive times per each parameter setting.

The needed input for the field test is the “on” and “off” time of the time sliced burst, PID (Packet Identifier) of the investigated burst, the number of FEC rows and the radio parameter values (frequency, modulation, code rate and bandwidth). The terminal stores the measurement results to a log file after the end of each burst until the field test execution is terminated.

According to the DVB-H implementation guide-lines [2], the target quality of service is the following:

- For the bit error rate after Viterbi (BA), the reception should comply at least DVB-H specific QEF (quasi error free) point $2 \cdot 10^{-4}$.
- The frame error rate should be less than 5%. In case of FER, i.e. DVB-T, this criterion is called FER5, and for the DVB-H specific MPE-FEC, its name is MFER5.

The field test software of N-92 is capable of collecting the RSSI (received power levels in dBm), FER (Frame Error Rate) and MFER (MPE Frame Error Rate, i.e. FER after MPE-FEC correction) values. In addition, there is possibility to collect information about the packet errors.

Figure 5 shows a high-level block diagram of the DVB-H receiver [2]. The reception of the Transport Stream (TS) is compatible with DVB-T system, and the demodulation is thus done with the same principles also in DVB-H. The additional DVB-H specific functionality consists of Time Sliced burst handling, MPE-FEC module and the DVB-H de-encapsulation.

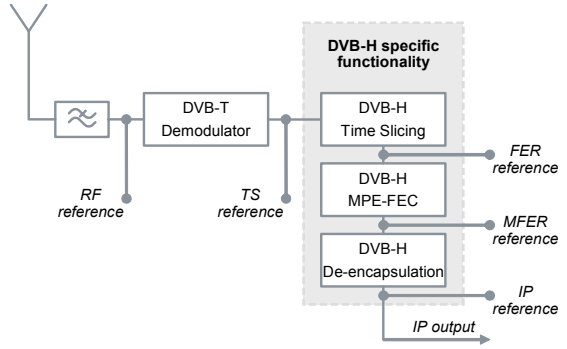


Figure 5. A principle of the reference DVB-H terminal.

As can be seen from Figure 5, the FER information, i.e. frame errors before MPE-FEC specific analysis, is obtained after the Time Slicing process, and the MFER is obtained after the MPE-FEC module. If the data after MPE-FEC is free of errors, the respective data frame is de-encapsulated correctly and the IP output stream can be observed without disturbances.

The measurement point for the received power level is found after the antenna element and the optional GSM interference filter. In addition, there might be optional external antenna connector implemented in the terminal before the RF reference point. The presence of the filter and antenna connector has thus frequency-dependent loss effect on the measured received power level in the RF point.

Figure 6 shows an example of the measurement data file. The example shows 4 consecutive test results. Each measurement field contains information about the occurred frame error (FER), frame error after MPE-FEC correction (MFER), bit error rate before Viterbi (BB) and after Viterbi (BA), packet errors (PA) and received power level (RSSI) in dBm. In this case, there was a frame error in the reception of the first measurement sample because the value of FER was "1". The FER value is either "0" for non-erroneous or "1" for erroneous frame. The MPE-FEC procedure could still recover the error in this case, because the MFER parameter is showing a value of "0". The second sample shows that there were no frame errors before or after the MPE-FEC. The third sample shows again frame error that could be corrected with MPE-FEC. The fourth sample shows an error that could not be corrected any more with MPE-FEC. In the latter case, the bit error information could not be calculated either. It seems that in this specific case, the RSSI value of about -87 dBm to -89 dBm has been the limit for the correct reception of the frames with MPE-FEC.

```
 FER:  1 MFER:  0
BB:2.5e-02 BA:1.2e-03
PE:  75
RSSI: -89
 FER:  0 MFER:  0
BB:4.5e-02 BA:1.7e-03
PE:  25
RSSI: -89
 FER:  1 MFER:  0
BB:3.6e-02 BA:8.2e-04
PE:   7
RSSI: -88
 FER:  1 MFER:  1
BB:0.0e+00 BA:0.0e+00
PE:   0
RSSI: -87
```

Figure 6. Example of the measured objects with four consecutive results for FER, MFER, BB, BA, PE and RSSI.

The plain measurement data has to be post-processed in order to analyze the breaking points for the edge of the performance. Microsoft Excel functionality was utilized in order to arrange the data in function of the received power levels.

According to the first sample of Figure 6, the bit error level before Viterbi (BB) was close to the QEF point, i.e. $2.5 \cdot 10^{-2}$. The bit error level after the Viterbi (BA) was $1.2 \cdot 10^{-3}$ which is already better than the QEF point for the acceptable reception. The bit error rate had been thus low enough for the correct reception of the signal. In this example, the amount of packet errors (PE) was between 7 and 75, and the averaged received power level was measured and averaged to -87…-89 dBm. It is worth noting that the RSSI resolution is 1 dB for single measurement event in the used version of the field test software.

Figure 7 shows the measured RSSI values during the complete test route. There were two rounds done during each test. The back lobe area of the test route can be seen in the middle of Figure, with fast momentarily drop of received power level. The received power level was about -50 dBm close to the site, and about -90 dBm in the cell edge. The duration of the single test route was approximately 25 minutes, and the total length of the route was 22.4 km.

The maximum speed during the test route was about 90 km/h, and the average speed was measured to 50 km/h (excluding the full stop periods). The speed is sufficient for identifying the effect of the MPE-FEC.

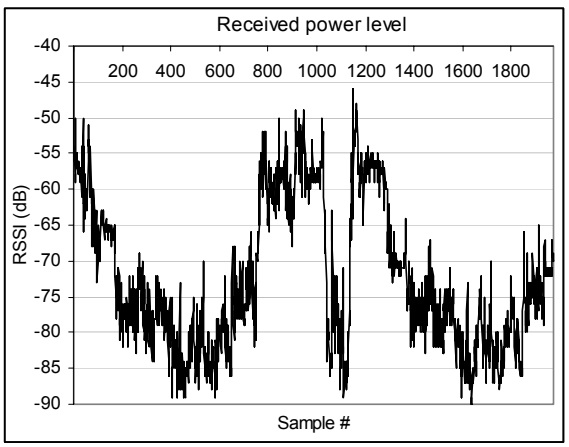


Figure 7. The RSSI values measured during the test route.

Figures 8 and 9 present the estimated coverage area for QPSK and 16-QAM cases with the code rate of 1/2 and MPE-FEC rate of 3/4. The Okumura-Hata based propagation model [10] was used with a digital map that contains the elevation data and cluster type information. The minimum received outdoor power level limit for QPSK was estimated as -84 dBm and for 16-QAM as -78 dBm.

The 80 % area location probability criterion was used in these coverage plots. The grid size is 1.6 km. (Background map source: Google Map).

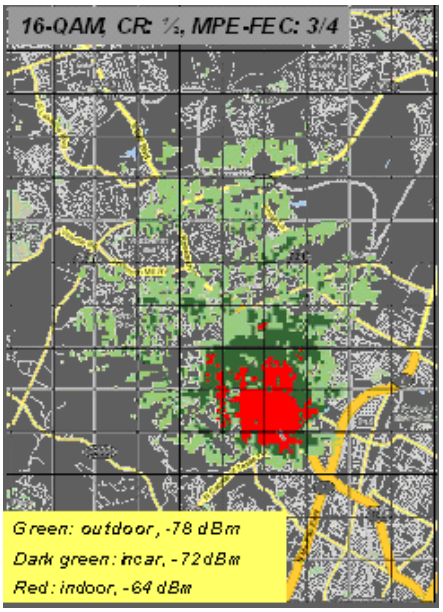


Figure 8. The predicted coverage area for 16-QAM. The raster shown in the map is 1.6 km. The main lobe and the test route is to north-west direction form the site.

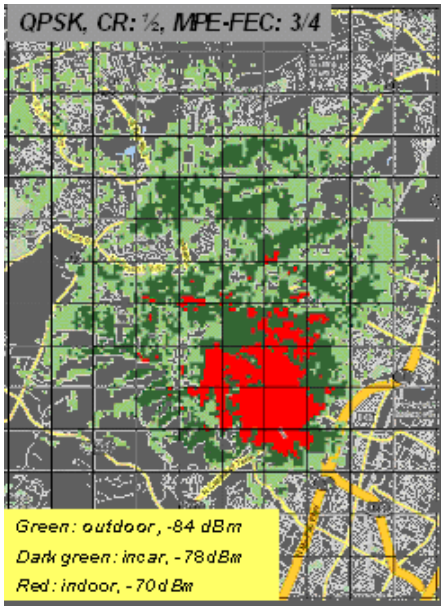


Figure 9. The predicted coverage area for QPSK, with the raster size of 1.6 km.

Based on the coverage plots, the test route was selected accordingly. The coverage plots correlated roughly with the drive tests, although the in-depth location-dependent signal level measurements were not carried out during this specific study.

5. Method for the analysis

The collected data was processed accordingly in order to obtain the breaking points, i.e. the QEF of $2 \cdot 10^{-4}$ and FER / MFER of 5% in function of the RSSI values for each test case. The processing was carried out by arranging the occurred events per RSSI value. For the BB and BA, the values were averaged per RSSI resolution of 1 dB. For the FER and MFER, the values represent the percentage of the erroneous frames compared to the total frame count per each individual RSSI value (with the resolution of 1 dB).

Figure 10 shows an example of the processed data for the bit error rate before and after the Viterbi for 16-QAM, CR 2/3, MPE-FEC 2/3 and FFT 2k. The results represent the situation over the whole test route in location-independent way, i.e. the results show the collected and averaged BB and BA values that have occurred related to each RSSI value in varying radio conditions.

As can be noted in this specific example, the bit error rate before Viterbi does not comply with the QEF criteria of $2 \cdot 10^{-4}$ even in relatively good radio conditions, whereas the Viterbi clearly enhances the performance. The resulting breaking point for the QEF

with Viterbi can be found around -78 dBm of RSSI in this specific case.

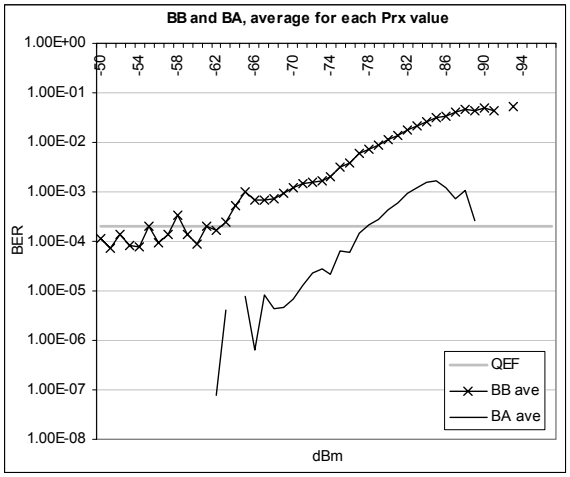


Figure 10. Post-processed data for the bit error rate before and after the Viterbi presented in logarithmic scale.

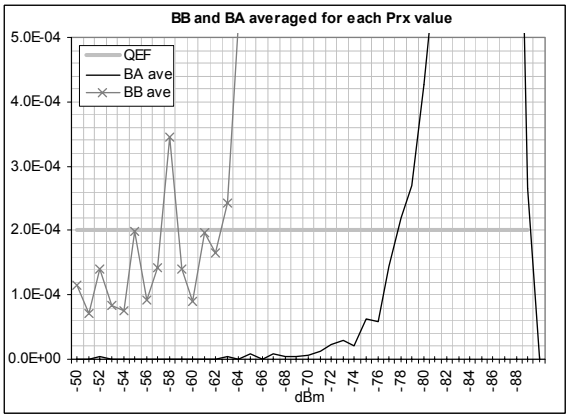


Figure 11. Processed data for the bit error rate before and after the Viterbi with an amplified view around the QEF point in linear scale.

For the frame error rate, the similar analysis yields an example that can be observed in Figure 12. Figure shows the occurred frame error counts (FER and MFER) as well as the amount of error-free events analyzed separately for each RSSI value. In this format, Figure 12 shows the amount of occurred samples in function of RSSI in 1 dB raster arranged to error free counts ("FER0, MFER0"), to counts that had error but could be corrected with MPE-FEC ("FER1, MFER0"), and to counts that were erroneous even after MPE-FEC ("FER1, MFER1").

It can be noted that the amount of the occurred events is relatively low in the best field strength cases and does not necessarily provide with sufficient statistical reliability in that range of RSSI values. Nevertheless, as the idea was to observe the performance especially in the limits of the coverage area, it is sufficient to collect reliable data around the critical RSSI value ranges.

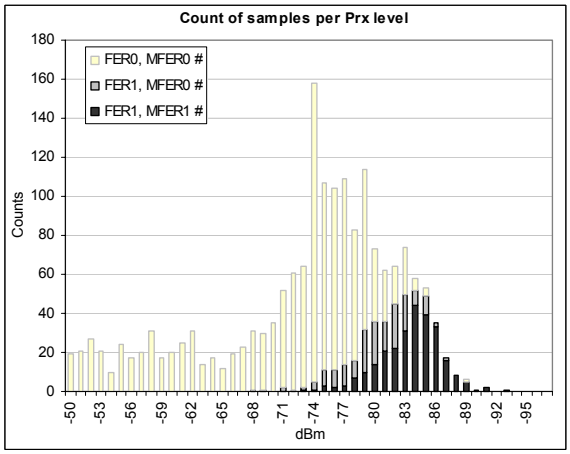


Figure 12. Example of the analyzed FER and MFER levels of the signal.

In this type of analysis, the data begins to be statistically sufficiently reliable when several tens of occasions per RSSI value are obtained, preferably around 100 samples as stated in [3]. In practice, though, the problem arises from the available time for the measurements, i.e. in order to collect about 100 samples per RSSI value in large scale it might take more than one hour to complete a single test case. There were a total of 32 test drive rounds carried out, 25 minutes each. In this case, the post-processing and analysis was limited to 2 terminals though due to the extensive amount of data.

The corresponding amount of total samples was normalized, i.e. scaled to 0-100% separately for each RSSI value. An example of this is shown in Figures 13 and 14.

By presenting the results in this way, the percentage of FER and MFER per RSSI and thus the breaking point of FER / MFER can be obtained graphically.

The 5% FER and MFER levels, i.e. FER5 and MFER5, can be obtained graphically for each case observing the breaking point for the respective curves. The corresponding MPE-FEC gain can be interpreted by investigating the difference between FER5 and MFER5 values (in dB). The graphics shows the observation point directly along the 5% error line. The parameter values of the following examples are still 16-QAM, CR 2/3, MPE-FEC 2/3 and FFT 2k.

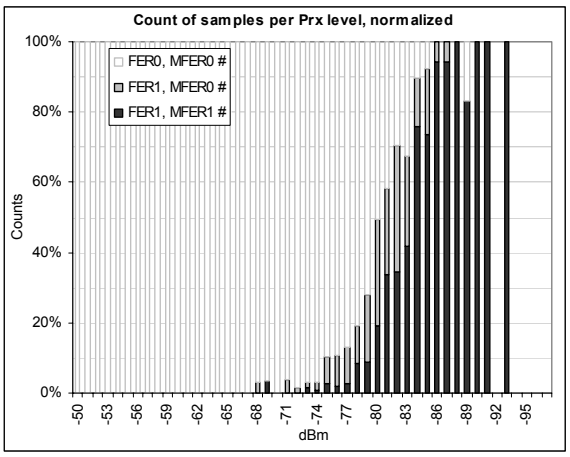


Figure 13. The post-processed data can be presented in graphical format with FER and MFER percentages for each RSSI value.

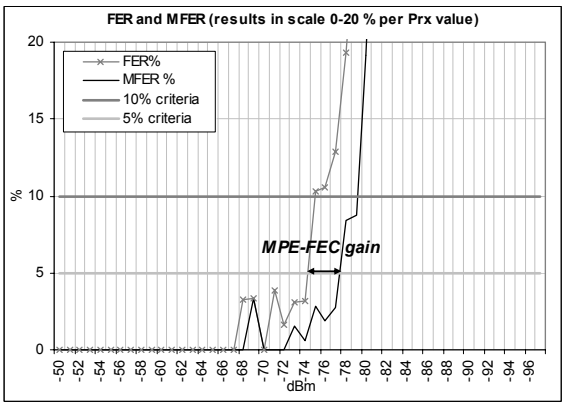


Figure 14. An amplified view to FER5 and MFER5 criteria shows the respective RSSI breaking points.

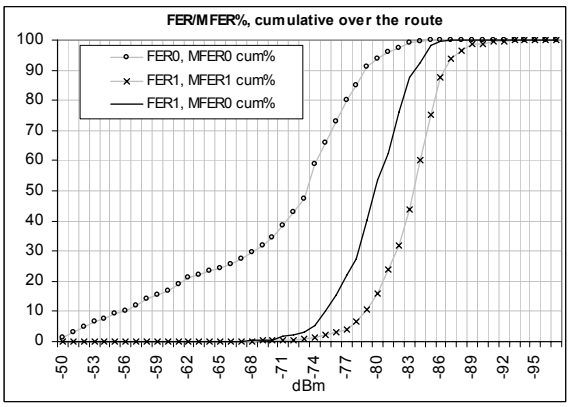


Figure 15. The processed data can also be presented in cumulative format over the whole route. This presentation gives a rough estimate about the RSSI range of the correction.

As additional information about the FEC and MPE-FEC performance in the whole scale of 0-100%, the cumulative presentation can be observed as shown in Figure 15. This format gives indication about the RSSI range where the MPE-FEC starts correcting.

The results presented in this case can be post-processed further in order to fragment the test routes into the more specific area and radio channel types. Figure 16 shows the segments during the test drive.

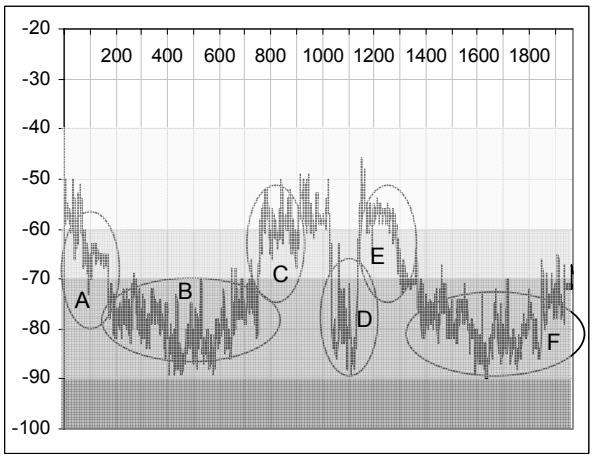


Figure 16. The segments of the test route.

Segments A and E represents the terminal moving away from the site in LOS within the main lobe of the antenna beam, with a maximum speed of about 70 km/hour and with a good field strength. Segments B and F represents the LOS or nearly LOS with the terminal moving with a maximum speed of 70-80 km/h, and the area represents the cell edge or near the cell edge, depending on the selected modes. Segment C represent the terminal moving towards the site with a maximum speed of 90 km/h, with a good field strength and LOS. Segment D represents the situation in back lobe of the antenna with N-LOS situation due to the shadowing of the site building, with the terminal speed varying between 0 and 80 km/h.

Having the segmentation done according to Figure 16, it can be seen that in the good field, as expected, the occurred FER and MFER instances are minimal and they do not affect on the reception of the DVB-H audio / video streams. Furthermore, the MPE-FEC does not bring enhancements for the performance within such a good field.

The most interesting parts of the segments are thus the ones that represents the situation nearer to the cell edge, both in main lobe (B and F) and in back lobe (D).

As an example, Figure 17 shows the analysis for the case QPSK, CR 2/3, MPE-FEC 1/2, and FFT 4k, indicating controlled behaviour of the FER and MFER until the breaking point, when the results from the segments B and F are combined and analyzed as a one complete block.

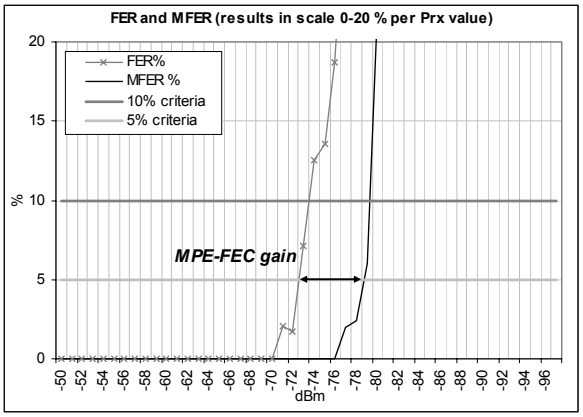


Figure 17. An example of the main lobe analysis with LOS. The curves are in general more controlled compared to the analysis of the whole route.

It is also interesting to investigate where in the RSSI scale the FER and MFER occasions have been occurred during these segments. Figure 18 shows the cumulative presentation of different FER and MFER occasions with above mentioned parameter values.

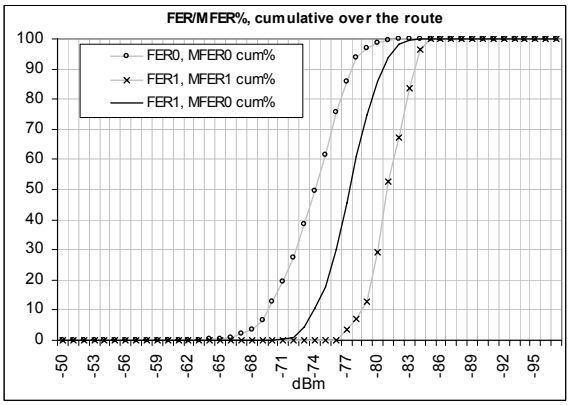


Figure 18. The cumulative presentation of FER and MFER gives a rough indication about the performance within the investigated area. Compared to Figure 15, the main lobe analysis produces clearer picture about single radio channel type.

It is worth noting, though, that this format gives only an indication about the RSSI range where the different modes of error correction tends to occur, i.e. the clean samples (FER0, MFER0), the successful MPE-FEC corrections (FER1, MFER0), and when the MPE-FEC is not able to correct the data (FER1, MFER1).

As a comparison, the N-LOS segment D yields Figures 19 and 20. The behaviour of the curves is not as clear as it is in main lobe. In addition to the attenuated N-LOS, the multi-path propagated signals are not strong in this area. It is though worth noting that the amount of the collected data is quite low, in order of 100-150 samples, which reduces considerably the reliability of the back lobe analysis.

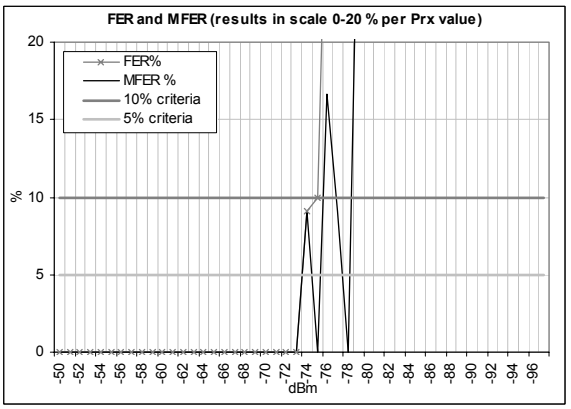


Figure 19. The back lobe analysis with N-LOS shows that the MPE-FEC is not able to correct the occurred frame error as efficiently as in main lobe with LOS. The respective segment has been selected from the same data file as shown in Figure 17 for the main lobe.

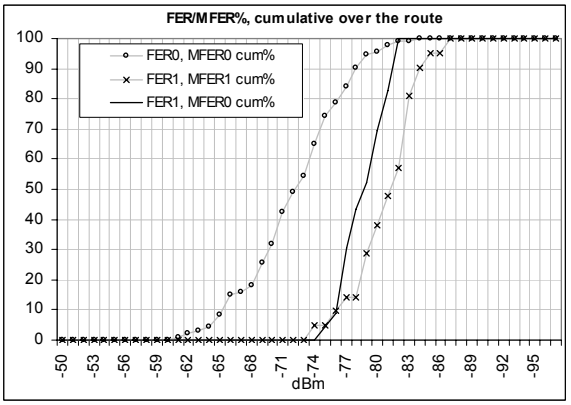


Figure 20. The cumulative presentation of back lobe analysis. It can be seen that the low amount of data affects clearly on the smoothness of the curves due to lack of data.

6. Terminal comparison

The terminal measures the received power level after the possible GSM interference suppression filter. There might also be external antenna connectors in either side of the filter. The terminal characteristics thus affects on the received power level interpretation. In order to obtain information about the possible differences of the terminal displays, separate comparison measurements were carried out.

There were a total of three terminals used during the testing. As the terminals were still prototypes, the calibration of the RSSI displays was not verified. This adds uncertainty factor to the test results.

Figure 21 shows a test case that was carried out in laboratory by keeping all the terminals in the same position and making slow-moving rounds within relatively good coverage area.

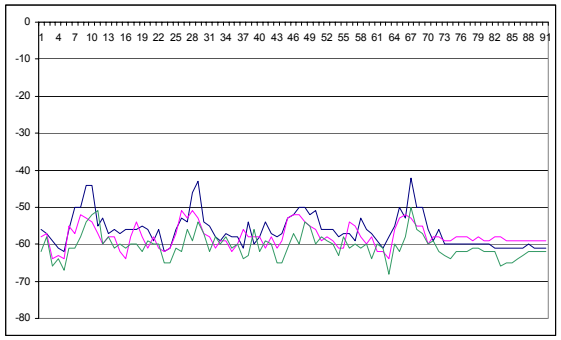


Figure 21. An example of the laboratory test case for the comparison of the RSSI displays of the terminals.

The systematic difference in RSSI displays can be noted, being about 2 dB between the extreme values. The same 2 dB difference between the terminals was noted in the field test analysis. The values obtained from the radio network tests cannot thus be considered accurate. Nevertheless, the idea of the testing was to investigate rather the methodology of the measurements than to obtain accurate values of the defined parameter settings.

In the in-depth analysis, in addition to the RSSI, also more specific differences between the terminals can be investigated. As Figures 22, 23 and 24 show, there is a systematic difference margin between the three utilized terminals as for the QEF, FER5 and MFER5. As a conclusion of Figures, in order to minimize the error margin that arises from the differences of the BER, FER and MFER interpretation of different terminals, it is important to calibrate the models accordingly.

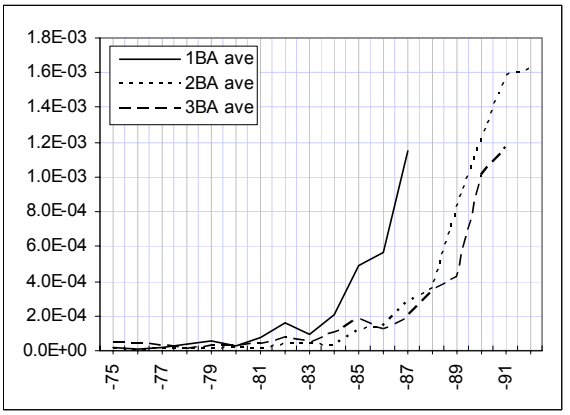


Figure 22. A comparison of 3 terminal models via BER measurement. It can be seen that two of the phones behave similarly whilst one is showing smaller bit error values.

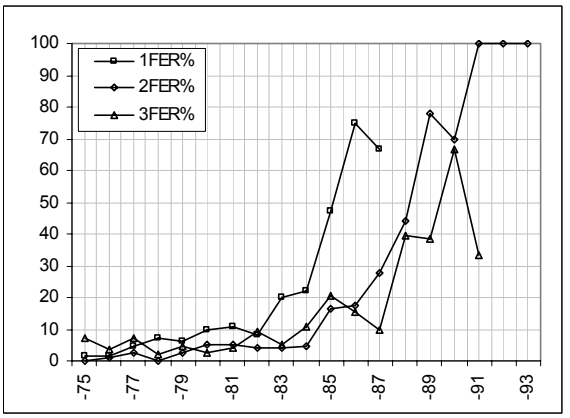


Figure 23. The systematic difference of the performance measurement values between the three terminals can be observed also via the FER curves.

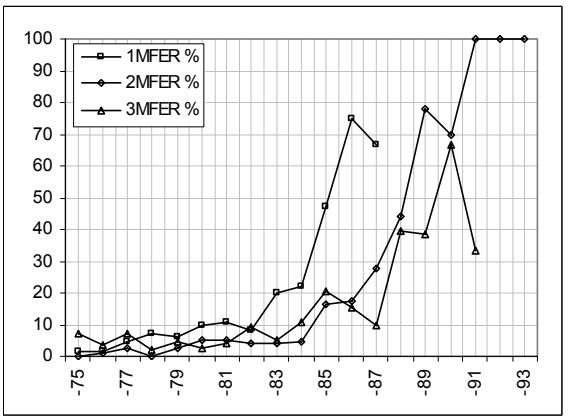


Figure 24. The differences of the terminals via the MFER analysis.

7. Building penetration loss

For the comparison purposes, there was also a set of test cases carried out in pedestrian environment with the same measurement methodology as described previously.

When designing the indoor coverage, the respective building loss should be taken into account. The loss depends on the building type and material. As an example, Figure 25 shows a relatively short measurement carried out in outdoor and indoor of a 10-floor hotel building within the coverage area of the investigated DVB-H network.

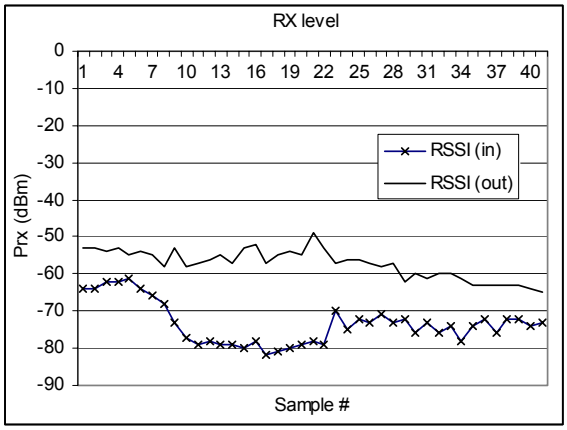


Figure 25. An example of short snap-shot type indoor and outdoor measurements carried out in the DVB-H trial network coverage area.

The building loss of Figure 25 can be obtained by calculating the difference of the received power values. The indoor average received power level is -73.4 dBm with a standard deviation of 5.7 dB. The outdoor values are -57.3 dBm and 3.9 dB, respectively. The building loss in this specific case is thus 16.1 dB. The value is logical as the building represents relatively heavy construction type, although the roof top was partially covered by large areas of class resulting relatively good signal propagation to interior of the building via the diffraction.

The test was repeated in selected spots inside and outside of the buildings, in variable field strengths both in main lobe and back lobe of the site antenna radiation pattern. As a general note, the static or low-speed pedestrian cases did not initiate the MPE-FEC functionality. Obviously more multi-path propagated signals and/or higher terminal speed would have been needed in order to "wake up" the MPE-FEC.

In the general case, the building loss should be estimated depending on the overall building type, height etc. in each environment type. In case of new areas, the best way for the estimation is to carry out sufficient amount of sample measurements for the most typical building types, although for the initial link budget estimations, the average of 12-16 dB could be a good starting point according to these tests.

8. Effects on the coverage estimation

Based on the measurement results, a tuned link budget can now be build up by using the essential radio parameters according to Figure 26.

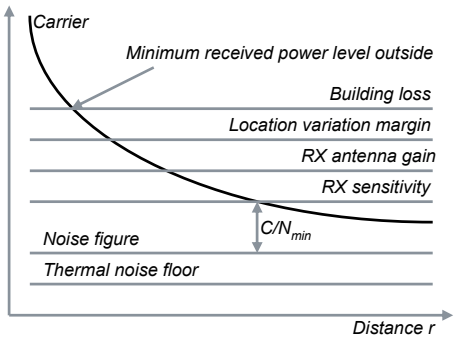


Figure 26. The main principle of the DVB-H link budget.

In order to estimate the useful cell radius of the DVB-H radio transmission, propagation prediction models can be used. One of the most used one in broadcast environment is the Okumura-Hata model, which is suitable as such for the macro cell type of environments [12]. The model functions sufficiently well in practice when the height of the transmitter site antenna is below 200 meters, and the frequency range is of 150-1500 MHz. The model is reliable for the cell ranges up to 20 km. The corrected Okumura-Hata prediction model for the distances over 20 km is defined in CCIR report 567-3. It is sufficient to estimate the cell radius for the high-power DVB-H transmitter sites with the radius up to 50 km. The most recent model that is especially suitable for the large variety of distances and transmitter antenna heights is ITU-R P.1546, which is based on the mapping of the pre-calculated curves. [13]

Based on the field test examples shown previously, the MPE-FEC gain can be included for the sensitivity Figure, compared to the DVB-T case which contains only FEC. The gain varies depending on the terminal speed and environment type, so the fine-tuning of the link budget should be considered depending on the local conditions.

9. Field test results

9.1. Complete route

As a result of the vehicle based field tests performed in this study, the following Tables 1-3 summarizes the RSSI thresholds for the QEF point of $2 \cdot 10^{-4}$ and FER / MFER of 5% criteria with different parameter values for the whole test route. In addition, the effect of MPE-FEC was obtained graphically for each parameter setting. The analysis was made for the post-processed data by observing the breaking points of BA, FER and MFER of the averaged values of 2 terminals (the other one resulting 2 dB lower RSSI). The guard interval (GI) was set to 1/4 in each case.

The MPE-FEC gain was obtained for each studied case. The effect seems to be lowest for the 64-QAM modulation, which might be an indication of too small amount of collected data in the relatively good field that this mode requires, although 64-QAM seems to be in general very sensible for errors.

It should be noted, though, that the terminals were not calibrated especially for this study. The RSSI display might thus differ from the real received power levels with roughly 1-2 decibels. The calibration should be done e.g. by examining first the level of the noise floor of the terminal and secondly examining the QEF point, i.e. investigating the signal level which is just sufficient to provide the correct receiving.

Table 1. The results for QPSK cases. The values represent the RSSI in dBm, except for the MPE-FEC gain, which is shown in dB.

FFT	8k	8k	4k	2k
CR	1/2	2/3	2/3	2/3
5%, MPE-FEC1/2	-88,1	-83,8	-78,4	-86,6
5%, FEC 1/2	-84,0	-77,3	-72,7	-81,4
MPE-FEC 1/2 gain	4,1	6,5	5,7	5,2
5%, MPE-FEC 2/3	-87,3	-83,0	-76,7	-84,4
5%, FEC 2/3	-83,3	-72,9	-73,5	-81,0
MPE-FEC 2/3 gain	4,0	10,1	3,2	3,4
BA QEF average:	-85,8	-81,7	-78,4	-84,9

Table 2. The results for 16-QAM cases.

FFT	8k	8k	4k	2k
CR	1/2	2/3	2/3	2/3
5%, MPE-FEC1/2	-77,6	-61,8	-77,4	-77,7
5%, FEC 1/2	-69,8	-61,6	-74,2	-73,4
MPE-FEC 1/2 gain	7,8	0,3	3,7	4,3
5%, MPE-FEC 2/3	-77,0	-63,5	-75,5	-77,0
5%, FEC 2/3	-72,1	-59,0	-71,0	-74,7
MPE-FEC 2/3 gain	4,9	4,5	4,5	2,3
BA QEF average:	-78,3	-73,7	-77,2	-77,4

Table 3. The results for 64-QAM cases.

FFT	8k	8k	4k	2k
CR	1/2	2/3	2/3	2/3
5%, MPE-FEC1/2	-59,9	-51,6	-65,0	-67,3
5%, FEC 1/2	-59,7	-51,6	-57,5	-65,2
MPE-FEC 1/2 gain	0,2	0,1	7,5	2,1
5%, MPE-FEC 2/3	-61,0	-51,3	-59,5	-68,1
5%, FEC 2/3	-60,3	-50,6	-54,5	-65,8
MPE-FEC 2/3 gain	0,7	0,7	5,0	2,4
BA QEF average:	-60,7	-53,0	-61,9	-68,3

As stated in [2], the moving channel produces fast variations already in TU6 channel type (typical urban 6 km/h) in the QEF criterion making the interpretation of the bit error rate before Viterbi very challenging. This also leads to the uncertainty of the correct calculation of the bit error rate after Viterbi. For the bit error rate before and after Viterbi, it is not thus necessarily clear how the terminal calculates the BB and BA values especially in the cell edge with high error rates. It has been also stated e.g. in [14] that QEF is not suitable for the instantaneous measurement due to the high variation that occurs in the mobile channel. This phenomena can be noted in the field test results as the breaking points of the QEF does not necessarily map to the corresponding FER / MFER criteria of 5% or near of it. It can thus be assumed that the most reliable results are obtained by observing the FER / MFER of the data, because their error detection is carried out after the whole demodulation and decoding process.

Furthermore, especially the frame error rate reflects the practical situation as the user interpretation of the quality of the audio / video contents depends directly on the amount of correctly received frames.

Nevertheless, the results correlate with the theory of different parameter settings, as well as with the MPE-FEC gain although it varies largely in the obtained results depending on the mode. As the test route contained different radio channel types (different vehicle speeds, LOS, near-LOS and non-LOS behind the building), the mix of the propagation types causes this effect on the results. In order to obtain the values nearer to the theoretical ones, it is important to carry out the test cases in separate, uniform areas as the radio channel type is considered, but on the other hand, these results represent the real situation in the investigated area with a practical mix of radio channel types.

9.2. Segmented route in main lobe

The following Tables 4 and 5 show the analysis for selected parameter settings of the terminal 1 (with 2 dB lower RSSI display compared to others) in order to present the principle of the segmented analysis in the

main lobe with around -70…-90 dBm of RSSI and the terminal speed of about 80 km/s. In case the breaking points could not be interpreted explicitly from the graphics, N/A was marked to Tables.

Table 4. The results for the QPSK cases in the main lobe. The channel type is nearly LOS and the terminal speed is 80 km/h.

FFT CR	8k 1/2	8k 2/3	4k 2/3	2k 2/3
5%, MPE-FEC1/2	-74.4	-84.7	-78.7	-85.3
5%, FEC 1/2	-74.3	-77.3	-72.5	-77.6
MPE-FEC 1/2 gain	0.1	7.4	6.2	7.7
5%, MPE-FEC 2/3	N/A	-82.7	-76.9	-83.3
5%, FEC 2/3	N/A	-76.1	-74.9	-79.2
MPE-FEC 2/3 gain	N/A	6.6	2.0	4.1
BA QEF average:	-84.3	-79.5	-75.7	-80.4

Table 5. The results of the setup for 16-QAM.

FFT CR	8k 1/2	8k 2/3	4k 2/3	2k 2/3
5%, MPE-FEC1/2	N/A	N/A	-76.4	-76.6
5%, FEC 1/2	N/A	N/A	-72.1	-75.5
MPE-FEC 1/2 gain	N/A	N/A	4.3	1.1
5%, MPE-FEC 2/3	-79.2	N/A	-75.3	-76.7
5%, FEC 2/3	-72.5	N/A	-72.5	-75-2
MPE-FEC 2/3 gain	6.7	N/A	2.8	1.5
BA QEF average:	-76.0	-69.5	-73.4	-76.4

It is now interesting to observe the behaviour of the FEC and MFEC values in this single radio channel type. Figures 27-28 summarise the performance of the investigated parameter set in the main lobe in the graphical format.

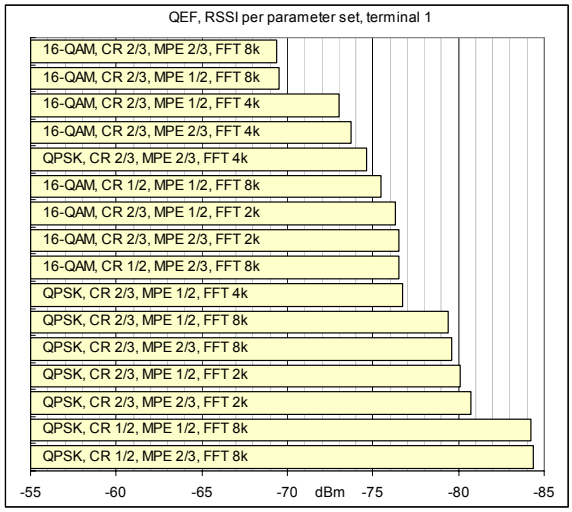


Figure 27. The summary of the QEF breaking points for the investigated parameter sets.

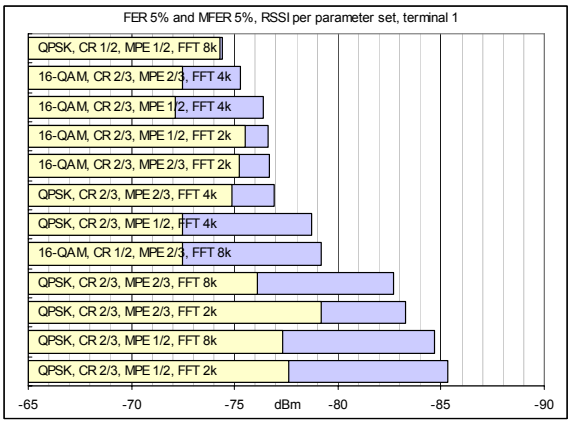


Figure 28. The summary of FER5 and MFER5 analysis.

It can be noted that the MPE-FEC functions more efficiently in this specific fragmented case, i.e. in the uniform radio channel of 80 km/h in cell edge area compared to the results obtained by analysing the whole test area with fixed radio channel types.

As a comparison, the difference of the QEF point and FER / MFER can now be obtained as shown in Figures 29 and 30.

In this case, it seems that the breaking point of QEF occurs somewhere between the FER5 and MFER5 values. As stated before, the QEF calculation is not necessarily as reliable as FER and MFER can show, but nevertheless, it is interesting to note that in this specific setup the QEF breaking point is within ±1.6 dB margin if we take simply the average of the FER and MFER values. QEF indicates thus the RSSI limits roughly in the same range as the FER and MFER does.

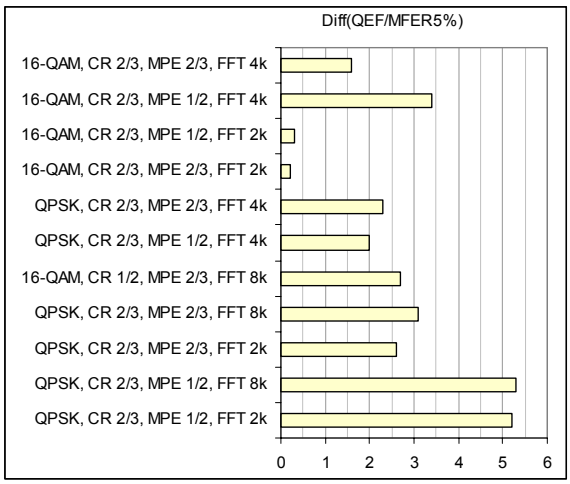


Figure 29. The difference of the QEF breaking points compared to the MFER5 results.

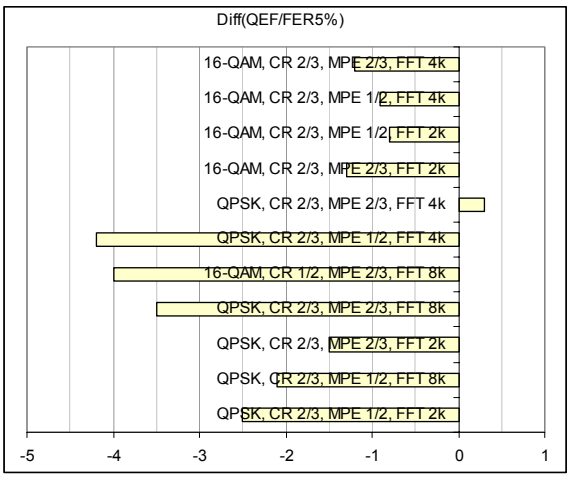


Figure 30. The difference of the QEF breaking points compared to the FER5 results.

10. Conclusions

The tests presented in this paper show that the realistic DVB-H measurement data can be collected with the terminals. The analysis showed correlation between the post-processed data and estimated coverage that was calculated and plotted separately with a network planning tool. The results correlate mostly with the theoretical DVB-H performance, although there was a set of uncertainty factors identified that affects on the accuracy of the results.

This study was meant to develop and verify the functionality of the methodology for measurements, post-processing and analysis, and as a secondary result, it also gave performance values for selected parameter set. The test environment consisted of multiple radio channel types, and the terminal displays were not calibrated specifically for these tests. An error of 1-2 decibels in RSSI values is thus expected.

The field test results show clearly the effect of the parameter values on radio performance in a typical sub-urban environment. Even if the hand-held terminal is not the most accurate device for the scientific purposes, it gives an overview about the general functioning and quality level of the network and the estimation of the effects of different network parameter settings.

The vehicle related test cases showed a logical functioning of MPE-FEC for different parameter settings. The results are in align with the theories especially when the analyzed data is limited to a single radio channel type in coverage edge of the main lobe with nearly LOS situation and with a vehicle speed of about 80 km/h. This case showed relatively significant effect of the MPE-FEC functionality giving a gain of up to 7 dB, which thus enhances the link budget and extends the coverage area compared to the basic FEC. On the other side, in the good field, the MPE-FEC is not needed as there are no frame errors present.

The whole test route can be analyzed also as one complete case, giving the overall information about the network performance in variable radio channels. It should be noted though that this case does not give comparable values for the performance measurements like the case is for the single radio channel.

The pedestrian test cases showed that the building loss can be obtained in easy way with the test setup. The analysis also showed that the MPE-FEC does not take place with such a low speeds even in low field strengths if there is lack of reflected multi-path propagated components. As the city center areas contains normally relatively good field and their main usage can be estimated to be pedestrian cases, the results indicates that the MPE-FEC setting could thus be relatively light in city areas, in order of 3/4 or 5/6, which gives still some MPE-FEC gain for sufficiently fast moving terminals, yet saving the capacity.

Regardless of the Excel sheet functionality that was developed for post-processing the data, there was a considerable amount of manual procedures in order to obtain the final results. As a possible future work item, basically all of the manual work can be automated by creating respective macro functionality for transferring the data from the terminal to the processing unit, to organize the data in function of the RSSI values, and to present the analysis in graphical and numerical format. Based on the methodology presented in this paper, the processing of the data is independent of the terminal type, but the special characteristics should be taken into account in the result tuning, as well as the proper calibration of the equipment. Furthermore, if the interface between the terminal and data processing unit supports proper protocols for fetching the data during the measurements, the post-processing and display can be performed in real time whilst carrying out the drive tests.

11. References

[1] Jyrki T.J. Penttinen. Field Measurement and Data Analysis Method for DVB-H Mobile Devices. The Third International Conference on Digital Telecommunications, 2008. IARIA, Published by the IEEE CS Press. 6 p.

[2] DVB-H Implementation Guidelines. Draft TR 102 377 V1.2.2 (2006-03). European Broadcasting Union. 108 p.

[3] Thibault Bouttevin (Editor). Wing TV. Services to Wireless, Integrated, Nomadic, GPRS-UMTS&TV handheld terminals. D8 – Wing TV Measurement Guidelines & Criteria. Project report. 45 p.

[4] Maite Aparicio (Editor). Wing TV. Services to Wireless, Integrated, Nomadic, GPRS-UMTS&TV handheld terminals. D6 – Wing TV Common field trials report. Project report, November 2006. 86 p.

[5] Maite Aparicio (Editor). Wing TV. Services to Wireless, Integrated, Nomadic, GPRS-UMTS&TV handheld terminals. D8 – Wing TV Country field trial report. Project report, November 2006. 258 p.

[6] Davide Milanesio (Editor). Wing TV. Services to Wireless Integrated, Nomadic, GPRS-UMTS&TV handheld terminals. D11 – WingTV Network Issues. Project report, May 2006. 140 p.

[7] Transmission System for Handheld Terminals (DVB-H). ETSI EN 302 304 V1.1.1 (2004-11). 14 p.

[8] Tero Jokela, Eero Lehtonen. Reed-Solomon Decoding Algorithms and Their Complexities at the DVB-H Link-Layer. IEEE 2007. 5 p.

[9] Teodor Iliev et al. Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television. ETSI EN 300 744 V1.5.1 (2004-11). IEEE 2008. 5 p.

[10] Heidi Joki, Jarkko Paavola. A Novel Algorithm for Decapsulation and Decoding of DVB-H Link Layer Forward Error Correction. Department of Information Technology, University of Turku, 2006. 6 p.

[11] Limits of Human Exposure to Radiofrequency Electromagnetic Fields in the Frequency Range from 3 kHz to 399 GHz. Safety Code 6. Environmental Health Directorate, Health Protection Branch. Publication 99-EHD-237. Minister of Public Works and Government Services, Canada 1999. ISBN 0-662-28032-6. 40 p.

[12] Masaharu Hata. Empirical Formula for Propagation Loss in Land Mobile Radio Services. IEEE Transactions on Vehicular Technology, Vol. VT-29, No. 3, August 1980. 9 p.

[13] Recommendation ITU-R P.1546-3. Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3000 MHz. 2007. 57 p.

[14] Gerard Faria, Jukka A. Henriksson, Erik Stare, Pekka Talmola. DVB-H: Digital Broadcast Services to Handheld Devices. IEEE 2006. 16 p.

Biography

Mr. Jyrki T.J. Penttinen has worked in telecommunications area since 1994, for Telecom Finland and it's successors until 2004, and after that, for Nokia and Nokia Siemens Networks. He has carried out various international tasks, e.g. as a System Expert and Senior Network Architect in Finland, R&D Manager in Spain and Technical Manager in Mexico and USA. He currently holds a Senior Solutions Architect position in Madrid, Spain. His main activities have been related to mobile and DVB-H network design and optimization.

Mr. Penttinen obtained his M.Sc. (E.E.) and Licentiate of Technology (E.E.) degrees from Helsinki University of Technology (TKK) in 1994 and 1999, respectively. He has organized actively telecom courses and lectures. In addition, he has published various technical books and articles since 1996. His main books are "GSM-tekniikka" ("GSM Technology", published in Finnish, Helsinki, Finland, WSOY, 1999), "Wireless Data in GPRS" (published in Finnish and English, Helsinki, Finland, WSOY, 2002), and "Tietoliikennetekniikka" ("Telecommunications technology", published in Finnish, Helsinki, Finland, WSOY, 2006).

On distributed SLA monitoring and enforcement in service-oriented systems

Nicolas Repp, Dieter Schuller, Melanie Siebenhaar, André Miede, Michael Niemann, Ralf Steinmetz
Technische Universität Darmstadt
Multimedia Communications Lab
Merckstr. 25, 64285 Darmstadt, Germany
firstname.lastname@KOM.tu-darmstadt.de

Abstract

For the integration of systems across enterprise boundaries, the application of Web service technology and the Service-oriented Architecture (SOA) paradigm have become state of the art. Here, the management of the quality delivered by third party services is crucial. In order to achieve high service quality, requirements therefore need to be initially specified using Service Level Agreements (SLAs), which are later on monitored during runtime. In case of SLA violations, appropriate countermeasures have to be executed.

The paper presents an integrated approach for SLA monitoring and enforcement using distributed autonomous monitoring units. Additionally, the paper presents strategies to distribute those units in an existing service-oriented infrastructure based on mixed integer programming techniques. Furthermore, appropriate framework support is given by the AMAS.KOM framework enabling distributed SLA monitoring and enforcement based on the developed distribution strategies. As a foundation for our approach, the WS-Re2Policy language is presented, which allows the specification of both requirements with respect to service quality and the necessary countermeasures at the same time.

Keywords: Monitoring; SLA enforcement; Location strategies; Service-oriented Architectures.

1. Introduction

In recent years, solution as well as implementation means for all sorts of complex communications systems, e.g., in telecommunications or business automation scenarios, are addressed by propagating Web service technology and the Service-oriented Architecture paradigm. Especially in business automation scenarios, Web services and SOAs are used for the realization of cross-organisational collaborations between enterprises by integrating the business processes and IT systems of the business partners. Here, the Business Process Execution Language (BPEL) allows the composition of services of different business partners to business processes as well as the execution across enterprise boundaries.

Nevertheless, a collaboration of business processes composed of several individual services over the boundaries of an enterprise bears several challenges enterprises have to cope with. In order to build dependable and trusted business relationships QoS and security aspects need to be addressed within the integration of third party services into an enterprise's business processes and IT system. Due to SOA's loose coupling, which permits the selection of third party services at runtime, flexible and permanently changing business relationships have to be considered in particular. Therefore, the participating parties need to define both business requirements and responsibilities of the partners by negotiating contracts and Service Level Agreements (SLA) respectively. From a technological point of view, normally defined XML-based policy documents are used for the definition of SLAs and other requirements.

However, recent approaches for the definition of such policies are insufficient as they do only address the actual requirements. But with regard to SLA enforcement, monitoring of the requirements during runtime is crucial. Therefore, also adequate countermeasures need to be defined within the policy documents in order to restore compliance with the policies in case of deviations from the specified values. Especially in distributed scenarios it is further helpful to provide several independent monitoring units with the information about requirements and countermeasures in order to enforce policies at different locations in an infrastructure.

In order to overcome both issues, we developed the Web service requirements and reactions policy language (WS-Re2Policy) presented in this article, which specifies requirements and reactions in a single policy file. Therefore, it can be deployed to different monitoring units for distributed SLA monitoring and enforcement. Additionally, we present a framework named Automated Monitoring and Alignment of Services (AMAS.KOM), which supports the implementation of WS-Re2Policy in the areas of Web services and SOAs. This article is an extension to our work [1] presented at the ICSNC 2008 conference as well as to our research presented in [2] and [3]. It enhances our previous work with respect to the underlying agent-based monitoring framework as well as to the distribution mechanisms, which are used in our approach to place monitoring units in an infrastructure.

The remaining part of this article is structured as follows. In the next section, the overarching scenario of distributed SLA monitoring and enforcement is discussed in more detail. Subsequently, the WS-Re2Policy language is discussed in depth and explained by the use of an example. Afterwards, the AMAS.KOM framework is presented, which allows the use of WS-Re2Policy for Web service-based SOAs. The following section presents our work towards the optimal distribution of monitoring units in a distributed service-based infrastructure. Here, a distribution strategy is developed and evaluated by the use of simulation. Before the article closes with a conclusion and outlook, the related work to our research is presented.

2. Scenario and approach

In cross-organisational collaborations on the basis of an integration of business processes and IT systems, the "classical" scenario consists of a single enterprise and different business partners, in which the enterprise wants to use different third party services provided by the partners in a *1 to n* client-server style [4].

Here, a centralised monitoring and deviation handling approach is performed and carried out by the service requester (SR) itself (cf. Figure 1a). Thus, all other components, i.e., the monitoring units (MU) and decision making components, are located at the service requesters', even if monitoring data is sometimes collected by distributed probes. However, in large-scale SOA scenarios a centralised approach is not applicable, due to a vast amount of service requesters and providers (e.g., *m to n*) which bears scalability and complexity issues. But also unclear responsibilities between participating partners or a lack of privacy lead to legal and governance issues. Furthermore, the collection and availability of monitoring data required for decisions is hindered due to the existence of different spheres of control representing domains which belong only to single partners. Consequently, there exists no sufficient quality and amount of monitoring data, so that adequate decision making and timely handling of SLA violations cannot be performed.

For reasons already stated, we propose a distributed approach to SLA monitoring and enforcement, which overcomes the information deficit and improves the speed of information provisioning. Our approach is based on the application of decentralised monitoring and alignment agents (MAAs) which obtain both monitoring requirements and the specification of countermeasures. The MAAs are placed within the infrastructure at various places (cf. Figure 1b). Here, a hybrid approach, i.e., a mixture of centralised and decentralised interaction styles, is taken, instead of a fully decentralised approach (i.e., Peer-to-Peer).

A WS-Re2Policy compliant policy file enables the monitoring and alignment agents to manage single services as well as service compositions in a semi-autonomous way

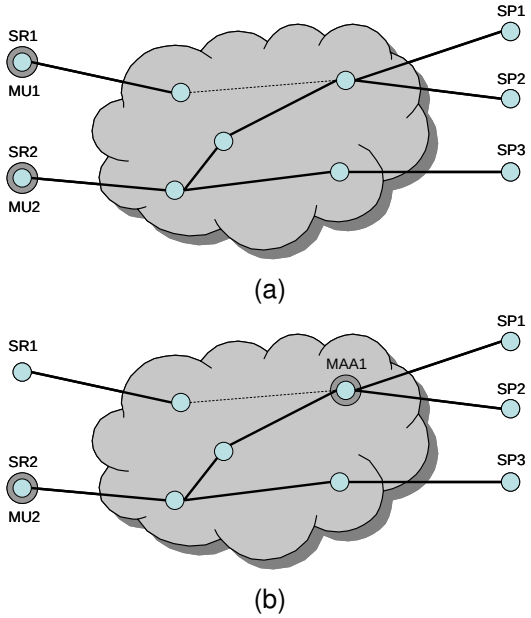


Figure 1. Monitoring styles (cf. [2])

according to the specified rules. Within a policy file agreed countermeasures in case of SLA violations are defined representing boundaries of the MAA's behaviour. Given that policies can be split into sub-policies, the corresponding subdivision of the MAA's behaviour forms the basis for new MAAs. Policies and MAAs can also be recombined in order to reduce the amount of MAAs up to a single instance. Using specialised communication mechanisms, MAAs can interact with each other, so that tasks can also be delegated between MAAs. Here, the communication protocols of the selected agent platform are facilitated (e.g., the Agent Communication Language). An example of cooperating agents based on our AMAS.KOM framework can be found in [5].

3. Web service requirements and reactions policy language

This section addresses various aspects of the WS-Re2Policy language in its most recent version. Starting from a theoretical point of view, a basic example is then used to discuss and explain the core elements of the language. For a discussion of a preliminary version of the WS-Re2Policy language we refer to [3].

3.1. Theoretical foundation of the WS-Re2Policy language

The well-founded Event-Condition-Action (ECA) rules paradigm, first discussed in the area of active databases (e.g., [6]), represents the basis of the WS-Re2Policy language. The

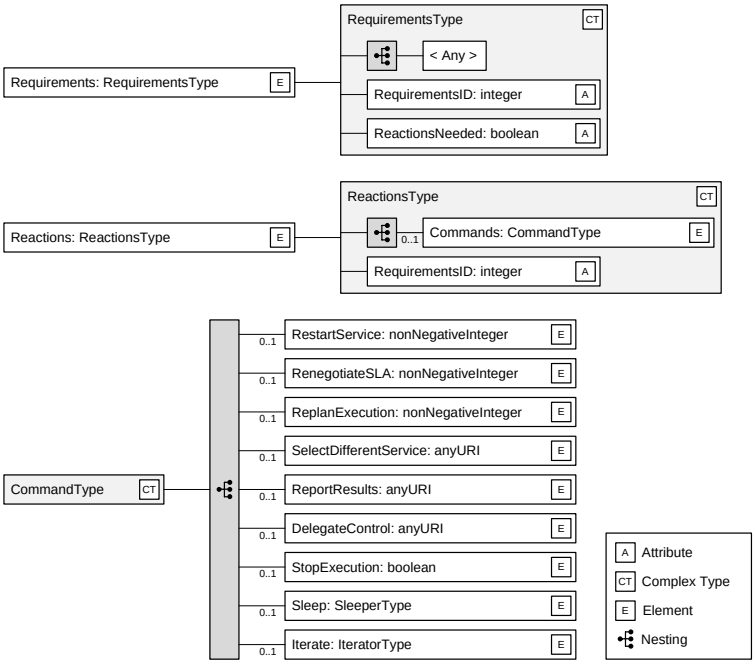


Figure 2. Core elements of the language

name ECA already states the main concepts of the ECA rules:

```
ON Event
IF Condition
DO Actions
```

Those concepts are directly used by the WS-Re2Policy language. Its elements can be mapped to those ECA concepts as follows:

- Event: current measure of a monitoring subject, e.g., the response time of a service composition.
- Condition: threshold for the monitoring subject, e.g., the upper bound of the service's response time.
- Actions: reactions to an SLA violation aiming at the enforcement of the SLA, e.g., the restart of a service after a failure or time-out.

The use of ECA-styled rules as a basis for our WS-Re2Policy language has different advantages. In the first instance, the application of ECA rules does not require a very deep understanding which allows the generation of policy files even by non-experts with tool support. Furthermore, using a rule-based system the separation of control logic from the real implementation of an MAA is supported, thus allowing adaptability, which is crucial for our approach. Finally, a broad theoretical foundation exists, ranging from possible optimisations of rule-based systems to distributed problem solving strategies of autonomous units in distributed systems using ECA in combination with π-calculus [7].

3.2. Core elements of the WS-Re2Policy language

The WS-Re2Policy language was designed as an extension to the World Wide Web Consortium's WS-Policy 1.2 framework in order to use existing standards and be compliant to it. Therefore, the WS-Re2Policy can be extended by other WS-Policy compliant languages, e.g., WS-SecurityPolicy.

Basically, a WS-Policy compliant policy consists of two main parts: the requirements and the reactions part as depicted in Figure 2. For the description of requirements any WS-Policy compliant language can be used. Currently, some basic QoS parameters, e.g., throughput and response time, are natively supported by our approach. In future versions of the language, further QoS parameters will follow.

In the WS-Re2Policy language, reactions are simple and easy to understand control structures describing possible countermeasures in case of deviations from SLAs. At this stage, the WS-Re2Policy language as well as the AMAS.KOM framework support the following reactions:

- Restarting of a service, which violated an SLA.
- Renegotiation of SLA parameters for a single service or composition.
- Replanning of an existing execution plan for the composition a unit is responsible for.
- Selection of different services, which offer comparable functions.
- Reporting of results to caller or third parties.
- Delegation of control to other units on the same level

```
<?xml version="1.0" encoding="UTF-8"?>
<re2:RequirementsReactionsSuite ... >
 <re2:Requirements ReactionsNeeded="true"
      RequirementsID="3428">
  <wsp:Policy wsu:Id="ID_236" Name="Security-QoS">
   <wsp:All>
    <wsp:All>
     <sp:EncryptedParts>
      <sp:Body/>
     </sp:EncryptedParts>
    </wsp:All>
    <wsp:All>
     <qos:Throughput>10</qos:Throughput>
     <qos:ResponseTime>23.55ms</qos:ResponseTime>
    </wsp:All>
   </wsp:All>
  </wsp:Policy>
 </re2:Requirements>
 <re2:Reactions RequirementsID="3428">
  <re2:Sleep time="10.0ms"/>
  <re2:Iterate time="0.0ms" count="2">
   <re2:RestartService/>
  </re2:Iterate>
  <re2:DelegateControl>caller</re2:DelegateControl>
 </re2:Reactions>
</re2:RequirementsReactionsSuite>
```

Figure 3. A WS-Re2Policy compliant example

or to the central control instance without raising exceptions.

- Interruption of execution and passing back control by raising exceptions.

Additionally, the WS-Re2Policy language supports further control structures, e.g., loops (so-called iterations).

Finally, with regard to the connection between the parts of the WS-Re2Policy language and the ECA structures, the requirements parts contain the events, i.e., the subjects to monitor, and their corresponding conditions. The reactions part of the policy contains the specified actions. A reference key is used to interconnect reactions and requirements, so that reactions can be reused in different requirement parts.

3.3. WS-Re2Policy – a basic example

A simple example of a WS-Re2Policy compliant policy document is depicted in Figure 3. The namespace declarations of both WS-SecurityPolicy and WS-Re2Policy were removed in order to improve readability.

Within the requirements part of the example, requirements in two different WS-Policy compliant policy languages are defined. The first requirements element contains WS-SecurityPolicy information (cf. the following tag: $< sp : EncryptedParts >$) representing the required security features for interaction with the service in this case. The second requirement element of the example defines QoS parameters specifying a minimum of 10 concurrent service calls and a maximum of 23.55 ms for the response time.

The reactions part of the document specifies the countermeasures in case of SLA violations. In the example in Figure 3, the MAA restarts the service twice 10 ms after an SLA violation occurred. If no normal service operation can be established, the MAA raises no exception. Instead, the control is passed back to the caller for further handling.

4. The AMAS.KOM framework

As a proof of concept, we designed the AMAS.KOM framework and its underlying architecture. By supporting all phases ranging from the modelling of requirements to the enforcement of SLAs by MAAs, AMAS.KOM aims towards a holistic SLA monitoring and enforcement approach. For this purpose, AMAS.KOM offers a transformation of a business process description and associated requirements into a monitored process. Within the transformation, an indirection of service calls to the MAA infrastructure is integrated by analysing and adapting an existing process description in form of a Web service composition.

The transformation process consists of the four steps *modelling and annotation*, *modification and splitting*, *generation*, and *distribution*. In the first step, *modelling and annotation*, the requirements concerning the complete business process are specified by manually enhancing a description of a business process with SLA assertions. Here, the business process is described in the Business Process Modelling Notation (BPMN – cf. [8]) which represents a graphical specification of a business process. Later on, the SLA assertions are transformed into policy documents describing the requirements and countermeasures for the complete process using the WS-Re2Policy language. In order to support the user with the annotation, AMAS.KOM provides a Web-based wizard. Within the second step called *modification and splitting*, separate policies for each service or sub-process, depending on the planned granularity of MAAs, are derived using the global policy document. Here, various QoS-aware planning algorithms can be used for planning purposes in order to generate feasible partitions, e.g., as discussed in [9], [10]. This process step is carried out automatically and results in policy documents and execution plans. Both contain simple Web service calls in combination with a policy document and execution plans for sub-processes in combination with the related policy documents. The third step, *generation*, includes the creation of MAAs on the basis of the predefined policies. Afterwards, the MAAs are distributed within the monitoring and alignment infrastructure during the *distribution* step. Due to a plug-in concept, MAAs are highly extensible. Therefore, only the configuration of the plug-ins needed, as specified by the requirements in the policy before their distribution, is necessary. For distribution purposes, various algorithms can be used in AMAS.KOM, e.g., a random distribution algorithm or the use of heuristics for the solution of the MAA location problem. An optimal

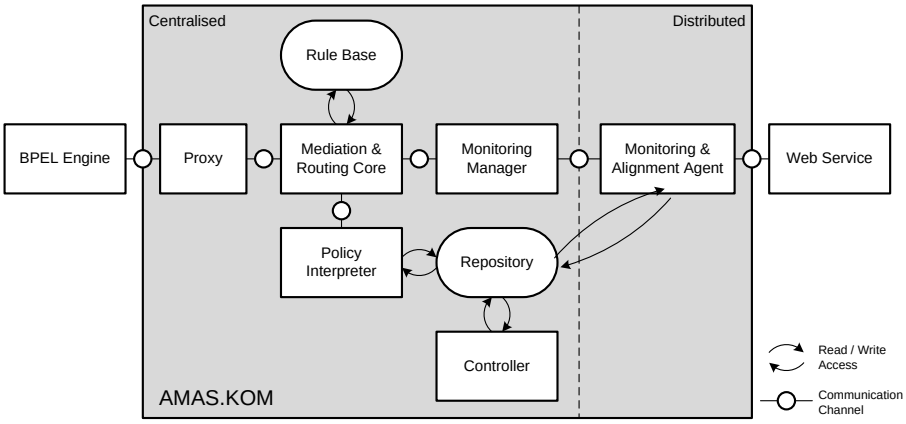


Figure 4. Overview of the AMAS.KOM framework

distribution strategy based on the analysis of given SLAs is presented in section 5.

The AMAS.KOM framework contains five core and various supportive components, like data storages for requirements, system configurations as well as monitoring data (cf. Figure 4). Subsequently, we will shortly discuss the core components:

- *Mediation & Routing Core*: combines both configuration information for MAAs as well as the applicable policies to specification sets and is furthermore responsible for routing both Web service call and specification set to Monitoring Managers.
- *Policy Interpreter*: calculates the effective policy, i.e., the policy out of a set of possible policies, which represents the current monitoring situation best.
- *Monitoring Manager*: generates MAAs based on a given specification set and distributes them in the infrastructure.
- *Controller*: provides the logic to build monitored processes by transforming complete requirement sets into service or sub-process specific policy documents.
- *Monitoring & Alignment Agent*: responsible for the actual monitoring as well as the execution of countermeasures.

The actual process of executing a monitored instance of a Web service call is described in the following. Generally, the Web service calls of the *BPEL Engine* are forwarded to a *Proxy* which redirects them to the *Mediation & Routing Core* component. For each Web service invocation, this component requests the effective policy from the *Policy Interpreter* which retrieves all applicable policies from the *Repository* and determines the effective one. Subsequently, the *Mediation & Routing Core* retrieves the associated configuration information of the service invocation from the *Rule Base*. Afterwards, the service invocation is routed to an appropriate *Monitoring Manager* which performs the generation of tailor-made monitoring units. The *Monitoring & Alignment Agent* performs the actual service call and tries to comply with the associated policy. For further evaluation purposes, the results are stored in the *Repository*. After the fulfilment of the policy, the result is passed back to the *BPEL Engine* via the *Mediation & Routing Core*. In case of an unsuccessful service invocation, appropriate alignment measures have to be accomplished.

We prototypically implemented the AMAS.KOM framework as a proof-of-concept for our distributed SLA monitoring and enforcement approach and the WS-Re2Policy language. Therefore, we used the JADE agent development framework to realise the MAAs, Apache Axis2 for Web service integration as well as the WSBPEL 2.0 standard for the specification of Web service-based collaborations. WS-Policy 1.5 and WS-SecurityPolicy 1.1 are also supported via the WS-Re2Policy language.

5. Distribution strategy for monitoring and alignment units

In this section, we introduce an approach to the distribution of MAAs in an infrastructure. As the installation of those monitoring and alignment components is associated with expenses, it should be avoided to distribute them randomly. Instead, we recommend to distribute the MAAs in a way that minimises the total costs, which can be divided in setup costs and costs for the communication between nodes.

5.1. Modelling basics and prerequisites

In order to model the distribution problem of MAAs, some assumptions have to be made.

The relationships between service requester, service providers, and potential intermediaries form a network, which can be modelled as an undirected graph. Intermediaries are all nodes in the communication path between service requester and providers. The communication between

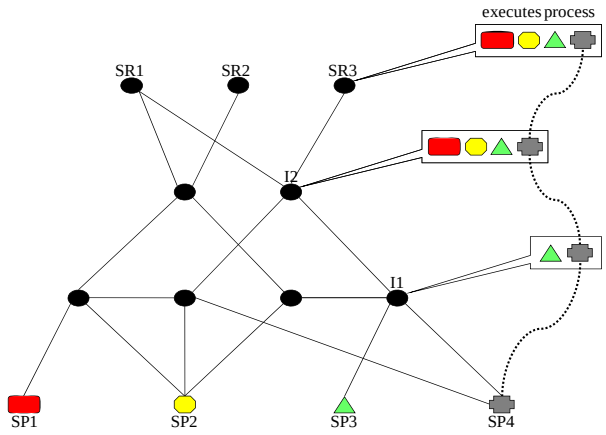


Figure 5. Example topology and routing

requester and providers forms a spanning tree, normally based on the shortest path from the requester to one provider. All intermediaries as well as the root itself can be seen as possible candidates for the execution of a monitoring and alignment unit.

The goal of our optimisation problem is the selection of those nodes, where the setup of monitoring and alignment units minimises the overall costs consisting of setup and communication costs. We call such an optimisation model *Monitoring Unit Location Problem* (MULP).

As constraints, we need to ensure that the alignment demands of all providers are satisfied. Here, the alignment demand defines the need for corrective actions with respect to a monitored service if an SLA is violated during service execution. The alignment demand is an abstract concept used for modelling, which can contain various elements like availability of a service or its error rate. As input for both the alignment demand calculation and the costs for communication as well as setup costs, the SLAs between the communicating parties are used.

In order to clarify the idea of the MULP as well as the concept of alignment demands, we present a simple example in Figure 5. As depicted in Figure 5, *Service Requester 3* (SR3) executes a process which is composed of services from different service providers – *Service Providers 1* to *4* (SP1 to SP4). If SR3 calls the service provided by SP4, the request and its corresponding response pass the *Intermediaries 1* and *2* (I1 and I2). Assume, that a MAA is installed on I1, which controls SP3 and SP4. In case the service provided by SP4 violates its current SLA, the MAA at I1 is able to detect this violation long before it is noted by the calling party. Furthermore, the MAA tries to correct the violation, e.g., by restarting the service, if the service was unavailable before. In the best case, the calling party does not even notice the problems. The alignment demand, which emerges from the SLA violation of the service, is satisfied by the MAA at I1, as it is responsible for the service provided by SP4. If I1 is unable to align the service, the control of the service execution is passed back to the calling party SR3.

Our optimisation problem, which determines the placement of MAAs in an infrastructure by reducing overall costs, can be mapped to a Warehouse Location Problem (WLP – cf. [11]). Depending on the existence of capacity restrictions with regard to the amount of supported alignment demands at an intermediary, the problem can be mapped to either a capacitated WLP or an uncapacitated WLP. In both cases, the corresponding optimisation problems are NP-hard.

Mixed integer programming techniques (cf. e.g., [12], [13]) can be applied for the purpose of solving the WLP-based *Monitoring Unit Location Problem* for arbitrary topologies. The resulting models can be solved using Branch-and-Bound algorithms afterwards.

5.2. Modelling the distribution strategy

Service requester, service providers and intermediaries are described by nodes of the network topology. Here, we can distinguish between the set of nodes that represent service providers (nodes $j \in M = \{1, ..., m\}$) and the set of nodes that represent service requesters and intermediaries (nodes $i \in N = \{1, ..., n\}$). By definition, only the service providers have alignment demands d_j, as already stated in the section before. Service providers are not allowed to directly execute monitoring and alignment units, because we only consider the requester perspective in our model. Service providers probably will carry out their own monitoring. For this, it is sufficient only to consider the nodes $i \in N$ when assigning setup costs. The actual costs for installing a MAA on node i are indicated by c_i^s. The cost of communication between node i (service requester or intermediary) and node j (service provider) are labelled with c_{ij}^c. In this context it is important to mention that the existence of an edge between two arbitrary nodes is not mandatory. Therefore, c_{ij}^c is defined as follows:

$$c_{ij}^c := \begin{cases} c_{ij}^{c*}, & \text{in case } (i,j) \text{ exists} \\ \min(c_{kj}^c + c_{ik}^{c**}), & \text{in case } (i,k) \text{ exists} \\ \infty, & \text{else} \end{cases}$$

Thereby, c_{ij}^{c*} describes the communication costs between intermediary i and provider j ($i \in N, j \in M$). Furthermore, c_{ik}^{c**} describes the communication costs between intermediary i and intermediary k ($i, k \in N$). These definitions ensure that c_{ij}^c specifies the minimum possible communication costs between intermediary i and provider j. The variables a_i constitute the capacities for nodes i. As already mentioned before, load restrictions could be seen as an example for such capacities. The decision variables y_i state, whether or not to install a MAA on node i, while the decision variables x_{ij} describe the communication at a given path assuring that the alignment

demands are satisfied. The definition of the required variables is also depicted in Table 1 and Table 2.

According to [11], the corresponding mathematical model of the MULP can be described as denoted in the following optimisation model:

Target function

$$\text{minimise } F(x,y) = \sum_{i=1}^{n} c_i^s y_i + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}^c x_{ij} \quad (1)$$

subject to

$$\sum_{i=1}^{n} x_{ij} = d_j \quad \forall j \in M \quad (2)$$

$$\sum_{j=1}^{m} x_{ij} \leq a_i y_i \quad \forall i \in N \quad (3)$$

$$y_i \in \{0,1\}, x_{ij} \geq 0 \quad \forall i \in N, \forall j \in M \quad (4)$$

Finally, it is important that the MAA setup costs at the service requesters' are zero, i.e., monitoring and alignment can be done at the service requester's process engine without additional costs. Here, centralised monitoring is always enabled and taken into account.

As stated in the previous section, we use Branch-and-Bound algorithms to solve the WLP. For large topologies, this would require strong computational effort. In such a case, we propose to relax the integrity conditions and apply heuristics afterwards – as for example H1_RELAX_IP discussed in [9] – to get a valid solution with integer values for y_i. This heuristic does not perform significantly worse compared with the optimal solution with respect to its solution quality (cf. [9]).

In any case, the partitioning and distribution of the optimisation process improves the scalability of the complete system (e.g., the load situation at the service requester's QoS management system), because the computation of the distribution schemes is not only carried out by a single system but by all existing monitoring units for the areas they are responsible for. Furthermore, the distribution of the monitoring units and hence the computation of the distribution schemes into control spheres of third parties with no external access allows the application of our approach, e.g., in scenarios with high security demands.

6. Evaluation of the distribution strategy

In this section, we present the evaluation of the distribution strategy presented before. For this, different infrastructure types and configurations are simulated. As prerequisites, we present the scenarios used for simulation as well as the simulation setup in the following sections.

Table 1. Definition of variables

i:	service requester, intermediary
j:	service provider
d_j:	alignment demand of provider j
c_s^i:	costs for installing a MAA on node i
c_{ij}^c:	communication costs between i and j

Table 2. Definition of decision variables

y_i:	whether or not to install a MAA on i
x_{ij}:	alignment demand of j satisfied by i

6.1. Simulation scenarios and assumptions

The overall goal of the simulation is to show that the distribution of MAAs can lead to cost savings for a given distributed scenario. The following aspects were taken into account during the evaluation:

- A cost comparison of the optimal distributed solution with the centralised solution.
- The comparison of the number of MAAs with the overall nodes and service providers.
- The analysis of the execution time behaviour of our distribution strategy.

During simulation different parameters were adjusted and their impact on the overall performance was measured. The following parameters were used in the evaluation:

- *Topology parameters*, containing the number of nodes and connectivity parameters of the network, which are used to configure the Waxman algorithm.
- *Percentage of service providers*, defining the amount of service providers in relation to the overall nodes. The service providers are randomly distributed in the given topology.
- *Cost ratio*, which is defined as the ratio of setup costs for installing a MAA on a node to the maximum of the communication costs per link.

Based on those simulation parameters, several simulation scenarios can be specified. For this article, we focus on scenarios in which the topologies (as well as the related matrices) are sparse, i.e., they have a high degree of unconnected links. We assume different cost ratios in order to test extreme configurations of the simulation. Here, cost ratios of setup costs and maximum communication costs of 1:4, 1:1, and 4:1 were used. Furthermore, every scenario contains random topologies including 10, 20, ..., 100 nodes. Every type of scenario is executed 10 times with different configurations. For the analysis of the results the mean of all calculated values is used.

Additionally, we assume the capacity of the nodes to be infinite for the evaluation, so we apply an uncapacitated WLP. Furthermore, we only take one service requester into account. In addition, it is possible to imagine a completely

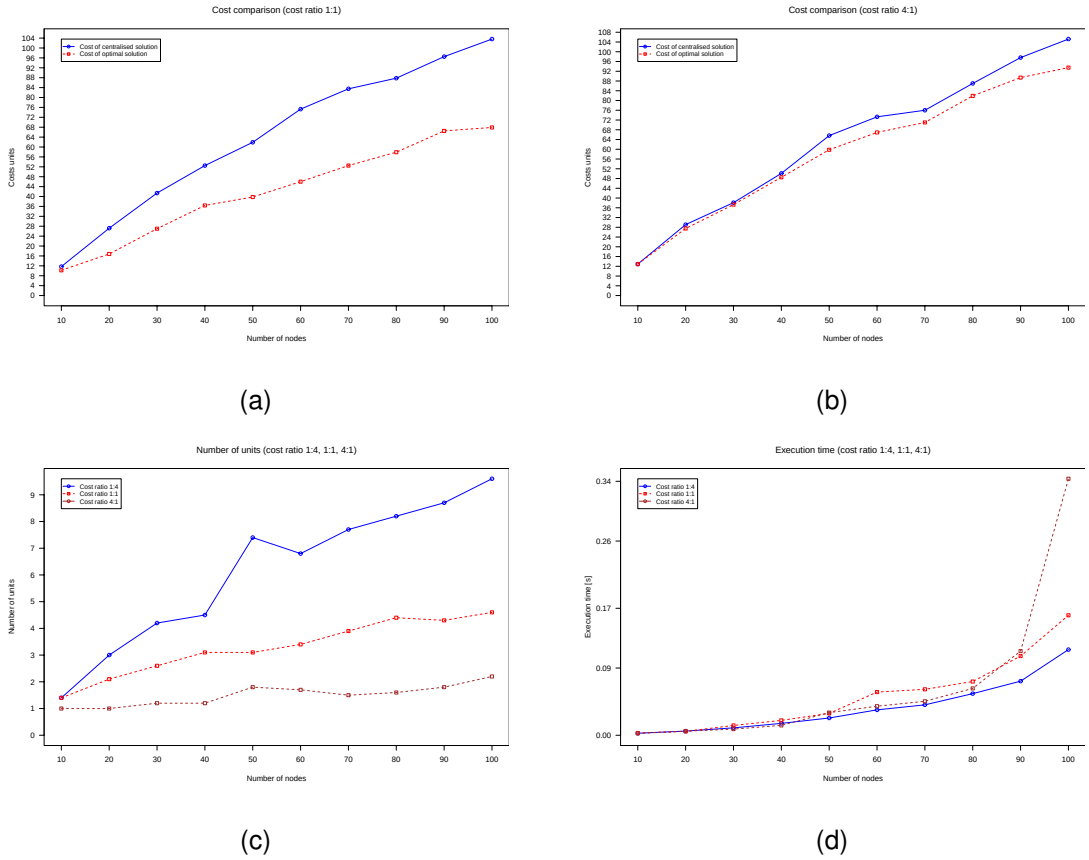


Figure 6. Cost savings, number of MAAs, and execution time samples

decentralised scenario in order to reach a high degree of scalability and robustness. However, concerning our simulation, we postulate that the alignment demand of one provider j has to be satisfied by a single node i having a MAA installed, which is located on the "shortest path" (with respect to our minimisation problem) from this provider j to the requester or by the requester itself. That means in the context of our work: $x_{ij} = d_j$.

6.2. Simulation setup

For simulation purposes, a workbench was set up which integrates topology and cost model generation with a solver capable of solving mixed integer programs as well as appropriate visualisation functionalities.

In detail, the topology generator BRITE (Boston University Representative Internet Topology Generator – cf. [14]) was modified in order to support the data formats needed by our model generator and the solver. BRITE supports the generation of topologies based on the Waxman- and the Barabasi-Albert methods – both of them well established in the research community. We used Waxman-generated topologies for the evaluation of our approach (cf. [15] for in-depth discussion of Waxman topologies), but our approach is not limited to those types of topologies. After topology generation, the topology description file is complemented by cost and demand vectors, which are generated randomly by our .NET based model generator. In order to generate real random values, the randomiser methods of the .NET cryptography API are used. The solution to the mixed integer program is calculated with the commercial XPress-MP solver (release 2007B). Results of the solver are prepared and visualised using the R statistics package (version 2.7.1). In order to automate the simulation runs, all components were linked together using shell scripts.

The simulation itself was run on an Intel Core 2 Duo 2.33 GHz computer with 2 GByte RAM running the Windows Vista operating system.

6.3. Discussion of results

In this section, we will discuss some of the results which were generated during the simulation runs.

Figure 6a shows a comparison of the overall costs of monitoring and alignment for different topologies ranging from 10 to 100 nodes. Here, 20 % of the nodes are service

providers and the cost ratio of setup costs to maximum communication costs is 1:1. As Figure 6a shows, almost every configuration has some optimisation potential by distributing monitoring and alignment units. The maximum is at 60 nodes with a 38.9 % cost reduction. At this point, 6 to 7 MAAs are needed on average to manage 12 service providers, which is depicted in Figure 6c.

Figure 6b again shows a comparison of overall costs, but in this case for a cost ratio of 4:1. Again, 20 % of the nodes are service providers and 100 random topologies ranging from 10 to 100 nodes were investigated. Figure 6b shows that cost savings are only marginal in this scenario. A maximum cost saving of 11.1 % on average can be achieved at 100 nodes. For this, at most two MAAs are needed as Figure 6c shows. Scenarios, which only place one MAA are centralised ones.

A drawback of our approach is the runtime behaviour of the mixed integer program. As noted before, a WLP like optimisation problem is NP-hard. As we can see in Figure 6d, the runtime of the solver for the mixed integer program grows exponentially – even for relatively small problem sets. Therefore, an application to real-time scenarios is only possible for small topologies – e.g., a topology of 1000 nodes takes more than 30 minutes to solve on our simulation system – or by applying heuristics, which are in scope of our further research.

7. Related work

Due to the amount of research areas touched in this article, related work out of different areas has to be taken into account. First, we discuss the area of specification languages as well as the corresponding monitoring systems, followed by the presentation of a selection of distributed monitoring approaches. Finally, we discuss the related work to the distribution of monitoring units in an infrastructure.

There are various approaches to specify monitoring requirements with respect to SLAs in the area of Web services and SOAs, almost all with appropriate system support. Robinson specifies functional monitoring requirements in temporal logic, without any support for the specification of countermeasures [16]. Again, using logic to specify functional monitoring requirements, Spanoudakis and Mahbub present a transformation of BPEL into event calculus, in which the requirements can be specified [17]. The specification of countermeasures is again not part of their approach. Sen et al. also use past time linear temporal logic for the description of monitoring requirements, without any support for countermeasure specification [18]. Furthermore, all of the logic-based approaches lack an easy readability by non-expert users.

Monitoring assertions, which are integrated in the form of pre- and post-conditions in BPEL, are discussed by Baresi and Guinea [19]. Also an approach by Baresi et al. is the Web service constraint language for the specification of functional and non-functional requirements [20]. It uses the WS-Policy language to specify requirements of users, providers, and third parties. Both approaches do not cover the handling of deviations or the specification of countermeasures, but offer framework support for integration. Lazovik et al. use business rules for the same purpose and with the same limitations [21].

The approaches presented above primarily focus on the specification of monitoring aspects. All of the examined policy and requirements languages above do not support the specification of countermeasures and therefore are not fully applicable to our scenario. An approach, which also covers basic policy enforcement aspects, is discussed by Ludwig et al. [22]. The authors use WS-Policy as a part of WS-Agreement to specify requirements in their CREMONA architecture. A focus of their work is on the initial creation and subsequent adaptation of agreements between different parties (i.e., during the negotiation of parameters), which include policy elements and their enforcement or renegotiation. A different policy language named CIM-SPL is presented in [23], which also supports the specification of countermeasures and therefore enables an enforcement of policies. CIM-SPL integrates the elements of the Common Information Model (CIM), an industry standard provided by the Distributed Management Task Force, into a policy language. The application of a heavy-weight standard like CIM as the foundation of a distributed decision making approach is currently under research.

An additional area of application for policy enforcement is the area of security. Sattanathan et al. discuss an architecture, which allows the securing of Web services by the use of adaptive security policies defining e.g., the security levels of incoming and outgoing Web service messages [24]. Here, security policies can be changed during execution time without changing the implementation. Furthermore, their architecture allows negotiation and reconciliation of security policies. Ardagna et al. also address policy enforcement issues with respect to the security of Web services [25]. In their work, an approach for the access control of Web services based on policies is presented, which also supports basic policy enforcement strategies.

Of further interest is the approach followed by Oracle with their Web Services Manager [26]. The Oracle Web Service Manager integrates centralised monitoring and policy enforcement with distributed information gathering of basic QoS parameters (e.g., response time), exceptions, and security aspects. Therefore, an architecture containing non-intrusive (i.e., gateways) and intrusive elements (i.e., agents running at the same application server as the Web service) was developed, which allows both basic reporting of monitoring results as well as the automatic selection and activation of policies based on given measurements.

With respect to distributed monitoring, there are different

approaches, which are in some parts comparable to our work, especially in their application of agent technology for the distribution of the monitoring logic. The integration of SNMP into an agent-based architecture for network management is discussed in [27]. Here, agents are responsible to collect the monitoring data from SNMP-capable data sources. Again for network management purposes, the work of [28] discusses a scalable framework based on mobile software agents. The approaches differ from our AMAS.KOM approach by not supporting deviation handling mechanisms. A further representative of agent-based monitoring approaches is discussed in [29], in which software agents act as area managers responsible for the monitoring and control of dedicated parts of a network. Area assignment is dynamic, allowing agents to adapt their zones during runtime and to migrate into the selected zone using agent mobility features.

Finally, some approaches to the distribution of monitoring units exist, which are related to our work presented in this article. All of the existing approaches are focusing on overall network monitoring, but not on monitoring of services in a SOA. An overview presenting applicable models and problem definitions for the location of network monitoring units is discussed in [30]. The authors analyse in detail what types of monitors can be matched to what kind of optimisation problems. Possible solutions to those problems are presented in addition. Furthermore, [31] present methods for the optimal positioning of monitoring units for network performance assessment. Hereby, the authors focus on the minimisation of the number of devices as well as finding their optimal location. Another approach to calculate the optimal number of monitoring units based on BGP (Border Gateway Protocol) data is presented by [32], where the authors give some theoretical foundation to define boundaries for the number of monitoring units needed in an Internet-like scenario. As a result, the authors claim that one third of all nodes in an arbitrary topology should execute monitoring functionalities in order to manage the complete infrastructure.

8. Conclusion and outlook

In this article, we presented an integrated approach for distributed SLA monitoring and enforcement in service-oriented systems. For this, we introduced the policy language WS-Re2Policy to specify requirements and countermeasures to SLA violations simultaneously. Additionally, the AMAS.KOM framework, supporting distributed monitoring and enforcement of SLAs in Web service-based cross-organisational collaborations as well as a distribution strategy to find the optimal locations of monitoring units in distributed scenarios were presented.

The AMAS.KOM framework utilises the mobile software agent paradigm to define autonomous monitoring and alignment units, which are distributed in a service-oriented system by applying our optimisation approach. The corresponding monitoring and alignment units are pre-configured using the WS-Re2Policy language. First performance evaluations showed that the overhead introduced by our agent-based approach is almost insignificant in a Web service scenario.

The evaluation of the distribution strategy proposed in this article, which calculates the optimal position of a monitoring unit with respect to the total cost, shows that cost improvements can be reached by distributing monitoring units in almost every scenario we investigated. Here, a realistic cost model is crucial because the cost ratio of setup costs to communication costs can limit potential benefits.

Currently, the WS-Re2Policy language exists in its second version and is implemented in a prototypical implementation of our AMAS.KOM framework. Nevertheless, the language is under continuous development. One of the planned major enhancements of WS-Re2Policy is the native support of various additional QoS-related parameters, as a common definition of a QoS policy is currently missing. Another focus of our ongoing work is on the improvement of the distribution strategy. As noted before, the current strategy is not applicable to large topologies under real-time conditions. At the moment, we are working on heuristics to improve the planning process.

Acknowledgements

This work is supported in part by E-Finance Lab Frankfurt am Main e.V. (http://www.efinancelab.com). In addition, parts of the research are carried out in the THESEUS TEXO project funded by means of the German Federal Ministry of Economy and Technology under the promotional reference ”01MQ07012”. The authors take the responsibility for the contents.

References

[1] N. Repp, A. Miede, M. Niemann, and R. Steinmetz, “Ws-re2policy: A policy language for distributed sla monitoring and enforcement,” in *Proceedings of the 3rd International Conference on Systems and Networks Communications*, October 2008, pp. 256–261.

[2] N. Repp, “Monitoring of services in distributed workflows,” in *Proceedings of the Third International Conference on Software and Data Technologies*, July 2008.

[3] N. Repp, J. Eckert, S. Schulte, M. Niemann, R. Berbner, and R. Steinmetz:, “Towards automated monitoring and alignment of service-based workflows,” in *Proceedings of the IEEE International Conference on Digital Ecosystems and Technologies 2008*, 2008, pp. 235–240.

[4] M. Bichler and K.-J. Lin, “Service-oriented computing,” *IEEE Computer*, vol. 39, no. 3, pp. 99–101, March 2006.

[5] A. Miede, J.-B. Behuet, N. Repp, J. Eckert, and R. Steinmetz, "Cooperation mechanisms for monitoring agents in service-oriented architectures," in *Proceedings of the 9th International Conference Wirtschaftsinformatik*, Februar 2009, pp. 749–758.

[6] J. Widom and S. Ceri, *Active Database Systems*, 1st ed. Morgan-Kaufmann, 1995.

[7] Y. Wei, S. Zhang, and J. Cao, "Coordination among multi-agents using process calculus and eca rule," in *Proceedings of the First International Conference on Engineering and Deployment of Cooperative Information Systems*, 2002, pp. 456–465.

[8] S. A. White and D. Miers, *BPMN Modeling and Reference Guide*, 1st ed. Future Strategies, 2008.

[9] R. Berbner, M. Spahn, N. Repp, O. Heckmann, and R. Steinmetz, "Heuristics for qos-aware web service composition," in *Proceedings of the 4th IEEE International Conference on Web Services*, 2006, pp. 72–79.

[10] G. Canfora, M. D. Penta, R. Esposito, and M. L. Villani, "Qos-aware replanning of composite web services," in *Proceedings of the IEEE International Conference on Web Services*, July 2005, pp. 121–129.

[11] W. Domschke and G. Krispin, "Location and layout planning: A survey," *OR Spektrum*, vol. 19, pp. 181–194, 1997.

[12] Y. Pochet and L. A. Wolsey, *Production Planning by Mixed Integer Programming*, 1st ed. Springer, 2006.

[13] R. Sridharan, "The capacitated plant location problem," *European Journal of Operational Research*, vol. 87 (2), pp. 203–213, December 1995.

[14] A. Medina, A. Lakhina, I. Matta, and J. Byers, "Brite: An approach to universal topology generation," in *Proceedings of the International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunications Systems*, 2001.

[15] B. M. Waxman, "Routing of multipoint connections," *Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 347–352, December 1988.

[16] W. Robinson, "A requirements monitoring framework for enterprise systems," *Journal of Requirements Engineering*, vol. 11, no. 1, pp. 17–41, 2005.

[17] G. Spanoudakis and K. Mahbub, "Non intrusive monitoring of service based systems," *International Journal of Cooperative Information Systems*, vol. 15, no. 3, pp. 325–358, 2006.

[18] S. Sen, A. Vardhan, G. Agha, and G. Rosu, "Efficient decentralized monitoring of safety in distributed systems," in *Proceedings of the 26th International Conference on Software Engineering*, 2004, pp. 418–427.

[19] L. Baresi and S. Guinea, "Towards dynamic monitoring of ws-bpel processes," in *Proceedings of the 3rd International Conference on Service oriented computing*, 2005, pp. 269–282.

[20] L. Baresi, S. Guinea, and P. Plebani, "Ws-policy for service monitoring," in *Proceedings of the 6th Workshop Technologies for E-Services*, 2006, pp. 72–83.

[21] A. Lazovik, M. Aiello, and M. Papazoglou, "Planning and monitoring the execution of web service requests," *International Journal on Digital Libraries*, vol. 6, no. 3, pp. 235–246, 2006.

[22] H. Ludwig, A. Dan, and R. Kearney, "Cremona: An architecture and library for creation and monitoring of ws-agreements," in *Proceedings of the 2nd International Conference on Service oriented computing*, 2004, pp. 65–74.

[23] D. Agrawal, S. Calo, K.-W. Lee, and J. Lobo, "Issues in designing a policy language for distributed management of it infrastructures," in *Proceedings ot the 10th IFIP/IEEE International Symposium on Integrated Network Management*, 2007, pp. 30–39.

[24] S. Sattanathan, N. C. Narendra, Z. Maamar, and G. K. Mostéfaoui, "Context-driven policy enforcement and reconciliation for web services," in *Proceedings of the Eighth International Conference on Enterprise Information Systems: Databases and Information Systems Integration*, 2006, pp. 93–99.

[25] C. A. Ardagna, E. Damiani, S. D. C. di Vimercati, and P. Samarati, "A web service architecture for enforcing access control policies," in *Proceedings of the First International Workshop on Views on Designing Complex Architectures*, 2006, pp. 47–62.

[26] K. Chu, O. Cordero, M. Korf, C. Pickersgill, and R. Whitmore, *Oracle SOA Suite Developer's Guide*, Oracle, 2006.

[27] M. Zapf, K. Herrmann, and K. Geihs, "Decentralized snmp management with mobile agents," in *Proceedings of the 6th IFIP/IEEE International Symposium on Integrated Network Management*, 1999, pp. 623–635.

[28] D. Gavalas, D. Greenwood, M. Ghanbari, and M. O'Mahony, "Using mobile agents for distributed network performance management," in *Proceedings of the 3rd International Workshop on Intelligent Agents for Telecommunication Applications*, 1999, pp. 96–112.

[29] A. Liotta, G. Pavlou, and G. Knight, "A self-adaptable agent system for efficient information gathering," in *Proceedings of the 3rd International Workshop on Mobile Agents for Telecommunication Applications*, 2001, pp. 139–152.

[30] K. Suh, Y. Guo, J. Kurose, and D. Towsley, "Locating network monitors: Complexity, heuristics and coverage," *Computer Communications*, vol. 29, no. 10, pp. 1564–1577, June 2006.

[31] C. Chaudet, E. Fleury, I. G. Lassous, H. Rivano, and M.-E. Voge, "Optimal positioning of active and passive monitoring devices," in *Proceedings of the 2005 ACM conference on Emerging network experiment and technology*, 2005, pp. 71–82.

[32] J. D. Horton and A. López-Ortiz, "On the number of distributed measurement points for network tomography," in *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003, pp. 204–209.

System-on-Chip Implementation of Neural Network Training on FPGA

Ramón J. Aliaga, Rafael Gadea, Ricardo J. Colom, José M. Monzó,
Christoph W. Lerche, and Jorge D. Martínez
Institute for the Implementation of Advanced Information and Communication Technologies (ITACA)
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
E-mail: *{raalva, rgadea, rcolom, jmonfer, chler, jdmartinez}@upvnet.upv.es*

Abstract—*Implementations of Artificial Neural Networks (ANNs) and their training often have to deal with a trade-off between efficiency and flexibility. Pure software solutions on general-purpose processors tend to be slow because they do not take advantage of the inherent parallelism, whereas hardware realizations usually rely on optimizations that reduce the range of applicable network topologies, or attempt to increase processing efficiency by means of low-precision data representation. This paper describes a mixed approach to ANN training, based on a system-on-chip architecture on a reconfigurable device, where a coprocessor with a large number of parallel neural processing units is controlled by software running on an embedded processor. Software control and the use of floating-point arithmetic guarantee system generality, and replication of processing logic is used to exploit parallelism. Implementation of the proposed architecture on a low-cost Altera FPGA achieves a performance of 431 MCUPS (millions of connection updates per second).*

Keywords: *artificial neural networks (ANN), backpropagation, field-programmable gate array (FPGA), multilayer perceptron (MLP), system-on-chip (SoC).*

1. Introduction

Artificial neural networks (ANNs) are bio-inspired architectures that implement parameterized non-linear functions of several variables, according to a computational structure based on mathematical models of the human brain [2][3]. The most important characteristics typically associated with them are parallelism, modularity and generalization capability [4].

Parallelism and modularity are given by the logical structure of the networks. ANNs are organized as a series of sequential layers consisting of several simple, identical computational elements, called neurons, which process the outputs from the previous layer in parallel. A sequential general-purpose processor is unable to take advantage of the high degree of parallelism in neural networks, hence hardware implementations on ASIC or reconfigurable devices are much more efficient [5].

"Generalization capability" refers to the fact that ANNs learn from example, i.e., after adjusting its parameters according to a given set of sample input-output pairs, they have good interpolation properties when presented with new, different inputs. This ability makes neural networks a popular choice for the implementation of function interpolators, estimators or predictors in real-time systems. Sample applications of ANNs include forecasting in economics [6], speech recognition [7], and medical imaging [8].

A large number of hardware architectures have been proposed for the implementation of ANNs, ranging from early analog proposals [9] to modern network-on-chip (NoC) platforms [10]. Most of the efforts on optimization of hardware ANN architectures have been concentrated on the implementation of the recall phase, i.e., of already-trained neural networks, relying on the training phase being performed off-chip using a software algorithm on a different platform. However, network training algorithms receive the same benefits from hardware parallelization.

As ANN training is much more expensive computationally, hardware realizations tend to resort to heavy optimization and simplification procedures in order to increase processing speed. This usually

implies a loss in generality of application, as the optimizations rely on the restriction of certain network parameters, especially network topology and arithmetic representation format. Fixed-point arithmetic is most common, especially 16-bit representation, which is considered the minimum precision that guarantees network generalization [11]. Studies such as [12] indicate that floating-point implementations on FPGA may be impractical in terms of resource usage; however, their ability to represent very small values with high precision translates into faster network convergence [12]. We believe that an ANN training system with wide applicability should be using floating-point arithmetic.

On the other hand, efficient implementations that focus on maximizing throughput, such as pipelined systolic arrays [13], leave little room for reconfiguration of the network topology. Most implementations can only train networks with a fixed structure; in other cases, the number of layers is fixed and only the number of neurons in each layer can be selected up to a maximum number. Changing the network topology requires system regeneration and device reconfiguration. This is a major drawback, because it is extremely difficult to determine an appropriate topology for a given problem prior to the actual training, except for very simple applications with a reduced number of inputs [14]. There exist procedures for the selection of the optimal network architecture for a problem, such as network pruning [15] or genetic algorithms [16], but all of them involve the execution of the base training algorithm on different network topologies at some point [17].

It follows that a flexible ANN training system should be able to train arbitrary network topologies with floating-point precision. One possible approach is the use of a distributed multiprocessor system with a job partitioning scheme. This is evaluated in [18] in the context of a LAN implementation, and it is shown that the optimal parallelization scheme for small networks with large training sets is the exploitation of training-set parallelism, i.e., having each processor implement the whole network functionality but work on a different subset of input data. In that case, the major cause for efficiency loss in the system is communication overhead between processing nodes.

In [1], we proposed the implementation of a similar multiprocessor system in a single FPGA, using embedded processors modified with custom parallel logic to accelerate neural computations. However, this approach resulted in limited efficiency, due to a high communication overhead given by the need of software-driven data distribution between processors, and restrictions on the custom logic, imposed by the embedded processors' architecture. In this paper, we present a refinement of the system where all neural processing is integrated in a single hardware coprocessor with a high number of parallel processing units. Data transmission and partial result combination is handled directly by dedicated hardware, and high efficiency is achieved through data and instruction pipelining and careful coding of the training algorithm, exploiting instruction-level parallelism. Training speed is significantly improved, up to 20 times faster than our previous implementation.

The paper is structured as follows. We begin by establishing the theoretical background behind ANN training in Section 2. The next section discusses our proposed hardware system architecture, describing the designed coprocessor and its integration in the whole system-on-chip. Section 4 describes software programming issues for both the master controller and the custom coprocessor. In the next section, implementation results on a specific Altera development board are presented. Finally, training performance is evaluated and conclusions are drawn.

2. Network training

Our proposed system architecture can be applied to a variety of ANN types that allow batch training operation, but our current implementation is constrained to one of the most widely used, the Multilayer Perceptron (MLP). In this section, we will describe this particular type of neural network and the training algorithms we have considered.

2.1. Multilayer Perceptron

A multilayer perceptron [19] comprises several layers (typically two or three) of similar simple processing elements, called *neurons*, which take the previous layer's neurons' outputs as inputs, as illustrated in Figure 1. Each neuron computes a weighted sum of its inputs and modifies the result by means of a bounded non-linear *activation function* φ whose purpose is to limit the range of the neuron's output. The transfer function for a neuron k is thus given by

$$v_k = b_k + \sum_j w_{jk} \cdot o_j$$
$$o_k = \varphi(v_k) \qquad (1)$$

where the sum runs over all neuron inputs, and o_j denotes the output of neuron j. The network's free

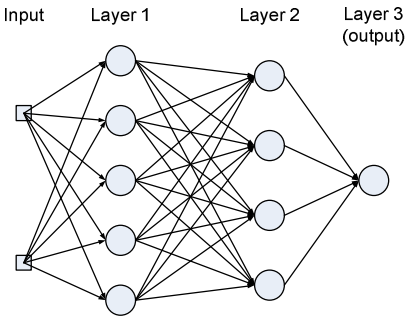


Fig. 1. A fully connected 2/5/4/1 multilayer perceptron.

parameters are the *weights* w_{jk} and *biases* b_k in each neuron.

The numbers of layers and of neurons in each layer are enough to completely describe the topology of a fully connected MLP, i.e., a network where all connections between neurons in consecutive layers are present. We shall use the notation N_0 / N_1 / … / N_M to refer to a MLP with N_0 inputs and M layers of neurons, where the i-th layer has N_i neurons. Partially connected MLPs can be thought of as special cases of MLP where some weights are forced to 0.

2.2. Backpropagation

MLP training is the process of adaptation of the free parameters in such a way that the network's global transfer function approaches some specific target behavior. This target function is defined by means of a set of V training vectors $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{V}$, representing sample inputs $\mathbf{x}_i$ and their associated desired network outputs $\mathbf{y}_i$. The set of network weights $\mathbf{W}$ is adjusted iteratively with the goal of minimizing the mean square error (MSE)

$$E(\mathbf{W}) = \frac{1}{2V}\sum_{i=1}^{V}\left\|F(\mathbf{x}_i;\mathbf{W}) - \mathbf{y}_i\right\|^2 \qquad (2)$$

where F is the ANN's transfer function, dependent on the parameters $\mathbf{W}$, and $\|\ \|$ denotes the Euclidean norm in N_M–dimensional space. In order to minimize E, computation of the gradient ∇E is needed. The most popular way to do this is the *error backpropagation* algorithm, or simply backpropagation [20], because of its efficient parallel, distributed implementation. This method of obtaining the network gradient consists of propagating neuron errors through the network layers in reverse order as follows:

Fix a training vector $(\mathbf{x}_i, \mathbf{y}_i)$. Starting at the output neurons, a *local gradient* is calculated as

$$\delta_j = \varphi'(v_j)\cdot\varepsilon_j \qquad (3)$$

where ε_j is the neuron's error, i.e., the difference between the estimated output $\varphi(v_j)$ and the desired output from the training vector. The local gradients in other layers are computed iteratively following the formula

$$\delta_j = \varphi'(v_j)\cdot\left\{\sum_k w_{jk}\cdot\delta_k\right\} \qquad (4)$$

where the sum runs over all neurons k in the *next* layer. Finally, the gradient for weight w_{jk}, connecting neuron/input j with neuron k in the next layer, is given by

$$\left.\frac{\partial E}{\partial w_{jk}}\right|_{\text{vector } i} = o_j\cdot\delta_k \qquad (5)$$

The gradient for the bias b_k is equal to δ_k.

Thus, the computations involved in the backpropagation algorithm for each training vector can be structured into three distinct phases (most authors only mention two phases; the last one is either ignored or merged with the second one):

- *Forward phase:* The outputs (and derivatives) in each neuron are computed recursively, from the first to the last layer.
- *Backward phase:* The local gradients δ_j in each neuron are computed recursively, backwards from the last to the first layer.
- *Gradient phase:* Gradients for each free parameter are computed using (5). Computations for this phase can be organized in any order.

Each individual phase may be carried out with parallel and distributed processing, however the forward and backward phases must be executed sequentially due to data dependence; failure to do so leads to a modified training algorithm [13].

The complete gradient for one *epoch*, i.e. presentation of the whole training set, is obtained by averaging the partial contributions from all training vectors:

$$\frac{\partial E}{\partial w_{jk}} = \frac{1}{V}\sum_{i=1}^{V}\left.\frac{\partial E}{\partial w_{jk}}\right|_{\text{vector } i}. \qquad (6)$$

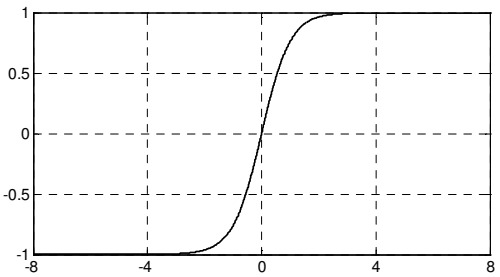


Fig. 2. Hyperbolic tangent.

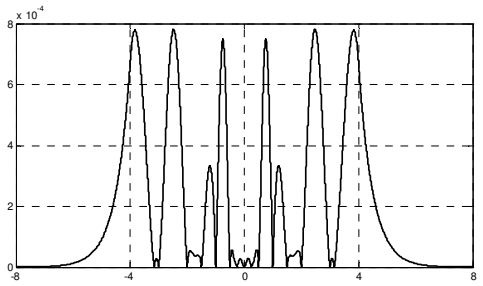


Fig. 3. Approximation error for the modified activation function.

2.3. Resilient Propagation

One common approach to weight adjustment is MSE minimization by standard gradient descent, i.e. the weights are updated by subtracting a multiple of their respective gradients from them after each epoch:

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \mu \cdot \nabla E\left(\mathbf{W}^{(n)}\right) \tag{7}$$

However, this procedure may provide very slow global network convergence due to a few neurons becoming *saturated*, i.e. having outputs close to the bounds given by the activation function and very small derivatives, leading to small weight updates between epochs, even if said weights are still far from their optimal values.

A number of modifications of the weight update mechanism have been proposed in order to address this issue, including conjugate gradient algorithms [21] and quasi-Newton methods [22]. We have selected the Resilient Propagation (RPROP) algorithm [23], where the magnitude of each weight update is kept independent of the gradient; instead, the last weight update is stored as reference and amplified or reduced depending on whether the gradient maintained or changed its sign. RPROP is reportedly faster than gradient descent by an order of magnitude, and allows a very efficient hardware implementation, in terms of both execution time and resource occupation.

2.4. Activation Function

It is a well-known fact that MLPs with at least two layers are universal approximators, i.e. they can be used to approximate any given continuous mapping with arbitrary accuracy on a bounded domain, as long as the activation function φ is bounded, monotone, and continuously differentiable [24]. Besides, convergence has been shown to be faster if φ is an odd bipolar function [25]. The most common activation function, shown in Figure 2, is the hyperbolic tangent

$$\varphi(t) = \frac{e^{t} - e^{-t}}{e^{t} + e^{-t}} \tag{8}$$

which has all of the aforementioned properties. However, this function does not lend itself to an efficient digital implementation, requiring large operators to implement exponentials and division.

Traditionally, ANN implementations have resorted to either look-up tables (LUT) or low-order approximations of φ, such as piecewise linear approximations [26], but these approaches are not viable in our situation: a LUT with floating-point precision would be too big, and piecewise linear approximations, while useful for hardware realizations of the recall phase of MLPs (i.e. of pre-trained networks with fixed weights), are inadequate for the implementation of the training phase, since they don't satisfy the hypothesis of the universal approximation theorem, thus hurting network convergence.

Our solution has been to implement a modified activation function $\tilde{\varphi}$, which is an odd cubic spline approximation of φ, with fixed exact values at abscissae 0, 0.25, 0.5, 1, 1.5, 2 and 3, saturation at 4, and fixed derivatives at the extreme points. This is a valid activation function since it satisfies all conditions stated previously, so it provides valid ANNs with correct training. This modified function allows an efficient implementation in our system architecture, based on repeated multiply-and-accumulate (MAC) operations. It can also be approximated by the hyperbolic tangent if needed, with an absolute error lower than 10^{-3}, as shown in Figure 3.

3. System architecture description

An overview of the designed system architecture and included components is presented in Figure 4. The core of the system is the neural coprocessor, with an embedded microcontroller acting as master processor,

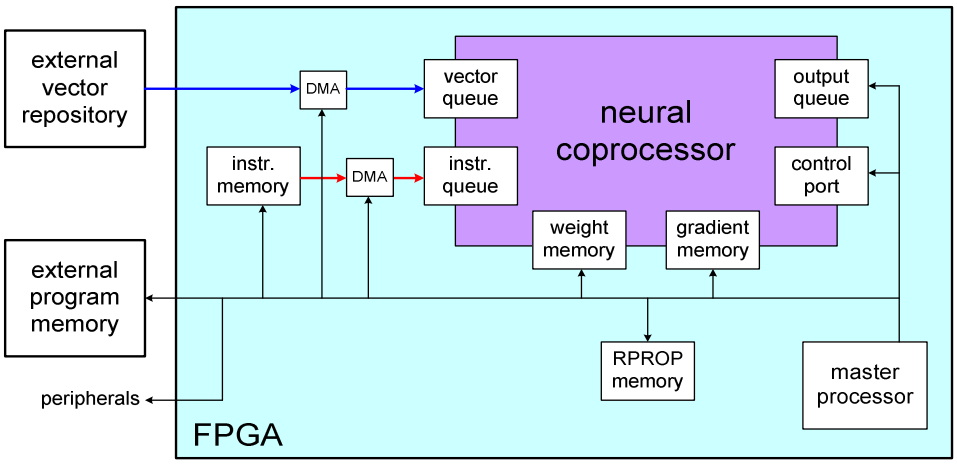


Fig. 4. System architecture.

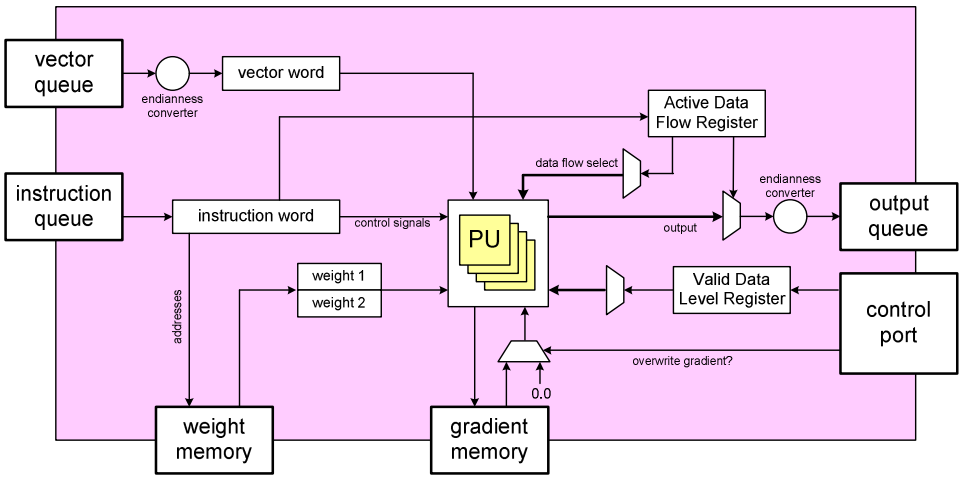


Fig. 5. Contents of the neural coprocessor.

running from a program memory block which may be external to the FPGA. Input data flows to the coprocessor, i.e. training vectors and coprocessor instructions, are fed to the coprocessor input queues using DMA devices, controlled by the master processor. Memory blocks containing current network weights and resulting gradients are integrated into the coprocessor, but are also externally accessible. All coprocessor ports are implemented as Altera Avalon slave interfaces. An external training set repository is assumed, as well as memory blocks for the storage of the coprocessor subprogram and the variables of the RPROP algorithm (weight update magnitudes and previous network gradient). The two latter should be internal to the FPGA to increase performance.

3.1. Coprocessor architecture

Figure 5 depicts the components of the neural coprocessor. It consists of a parameterized number P of arithmetic processing units (PU), restricted to a power of two to simplify logic design. All PUs execute the same operations simultaneously, and are capable of processing up to 8 different time-multiplexed data flows thanks to datapath pipelining. Each of the $8P$ supported data flows has an associated internal memory block for the storage of temporary variables; these memories can be individually read or written in order to set up the training session or retrieve results, according to an Active Data Flow Register.

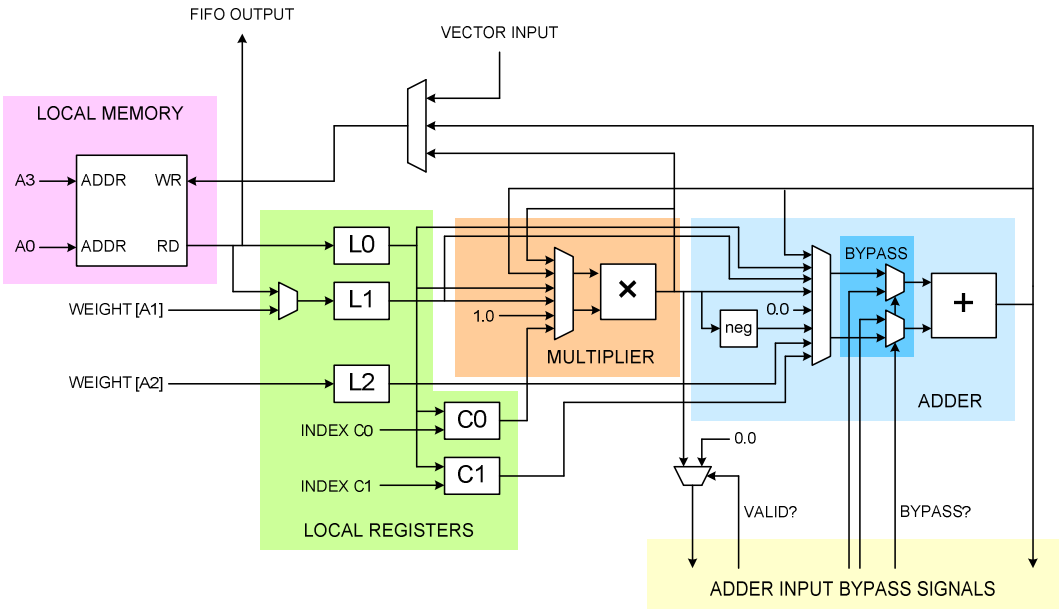


Fig. 6. RTL description of a processing unit.

The coprocessor's control port allows the master processor to access internal control registers in order to modify the coprocessor's behavior. Options include automatic endianness conversion for training vectors and results, to support communication with different remote hosts, and clearing of the gradient memory, in order to set up new training epochs. Additionally, a Valid Data Level Register specifies the number of data flows actually in use, so that results from invalid data flows may be ignored. Coprocessor status signals can also be read from the control port; they include input/output FIFO status and a "user bit" configurable through coprocessor instructions. These signals are used for execution flow control.

Figure 6 shows the schematic for the contents of each PU. It consists of an adder unit, a multiplier unit, a local memory block and a reduced number of local registers.

The adder and multiplier units are based on Altera 32-bit floating-point pipelined IP cores. Both of them implement an 8-stage pipeline and accept a new operation each clock cycle. Their inputs can be chosen from the local registers, their current outputs, and hardwired constants (1 for the multiplier and 0 for the adder) that allow them to behave as shift registers. The set of operand choices for both units has been kept small in order to reduce both the logic usage and the size of the instruction word; the backpropagation algorithm was analyzed in order to determine the smallest possible set of operand choices.

Arithmetic operators are not disabled during instructions that don't need to use them; instead, they are used as shift registers by multiplying its current value by 1 or adding 0, respectively. This is necessary because a different functional unit might need to read their outputs in successive clock cycles, corresponding to different data flows, hence these outputs must change every clock cycle.

A dual-port memory block (one read and one write port) is used as a large register bank to store variables for each data flow; the top 3 bits from each port's address input are used to select the correct data flow. A previous analysis of the backpropagation algorithm revealed that one read and one write operation per cycle are enough except in some parts of the backward phase, where two reads may be needed; however, allowing them in a single clock cycle would require a triple port memory which would have to be implemented as two dual port memory blocks in parallel, thus halving the amount of memory available for each data flow.

Each PU has three general-purpose local registers named L0, L1 and L2, implemented as 8-stage 32-bit shift registers. They are used as intermediate storage between large memories (global weight memory or local register banks) and the arithmetic units, in order to allow reuse of commonly accessed variables. Additionally, two registers C0 and C1 are used to load polynomial coefficients from a small 512 bit LUT, for the computation of the spline function $\tilde{\varphi}$ and its

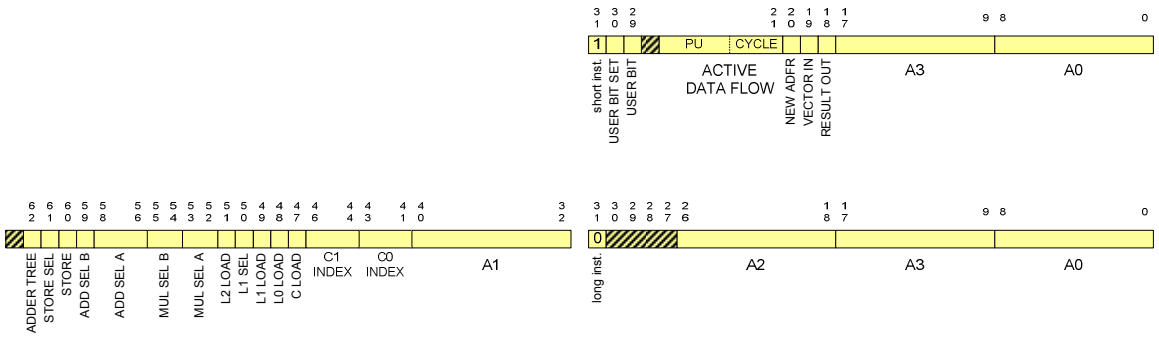


Fig. 7. Coprocessor instruction word format.

derivative; the value of the abscissa is always taken from L0. C0 is used for multiplicative and C1 for additive coefficients.

As illustrated in Figure 6, the inputs to the adder subsystem in each PU can be bypassed using external control signals, allowing the coprocessor to rearrange the existing adder logic into a multistage adder tree, which can be used to combine results from different data flows.

3.2. Coprocessor programming model

Figure 7 describes instruction word format for the neural coprocessor. There are two different types of instructions: "short" and "long" instructions, with a length of 32 and 64 bits respectively.

The purpose of short instructions is to implement communication with the coprocessor's input and output data streams by accessing individual addresses in local memory. Each instruction specifies up to three sequential tasks: updating the Active Data Flow Register, writing the contents of internal memory address A0 from the (new) active data flow into the output queue, and loading the first word from the FIFO vector queue into address A3. Additionally, the value of the "user bit" available through the control port may be changed. These instructions are consumed at a rate of one every clock cycle and implement a three stage instruction pipeline.

Long instructions are used to perform arithmetic operations on values in local memory, possibly using weights from global memory as parameters. Each instruction specifies four different operations that are executed sequentially, if enabled:

- Loading local registers with values from memory. Addresses A0 from local memory and A1 and A2 from global memory are available. Also, coefficient register C0 is loaded.
- Multiplication. Four bits are used to select input operands; a value of zero indicates "no operation", i.e., multiplying the current output with 1. Register C1 is also loaded in this stage.
- Addition. Again, four bits select the values to be added, with zeros indicating addition of the current adder output with 0.
- Storing arithmetic results (from either the adder or the multiplier) into address A3 of local memory.

This specifies four instruction pipeline stages, each one taking 8 clock cycles to complete. A new instruction is accepted every 8 clock cycles; each cycle, a new input is taken from a different data flow. Hence the coprocessor works as a Single-Instruction Multiple-Data (SIMD) machine, with the same instruction executing simultaneously on all PUs and acting on 8 time-multiplexed data flows within each PU.

The execution of long instructions follows the VLIW (Very Long Instruction Word) paradigm, in that a sequence of simple sequential operations on the execution units (adder, multiplier, registers) is specified by a single instruction, so that instruction-level parallelism is achieved at compilation time. The coprocessor hardware provides no forwarding or scheduling logic; instead, the compiler that generates coprocessor code is responsible for the optimization of the algorithm and the prevention of data hazards between consecutive instructions with data dependence, inserting NOP instructions or ad-hoc data forwarding as needed. Thus, hardware complexity is reduced at the expense of increased compilation time.

An extra bit in the instruction word allows the coprocessor to enter an extended instruction pipeline that implements an adder tree. Adder input selection is ignored in this case; instead, the bypass signals shown in Figure 6 are used to implement the accumulation of results from the multipliers in each data flow. Figure 8

schematizes the connections between multipliers and adders in both operating cases. The organization of the resulting adder tree into pipeline stages is shown in Figure 9. Accumulation of the outputs of the multipliers is divided in two different phases. The first phase uses $P-1$ adders to implement a $\log_2 P$-stage adder tree where spatial accumulation is performed, i.e., results from the same time slot in different PUs are combined. Its output is a stream of 8 partial sums coming out in consecutive clock cycles, which are accumulated in the second phase using the remaining adder and delay lines. There are eight available time slots; seven of them are used to combine of the partial sums using a 3-stage temporal tree structure, obtaining the sum of all multiplier outputs. The last time slot is used to accumulate that value with the previous value in address A2 of gradient memory. The second phase is independent of P and takes 5 eight-cycle stages to complete.

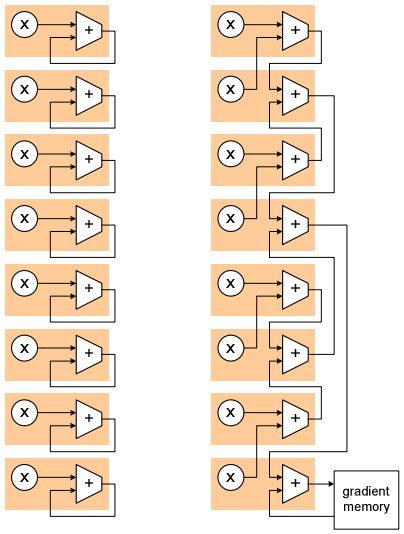


Fig. 8. Left: normal configuration. Right: Adder tree configuration.

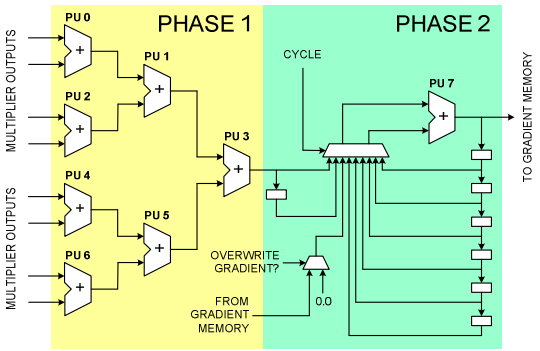


Fig. 9. Adder tree pipeline stages.

4. System software

This implementation of MLP training takes advantage of training-set parallelism, with different training vectors being handled in different coprocessor data flows using the same instructions. This obviously restricts our implementation to batch-mode training.

4.1. Backpropagation algorithm

The coprocessor code necessary for ANN training is divided into two distinct sections. The first one is a series of short instructions that read values from the vector queue and store them in the correct local memory addresses for each data flow. This way, up to $8P$ whole training vectors are loaded into the same local addresses. The second section is formed by long (arithmetic) instructions that execute both the backpropagation algorithm and the gradient accumulation process. All data flows share the same local memory map, with their input values (training vectors) being pre-loaded by the first part of the program. An extensive analysis of the backpropagation algorithm has been done in order to achieve an optimal translation into coprocessor instructions, following the three phase structure described in Section 2.2.

In the forward phase, the outputs and derivatives from all neurons in each layer are computed sequentially, according to (1). For each layer, the value of v_k is obtained first for all neurons k in that layer, using as many MAC instructions as layer inputs for each neuron. The coprocessor ability to load two different network weights is exploited to include the addition of bias b_k with no need for an extra instruction. After all v_k have been computed, the activation function and its derivative are evaluated using the Horner scheme [27], organizing arithmetic operations as a sequence of n MACs, where n is the polynomial degree:

$$\begin{aligned} \tilde{\varphi}(x) &= a_3x^3 + a_2x^2 + a_1x + a_0 \\ &= \left(\left(a_3 \cdot x + a_2\right)\cdot x + a_1\right)\cdot x + a_0 \\ \tilde{\varphi}'(x) &= b_2x^2 + b_1x + b_0 \\ &= \left(b_2 \cdot x + b_1\right)\cdot x + b_0 \end{aligned} \tag{9}$$

Both polynomials can be computed in parallel by alternating the use of the adder and the multiplier, such that in a given 8-cycle stage, the adder is evaluating one of the functions while the multiplier is evaluating the other one. This is outlined in Figure 10, where each horizontal line represents concurrent operations. This

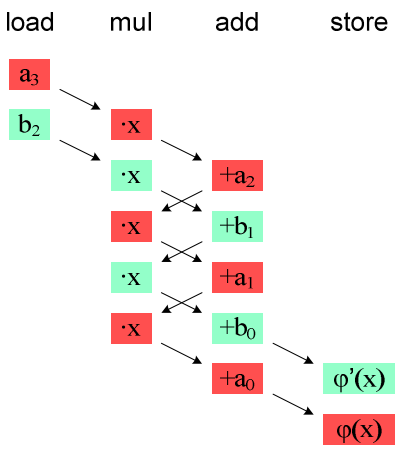


Fig. 10. Evaluation of the activation function and its derivative.

way, only 5 instructions are necessary for the computation of the activation function and its derivative.

The backward phase begins with simple multiplication operations to obtain local gradients in the output layer according to (3). For other layers, the bracketed sum in (4) needs to be computed first using as many MAC operations as neurons in the next layer; if the next layer is small enough, data forwarding has to be planned. Finally, the gradient phase makes use of the adder tree configuration of the coprocessor, computing gradients using (5) and then adding all multiplier results and accumulating them with the previous partial sum in gradient memory. Aggregated square error is also computed and accumulated; it must be divided by the total number of vectors afterwards to obtain the epoch's MSE (2).

4.2. Master processor program

Execution of whole training sessions is controlled by the master processor. The first action is initialization of network weights using the Nguyen-Widrow rule [28]; initial weights are then stored in the coprocessor's weight memory. Afterwards, coprocessor code is compiled for both code sections specified in Section 4.1 and stored in memory; this code loads $8P$ vectors, executes backpropagation on all of them simultaneously and accumulates the results in gradient memory. Coprocessor code needs to be recompiled each time because it is dependent on network topology, which is a parameter for each training session. Arbitrary MLP topologies are supported, as long as each data flow's local memory is large enough to contain all associated temporary variables.

After these initialization steps, the actual training is performed. New epochs are issued until a stop condition is fulfilled (either a maximum number of epochs is reached, or the network's MSE falls below a given threshold). For the master processor, each epoch consists of a series of DMA transfers. A transfer of the whole training set into the vector queue is first set up. After that, a number of consecutive transfers of the whole coprocessor code into the instruction queue are issued, as many times as needed to exhaust the whole training set, since only $8P$ vectors are processed each time. Before the first transfer, the control port must be used to tell the coprocessor to overwrite gradient memory instead of accumulating previous results; this option must be turned off after the first iteration. Similarly, for the last transfer, the value of the Valid Data Level Register must be updated to reflect the actual number of vectors left, so that outputs from inactive data flows are not accumulated when computing gradients. After the last instruction transfer is processed, the gradient memory contains the final network gradient of (6), except for the factor $1/V$, as well as the epoch's aggregated square error. The RPROP algorithm is now executed by the master processor to obtain the new network gradients; the multiplicative factor $1/V$ is irrelevant since only the sign of the gradient components is used.

5. Implementation

The system has been implemented on an Altera DE2-70 development board, with a Cyclone II EP2C70F896C6 reconfigurable device and external RAM memory for both the master processor program and storage of the training set; up to 64 MB are available for training vectors. The master controller is implemented as a Nios II/f embedded processor. A 16 KB internal memory for coprocessor instructions is necessary to contain the program for all applicable MLP topologies.

The limitation on the size of the coprocessor fitting into the FPGA comes from the amount of available on-chip memory blocks for the realization of local memories, hence the minimum amount of necessary memory was determined first. It was established that 128 words (4 kbit) were enough for each data flow, allowing the training of networks with up to approximately 40 neurons. Also, weight and gradient memories were limited to 2 KB, imposing a limit of 511 network weights.

Each coprocessor PU has an occupation of approximately 2970 logic elements and 10 4-kbit memory blocks, as well as four 18x18 multiplier

blocks, with full-capability floating-point operators, i.e. supporting denormal numbers and full-range rounding. It is possible to fit up to $P = 16$ processing units in the coprocessor for this FPGA, allowing the simultaneous processing of up to 128 training vectors. The coprocessor works with an internal 100 MHz clock, while the rest of the system uses a 50 MHz clock; the use of dual-port memories for every coprocessor interface makes it possible to implement independent clock domains. Thus, the coprocessor has a maximal computational performance of 3.2 GFLOPS; this value is maintained during most of the execution of the backpropagation algorithm and gradient accumulation.

The system makes use of the on-board Ethernet NIC to allow the board to be connected to a local IP network, so that a remote computer can use the FPGA system to perform any network training. The remote host needs to provide the MLP topology, the whole training set, and training stop conditions (maximum epoch number or MSE threshold). On training completion, the board returns the final network weights, the MSE evolution for all epochs, and information about total training time. Matlab has been used to implement the connection protocol and test the training results on the client PC side.

6. Performance

Training performance was evaluated for the well-known Iris plants problem [29], using a data set consisting of 150 four-dimensional inputs and three different outputs representing membership to three different classes with the values 1 and –1. The number of training epochs was fixed to 100, and the size of the training set V was modified by replicating the base data set. The number of CUPS (Connection Updates Per Second) was selected as a metric for system performance. This quantity is defined as the amount of network parameters divided by the time needed to process each training sample and update network parameters accordingly; for batch training, this is equal to

$$\mathrm{CUPS} = \frac{WV}{T_{\mathrm{epoch}}} = \frac{WVN_{\mathrm{epochs}}}{T_{\mathrm{training}}} \qquad (10)$$

where W is the number of network weights, N_{epochs} is the number of training epochs, and T_{epoch} and T_{training} are the time length of each epoch and the whole training session, respectively.

The largest topology supported by the system for the Iris problem was found to be the 4/18/18/3 MLP. Training time was measured for this network and the smaller 4/5/5/3, 4/9/8/3 and 4/12/12/3 topologies.

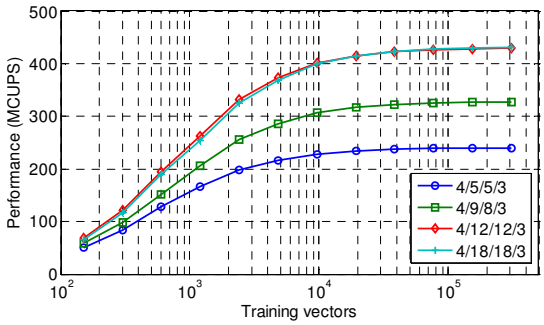


Fig. 11. System performance for different topologies and training set sizes.

Results are plotted in Figure 11. Training performance increases approaching a limit value as the training set grows larger, as expected from a system where parallelization speedup stems primarily from training set parallelism; this limit value is dependent on network topology. Since vector loading time is constant for a given problem, the fraction of execution time spent on arithmetic computations (backpropagation and gradient accumulation) is higher for more complex topologies. Hence, system performance increases as the MLP size grows, converging to a peak value for sufficiently complex networks (4/12/12/3 and higher). For our implementation, this peak value is 431 MCUPS.

7. Conclusion and future work

A system-on-chip architecture for MLP training has been proposed, where a high level processor controls the system execution flow and data transfers and generates parameterized machine code, depending on network topology, for a low level custom hardware coprocessor where the backpropagation algorithm is carried out. A SIMD architecture for the neural coprocessor is described, with a large number of replicated pipelined arithmetic operators in order to support the simultaneous processing of hundreds of training vectors. Optimized coprocessor code allows the arithmetic operators to reach near 100% utilization.

Implementation of the proposed architecture on a low-cost Altera FPGA reaches a training performance exceeding 430 MCUPS. This is a competitive value compared to other state-of-the-art FPGA implementations of MLP training with fixed topology and fixed point precision. As far as we know, no other FPGA implementation exists with floating point precision and arbitrary topology training without device reconfiguration; such features are exclusive of either software implementations on general-purpose

processors, which are slower, or ASIC implementations (so-called *neurochips*), which are much more costly.

Scalability of this architecture on larger FPGAs is constrained mainly by the amount of available on-chip memory for the realization of local memory blocks; it is probably not practical to relocate this resource to external memory because of frequent, wide access (512 bits every clock cycle for our current implementation). The addition of custom superpipelined floating-point operators might allow higher clock frequencies, although they will be ultimately be limited by FPGA routing resources. We intend to explore the possibility of using floating-point representations with less than 32 bits, as well as the adaptation of the neural coprocessor architecture to allow other types of ANN such as Radial Basis Networks.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under FPU Grant AP2006-04275 and the Education Department of the Valencian Government under Grant GVPRE/2008/082.

References

[1] R. J. Aliaga, R. Gadea, R. J. Colom, J. M. Monzó, C. W. Lerche, J. D. Martinez, A. Sebastiá, F. Mateo, "Multiprocessor SoC implementation of neural network training on FPGA", *2008 International Conference on Advances in Electronics and Micro-electronics (ENICS)*, pp. 149-154, 2008.

[2] W. S. McCulloch and W. H. Pitts, "A logical calculus of the ideas imminent in nervous activity", *Bulletin of Mathematical Biophysics*, no. 5, pp. 115-133, 1943.

[3] B. Widrow and M. E. Hoff, "Adaptive switching circuits", *IRE WESCON Convention Record*, part 4, pp. 96-104, 1960.

[4] J. Zhu and P. Sutton, "FPGA implementations of neural networks – a survey of a decade of progress", *Proceedings of the International Conference on Field Programmable Logic*, pp. 1062-1066, 2003.

[5] M. R. Zargham, *Computer Architecture: Single and Parallel Systems*, p. 346, Prentice Hall, 1996.

[6] N. L. D. Khoa, K. Sakakibara, and I. Nishikawa, "Stock price forecasting using backpropagation neural networks with time and profit based adjusted weight factors", *Proceedings of the SICE-ICASE International Joint Conference*, pp. 5484-5488, 2006.

[7] R. P. Lippmann, "Review of neural networks for speech recognition", *Neural Computation*, no. 1, pp. 1-38, 1989.

[8] R. J. Aliaga, J. D. Martinez, R. Gadea, A. Sebastiá, J. M. Benlloch, F. Sánchez, N. Pavón and C. W. Lerche, "Corrected position estimation in PET detector modules with multi-anode PMTs using neural networks", *IEEE Transactions on Nuclear Science*, vol. 53, no. 3, pp. 776-783, 2006.

[9] D. K. McNeill, C. R. Schneider and H. C. Card, "Analog CMOS neural networks based on Gilbert multipliers with in-circuit learning", *Proceedings of the 36th Midwest Symposium on Circuits and Systems*, vol. 2, pp. 1271-1274, 1993.

[10] T. Theocharides, G. Link, N. Vijaykrishnan, M. J. Irwin and V. Srikantam, "A generic reconfigurable neural network architecture as a network on chip", *Proceedings of the IEEE International SoC Conference*, pp. 191-194, 2004.

[11] J. L. Holt and T. E. Baker, "Backpropagation simulations using limited precision calculations", *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, pp. 121-126, 1991.

[12] A. W. Savich, M. Moussa, and S. Areibi, "The impact of arithmetic representation on implementing MLP-BP on FPGAs: A study", *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 240-252, 2007.

[13] R. Gadea, J. Cerdá. F. Ballester, and A. Mocholí, "Artificial neural network implementation on a single FPGA of a pipelined on-line backpropagation", *Proceedings of the 13th International Symposium on System Synthesis*, pp. 225-230, 2000.

[14] C. Xiang, S. Q. Ding and T. H. Lee, "Architecture analysis of MLP by geometrical interpretation", *2004 International Conference on Communications, Circuits and Systems*, vol. 2, pp. 1042-1046, 2004.

[15] B. Hassibi, D. G. Stork and G. J. Wolff, "Optimal Brain Surgeon and general network pruning", *Proceedings of the International Conference on Neural Networks*, vol. 1, pp. 293-299, 1993.

[16] G. F. Miller, P. M. Todd and S. U. Hedge, "Designing neural networks using genetic algorithms", *Proceedings of the 3rd International Conference on Genetic Algorithms*, pp. 379-384, 1989.

[17] X. Yao, "Evolutionary artificial neural networks", *International Journal of Neural Systems*, vol. 4, no. 3, pp. 203-222, 1993.

[18] S. Babii, V. Cretu, and E. Petriu, "Performance evaluation of two distributed backpropagation implementations", *Proceedings of the International Joint Conference on Neural Networks*, Orlando, 2007.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd Edition, pp. 156-255, Prentice Hall, 1999.

[20] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning internal representations by error propagation", *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol I: Foundations*, Chapter 8, MIT Press, Cambridge, Massachusetts, 1986.

[21] C. Charalambous, "Conjugate gradient algorithm for efficient training of artificial neural networks", *IEE Proceedings – Circuits, Devices and Systems*, vol. 139, no. 3, pp. 301-310, 1992.

[22] R. Fletcher, *Practical Methods of Optimization*, 2nd Edition, pp. 49-57, John Wiley & Sons, 1987.

[23] M. Riedmiller and M. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm", *Proceedings of the IEEE International Conference on Neural Networks*, vol. 1, pp. 586-591, 1993.

[24] K. Hornik, M. Stinchcombe and H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks*, vol. 2, pp. 359-366, 1989.

[25] Y. Le Cun, I. Kanter and S. A. Solla, "Second order properties of error surfaces: learning time and generalization", *Proceedings of the Conference on Advances in Neural Information Processing Systems*, vol. 3, pp. 918-924, 1990.

[26] V. Havel and K. Vlcek, "Computation of a nonlinear squashing function in digiral neural networks", *11th IEEE Workshop on Design and Diagnostics of Electronic Circuits and Systems*, pp. 1-4, 2008.

[27] D. Knuth, *The Art of Computer Programming, vol. 2: Seminumerical Algorithms*, 3rd Edition, pp. 486-487, Addison-Wesley, 1997.

[28] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights", *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 21-26, 1990.

[29] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals Eugenics*, pp. 179-188, vol. 7, 1936.

MODIFIED SRF-QRD-LSL ADAPTIVE ALGORITHM WITH IMPROVED NUMERICAL ROBUSTNESS

Constantin Paleologu, Felix Albu, Andrei Alexandru Enescu, and Silviu Ciochină

Telecommunications Department, University Politehnica of Bucharest, Romania
e-mail: {pale, felix, aenescu, silviu}@comm.pub.ro

ABSTRACT

The QR-decomposition-based least-squares lattice (QRD-LSL) algorithm is one of the most attractive choices for adaptive filters applications, mainly due to its fast convergence rate and good numerical properties. In practice, the square-root-free QRD-LSL (SRF-QRD-LSL) algorithms are frequently employed, especially when fixed-point digital signal processors (DSPs) are used for implementation. In this context, there are some major limitations regarding the large dynamic range of the algorithm's cost functions. Consequently, hard scaling operations are required, which further reduce the precision of numerical representation and lead to performance degradation. In this paper we propose a SRF-QRD-LSL algorithm based on a modified update of the cost functions, which offers improved numerical robustness. Simulations performed in fixed-point and logarithmic number system (LNS) implementations support the theoretical findings. Also, in order to outline some practical aspects of this work, the proposed algorithm is tested in the context of echo cancellation. It is shown that this algorithm outperforms by far the normalized least-mean-square (NLMS) algorithm (which is the most common choice for echo cancellation), especially in terms of double-talk robustness.

Index Terms— Adaptive filters, echo cancellation, fixed-point arithmetic, logarithmic number system (LNS), QR-decomposition-based least-squares lattice (QRD-LSL).

1. INTRODUCTION

The QR-decomposition-based least-squares lattice (QRD-LSL) adaptive algorithm [1]–[3] combines the good numerical properties of QR-decomposition and the desirable features of a recursive least-squares lattice. Whereas its transversal counterpart, i.e., the recursive QR-decomposition-based recursive least-squares (QRD-RLS) algorithm [2], requires a high computational load on the order of M^2 (where M is the adaptive filter order), the QRD-LSL algorithm is "fast" in the sense that the computational complexity is reduced to a linear dependence on M. This algorithm exploits the shifting property of serialized input data, i.e., the Toeplitz structure of the data matrix, to perform joint-process estimation in a fast manner. Due to its features, the QRD-LSL algorithm proves to be a very attractive choice for many applications [4]–[6].

The standard version of the QRD-LSL algorithm uses Givens rotations for implementing the QR-decomposition, which implies the use of square-root operations. In general, these operations are expensive and awkward to calculate in practice, constituting a bottleneck for overall performance. It has to be noticed that the square root operations require a large computing time, as they must be approximated by another technique, e. g., Taylor series. Consequently, the computing time increases significantly and the application areas of these algorithms become restricted. Thus, the main goal is to minimize the number of instructions within the implemented algorithm. For these reasons, square-root-free QRD-LSL (SRF-QRD-LSL) algorithms have been formulated, using special methods for performing Givens rotations without square roots [7].

Following these issues, a crucial aspect is the behavior of the adaptive algorithm in finite precision implementations. In this context, due to cost considerations, fixed-point digital signal processors (DSPs) could be preferred over floating-point ones. Quantization effects in the former result in a deviation of the adaptive filter performance from that observed in infinite precision. This deviation, which becomes more apparent as the number of representation bits is reduced, may take the form of an increased residual error after convergence (i.e., loss of precision), or more dramatically, of an unbounded accumulation of quantization errors over time (i.e., numerical instability). Moreover, there are some limitations related to the dynamic range of the algorithm's parameters. It is well known that in a fixed-point implementation context the absolute values of all involved parameters have to be smaller than one. In the case of the classical QRD-LSL algorithm, the cost functions asymptotically increase;

This work was supported by the UEFISCSU Romania under Grants PN-II-"Idei" no. 65/01.10.2007 and no. 331/01.10.2007.

they are upper bounded by the value $1/(1-\lambda)$, where λ is the exponential weighting factor (with $0 < \lambda \leq 1$). When dealing with a value of this parameter very close to one, (which is highly preferred due to stability reasons [2]) very large values of the cost functions are expected. Therefore, in order to prevent the overflow phenomena, it is necessary to perform hard scaling operations; this could lead to significant degradation of the algorithm's performance due to the decreased precision of numerical representation.

Most contemporary microprocessors perform real arithmetic using the floating-point system. Floating-point circuits are large, complex and much slower than fixed-point units; they require separate circuitry for the add/ subtract, multiply, divide, and square-root operations. All floating-point operations are liable to a maximum half-bit rounding error. A recent alternative to the floating point is the logarithmic number system (LNS). The LNS performs the real multiplication, division, and square-root at fixed-point speed [8].

In this paper, we propose a version of the SRF-QRD-LSL algorithm based on a modified update formula of the cost functions. The main issue is that the maximum value of the cost functions will be at initialization and then they will asymptotically decrease to a lower bound. Therefore, the scale factors are less critical, leading to an increased precision of numerical representation [9].

In [10] we have analyzed some versions of the QRD-LSL algorithm according to some ITU standard requirements concerning echo cancellers. The experiments indicated that the QRD-LSL algorithms fulfill by far the requirements of the ITU G.168 recommendation [11] concerning the steady-state echo return loss enhancement and convergence speed. A special interest was given to the problem of the behavior of the algorithms during the double-talk periods. Two distinct effects were identified, i.e., 1) the incomplete attenuation of the far-end signal and 2) the unwanted attenuation of the near-end signal, as a result of the near-end signal leakage to the output of the adaptive filter through the error signal. Using a value of λ very close to 1 can reduce the last effect. The experiments prove that it is possible to work with such a high value of λ by preserving in the same time a convergence speed that fulfils the requirements of the ITU recommendation. Generally, one can assert that this class of algorithms is much more robust to double-talk than the normalized least-mean-square (NLMS) type algorithms. The QRD-LSL algorithms could satisfactorily operate even in the absence of a double-talk detector (DTD). Since the proposed SRF-QRD-LSL algorithm is suitable for fixed-point implementation we present a network echo canceller based on this algorithm, implemented on a fixed-point DSP.

The rest of the paper is organized as follows. In Section 2, we briefly present the original SRF-QRD-LSL algorithm and we discuss some fixed-point and LNS implementation aspects. The proposed algorithm is developed in Section 3. Some backgrounds of echo cancellation are given in Section 4. The experimental results are given in Section 5. The simulations performed in both fixed-point DSP and LNS implementations prove the theoretical findings; also, the experiments performed in echo cancellation context outline the practical aspects. Finally, Section 6 concludes this work.

2. NUMERICAL ISSUES OF THE SRF-QRD-LSL ALGORITHM

Among the versions of the SRF-QRD-LSL algorithm presented in the literature we have chosen the most frequently used one [12]. This version is described below.

Initialization:

$J_m^f(0) = J_m^b(0) = \delta$ - positive constant

$k_m^f(0) = k_m^b(0) = k_m^c(0) = 0$

$e_0^f(n) = e_0^b(n) = x(n)$ - input signal

$e_0(n) = d(n)$ - desired signal, $\alpha_0(n) = 1$

For time moment n and filter order m compute:

- *Lattice part*:

$$J_m^f(n) = \lambda J_m^f(n-1) + \alpha_m(n-1)\left|e_m^f(n)\right|^2$$

$$c_m^f(n) = \lambda \frac{J_m^f(n-1)}{J_m^f(n)}$$

$$s_m^f(n) = \frac{\alpha_m(n-1)e_m^f(n)}{J_m^f(n)}$$

$$J_m^b(n) = \lambda J_m^b(n-1) + \alpha_m(n)\left|e_m^b(n)\right|^2$$

$$c_m^b(n) = \lambda \frac{J_m^b(n-1)}{J_m^b(n)}$$

$$s_m^b(n) = \frac{\alpha_m(n)e_m^b(n)}{J_m^b(n)}$$

$$e_{m+1}^f(n) = e_m^f(n) + k_m^{f*}(n-1)e_m^b(n-1)$$

$$k_m^f(n) = c_m^b(n-1)k_m^f(n-1) - s_m^b(n-1)e_m^{f*}(n)$$

$$e_{m+1}^b(n) = e_m^b(n-1) + k_m^{b*}(n-1)e_m^f(n)$$

$$k_m^b(n) = c_m^f(n)k_m^b(n-1) - s_m^f(n)e_m^{b*}(n-1)$$

- *Ladder part*:

$$e_{m+1}(n) = e_m(n) - k_m^{c*}(n-1)e_m^b(n)$$

$$k_m^c(n) = c_m^b(n)k_m^c(n-1) + s_m^b(n)e_m^*(n)$$

$$\alpha_{m+1}(n) = \alpha_m(n-1) - \frac{\alpha_m^2(n-1)\left|e_m^f(n)\right|^2}{J_m^f(n)}$$

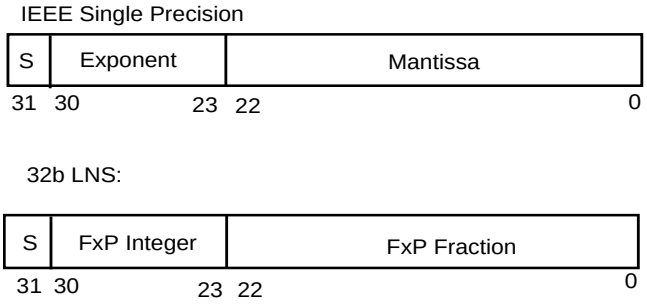


Fig. 1. IEEE standard single precision floating point representation and the 32-bit LNS format.

The superscript * denotes complex conjugation. The parameters involved in the algorithm are denoted as follows (for prediction order m and time index n):

- $J_m^b(n), J_m^f(n)$ - sum of weighted backward/forward prediction error squares (i.e., the cost functions);
- $e_m^b(n), e_m^f(n)$ - backward/forward prediction error;
- $k_m^b(n), k_m^f(n)$ - lattice structure coefficients;
- $c_m^b(n), c_m^f(n), s_m^b(n), s_m^f(n)$ - Givens parameters;
- $e_m(n)$ - joint-process a posteriori estimation error;
- $\alpha_m(n)$ - joint-process a priori estimation error;
- $k_m^c(n)$ - ladder structure coefficients;
- λ - exponential weighting factor.

Next, let us assume the context of a fixed-point DSP implementation, with a word length of B bits. The absolute values of the algorithm's parameters have to be less than one. For simplicity, we will take into consideration only the forward prediction part of the algorithm; the same analysis could be straightforwardly extended for the backward prediction part. According to the update equation of the cost function, the upper bound limit is

$$J_m^f(n)\Big|_{n\to\infty} = \frac{1}{1-\lambda} \tag{1}$$

For a value of λ very close to one, the cost functions reach to very large values. Therefore, a scaling operation (i.e., right shift) has to be applied, such that

$$J_m^f(\infty)2^{-Q} = 1 \tag{2}$$

The parameter Q is the scaling factor that is computed as

$$Q = \left\lceil -\log_2(1-\lambda) \right\rceil \tag{3}$$

where $\lceil \bullet \rceil$ denotes superior integer part.

Table 1. LNS arithmetic operations.

x + y	ADD	Lz = Lx + log(1+2^(Ly-Lx)), Sz depends on sizes of x,y
x - y	SUB	Lz = Lx + log(1-2^(Ly-Lx)), Sz depends on sizes of x,y
x * y	MUL	Lz = Lx + Ly, Sz = Sx OR Sy
x / y	DIV	Lz = Lx - Ly, Sz = Sx OR Sy
x^2	SQU	Lx << 1, Sz = Sx
x^0.5	SQRT	Lx >> 1, Sz = Sx
x^-1	RECIP	Lz = Lx, Sz = -Sx
x^-0.5	RSQRT	Lz = Lx >> 1, Sz = -Sx

For example, a value of $\lambda = 0.99$ implies $Q = 7$ bits. The numerical precision loss in this case is important. Assuming a word length $B = 16$ bits we can notice that the numerical precision representation is almost half reduced. Moreover, the small parameters from the algorithm (e.g., estimation errors) will be affected because they are "pushed" to the lowest limit 2^{-B+1}, increasing the stalling probability.

We should note that a value of λ very close to one is not unusual in practice. In general, mainly due to stability reasons, the weighting factor for this type of algorithms is greater than $\lambda = 1 - 1/3M$ [12]. Accordingly, a filter with the order $M = 33$ is sufficient to produce such a value of λ.

As an alternative to floating-point, the LNS offers the speed advantages when implementing algorithms with many multiplication, division, and square-roots; in the case of multiplication and division operations there are not rounding errors at all. These advantages are, however, offset by the problem of performing logarithmic addition and subtraction.

The 32-bit floating-point representation consists of a sign, 8-bit biased exponent, and 23-bit mantissa. The LNS format is similar in structure, as shown in Fig. 1. The 'S' bit again indicates the sign of the real value represented, with the remaining bits forming a 31-bit fixed-point word in which the size of the value is encoded as its base-2 logarithm in 2's complement format. Since it is not possible to represent the real value zero in the logarithmic domain, the 'spare' (most negative) code in the 2's complement fixed-point part is used for this purpose, which is convenient since smaller real values are represented by more negative log-domain values. The chosen format compares favorably against its floating-point counterpart, having greater range and slightly smaller representation error [8]. A 20-bit LNS format is similar. It maintains the same range as the 32-bit, but has precision reduced to 11 fractional bits. This is comparable to the 16-bit formats used on commercial DSP devices. The LNS arithmetic operations are presented in Table 1. More details about the LNS and some of its applications are available in [13], [14].

3. MODIFIED SRF-QRD-LSL ALGORITHM

In order to overcome the large dynamic range of the algorithm parameters we propose to update the cost functions in sort of "reverse" manner. Let us focus on the forward prediction part of the algorithm, where we can rewrite the cosine Givens rotation parameter as

$$c_m^f(n)=\frac{\lambda J_m^f(n-1)}{J_m^f(n)}=\frac{\lambda J_m^f(n-1)}{\lambda J_m^f(n-1)+\alpha_m(n-1)\left|e_m^f(n)\right|^2}=$$

$$=\frac{1}{1+\alpha_m(n-1)\left|e_m^f(n)\right|^2\bar{J}_m^f(n-1)} \quad (4)$$

where

$$\bar{J}_m^f(n-1)=\frac{1}{\lambda J_m^f(n-1)} \quad (5)$$

Similarly, the sine Givens rotation parameters can be expressed as

$$s_m^f(n)=\frac{\alpha_m(n-1)e_m^f(n)}{\lambda J_m^f(n-1)\left[1+\frac{\alpha_m(n-1)\left|e_m^f(n)\right|^2}{\lambda J_m^f(n-1)}\right]}=$$

$$=\alpha_m(n-1)e_m^f(n)\bar{J}_m^f(n-1)c_m^f(n) \quad (6)$$

Finally, according to (4), the update of the modified cost function from (5) becomes

$$\frac{1}{\lambda J_m^f(n)}=\frac{1}{\lambda}\cdot\frac{1}{\lambda J_m^f(n-1)}\cdot\frac{1}{1+\frac{\alpha_m(n-1)\left|e_m^f(n)\right|^2}{\lambda J_m^f(n-1)}} \quad (7)$$

so that

$$\bar{J}_m^f(n)=\frac{1}{\lambda}\bar{J}_m^f(n-1)c_m^f(n) \quad (8)$$

The initial value of the modified cost function will be chosen as

$$\bar{J}_m^f(0)=\frac{1}{\lambda\delta} \quad (9)$$

The backward prediction part of the algorithm can be modified in a similar manner. In the ladder part of the algorithm the update of the a priori estimation error has to be rewritten as

$$\alpha_{m+1}(n)=\alpha_m(n-1)-\lambda\alpha_m^2(n-1)\left|e_m^f(n)\right|^2\bar{J}_m^f(n) \quad (10)$$

Concluding, the first six relations from the lattice part of the original SRF-QRD-LSL algorithm described in Section 2 have to be changed as follows:

$$c_m^f(n)=\frac{1}{1+\alpha_m(n-1)\left|e_m^f(n)\right|^2\bar{J}_m^f(n-1)}$$

$$\bar{J}_m^f(n)=\frac{1}{\lambda}\bar{J}_m^f(n-1)c_m^f(n)$$

$$s_m^f(n)=\lambda\alpha_m(n-1)e_m^f(n)\bar{J}_m^f(n)$$

$$c_m^b(n)=\frac{1}{1+\alpha_m(n)\left|e_m^b(n)\right|^2\bar{J}_m^b(n-1)}$$

$$\bar{J}_m^b(n)=\frac{1}{\lambda}\bar{J}_m^b(n-1)c_m^b(n)$$

$$s_m^b(n)=\lambda\alpha_m(n)e_m^b(n)\bar{J}_m^b(n)$$

We may notice that a slight modification was performed in (6) according to (8). In addition, the last equation from the ladder part of the original algorithm (see Section 2) has to be replaced by (10); the initial value of the modified cost functions is given by (9). In this manner, it results a modified SRF-QRD-LSL (MSRF-QRD-LSL) algorithm.

The proposed algorithm is mathematical equivalent with the original one, so that they will have the same behaviour in infinite precision. Nevertheless, in a fixed-point arithmetic context they will behave differently, as we will demonstrate in the following.

As we have shown in Section 2, the cost functions of the original algorithm asymptotically grow to very large values, when the value of λ is close to one. Thus, hard scaling operations are required in order to avoid overflows. On the other hand, the modified cost functions of the proposed MSRF-QRD-LSL algorithm are updated in a reverse manner, so that they will asymptotically decrease to a lower bound, which can be computed as

$$\bar{J}_m^f(n)\Big|_{n\to\infty}=\frac{1-\lambda}{\lambda} \quad (11)$$

The maximum value of these functions will be the initial one given in (9). Consequently, we have to impose a scale factor such that

$$\frac{1}{\lambda\delta}2^{-S}=1 \tag{12}$$

which leads to

$$S=\left\lceil -\log_2\left(\lambda\delta\right)\right\rceil \tag{13}$$

The value of the initialization parameter δ slightly influences only the initial convergence of the algorithm [15], so that it is less important than the parameter λ. Typically, we can set $\delta = 1$. Consequently, for a value of the weighting factor $\lambda = 0.99$, the value of the scale factor from (13) is $S \cong \lceil -0.015 \rceil = 1$ bit, which is insignificant. Thus, in practice we may set the initial value of the modified cost functions to one, so that there is almost no need for the scale factor S any more.

A second aspect that we have to take into account is related to the lower bound from (11). As we may notice from the update (8), the algorithm stalls when the modified cost functions decrease under the lowest limit 2^{-B+1}. In order to prevent this problem we have to impose

$$\frac{1-\lambda}{\lambda}\geq 2^{-B+1} \tag{14}$$

From (14), the condition for the maximum value of the weighting factor results as

$$\lambda \leq \frac{1}{1+2^{-B+1}} \tag{15}$$

For example, if the word length is $B = 16$ bits, we have to choose $\lambda \leq 0.999969$, which is not a severe limitation. Concluding, if we initialize the cost functions with the value one and if we take into account condition (15), there is almost no need for scaling operations in the proposed MSRF-QRD-LSL algorithm. Consequently, its numerical robustness is improved as compared to the original SRF-QRD-LSL algorithm.

4. BACKGROUNDS OF ECHO CANCELLATION

In practice, there is a need for network echo cancellers for echo paths with long impulse response. Therefore, long FIR adaptive filters (e.g., $M \geq 256$) are required. It is well known that the longer impulse response implies slower convergence rate, thus rendering traditional algorithms like NLMS inadequate. Based on convergence performance alone, a RLS-based algorithm is clearly the algorithm of choice. However, the requirements of an echo canceller are for both rapid convergence and a low computational cost. Thus, a highly desirable algorithm is a low cost (i.e., fast) RLS algorithm. On the other hand, another consideration for this application is the algorithm stability, because it is unacceptable for the algorithm to diverge unexpectedly from the true solutions. Taking these two aspects into account, the proposed MSRF-QRD-LSL algorithm could be feasible for real-time applications, like network echo cancellation.

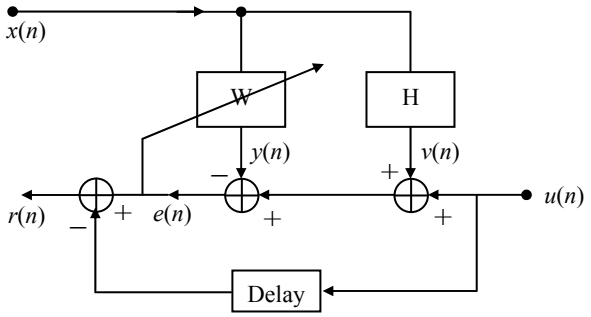


Fig. 2. Echo cancellation configuration.

Besides convergence rate and complexity issues, an important aspect of an echo canceller is its performance during double-talk. In the case of NLMS-based algorithms, the presence of near-end signal considerably disturbs the adaptive process. To eliminate the divergence of echo cancellers the standard procedure is to inhibit the weight updating during the double-talk. The presence of double-talk is detected by a DTD. A number of samples are required by the DTD to detect the double-talk presence. Nevertheless, this very small delay can generate a considerable perturbation of the echo estimate.

Let us consider the "interference cancellation" configuration from Fig. 2, which is the basis of echo cancellation. The purpose of the scheme is to extract the signal $u(n)$ from the mixture $u(n) + v(n)$. In the case of an echo canceller, $x(n)$ is the far-end signal, $u(n)$ is the near-end, H is the echo path equivalent to a FIR filter with the impulse response $\mathbf{h}(n)$, and W is an adaptive filter, having the coefficients $\mathbf{w}(n)$. As was suggested in [16], to make more apparent in results, it is convenient to subtract out the direct near-end component from the error signal $e(n)$. In this manner, the residual error $r(n)$ cumulates the undesired attenuation of the near-end signal $u(n)$ and the imperfect rejection of the echo path response $v(n)$. In a real application such a subtraction can never be done because the signal $u(n)$ is not available.

In the real case of any adaptive algorithm the coefficients $\mathbf{w}(n)$ depend on the signal $u(n)$. As a consequence, two effects appear [6], [10]:

- $\mathbf{w}(n)$ differs to $\mathbf{h}(n)$ in a certain extent and this may be viewed as a divergence of the algorithm; as a direct consequence, will result a decrease of the echo return loss enhancement (ERLE).
- $y(n)$ will contain a component proportional to $u(n)$, that will be subtracted from the received signal. This phenomenon is in fact a leakage of the $u(n)$ in $y(n)$,

through the error signal $e(n)$; the result consists of an undesired attenuation of the near-end signal.

In the case of the RLS-based algorithms, this leakage process is important for lower values of λ, where $y(n) \approx v(n) + u(n)$, and is practically absent for $\lambda \approx 1$, where $y(n) \approx v(n)$. On the other hand, when λ is very close to one, we deal with the finite-precision effects presented in Section 2. These practical aspects motivate the use of the proposed MSRF-QRD-LSL instead of the classical algorithm.

5. SIMULATION RESULTS

- *Fixed-point and LNS implementations*

For the first set of experimental results we consider a "system identification" configuration. In this class of applications an adaptive filter is used to provide a linear model that represents the best fit (in some sense) to an unknown system. The adaptive filter and the unknown system are driven by the same input. The unknown system output supplies the desired response for the adaptive filter. These two signals are used to compute the estimation error, in order to adjust the filter coefficients.

In a first experiment, the input signal is a random sequence with an uniform distribution on the interval (–1;1). The order of the adaptive filter is $M = 32$ and it is equal to the order of the system that has to be identified. We compare the original SRF-QRD-LSL algorithm presented in Section 2 with the proposed MSRF-QRD-LSL algorithm. The parameters of the algorithms were fixed to $\lambda = 0.99$ and $\delta = 1$. This set of simulations was run on a fixed-point DSP, using a word length $B = 16$ bits. The results are presented in Fig. 3. Following the discussions from Sections 2 and 3, it can be seen that the precision loss (because of the scale factors) disturbs the behavior of the SRF-QRD-LSL algorithm (Fig. 3 - upper; the error starts to grow after 3000 iterations). In the same context, the proposed MSRF-QRD-LSL algorithm achieves good performances (Fig. 3 - lower; the algorithm is stable after 3000 iterations) due to its improved numerical robustness.

A comparison of the algorithms performance on 32-bit floating point, 32-bit LNS, and 20-bit LNS implementations is performed in a similar "system identification" scheme. The input signal was generated as a first-order AR process with a correlation matrix eigenvalue spread of 20. The standard deviation of the input was 0.1 and the standard deviation of measurement noise was 0.001. The parameters of the algorithms were identical to those of the previous example. An accurate standard for comparison of the outputs of both algorithms LNS implementation was obtained by presenting the input data to their double precision versions and compute the absolute sum of errors of the 20-bit or 32-bit LNS outputs.

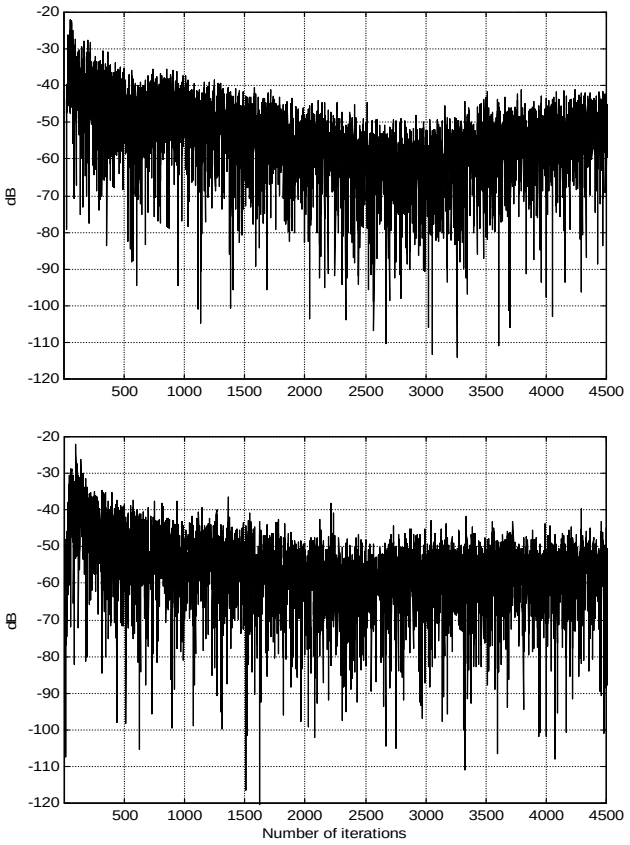


Fig. 3. Square error [dB], fixed-point $B = 16$ bits, $\lambda = 0.99$, δ=1, algorithms: (*upper*) SRF-QRD-LSL, (*lower*) MSRF-QRD-LSL.

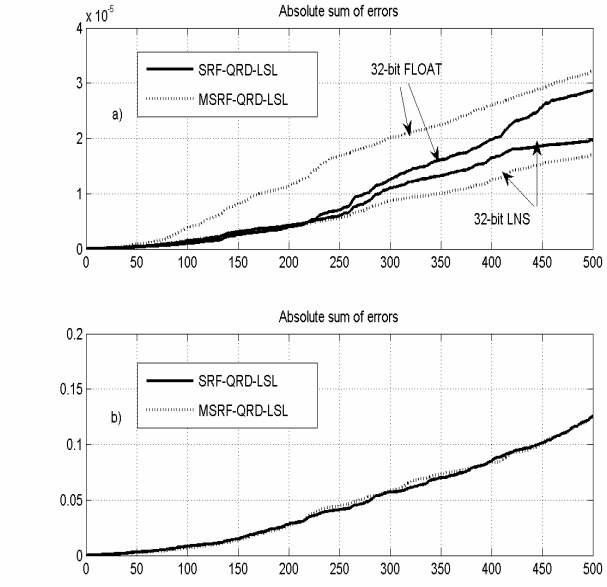


Fig. 4. (*upper*) The absolute sum of errors for 32-bit LNS and FLOAT implementations of the investigated algorithms; (*lower*) The absolute sum of errors for 20-bit LNS implementations of the investigated algorithms.

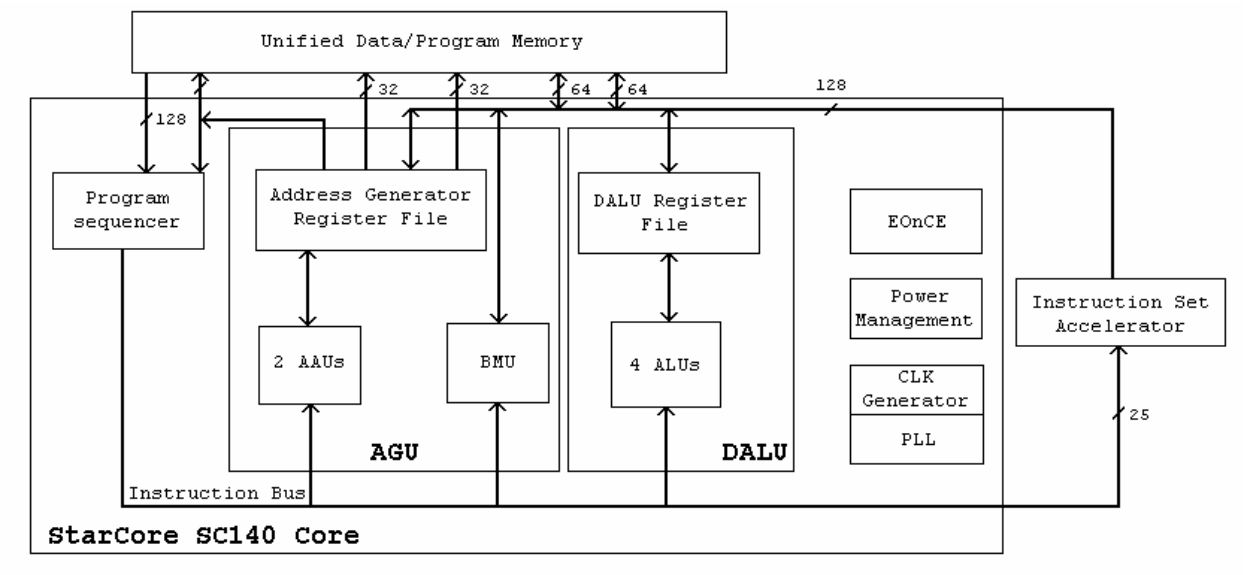


Fig. 5. Block Diagram of SC140 Core.

The 32-bit LNS and 32-bit floating-point results were virtually identical (see the amplitude of the error in Fig. 4 - upper). The absolute sum of errors for the 32-bit LNS implementation (solid line in Fig. 4 - upper) is smaller than that of the 32-bit floating-point implementation (dotted line). This is consistent with other results reported on [13] and [14] using a similar class of algorithms. It can be seen from Fig. 4 that the LNS implementations of both algorithms have only slightly different performances. As expected, the absolute sum of errors for the 20-bit LNS implementation is much higher than that of the 32-bit LNS implementation. However, for some applications that do not require much precision, the 20-bit LNS version could be an attractive alternative. As expected, the proposed algorithm does not offer advantages over the classical one when floating-point implementation is used.

- *DSP implementation of a network echo canceller*

For the practical implementation of the echo canceller we chose a well-known processor, i.e., Motorola SC140 DSP, with a large number of million instructions per second (MIPS) and a parallel architecture that allows several instructions to be executed simultaneously. In addition, we have structured the algorithm in a way to allow a high complexity algorithm to be run in real-time in a specific application. The specific features of this architecture [17] are the following (see Fig. 5):

- High level abstraction of the Application Software: applications development in C language; hardware supported integer and fractional data.
- Scalable performance: 4 Arithmetic logic Units (ALUs) and 2 Address Arithmetic Units (AAUs); 4 million multiply and accumulate operations per second (MMACS) for each megahertz of clock frequency.

The core important features are:

- Up to 10 RISC MIPS for each megahertz of clock frequency;
- A true (16*16) + 40 → 40-bit MAC unit in each ALU;
- A true 40-bit parallel barrel shifter in each ALU;
- 16 x 40-bit data registers for fractional and integer data operand storage;
- 16 x 32-bit address registers (8 can be used as 32-bit base address registers);
- 4 address offset registers and 4 modulo address registers;
- Unified data and program memory space (Harvard architecture);
- Byte addressable data memory.

However, the main feature that we have already mentioned is the C compiler and the ability to convert C source code into assembly code. The complexity of the MSRF QRD-LSL algorithm is quite large and therefore the need for flexibility is important, since programming in C code is much easier than implementing the algorithm direct in assembly code. The C compiler supports ANSI C standard and also intrinsic functions for ITU/ETSI primitives. Assembly code integration is also possible, which optimizes supplementary the code.

One of our main goals is to minimize the number of cycles needed by the algorithm per iteration, in order to lower the computational time per iteration under the sampling time of the CODEC. If we take advantage of the fact that the structure of the algorithm is symmetrical (i.e., similarities between the forward prediction structure and the backward prediction structure) then we can use two identical blocks for each lattice cell; thus, we can call twice a function in C language during one iteration. The filtering part is included in backward prediction part and is

performed if a flag is set. This flag is set before the backward prediction and reset before the forward prediction. Another optimization technique accomplished using this procedure is that all the transformations are made in-place, regardless of the iteration (i.e., the moment of time), saving a large amount of memory. Choosing an appropriate level of optimization from the C compiler, i.e., Code Warrior (0 – 3), makes further optimization. As well, the proper use of intrinsic functions from C compiler can further reduce the number of cycles.

The standard ITU-T G.168 [11] recommends certain test procedures for evaluating the performance of echo cancellers. Test signals used are so-called composite source signals (CSS) that have properties similar to those of speech with both voiced and unvoiced sequence as well as pauses. Moreover, we choose a long network echo path (i.e., 64 ms) according to the same above recommendation. The impulse response and the corresponding magnitude function of the hybrid are shown in Fig. 6. The echo return loss (ERL) of this echo path is about 10 dB and it is considered typical.

The first experiment refers to the convergence rate and echo return loss enhancement. In Figs. 7 and 8, we present the convergence results obtained by the NLMS (using a normalized step-size $\mu = 1$) and the MSRF-QRD-LSL (with $\lambda = 0.9999$) algorithms. For ITU recommendation G.168 testing purposes, the method defined for measuring the level of the signals is a root mean square (RMS) method, using a 35 ms rectangular sliding window. The measurement device comprises a squaring circuit and an exponential filter (35 ms, 1-pole). The following conclusions are obvious:

- the convergence speed and ERLE are clearly superior in the case of the MSRF-QRD-LSL algorithm;
- test 1 (steady-state and residual echo level test) is by far fulfilled in the case of the MSRF-QRD-LSL algorithm;
- the requirements of the recommendation test 2B (convergence speed) are accomplished in the case of the MSRF-QRD-LSL algorithm for time far less than one second.

The second experiment was performed using speech as excitation signals in order to simulate a real-world conversation. It evaluates the performances of the echo canceller for a high level double-talk sequence (similar to test 3B). The double-talk level in this case is about the same as that of the far-end signal. The results are presented in Figs. 9 and 10. In the case of SRF QRD-LSL algorithm one can see that the near-end signal $u(n)$ is recovered in $e(n)$ with slight distortions. The NLMS based echo canceller (using Geigel DTD [18]) fails in this situation because double-talk appears during initial convergence phase so that the adaptive process is prematurely inhibited. Let us remind that we do not use any DTD in our echo canceller based on the MSRF-QRD-LSL algorithm.

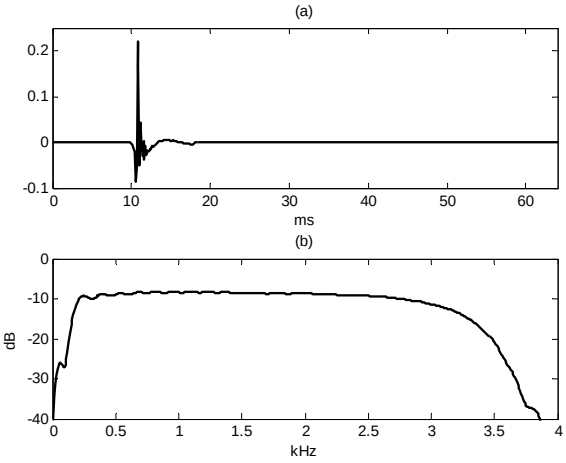


Fig. 6. Network echo path characteristics: (a) impulse response; (b) frequency response.

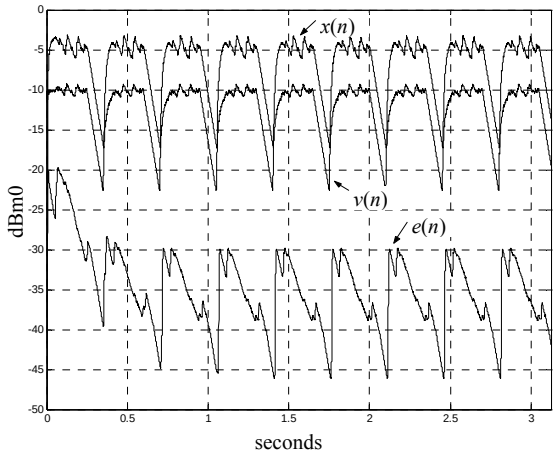


Fig. 7. Power levels [dBm0] for the far-end signal $x(n)$ (CSS), the echo signal $v(n)$, and the error signal $e(n)$, in the case of the NLMS algorithm.

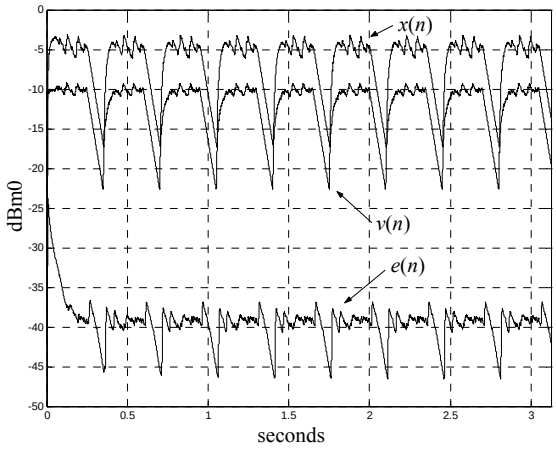


Fig. 8. Power levels [dBm0] for the far-end signal $x(n)$ (CSS), the echo signal $v(n)$, and the error signal $e(n)$, in the case of the MSRF-QRD-LSL algorithm.

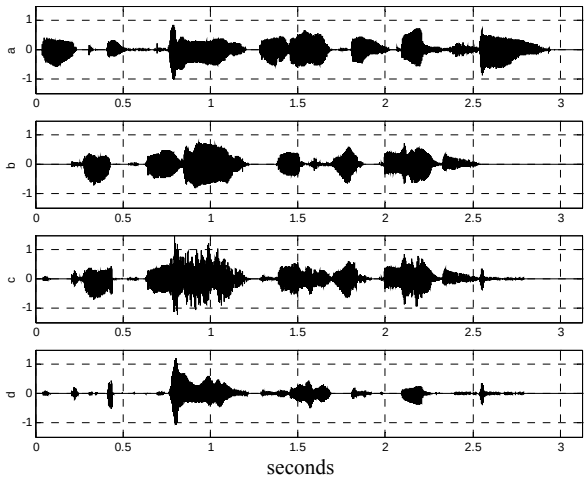


Fig. 9. Performances of the NLMS algorithm for speech signals, during double-talk: (a) far-end signal $x(n)$; (b) near-end signal $u(n)$; (c) recovered near-end signal in $e(n)$; (d) residual error $r(n)$.

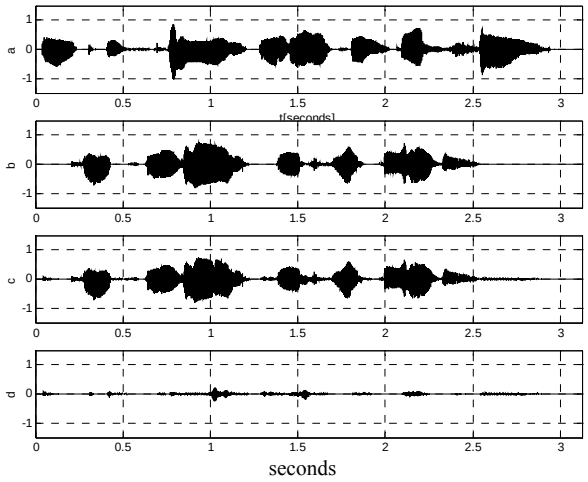


Fig. 10. Performances of the MSRF-QRD-LSL algorithm for speech signals, during double-talk: (a) far-end signal $x(n)$; (b) near-end signal $u(n)$; (c) recovered near-end signal in $e(n)$; (d) residual error $r(n)$.

Also, in our experiments, we have not used any non-linear processor (NLP) [11]. A NLP placed in the send path of the echo canceller will produce an additional attenuation of the residual echo level during the silent periods of the near-end talker, improving the overall performances.

Other experiments (which are not shown here) were performed using more "difficult" echo paths (ERL up to 6dB). It should be noted that 6 dB is a typical worst case value encountered for most networks, and most current networks have typical ERL values better than this. Nevertheless, it was obvious the superior convergence rate and double-talk robustness of the echo canceller based on the MSRF-QRD-LSL adaptive algorithm.

6. CONCLUSIONS

The SRF-QRD-LSL algorithm is an attractive choice in many adaptive systems, due to its fast convergence rate and good numerical properties. Nevertheless, it faces some limitations in fixed-point implementations, where the scaling operations required by the large values of the cost functions affect its performances.

In this paper, we have proposed a modified version of the SRF-QRD-LSL algorithm, based on a modified update of the cost functions. In our approach, these parameters are computed in a reverse manner, preventing the use of the scaling procedures. The modified cost functions decrease to a lower bound, so that the maximum value is the initial one. In this case, milder scaling conditions have to be imposed as compared to the case of the original algorithm. Using a proper initialization, the proposed algorithm does not need for scaling operations any more. Consequently, its numerical robustness is significantly improved.

The simulations performed in a fixed-point DSP context support the theoretical findings. Moreover, the proposed algorithm proved to have efficient LNS implementation as well, due to the important number of multiplications and divisions.

The proposed MSRF-QRD-LSL algorithm was tested in the context of network echo cancellation. The experimental results showed that the requirements of the ITU-T G.168 recommendation concerning the steady-state echo return loss enhancement and convergence speed are fulfilled. The most relevant result is that the MSRF-QRD-LSL algorithm could satisfactorily operate even in the absence of the DTD, which is a very important issue in echo cancellation.

7. REFERENCES

[1] I. K. Proudler, J. G. McWhirter, and T. J. Shepherd, "QRD-based lattice filter algorithms," *Proc. SPIE*, 1989, vol. 1152, pp. 56-67.

[2] S. Haykin, *Adaptive Filter Theory - Fourth Edition*, Prentice Hall, Upper Saddle River, NJ, 2002.

[3] Ph. Regalia, "Numerical stability properties of a QR-based fast least squares algorithm," *IEEE Trans. Signal Process.*, vol. 41, no. 6, pp. 2096-2109, June 1993.

[4] L. M. Davis, "Scaled and decoupled Cholesky and QR decompositions with application to spherical MIMO detection," *Proc. IEEE WCNC*, 2003, vol. 1, pp. 326-331.

[5] S. R. Muruganathan and A. B. Sesay, "A QRD-RLS-based predistortion scheme for high-power amplifier

linearization," *IEEE Trans. Circ. Syst-II*, vol. 53, no. 10, pp. 1108-1112, Oct. 2006.

[6] S. Ciochină and C. Paleologu, "On the performances of QRD-LSL adaptive algorithm in echo cancelling configuration," *Proc. IEEE ICT 2001*, Bucharest, Romania, 2001, vol.1, 563-567.

[7] E. N. Frantzeskakis and K. J. R. Liu, "A class of square root and division free algorithms and arhitectures for QRD-based adaptive signal processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2455-2469, Sep. 1994.

[8] J. Coleman, S. Christopher, J. Kadlec, R. Matousek, M. Tichy, Z. Pohl, A. Hermanek, and N. Benschop, "The European Logarithmic Microprocessor," *IEEE Trans. Computers*, vol. 57, no. 4, pp. 532-546, Apr. 2008.

[9] C. Paleologu, F. Albu, A. A. Enescu, and S. Ciochină, "Square-root-free QRD-LSL adaptive algorithm with improved numerical robustness," *Proc. ICN 2008*, Cancun, Mexic, 2008, pp. 572-577.

[10] C. Paleologu, S. Ciochină, and A. A. Enescu, "A simplified QRD-LSL adaptive algorithm in echo cancelling configuration," *Proc. IEEE ICT 2002*, Beijing, China, 2002, vol. 1, pp. 240-244.

[11] ITU-T Recommendation G.168, *Digital Network Echo Cancellers*, 2002.

[12] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processsing*, Artech House, Boston, U.S.A., 2005.

[13] F. Albu, J. Kadlec, N. Coleman, and A. Fagan, "Pipelined Implementations of the Modified EF-LSL Algorithm," *Proc. IEEE ICASSP2002*, Orlando, U.S.A, 2002, pp. 2681-2684.

[14] F. Albu, C. Paleologu, and S. Ciochină, "Analysis of LNS implementation of QRD-LSL algorithms," *Proc. IEEE CSNDSP 2002*, Stafford, U.K., 2002, pp. 364-367.

[15] S. Ciochină, C. Paleologu, and A. A. Enescu, "On the behaviour of RLS adaptive algorithm in fixed-point implementation," *Proc. IEEE ISSCS 2003*, Iaşi, Romania, 2003, vol. 1, pp. 57-60.

[16] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Speech and Audio Proc.*, vol. 8, no. 5, pp. 508-518, Sept. 2000.

[17] *Motorola SC140 DSP Core*, Reference Manual, Revised 1, 6/2000.

[18] D. L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Commun*., vol. 26, no. 5, pp. 647-653, May 1978.

Temperature Distribution Analysis of Ultrasound Therapy Transducers by Using the Thermochromatic Liquid Crystal Technique

G. A. López, A. Valentino
Department of Bionics
UPIITA-IPN, Mexico
glopezm0300@ipn.mx, avalentino@ipn.mx

A. Vera, L. Leija
Department of Electrical Engineering
CINVESTAV-IPN, Mexico
arvera@cinvestav.mx, lleija@cinvestav.mx

Abstract

Several methods for the acquisition of bidimensional thermal images of acoustic fields produced by ultrasonic transducers for therapy have been described. A low cost technique used for the characterization of ultrasonic therapy equipment consists in placing a thermochromatic liquid crystal (TLC) film in a cross-section of the ultrasonic beam, forming a colorful image of the temperature distribution. This method gives information about the whole thermal pattern, which is related to the intensity of ultrasound, offering a qualitative measurement. The present paper describes the acquisition of a sequence of thermal images using the TLC technique. Diathermia temperatures were induced by using physiotherapy ultrasound equipment at 1 MHz in polyvinyl chloride and polyurethane copolymer films. Digital image processing and computer graphics were used in order to get the mathematical model of the monotonic relation between temperature and color in TLC films for the temperature range from 35ºC to 40°C. 3D thermal beam-shape reconstruction for the ultrasound therapy transducer was obtained for a future medical parameter analysis such as radiation beam pattern, penetration depth and effective field size. A numerical method based on RGB color model and Euclidean distance for getting a quantitative measurement of temperature rise in thermal images was developed.

Keywords — image processing, ultrasound, thermochromic liquid crystal

1. Introduction

As a coherent ultrasonic wave propagates through biological tissue, it is attenuated due to absorption and scattering. Absorption results from the irreversible conversion of acoustic energy to local heat, and it is the primary mode of attenuation in tissue. Ultrasound ability to interact with tissue to produce local heating has been known for a long time and nowadays ultrasound physiotherapy is widely used in health care to treat tissue injuries [13]. However, a large number of therapeutic equipment does not meet international standards. Rigorous quality control to verify whether physiotherapy ultrasound equipment performance is within acceptable range of acoustic intensity output plays a very important role in this context [2]. Thermographic method of color analysis with TLC films is a fast and simple way to evaluate the bidimensional thermal distribution in an acoustic field produced by the ultrasound physiotherapy equipment [3].

TLC films temperature visualization is based on the properties of some cholesteric liquid crystal materials that reflect definite colors at specific temperatures and viewing angles. These properties depend on their molecular organization; they are composed of molecular layers, where each one has a light rotation respect to the closest adjacent plane around an axis. This propriety generates a helicoidal structure that reflexes the incident white light [4]. Correspondence between color and temperature is possible since liquid crystals emit narrow centered bands around one wave length and these bands change regularly for others colors with temperature rise [5].

Color changes in TLC films are repeatable and reversible as long as the TLC films are not physically or chemically damaged. The time response of TLC films is approximately 10 ms. Due to the reversibility and the repeatability of color changes, TLC films can be calibrated accurately and used in this way as a temperature indicators [6].

This work was partially supported by the European Project ALFA - Contract Nº AML/B7-311/97/0666/II-0343-FA-FCD-FI, and the Mexican projects Conacyt 45041-60903-68799 and ICyTDF.

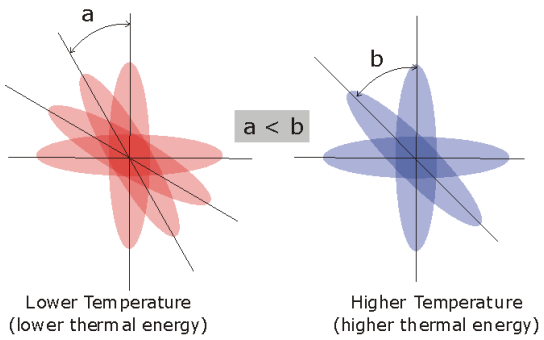


Figure 1. Relative angles between the molecules with the change of temperature [7].

The use of thermochromatic materials for thermal distribution mapping is based on the fact that the acoustic energy absorption by one medium and the temperature rise, according to the temporal average intensity of the ultrasound, is given by [8]:

$$A = 2\alpha_a I, \quad (1)$$

where A is the energy absorption rate in the material, α_a is the absorption coefficient and I is the temporal average intensity of ultrasound.

As the intensity distribution is non-uniform in a cross-section of the ultrasonic beam [9], it is possible to observe changes in the intensity distribution or temperature in a plane where the beam is intercepted by the thermochromatic film [10].

Digital image Processing (*DIP*) allows visualizing and extracting information from an image and it has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data, and can avoid problems such as the noise and signal distortion during processing [11]. Three-dimensional reconstruction is the process in which a sequence of two-dimensional images taken from a common scene is processed to create the planar projections of a volume [12]. Changes produced by the acoustic wave passing through a medium might be interpreted for obtaining the geometrical and mechanical characteristics of the medium.

Temperature qualitative measurements using TLC films have been described. They use DIP algorithms based on the conversion of color images to gray-scale images [12]. By using color models, an abstract mathematical model describing the way colors can be represented as tuples of numbers; it is possible to obtain a relation between the temperature rise and the color change in TLC films for a temperature qualitative measurement [1].

This paper describes the acquisition of a sequence of thermal images of physiotherapy ultrasound equipment with a thermographical system; a chromatic modeling of TLC films in the temperature range from 35ºC to 40ºC; DIP for extracting a quantity that represents color and obtaining a mathematical relation between color-temperature for a quantitative evaluation of temperature rise in ultrasonic thermal images and finally, a tree-dimensional reconstruction for a future medical parameters evaluation.

2. Methodology

2.1. Thermochromatic liquid crystal films

These materials reflect incident white light selectively due to their layered-molecular structure to show bright iridescent color. Their color change, due to temperature increases, usually from colorless to red at low temperature, through the colors of the visible spectrum to blue and colorless again (Figure 2)[14]. They are viewed normally against a black background, and the materials have good long-term stability.

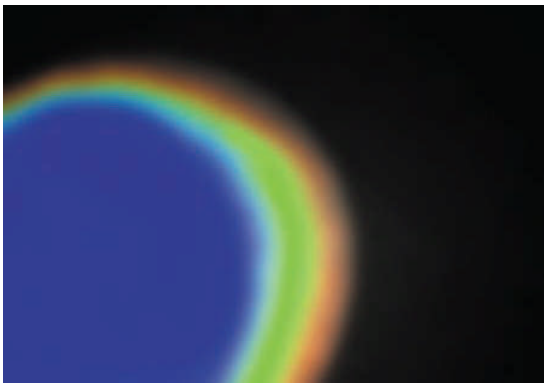


Figure 2. Color changes in TLC films [15].

The TLC film used in this work (R35C5W, Hallcrest®) has a start temperature of 35°C and a bandwidth of 5°C (Table 1) [15].

Table 1. Color-temperature relation for the TLC film.

Red Start (black to red)	Green Start	Blue Start	Clear Point (blue to black)
ºC	ºC	ºC	ºC
35±0.5	36±0.5	40±0.5	49±0.5

2.2. Practical implementation

2.2.1 Chromatic modeling

The experimental setup for obtaining the sequence of thermal images is depicted in Figure 3.

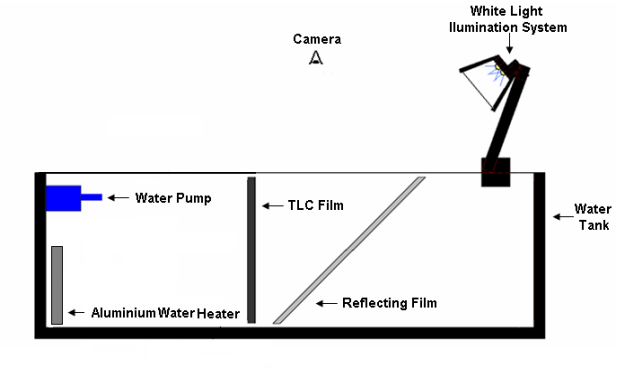


Figure 3. Experimental setup.

A TLC film was immersed in a black painted water tank (dimensions: 70 cm x 20 cm x 20 cm). In order to avoid image distortion due to bubble formation, degasified distilled water was used. The water temperature was increased up to 45ºC by using an aluminum water heater. A water pump was used to avoid the temperature gradients formation. To get thermal images, the heater was removed and the temperature was monitored with an Hg thermometer (Brannan®, 0.1ºC resolution). The sequence of images started at 40ºC, with a temperature increase between acquisitions of 0.5ºC, and finished at 35ºC. The thermal images were viewed by using a metallic reflecting film at 45º from the TLC film and recorded by a commercial camera (Sony®, DSC-W55). A white light illumination system (color temperature of 6500º K) was used as floodlighting to control the white light intensity.

Numerical methods were applied to each thermal image to obtain the chromatic components in the RGB color model. The values of the components were between 0 and 255. The color was digitally represented with 1 byte. The algorithm is illustrated in Figure 4.

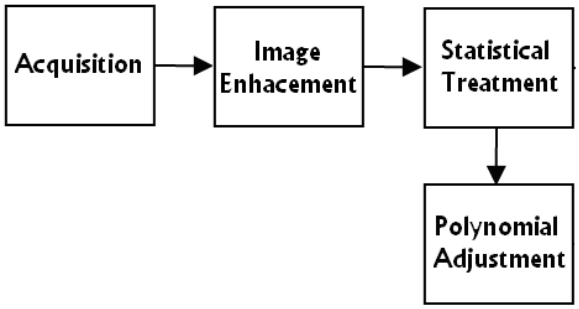


Figure 4. Flow diagram of the algorithm for image processing.

The image enhancement consists of spatial filtering to reduce noise of the images originated by electronic noise in the input device (digital camera) sensor and circuitry, or particles in the medium. For doing that a median filter was implemented.

The median filter, instead of the replacement of the pixel value with the mean values of neighboring pixels, replaces the pixel with the median of the neighboring values. Neighborhood averaging can suppress isolated out-of-range noise, but the side effect is that it also blurs sudden changes (corresponding to high spatial frequencies) such as sharp edges. The median is calculated by first sorting all the pixel values from the surrounding neighborhood into numerical order and then replacing the pixel being considered with the average pixel value. This filter is applied to each matrix (3) that represents the image in the RGB color model. The arithmetic average of the pixel values in filtered images is obtained for each chromatic component.

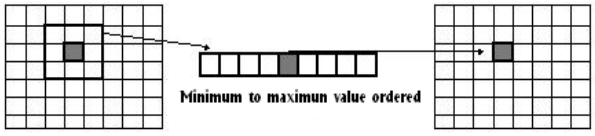


Figure 5. Median filter implementation.

Statistical treatment consists of obtaining the median value of the chromatic components on different thermal image sequences to describe the tendency in the color components with the change of temperature.

To obtain the polynomials that relate the change of the chromatic components and the temperature, the minimum quadratics adjustment is applied to the median of the pixel values in the color components.

2.2.2 Thermal mapping

The experimental setup for exposing and viewing the thermal imaging system is illustrated in Figure 6.

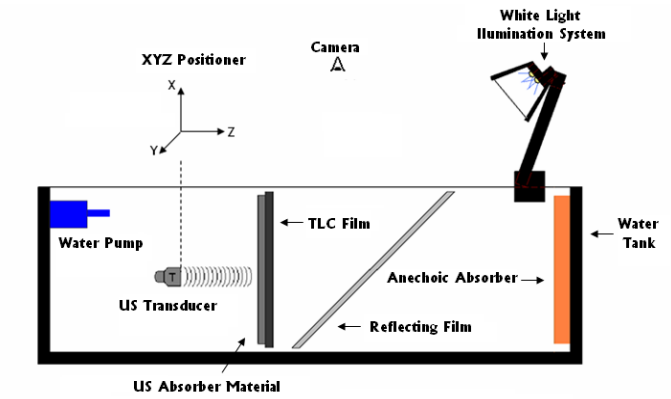


Figure 6. Schematic of the experimental setup.

Thermography system sensibility is related with the temperature rise induced by the ultrasound intensity and is determined by the absorbed ultrasonic energy [8].

The use of low reflection and high absorption attenuation coefficient materials enhances the heating of the TLC film. Polyvinyl chloride and polyurethane copolymer film with a water sonic wave reflection coefficient of 0,06 ±0,01 and a sonic attenuation coefficient of 23,5 ±1,02 dB•cm-1•MHz was used[10]. These values agree with the IEC 61161 regulations for absorber materials [16].

A physiotherapy transducer (1MHz) and TLC film coupled to copolymer film with acoustic gel; were immersed in the black painted water tank filled with degasified distilled water. To minimize standing wave effects, ultrasound transmitted through the imaging system was absorbed and scattered by an anechoic rubber absorber positioned at the end of the system (Ham A, National Physical Laboratory) [17]. For mapping the temperature distribution on the copolymer film, the temperature of water was fixed at 34.5°C, below the start point of the TLC film. Circulation from water pump was added to system for a homogeneous water temperature.

Transducer was excited with a physiotherapy equipment (Ibramed®, Sonopulse) in continuous mode at 1 MHz and nominal ultrasonic intensity was set up to 2W•cm-2.

An algorithm for processing the bi-dimensional thermal images sequence and for enhancing the visualization of the registered thermal phenomenon was developed. Figure 7 illustrates the processes of the digital image treatment.

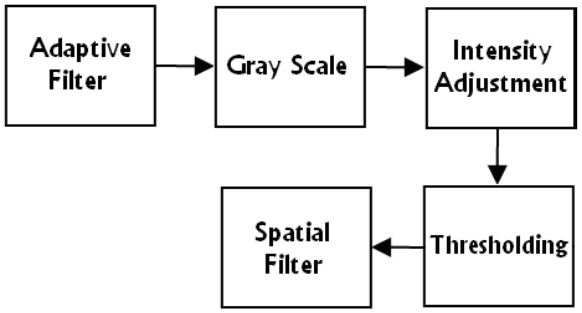


Figure 7. Algorithm for image processing.

Wiener adaptive filter implementation allows preserving edges and reducing noise in images facilitating three-dimensional reconstruction and temperature detection. Pixelwise adaptive Wiener method based on statistics estimated from a local neighborhood of each pixel was applied to the three layers that represent the image considering that the software uses RGB model for representing an image.

Adaptive filter estimates the local mean and variance around each pixel by using the following equations[18]:

$$\mu = \frac{1}{NM} \sum_{n_1, n_2 \in \eta} a(n_1, n_2), \quad (2)$$

$$\sigma^2 = \frac{1}{NM} \sum_{n_1, n_2 \in \eta} a^2(n_1, n_2) - \mu^2, \quad (3)$$

where μ is the local mean and σ^2 the variance in a *N*-by-*M* local neighborhood matrix for each pixel in the evaluated image. Then a pixelwise is created using the following expression:

$$b(n_1, n_2) = \mu + \frac{\sigma^2 - v^2}{\sigma^2}(a(n_1, n_2) - \mu), \quad (4)$$

where v^2 is the noise variance; the algorithm uses the average of all the local estimated noise variances.

Once the images are filtered, the image sequence is converted from color to gray-scale; three-dimensional reconstruction cannot be used with images represented with more than one matrix. The expression for the conversion is given by [18]:

$$I = 0.2989R + 0.5870G + 0.1140B, \quad (5)$$

where *I* is the gray intensity, *R,G,B* represent red, green and blue intensities for the transformed pixel.

Afterward, intensity adjustment for improving image contrast based on gamma correction was carried out. Gamma correction modifies middle values. Contrast in clear or dark areas is enhanced without affect neither white (255) nor black (0). Figure 8 illustrates image contrast improvement by applying gamma correction.

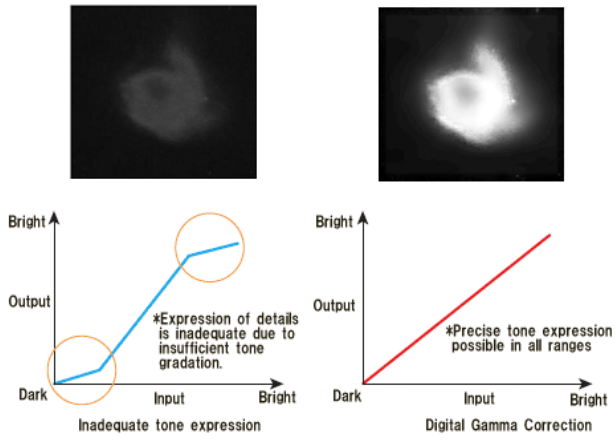


Figure 8. Image intensity adjustment.

Finally, the spatial filter consists of image thresholding and the resulting binary mask multiplied with the intensity adjusted image. Spatial filter eliminates environmental noise that distorts

temperature distribution in thermal images enhancing visualization and three dimensional reconstruction.

2.2.3 Temperature detection

Euclidean distance definition for two-point separation was used for the temperature detection taking into account color spatial point representation in the RGB color model. A numerical method for calculating Euclidean distance between evaluated pixel chromatic components in filtered thermal image and chromatic components for each one of the obtained temperatures in TLC characterization was developed. Distance is given by the following expression:

$$D(t) = \sqrt{(Rp - Rt)^2 + (Gp - Gt)^2 + (Bp - Bt)^2} \quad (6)$$

where D is the Euclidean distance, t the evaluated temperature, Rp, Gp, Bp represent chromatic component values for the evaluated pixel and Rt, Gt, Bt chromatic component values for the evaluated temperature.

Subsequently, the minimum distance was calculated and was related with the temperature for the evaluated pixel.

2.2.4 Three-dimensional thermal pattern reconstruction

Gray scale filtered thermal images for volumetric thermal distribution reconstruction are used. The goal is to create a three dimensional matrix with all the processed images and to interpolate similar gray intensity values.

Three dimensional reconstruction is divided in the following four steps:

- Volumetric data organization and processing.
- Isosurfaces creation (interpolation).
- Reconstruction configuration: color, illumination and graphic texture.
- Vision angle and perspective.

First, images in a three dimensional matrix $V(x,y,z)$ were organized, where x,y represent the coordinate and gray intensity level for the image and z represents the image number of the sequence. Manipulating the matrix for just taking a part of it was possible; that implies that it is possible to obtain cuttings of the temperature pattern in any transversal section for viewing the thermal distribution inside the volume.

Thermal image edge filtering was applied for a smooth volume isosurface. Then, image outline pixel values are interpolated with their respective gray intensity level. This procedure allows creating the volume geometry and giving a specific color to each of the interpolated gray levels. Finally, perspective, texture and illumination are configured. Figure 9 illustrates edge interpolation carried out by the algorithm for reconstructing the thermal distribution.

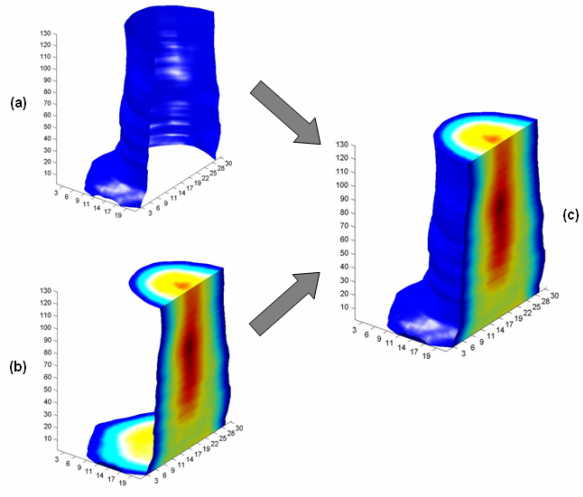


Figure 9. (a) External edges interpolated. (b) Internal edges interpolated. (c) Reconstructed volume [12].

3. Results

3.1 Chromatic modeling

The images were taken at different times. Nine sequences of thermal images were taken. A median filter with a 15 x 15 kernel was implemented to improve the images. The arithmetic mean was applied to get the tuples of values for the chromatic components in the images. The tuples for the nine measurements were treated statistically with the median and the standard deviation.

For the polynomial adjustment, the minimum quadratics adjustment and the median were used. The polynomial degree was adjusted to the best fitting and the lowest degree considering the behavior tendency of the color component values.

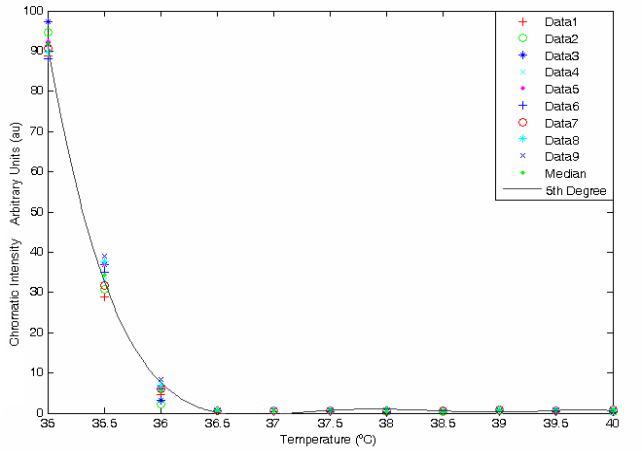


Figure 10. Polynomial adjustment of the chromatic red component.

Red components are higher in the range from 35ºC to 36ºC. Red components have a non-lineal behavior tendency (Figure 10).

The mathematical model for the chromatic red components is defined by the equation:

$$R=-4.30t^5+11.05t^4-4.75t^3-5.03t^2+2.66t+0.69, \quad (7)$$

where R is the intensity of the red components and t is the temperature defined between 35ºC and 40ºC.

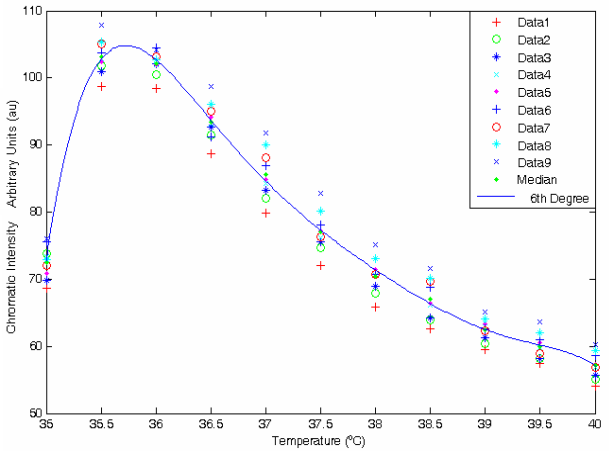


Figure 11. Polynomial adjustment of the chromatic green component.

Green components are present in the range from 35.5ºC to 37ºC. Green components have also a marked non-lineal behavior tendency (Figure 11).

The mathematical model for the chromatic green components is defined by the algebraic expression:

$$G=-3.12t^6+5.55t^5+1.43t^4-5.44t^3+7.37t^2-21.50t+77.3, \quad (8)$$

where G is the intensity of the green components and t is the temperature defined between 35ºC and 40ºC.

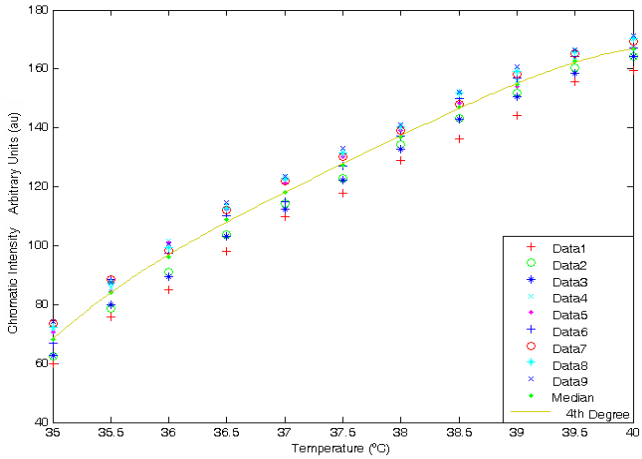


Figure 12. Polynomial adjustment of the chromatic blue component.

Blue components are present in the range from 37.5ºC to 40ºC. Blue component has an almost-lineal behavior tendency (Figure 12).

The mathematical model for the chromatic blue component is defined by the expression:

$$B=-1.63t^4+0.41t^3-0.82t^2+31.81t+127.72, \quad (9)$$

where B is the intensity of the blue component and t is the temperature defined between 35ºC and 40ºC.

Table 2. Statistical treatment for the chromatic components measurements

Temp.	Median			Standard Deviation		
ºC	Red	Blue	Green	Red	Blue	Green
40	0.56	167.20	56.92	0.19	3.60	2.03
39.5	0.63	163.92	59.85	0.19	3.60	2.04
39	0.54	155.87	62.68	0.16	5.19	1.78
38.5	0.47	148.34	66.37	0.10	5.13	3.15
38	0.44	139.03	70.67	0.14	4.03	2.75
37.5	0.52	130.34	76.33	0.13	5.20	3.15
37	0.51	121.16	84.83	0.14	5.14	3.87
36.5	0.64	11.20	93.11	0.13	5.82	2.93
36	6.06	98.25	102.11	1.90	5.88	1.86
35.5	35.10	85.76	102.85	3.46	4.78	2.69
35	91.75	70.76	72.90	2.87	5.39	2.50

3.2 Thermal mapping

Once the ultrasound beam was switched on, the formation of a stable thermal image on the TLC film took approximately 30 s. Measurements were made at 2 mm intervals up to 80 mm. Thermal images were taken every 30 s so that the image was steady. At the end a sequence of 40 images was obtained (Figure 13).

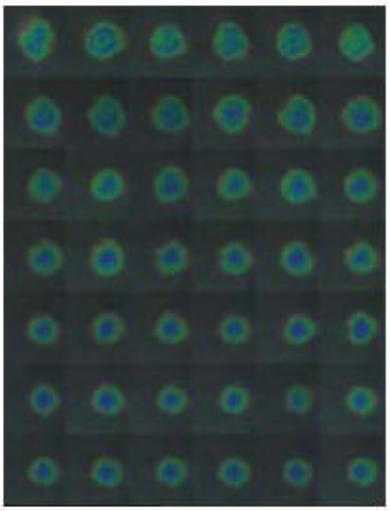


Figure 13. Thermal image sequence.

An adaptive Wiener filter with a 25 x 25 kernel was implemented to improve the images; binarization was applied with a 0.5 luminance level (Figure 14).

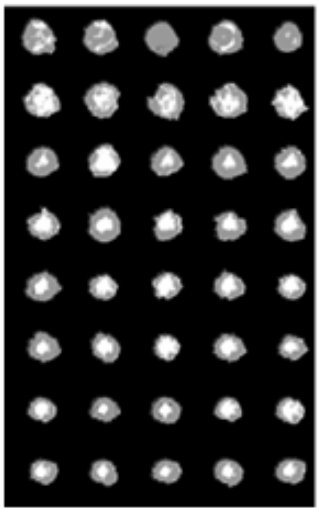


Figure 14.Gray scale thermal images.

3.3 Temperature detection

A graphic interface was developed. This graphic interface facilitates user interaction and avoids the necessity of working with the programming code. The algorithm is semiautomatic. Users have to take part in some image treatment process and parameters configuration (Figures 15 and 16).

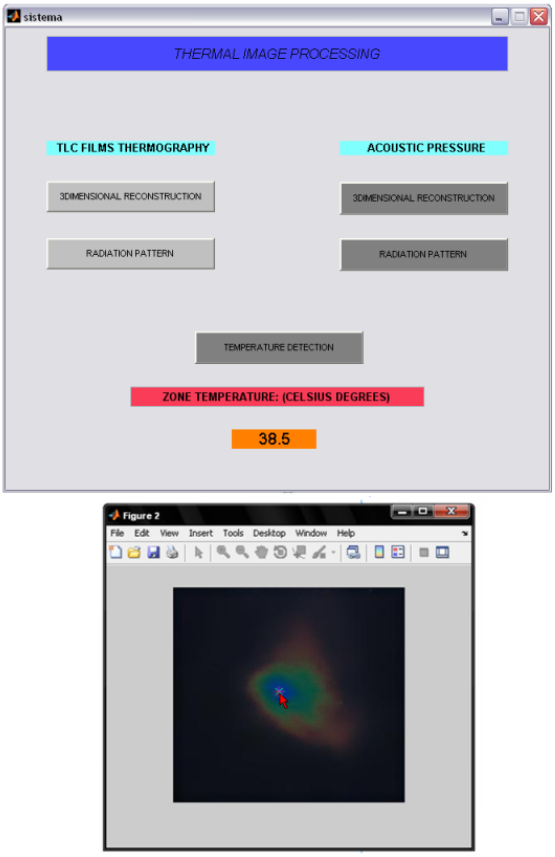


Figure 15.Temperature detection with graphic interface.

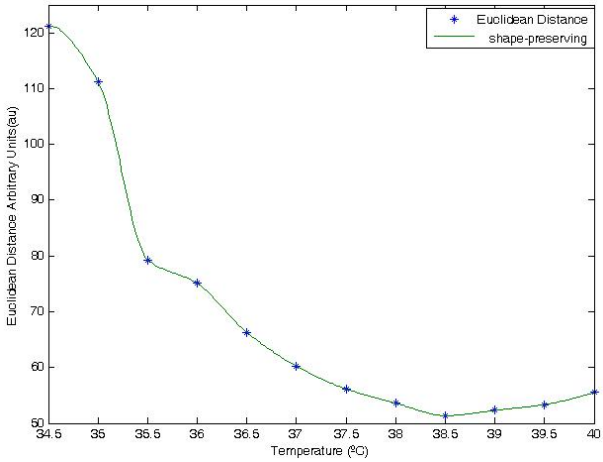


Figure 16.Euclidean distance for temperature evaluation.

3.4 Three-dimensional thermal pattern reconstruction

For making three-dimensional reconstruction, each image of the sequence was superposed one behind the

other in a three-dimensional array to interpolate pixel values corresponding to the edge of each image with the same level of intensity to create an isosurface. Color map, color bar, illumination and material were configured for the scene.

Radiation pattern gives information about the non-homogeneities of acoustic intensity when they are related to the gradients of temperature in thermal images (Figure 17).

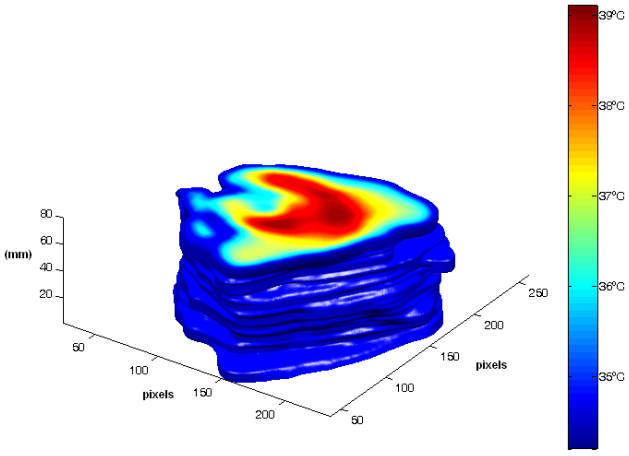


Figure 17. 3-D Temperature pattern reconstruction.

Data of the three-dimensional (3-D) array was used for taking a transversal section for viewing thermal distribution inside the volume.

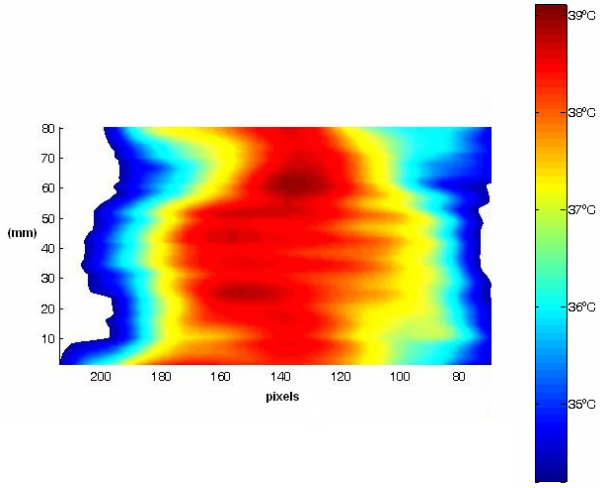


Figure 18. Temperature pattern reconstruction.

The thermal distribution pattern in the radiated copolymer shows a high temperature region distributed throughout *z* axis with two regions where the temperature reaches its maximum point, between 25 – 60 mm of depth (Figure 18).

4. Discussion

One of the main advantages of the technique based on TLC sheets is that the reflected light is within the visible spectrum allowing viewing and photography by using a rather conventional camera.

Chromatic modeling. DIP was used for a qualitative measurement of thermal distribution in gray-scale images [12]. In this work a sequence of thermal images between 30-35 Celsius degrees were obtained and a DIP algorithm was implemented for the mathematically modeling in the TLC film using the RGB color model. The use of a color model allows obtaining a mathematic relation between temperature and an abstract representation, as color is considered, to get a temperature quantitative measurement.

Thermal mapping. This method applied to ultrasound fields was proposed by Cook and Werchan [19] improved by Macedo *et al.* [10] and Gómez *et al.* [12]. Coupling polyvinyl chloride and polyurethane copolymer material to TLC film improves ultrasound absorption, delays thermal equilibrium and enhances system sensibility.

These techniques are dependent of the following factors:

Surrounding illumination – Uncontrolled light sources and possible reflections from surfaces in the TLC may cause errors in the measured temperatures; an illumination analysis is required in other applications of the method.

Viewing angle – The TLC colors depend on the angles between the camera and the TLC surface.

Controlled temperature conditions – It is important to control the temperature to achieve desired accuracy and to avoid the temperature gradients formation. Hysteresis is one of the problems when the crystals are heated above their clearing point temperature [7].

Digital resolution – The minimal number of pixels needed per unit length of the TLC surface depends on the magnitude of the temperature gradient: locations with large gradients needing more pixels to achieve a desired accuracy. There is a possible algorithm failure when the thermal image resolution is increased. Time of processing also increases when resolution increases.

Temperature detection. A classifier algorithm for temperature detection was developed by using the RGB color model and the chromatic modeling by relating TLC film temperatures with colors and calculating distance separation between each temperature and the evaluated pixel in thermal image.

Minimum separation distance was taken for estimating temperature in the evaluated pixel.

Three-dimensional thermal pattern reconstruction. The three-dimensional reconstruction of ultrasound thermal beamshape gives the possibility to know the radiation pattern (temperature distribution) and penetration depth which are important medical parameters for ultrasound physiotherapy treatment.

5. Conclusion

A thermographic technique based on TLC films to obtain two dimensional thermal images was developed; the practical implementation of the method is simple and low cost, it is a suitable way for chromatic and ultrasound therapy transducers characterization with TLC films.

DIP allows image improvement by reducing noise in thermal images and the possibility to manipulate images as a mathematical representation for an efficient obtainment of the required information in an image. Adaptive filter Wiener implementation allows reducing noise selectively enhancing definition in thermal images and tree-dimensional reconstruction.

The color change in TLC films is related with the chromic contribution of the components in the RGB color model and related with the changes in temperature. The statistical analysis shows the non-linear but regular change in the tendency for the RGB components. Tests with different TLC films in controlled environmental conditions are necessary trying to find a general mathematic model that describes thermochromic behavior allowing a higher resolution and extending measurement range.

An algorithm for chromatic modeling in TLC films based on RGB color model was developed obtaining a mathematical relation between color and temperature allowing a quantitative temperature measurement by using a classifier algorithm based on Euclidean distance. Neural networks are proposed for temperature detection algorithm alternative technique.

The future work is to develop tissue equivalent materials that mimic thermal and acoustic proprieties for obtaining the thermal distribution and to detect changes in beam uniformity for ultrasonic therapy equipment.

6. Acknowledgment

Authors acknowledge *the National Polytechnic Institute* UPIITA-IPN, *CONACyT (projects 68799*, 45041 and 60903), ICyTDF for their financial support. Authors also thank to *M. in C. Hugo Zepeda Peralta* and M. *in C. Rubén Pérez Valladares* for their technical support.

7. References

[1] G.A. López, A. Valentino, A. Vera, L. Leija, "Chromatic Modelling in Liquid Crystal Films for Ultrasound Thermography", *International Conference on Advances in Electronics and Micro-Electronics* ENICS 2008, Valencia, Spain, October 2008.

[2] P.A. Artho, J.G. Thyne, B.P. Warring, C.D. Willis, J.M. Brisme´e, N.S. Latman, "A Calibration Study of Therapeutic Ultrasound Units" *Physical Therapy*, v. 82, n. 3, p. 257-263, 2002.

[3] C.H. Jones, P. Carnochan, "Infrared Thermography and Liquid Crystal Plate Thermography"; *Physical Techniques in Clinical Hyperthermia*; J. W. Hand, J. R. James, Eds. Research Studies Press, pp. 507–547,1986.

[4] P.T. Ireland, T.V. Jones, "Liquid Crystal Measurement of Heat Transfer and Surface Shear Stress"; *Measurement Science and Technology*, vol. 11, no. 7, pp. 969-986, 2000.

[5] L. Cristoforetti, R. Pontalti, L. Cescatti, R. Antolini, "Quantitative Colorimetric Analysis of Liquid Crystal Films (LCF) for Phantom Dosymetry in Microwave Hyperthermia," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 11, pp. 1159-1165, 1993.

[6] Stasiek, J. A. & Kowalewski, T. A. Thermochromic liquid crystals applied for heat transfer research. *Opto-Electron*, Rev. 10, pp. 1–10, 2002.

[7] S. Baknaria, A. M. Anderson, "A Transient Technique for Calibrating Thermochromic Liquid Crystals: The Effects of surface Preparation, Lighting and Overheat", *ASME International Mechanical Engineering Congress ad Exposition*, Louisiana, USA, November 2002.

[8] J K. Martin, R. Fernández, "A Thermal Beam-Shape Phantom for Ultrasound Physiotherapy Transducers," *Ultrasound in Medicine and Biology*, vol. 23, no. 8, pp. 1267-1274, 1997.

[9] P.Fish, *Physics and Instrumentation of Diagnostic Medical Ultrasound*, Eds. John Wiley & Sons, New York, EUA, 1994.

[10] A. R. Macedo, A. V. Alveranga, W. C. A. Pereira, J. C. Machado, "Ultrasonic Beam Mapping Using the Chromothermic Proprieties of Cholesteric Liquid Crystals", *Brazilian Biomedical Engineering Magazine* vol. 19, no. 2, pp. 61-68, 2003.

[11] K. Castleman, *Digital Image. Processing*, Eds. Prentice Hall, New Jersey, USA, 1996.

[12] W. Gómez, W.C.A. Pereira, L. Leija, M.A.V. Krüger, A. Vera, "3D Thermal Mapping of Ultrasonic Beam for Ultrasonic Transducers", *XXIX National Biomedical Engineering Congress*, Guerrero, México, October 2006.

[13] G. Haar, "Therapeutic ultrasound," *European Journal of Ultrasound*", vol. 9, pp. 3–9, 1999.

[14] "Hallcrest Data Sheet Listing: *Handbook of Thermochromic Liquid Crystal Technology" Available* http://www.hallcrest.com/downloads/randtk_TLC_Handbook.pdf

Access on 5 April 2008.

[15] "Hallcrest Data Sheet Listing: *Thermacolor Thermochromic Sheets & Films" Available* http://www.hallcrest.com/downloads/ThermacolorThermochromicSheetsFilms_SS.pdf

Access on 10 February 2008.

[16] *Ultrasonic Power Measurement in Liquids in the Frequency Range 0.5 to 25 MHz*, International Electrotechnical Commision, IEC 61161, 1992.

[17] B. Zequiri, C. J. Bickley, "A New Anechoic Material for Medical Ultrasonic Applications," *Ultrasound in Medicine and Biology*, vol. 26, no. 3. pp. 481–485, 2000.

[18] Image Processing Toolbox (6.0) User's Guide; The Math Works, Natick, MA, 2008.

[19] B.D. Cook, R.E. Werchan, "Mapping Ultrasonic Fields with Cholesteric Liquid Crystals" *Ultrasonics*, v. 9, n. 2,pp. 101-102, 1971.

[20] W. Gómez, M. A. V. Krüger, W. C. A. Pereira, L. Leija, A. Vera, "Analysis of SAR with Thermochromic Liquid Crystal Sheets in Focused Ultrasound Beam", *XX Brazilian Biomedical Engineering Congress*, Sao Paulo, Brazil, October 2006.

A Cross-layer Mechanism Based on Dynamic Host Configuration Protocol for Service Continuity of Real-Time Applications

Luis Rojas Cardenas
Universidad Autonoma Metropolitana
Vicentina DF, Mexico
e-mail: lmrc@xanum.uam.mx

Mohammed BOUTABIA
dept. wireless networks and multimedia services
Telecom Sudparis
Evry, France
e-mail: Mohamed.boutabia@it-sudparis.eu

Hossam AFIFI
dept. wireless networks and multimedia services
Telecom Sudparis
Evry, France
e-mail: Hossam.afifi@it-sudparis.eu

***Abstract*—Most important frameworks supporting mobile communications are not capable of meeting real-time application requirements because of the service degradation appearing during the handover process. Such degradation is mainly noticed as an excessive blocking time and a non-negligible packet loss rate. This is due to slow procedures for address allocation, too much packets exchanged by signaling procedures, and the delay required to establish a new end-to-end delivery path. Although these problems have been widely analyzed, and a number of solutions have been proposed, better handover performances are still needed. In this paper, we propose the introduction of some functionalities into access point equipments to improve the handover performances. These functionalities are based on the reduction of both the address allocation delay and the number of exchanged signaling packets, as well as the parallel execution of certain procedures. Our approach is implemented over the signaling mechanism of Dynamic Host Configuration Protocol (DHCP), from which extended options are used to convey information related to each procedure allowing mobile communication to be maintained.**

Keywords-component; mobility; cross-layer; real-time; DHCP;

I. INTRODUCTION

All As new wireless technologies are deployed and *Mobile Nodes* (MN) such as mobile phones, PDA, Internet tablets etc. acquire more hardware capabilities in terms of processing speed, communication and storage space, it is expected that wireless communications will be more heterogeneous and commonly based on IP protocol. However, to operate in such scenario, mobile nodes must be equipped with multiple wireless cards such as WIFI, WIMAX (*Worldwide Interoperability for Microwave Access*), UMTS (*Universal Mobile Telecommunications System*), etc. and special communication protocols able to cope with mobility. These capabilities will allow Mobile Nodes not only to communicate through different network technologies, but also to choose the most convenient one in case of several available networks; this latter characteristic is known now as ABC (*Always Best Connected*) capability.

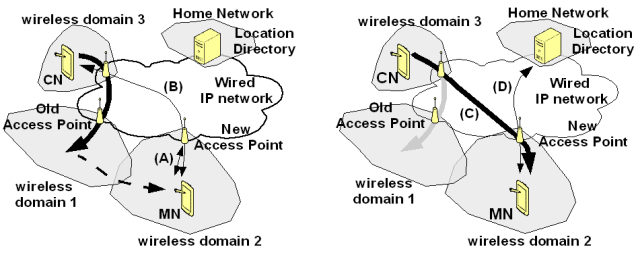
In this context, it seems that Internet Protocol will play an important role in this world of heterogeneity and mobility in spite of the fact that it was not designed to handle mobility. Indeed, Internet protocols are not suitable for supporting mobile communications because of its principles for handling addressing and routing. They establish that any host address must be derived from the network address where it is physically attached as well as they do not consider that a host can change its attachment address at the middle of a session. Under such scheme, when a MN moves from its original network to a *Foreign Network* (FN), it will experience at least the following problems: 1) when it reaches a new network, any communication becomes impossible. Given that its address is not valid in the context of the foreign network, it can not be accepted neither by foreign nodes nor corresponding routers. Obtaining a new valid address from the foreign network is then necessary. 2) The ongoing communication associations are lost due to address inconsistency i.e. at operating system level each communicating system represents a communication association by means of a 5-tuple {*protocol*, *local-address*, *local-process*, *foreign-address*, *foreign-process*} [13]. If one of these elements becomes inconsistent, for example when a mobile node reaches a foreign network and a new *local address* is obtained, ongoing communications are lost. Nevertheless, informing the corresponding node about the new local address can help to recover the lost communication. 3) Mobile hosts disappear from the global network. Normally, hosts are found in the network by means of the *Location Directory* (LD). It is a distributed database containing the host name and its corresponding IP

address which is known in the Internet world as DNS (*Directory Name System*). If one of the elements of this association changes without informing the LD, nobody in the network will be able to reach that host. That is what happens when a MN goes from one network to the other and MN changes its IP address. To keep in touch with the global net, MN must inform the LD each time it acquires a new IP address.

In order to cope with IP limitations in mobile communications, a number of approaches have been proposed. Although they tackle the problem from different perspectives [3][4][5], they agree on the way iit must be handled. Indeed, the main approaches relay on a number of procedures that can be classified on: 1) *network discovery and address allocation*, 2) *preservation of the ongoing communications* and 3) *update of the global location directory*. These three procedures and the problems they address are analysed in more detail in the following paragraphs.

First of all, when a node reaches a new network and discovers it by means of low level mechanisms, an *address allocation* phase starts. We consider that the MN starts this phase by sending an address request message to the FN. This phase finishes when the corresponding access point informs the MN about the allocated address by means of an acknowledge message (see label A in Figure 1.a).

In the second phase, as soon as MN obtains a new address and in order to maintain the ongoing communications, MN notifies the *correspondent node* (CN) about the new acquired address (Label B). Then, the CN immediately redirects the data flow to the new address (see Figure 1.b, label C). These two phases are the most critical ones. Actually they form what is known as *handover process*. In the third phase, the location tracking procedure is achieved to maintain the reachability of the MN at global network level. This is a less critical operation and it is achieved by updating the Location Directory (LD) with the new acquired address (see Figure 1.b, label D).



a) address allocation b) location tracking process

Fig. 1. Handover process

In this paper, we propose the introduction of new functionalities into access point equipments to improve the handover performances. Specifically, blocking times are minimized by reducing both the address allocation delay and the number of exchanged signaling packets. Additionally, parallel execution of certain procedures contributes to obtain a performance gain.

The remainder of this article is organized as follows. In Section II, we speak about the related work. The architecture of our approach is described in Section III and the corresponding performance analysis is presented in Section IV. Section V discusses security issues of our solution. The conclusion and future work are discussed in Section VI.

II. RELATED WORK

By tacking into account the classification of procedure stated above, we analyse the most representative approaches for handling node mobility, in particular Mobile IP and SIP. The global performances around the handover process are especially important for this analysis.

A. Network Layer Perspective: Mobile IP

The main goal of *Mobile IP* (MIP) is to avoid upper layers to be worried about address changing due to node mobility. The principle is as follows: when the CN sends packets to the MN, it employs the home address of the MN so that packets arriving to the home network are intercepted by the *Home Agent* (HA) and sent to the Care-of-Address (CoA) via a tunnel. As this latter is associated to the FA, it receives the packets and redirects them to the MN. This mechanism allows a transparent application operation. Recent standards of IETF propose more sophisticated mobility schemes like MIPv6 [9] and Fast MIPv6 [10] but these standards cannot be widely deployed and have to wait the transition to IPv6.

Address Allocation

The mechanism for CoA acquisition relies on the services of the new FA, which periodically broadcasts a *Router Advertisement* message containing CoA related information. This mechanism has a drawback: the minimum broadcast period is one second [3]. A faster mechanism is based on *Router Solicitation* message which explicitly requests a *Router Advertisement*. This operation takes $2t_s$ and corresponds to the round-trip time between the MN and FA (see Figure 2).

Preservation of the ongoing communication

When the new CoA is obtained, the MN must inform its HA about the obtained CoA by sending a REGISTRATION message. After this registration, the HA can forward the packets (originally sent by the CN to MN's home address) to the FA by tunneling and then to the MN. This scheme generates what is known as *triangle routing*, which is characterized by the introduction of additional end-to-end delay. To reduce this delay the *Route Optimization* (RO) [8] can be used so the CN encapsulates packets directly to the

current CoA without passing through the HA. This procedure is described in Figure 2.

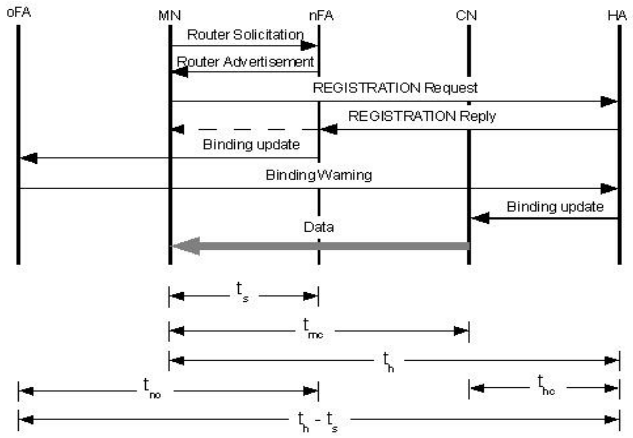


Fig. 2. Handover process in Mobile IP with RO.

The HA sends to the MN a REGISTRATION Reply message which is intercepted by the *new FA* (nFA) and then sent to the *old FA* (oFA) which sends a Binding Warning message to the HA. Finally the HA send a Binding update to the CN which starts sending data directly to the MN. Smooth Handoff [4] is an additional functionality that reduces the packet loss generated during the handover by means of a Binding Update between the nFA and the oFA. In accordance with [7], during the handover process, the service disruption time under MIP with RO is:

$$T_{mip\_inter} = t_{no} + 3t_h + t_{hc} + t_{mc} \quad (1)$$

Where t_{no} denotes the delay of a message between the new FA and the old FA, t_h is the delay between the MN and the HA, t_{hc} is the delay between the home network and CN, finally t_{mc} is the time that data takes to arrive from the CN to MN.
The smooth handoff starts after a delay of :

$$T_{mip\_smooth} = 2t_s + 2t_h + 2t_{no} \quad (2)$$

The Smooth Handoff avoids packets to be lost by redirecting packets from the old FA to the new FA before a handoff process is completely achieved. A tunnel created between these FAs undertakes this task.

Maintaining the global location tracking

There is no need to a global tracking registration in MIP since the HA is updated with the new CoA and the future CNs will use the original home address to reach the MN.

B. Application Layer Perspective: S IP

Handling mobility at transport and network layer requires considerable changes in the MN kernel. This is the main motivation for developing upper layers solutions, such as *Session Initiation Protocol* (SIP). SIP is capable of supporting terminal mobility, session mobility, personal mobility, and service mobility. Moreover, SIP has been widely accepted as the signaling protocol in emerging wireless networks. Therefore, SIP seems to be an attractive candidate for an application-layer mobility management protocol for heterogeneous all-IP wireless networks. However, SIP entails application-layer processing of messages, which may introduce considerable delay.

Address Allocation

After a MN discovers a new network by means of low level procedures, an address allocation phase starts. The procedure commonly used in this context is DHCP. Although this TCP/IP-based protocol was not designed to operate in mobile contexts, it is widely employed to support address allocation in access networks. This protocol relies on four different DHCP messages: DHCP Discover, DHCP Offer, DHCP Request, and DHCP Acknowledge, which are all UDP packets. DHCP satisfies most of non real-time applications but it appears to be unsuitable when it deals with real-time ones. The main problem here is related to the number of packets and the long delay that DHCP takes for address allocation. This latter is mainly caused by the address conflict checking mechanism based on ICMP Echo request and reply. A DHCP server has to send out an ICMP Echo Request to the address in question before responding to a Discover message. If nobody responds with an ICMP Echo Reply within a typical interval of 1 to 3 seconds, the DHCP server will send the Offer message. As far as the client is concerned, it performs a similar checking. In order to improve the performances of DHCP, there are some proposals to reduce the number of packets, from four to only two [2]. Others works suggest to remove the address conflict checking [17]. And finally, there are proposals to use new protocols for supporting address allocation more suitable for mobile applications, in particular, Dynamic Registration and Configuration Protocol (DRCP) [12].

Preservation of the ongoing communication

The procedure allowing MN to preserve its ongoing communications is known as *mid-call* procedure. The principle is the following: when the MN reaches a new network and a new address has been acquired, the MN sends a re-INVITE request to the CN. This operation is accomplished without intervention of any intermediate SIP proxies. This INVITE request contains an updated session description with the new IP address. The CN starts sending data to the MN's new location as soon as it gets the re-INVITE message.

In accordance with [6], the total handover delay in SIP must consider: the DHCP and ARP resolution delay, the updating delay of the LD or Home Registrar (HR), and the time the INVITE message takes from the MN to the CN and the time CN data take to reach the new MN location.

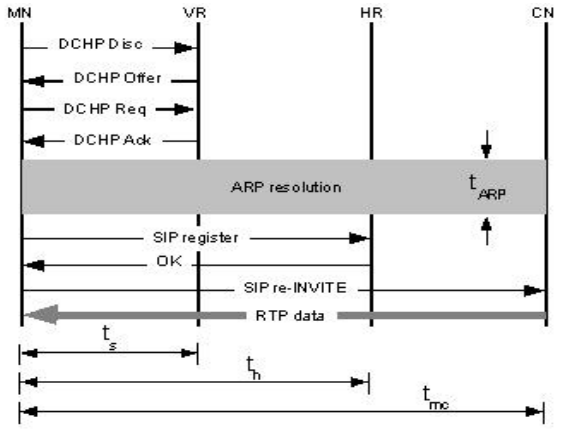


Fig. 3. Handover process in SIP

This total time T_{sip_inter} is given by:

$$T_{sip\_inter} = 4t_s + t_{arp} + 2t_h + 2t_{mc}. \quad (3)$$

Where $4t_s$ corresponds to the four messages exchanged by DHCP, t_{arp} is the time of address conflict checking, $2t_h$ is the time to update the HR and $2t_{mc}$ corresponds to the time the INVITE message takes, in addition to time for redirecting data from the CN to the new MN location.

One drawback of this approach is the packet loss rate during the handover process; which can also be seen as a period of disruption. While MIP solves this issue by implementing a smooth handoff, SIP lacks such a mechanism. Consequently all the packets transmitted during $2t_s + 2t_h + 2t_{no}$ will be lost.

Maintaining the global location tracking

The location tracking procedure is achieved by sending an update message to the home registrar which update the current location of the user agent allowing the future clients to reach the MN with the same URI.

III. THE FAST CROSS-LAYER HANDOFF

In this section, we describe our proposal called *Fast Cross-Layer Handoff* (FCLH) [12], which is capable of improving the performances of mobile communication with respect to the approach described above [6]. The handoff improvement is obtained by following these three operations: i) the reduction of the address allocation delay, ii) the minimization of the number of exchanged signaling packets, and iii) parallel execution of certain procedures. These operations are accomplished in order to support the quality of service requirements imposed by voice over IP (VoIP) applications type. We integrate FCLH scheme to SIP mobility which is the most convenient mobility protocol for real-time applications as it was proven in [14], but it can be integrated to other application level mobility protocols.

A. Overview

Our approach is based on the idea that the three main procedures required to support mobile communications (see section I) can be achieved in parallel and started by only one message. Parallel processing is possible because it relies on three different entities: the Correspondent Node, the Location Directory and the two access points involved in the handover process. More specifically, CN is involved in the *preservation of the ongoing communication* (POC), The LD supports the *maintaining of the global location tracking* (MGLT) and the APs are responsible of the *address allocation* (AA) service and the smooth handoff procedure (SHP). In principle, these tasks are more or less independent, so they can be achieved in parallel. The only condition for doing so is breaking the classical layered protocol stack. In fact, this operation is commonly called cross-layer which opens up the possibility to introduce parallelism on the different tasks to maintain ongoing communications while speeding up the global performance of the handover process. In order to explain this principle, consider a MN reaching a new network. To obtain a new address, it exchanges DHCP messages with the visited network and then informs its home registrar about the new location using a SIP-register message. In Figure 4.a, we can see that DHCP and SIP can only operate in a sequential way because SIP cannot start updating its home registrar without knowing the new allocated address. This update is possible only after receiving the DHCP ACK message. In contrast, a cross-layer approach has less restriction; therefore the MN achieves two different transactions with only one DHCP message as shown in Figure 4.b; It not only negotiates a new address but also informs the HR about the new location.

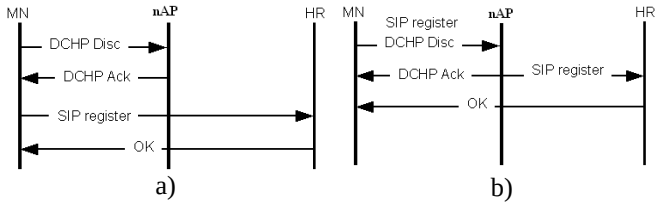


Fig 4. Classical vs Cross-layer protocol transactions.

Two advantages can be obtained from this approach: the number of exchanged messages is minimized and the handover delay is reduced. But this approach is possible at the expense of implementing some SIP functionalities in the different entities participating in the handoff process. In Figure 4.b, a SIP register message is built by MN and

included in discovery message. The nAP completes the received SIP message with the missing information which is the new acquired address. At this level, some questions must be asked: Does DHCP allow to convey such information? Is the DHCP payload capacity enough to carry messages like SIP-messages? The answers are: First, DHCP has the option fields which have been created to convey vendor-independent options between client and server, so we can use these fields to convey SIP-messages. Second, the payload capacity of a DHCP message depends on the Maximum Transfer Unit (MTU) of the visited network. For WiFi networks, the MTU is 1492 bytes. So, there is no problem with SIP messages, for example, a re-INVITE message is more or less 140 byte long [6]. Following the principle stated above, Figure 5 proposes the FCLH mechanism and the interaction between the different entities.

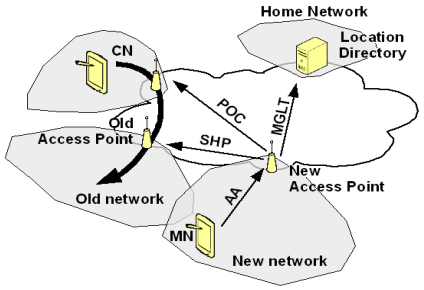


Fig. 5. A single message starts in parallel all the procedures required for handling mobility.

It should be noted that the cross-layer capacity is mainly supported by some functionalities installed on access points. The procedure is achieved as follows. Access points receive classical Discovery DHCP messages, which contain upper level information. The DHCP server installed on access points process and ask under the standard protocol but upper layer information contained in DHCP extended options is extracted and completed with the MN's new allocated address. Upper layer information corresponds to MGLT, SHP and POC procedures.

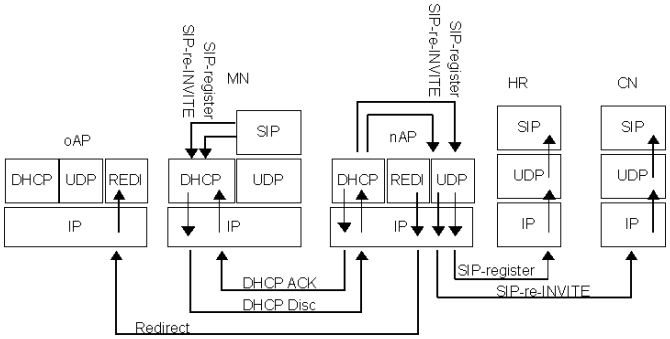


Fig 6, layered protocol stack.

A special procedure recovers this information and generates at least three data packets, which are sent to the corresponding network entities (see Figure 5 and Figure 6). As far as correspondent node and location directory are concerned, their protocol stack is not modified. Figure 6 represents the layered protocol stack of access point as well as the interaction with the MN and the other elements.

It should be noted that the cross-layer operation is mainly supported by MN and access points. Indeed, CN and LD do not require additional components or modifications to operate with FCLH. Moreover, a MN equipped with our approach can operate on networks that are not equipped with FCLH. In this case, MN can distinguish the absence of the FCLH infrastructure in the new visited network by means of the options included in DHCP ACK packets. In a similar way, a network equipped with FCLH can operate with any standard DHCP client.

Now, as in our approach a handover process is started by sending only one packet from MN to the discovered access point, this packet must to convey information corresponding to the POC, the MGLT and the SHP processes. The POC process is started by a SIP re-INVITE message, whereas the MGLT process requires a SIP register message. As far as SHP process is concerned, it is not related to SIP. It is an optimization mechanism similar to that proposed in low latency handoffs in MIPv4 [15]. It supports smooth handover by creating a tunnel between old and new access routers. This tunnel is used to convey the packets that were intended for MN when it was unreachable during the physical handover. The information that should be known to previous access router to perform an SHP is the new access router address so that the tunnel can be initiated. As far as the AA process is concerned, it is essentially supported by a DHCP procedure. The protocol stack of a MN with FCLH capacities is represented in Figure 7.

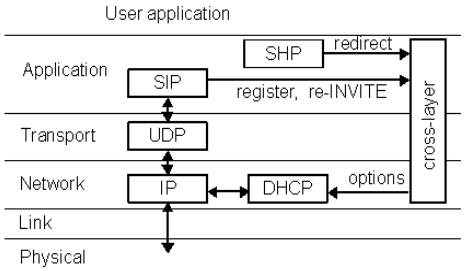


Fig 7. Protocol Stack at Mobile Node.

SHP and SIP modules insert the necessary information allowing the handover process to be started. This information is inserted in the DHCP-DISCOVER message by means of the cross-layer. All the information related to MGLT, POC and SHP processes are sent in a single packet. However, in the downlink, responses related to those procedures are processed normally and sent directly to the MN.

Figure 8 represents the protocol stack of the access point. This stack is responsible for dispatching the information contained in the DHCP-DISCOVER message, and then it

rebuilds each message by filling the destination address with the new allocated IP address. Finally, the nAP sends all the messages to the appropriate correspondent modules at once.

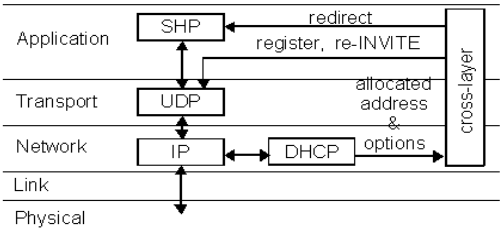


Fig 8. Protocol stack at Access Point.

Address allocation

After MN discovers a new network by means of low level procedures, an address allocation phase starts. In the context of FCLH, the protocol used is DHCP as proposed in [2]. This document proposes to reduce, from four to two, the number of messages required to allocate an IP address by DHCP server. To eliminate *duplicate address detection* (DAD) delay we implement a scheme of address reservation in advance. Under this scheme, a process running in the access point reserves a number of addresses and keeps them alive by running the DAD in the background. Moreover, our proposal is a full compatible approach: a MN with our solution can operate in classical DHCP context. A node can distinguish between a FCLH context and a classical context by the options contained in the DHCP ACK. If the MN realizes that DHCP ACK does not include the options it waits for, then it starts a classical procedure. On the other hand, when a classical DHCP server receives DHCP messages with extended options, it just drops the options it does not know and continues a classical procedure. Figure 9 shows the AA procedure in FLCH.

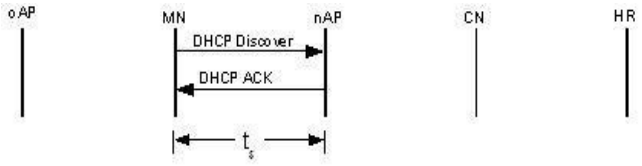


Fig. 9. Address allocation with DHCP-FCLH.

Under this address allocation scheme, an address can be obtained from the *new access point* (nAP) in only $2t_s$, where t_s is the delay of the channel connecting the MN and the access point.

Preservation of the ongoing communication

At the same time, the DHCP ACK message is sent to the MN, SIP re-invite message is sent to the CN. Indeed, the access point builds a SIP message by using the information contained in the extended option of the DHCP Discover message. To send this message, the AP acts as a router and emulates the SIP re-INVITE message as if it was sent by the MN. This is possible because the access point decides which address will be allocated to the MN, from the list of reserved addresses. Once the SIP re-INVITE message has been accepted by the CN, it finally sends an OK response to the MN. The different events of the handover process are described in Figure 10. It should be noted that this approach is cross-layer because in this case, a link layer message generates a SIP re-INVITE message without respecting the classical sequence of events neither the hierarchy of the protocol layers.

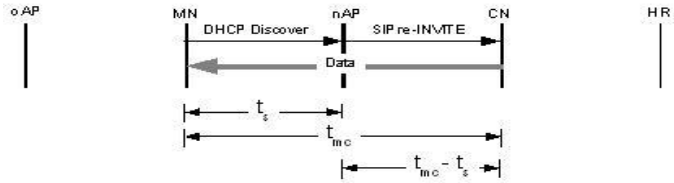


Fig. 10. Preservation of the ongoing communication: the handover process.

The service disruption time during the handover process is as follows:

$$T_{fclh\_inter} = t_s + (t_{mc} - t_s) + t_{mc}$$
$$T_{fclh\_inter} = 2t_{mc} \quad (5)$$

The Smooth handoff in FCLH is achieved by redirecting the data packets received by oAP to the nAP before the CN knows that MN has changed its network attachment point. The nAP requests this service to oAP by means of a special message which contains both the old and the new address of the MN. In contrast to MIP, our approach does not require the establishment of a tunnel and the encapsulation of the original data flow. This method improves the procedure performance and simplifies the implementation.

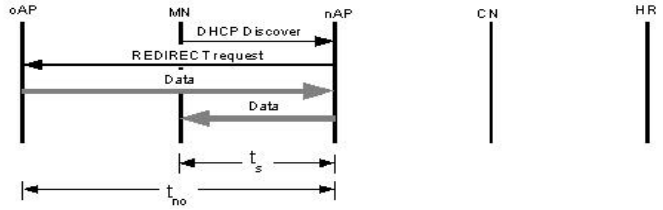


Fig. 11. Smooth handoff in FCLH.

More specifically, the access point has to change only the IP header of the packets, recalculate the CRC (Cyclic Redundancy Check) and finally redirect the data flow to the new MN address (Figure 11). The time required to obtain the smooth handoff is calculated as follows:

$$T_{fclh\_smooth} = t_s + t_{no} + (t_{no} + t_s)$$
$$T_{fclh\_smooth} = 2t_s + 2t_{no} \quad (6)$$

Maintaining the global location tracking

Once again, the SIP Register message is generated by the access point after the reception of the message DCHP

Discover. The Information contained in this message as well as the address chosen by the access point are used to generate a SIP Register message. The MGLT process is described in Figure 12.

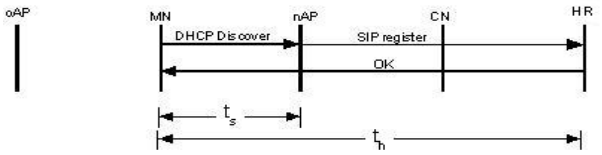


Fig. 12. Maintaining the global location tracking.

The delay required to update the LD, or HR in the context of SIP, is only t_h.

IV. PERFORMANCE EVALUATION

In this section, our discussion is based on the above analysis. More precisely, this work compares the performances of MIP and SIP for an application of voice over IP (VoIP). So, the test conditions used here are the same as those considered in [6]. We assume t_s = 10 ms which corresponds to a relative low bandwidth for the wireless channel. For the wired network connecting wireless access points, we consider a more important bandwidth, then a smaller delay t_{no} = 5 ms. On the other hand, we suppose that processing time of the different entities is negligible.

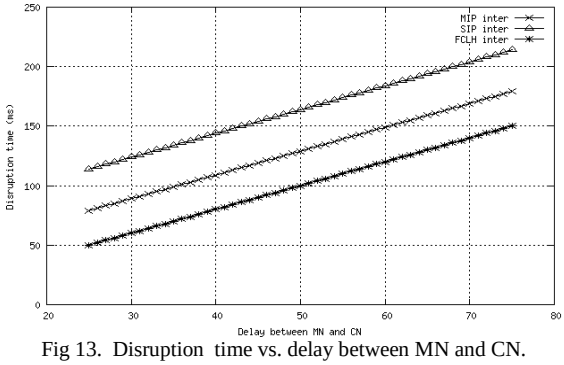


Fig 13. Disruption time vs. delay between MN and CN.

We take three different configurations. In the first one, the MN is connected to the network via a wireless channel and the distance of the CN varies. In the second configuration, the CN and the MN are close but the distance from the MN's home network varies. Finally, in the last configuration the delay of the wireless channel varies. In Figure 13, we can see that the disruption time increases as the delay t_{mc} between the MN and CN increases. In Figure 14, t_h increases, t_{mc} = 25 ms, and the wireless link delay is equal to 10 ms. Observe that the disruption time associated to SIP becomes smaller than MIP as the delay between MN and its home network increases. MIP disruption time increases because the handover delay depends on the registration within the HR. As far as our approach is concerned, the handover process depends on the delay between the CN and the MN only. That explains why the disruption time is constant.

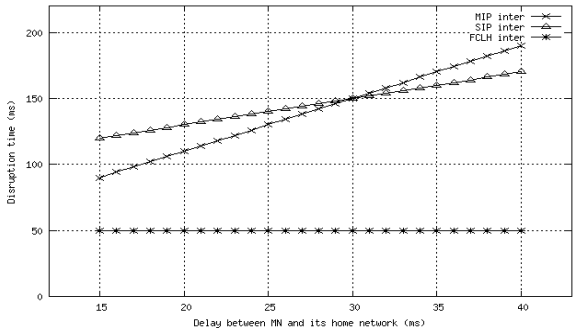


Fig 14. Disruption time vs delay between the MN and its home network.

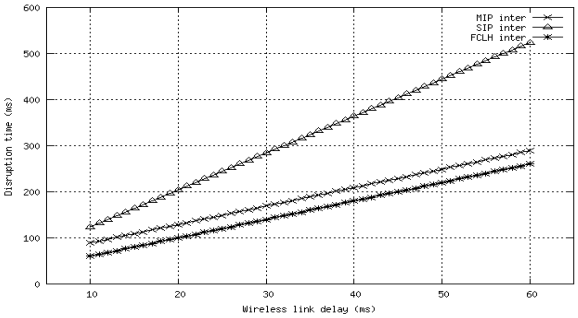


Fig. 15. Disruption time vs. wireless link delay.

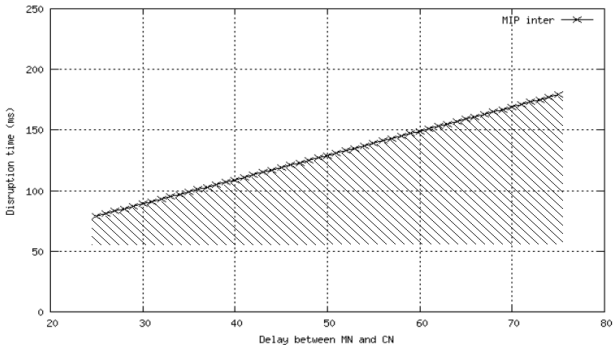


Fig. 16. Smooth handoff area for MIP.

Finally, the last scenario demonstrates the impact of the wireless delay on disruption time (see Figure 15). We can see that the impact of this parameter is limited on the total handoff delay in the case of FLCH. This result is due to minimization of signaling packet exchange over the wireless channel during the hand over.

As far as smooth handoff is concerned, MIP takes advantage over SIP which lacks this mechanism. By taking into account $t_{mip\_smooth}=2t_s+2t_h+2t_{no}$, the MN in MIP starts receiving data packets before the handover is accomplished. This period can be calculated as being $T_{mip\_inter}-T_{mip\_smooth}=T_{mip\_inte}-54$ ms (see Figure 16). In our approach $T_{fclh\_handoff} = 2t_s + 2t_{no}$, therefore the MN receives forwarded

packets for a period of time equal to T_{fclh_inter} -$T_{fclh\_handoff}$ = T_{fclh_inter} – 30 ms (see Figure 17).

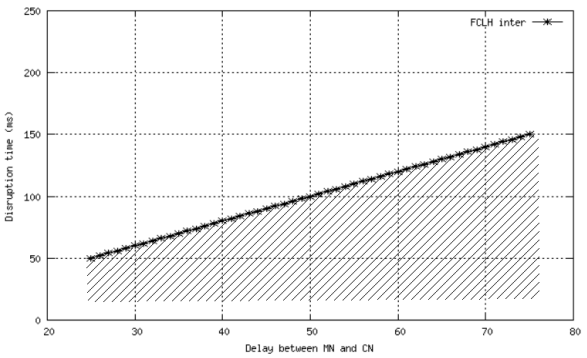


Fig. 17. Smooth handoff area for FCLH.

The smooth handoff period starts earlier in our approach than in MIP, that is to say, we can recover packets earlier and for more time before the handover is accomplished. On the one hand, the smooth handoff should start as soon as possible in order to avoid damaging the user perception. On the other hand, the faster the smooth handoff occurs the smaller the buffer size allocated to packet collection before redirection in the AP is.

V. SECURITY CONSIDERATIONS

Since our scheme rely on DHCP, it is mandatory to secure the access to DHCP service by authenticating the clients and encrypting the payload witch contains parameters related to the current session. Malicious client can generate a lot of address requests to prevent the legitimate users from acquiring their IP addresses. The second attack is when a malicious DHCP server in the network answers to client requests and provides bogus configuration that prevent them from using network resources normally. As we talk about mobility in wireless networks, the first thing that should be carefully secured is the physical medium with adequate authentication and encryption protocols so that only authorized client can access physically the network. To overcome DHCP weakness, IETF proposed an authentication option [11] for DHCP messages. This option allows only authorized client to request for address configuration and allows for the clients to authenticate the server identity. Moreover, the threat can still come from classical DHCP users since the critical point is maintaining a pool of addresses ready for use by mobile hosts. We propose to reduce the lease for these addresses compared to addresses intended for non mobile nodes. We also have to encrypt the DHCP option containing information related to the different operations to prevent man-in-the-middle attacks.

VI. CONCLUSION

In this paper, we proposed the introduction, into access point equipments, of some functionalities to improve the handover performances. These functionalities are based on the reduction of the address allocation delay, the number of exchanged signaling packets as well as the parallel execution of certain procedures. Our approach is implemented over the signaling mechanism of DHCP in such a way that the proposed functionalities can be used by means of extended options. The obtained results indicate that our approach can reduce the handover delay with respect to most popular and already improved approaches such as MIP with Route Optimization as well as SIP. Moreover, our proposal does not require the introduction of additional entities in the network neither modifications in the current protocol stack, and in contrary to certain approaches, which do not consider the address allocation delay in the calculation of their results, we consider this problem and solve it. Finally, we presented the main attacks that threaten our system, and proposed some methods to make it more secure. Future research will be orientated to the simulation of this approach and its study in more heterogeneous context.

REFERENCES

[1] L. Rojas-Cardenas, M. Boutabia, and H. Afifi “an infrastructure-based approach for fast and seamless handover”, 3rd International Conference in Digital Telecommunications ICDT 2008, Bucharest, 29 June -6 July 2008.

[2] Park, S., Kim, P., and Volz, B., "Rapid Commit Option for the Dynamic Host Configuration Protocol version 4 (DHCPv4)," RFC4039, IETF, March 2005.

[3] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg. SIP: Session Initiation Protocol. rfc 2543, IETF, March 1999.

[4] C. Perkins, "IP Mobility Support", RFC 2002, Internet Engineering Task Force, October 1996

[5] A. C. Snoeren and H. Balakrishnan, "An End-to-End Approach to Host Mobility", 6th IEEE/ACM Mobicon 2000. Boston, August 2000.

[6] N. Banerjee, W. Wu, and S. K. Das,"Mobility Support in Wireless Internet", IEEE Wireless Communications, October 2003.

[7] T. T. Kwon, M. Gerla, S. Das, S. Das, "Mobility Management for VoIP Service: Mobile IP vs. SIP", IEEE Wireless Communications, October 2002.

[8] Nen-Chung Wang and Yi-Jung Wu, “A Route Optimization Scheme for Mobile IP with IP Header Extension”, IWCMC'06, Canada, July 3-6, 2006.

[9] D. Johnson, C. Perkins, and J. Arkko, ”Mobility Support in IPv6”. RFC 3775, IETF, June 2004.

[10] R. Koodli, “Fast Handovers for Mobile IPv6”, RFC 5268, IETF, June 2008

[11] R. Droms, W. Arbaugh, “Authentication for DHCP Messages”, RFC 3118, IETF, June 2001.

[12] A. McAuley et al., "Dynamic Registration and Configuration Protocol (DRCP) for Mobile Hosts," Internet draft, draft-itsumo-drcp-01.txt, July 2000, work in progress.

[13] W. Stevens. UNIX Network programming. Prentice Hall

[14] S. Mohanty, F. Akyildiz “Performance Analysis of Handoff Techniques Based on Mobile IP, TCP Migrate, and SIP”, IEEE Transactions on Mobile Computing, VOL.6, NO.7, July 2007.

[15] K. El Malki “Low-latency Handoffs in Mobile IPv4”, RFC 4881, IETF, June 2007.

Retrieving Information from Hybrid Spaces Using Handhelds

Spyros Veronikis[1], Dimitris Gavrilis[2,3], Kyriaki Zoutsou[1,4], and Christos Papatheodorou[1,2]
[1]Department of Archives and Library Sciences, Ionian University
72, I. Theotoki, GR-49100
Corfu, Greece
[2]Digital Curation Unit, Athena Research Centre
6, Artemidos & Epidavrou, GR-15125
Athens, Greece
[3]Library and Information Service, Panteion University
136, Syggrou Av., GR-17671
Athens, Greece
[4]Library and Information Service, University of Patras
University Campus, Rio, GR-26504
Patras, Greece

***Abstract*—Hybrid spaces consist of information resources of both physical and electronic form. With the advent of electronic publishing and WWW hybrid libraries became popular and widely acknowledged for their high informative quality and anytime availability. On the other hand, modern computing handheld devices and wireless communication networks can support their users in accessing and using these information volumes wherever a need arises. Therefore, the user can query an information system about the electronic resources and simultaneously explore the nearby physical resources, in a way that enhances awareness of available information collections and relations among them, and also create a new experience while seeking in a hybrid space. In this paper we present the design methodology of creating such a service in an academic library, as well as the evaluation model, the procedure and the results from assessing satisfaction for the use of that service. Our findings imply that users believe that the unified search for physical and electronic resources is an important feature when seeking information in big physical and electronic collections.**

***Index Terms*— evaluation; library service; mobile and ubiquitous computing; personal digital assistant (PDA)**

I. Introduction

This paper extends our previous work on creating and evaluating new services for hybrid libraries, which are supported by mobile computing devices, presented in UBICOMM 2008 conference [1]. During the last two decades technological advancements in the fields of computer technology, communication networks and electronic authoring and publishing resulted in a tremendous growth of data and information available to the public. The Internet and the World Wide Web successfully accommodated the majority of available computer networks, thus enabling data exchange and information sharing in a much faster way. During the 1990s, it was estimated that the Internet grew by 100% per year, with a brief period of explosive growth in 1996 and 1997 and nowadays the access and growth of available data is constantly increasing by orders of magnitude every year [2]. The information content created, disseminated and used has changed from static text to live and static multimedia, including plain and hyperlinked texts, raw data, audio/video files, images, and documents with spatio-temporal attributes. As of March 31, 2009 1.59 billion people use the Internet according to Internet World Stats [3].

To benefit from the wealth of information available libraries enrich their print collections with supervised digital sources, held either locally or in remote information organizations, such as digital libraries. Digital Libraries are information systems capable of keeping information content in collections of digital format and accessible by computers. Some of the best-known digital libraries are Project Perseus [4], Project Gutenberg [5], and ibiblio [6].

With the electronic information available evolving from structured (e.g., database tables), to semi-structured (e.g., metadata for texts and multimedia files), and unstructured (WWW pages), organization of the content had been ineffective and therefore new powerful information retrieval techniques should be implemented. Since the goal is not to just produce more data but actually to use them towards some purpose, information retrieval techniques needed to be adapted to the content evolution in order to provide valuable information to the users.

Unlike book collections, which are well structured and organized and a certain book can be easily located using author and title indexes, electronic semi-structured data follow organization principles and rules that are not very strict. As a result, indexes and logic-based query languages do not have adequate power to retrieve precisely information from the new collections. In addition, a great portion of the content provided from electronic collections is stored in distributed repositories, with different organization and metadata schemes. Due to the nature and organization of the new collections made available, new approaches based on different principles evolved, such as interoperability protocols and data integration methods.

Besides the wealth of information available, people are also interested in insights, i.e., in relationships among different data items to understand the true nature of things. Databases and search engines are not capable of pointing out these relations and therefore these technologies were supplemented using visualization and similar approaches, often called On-Line Analytical Processing (OLAP) tools [7]. These tools are well suited to gain insight from a structured source. However, they cannot provide further exploration leading to insights. Dr. Ramesh Jain [8] proposes the utilization of new systems, suitable to explore unstructured data and capable of providing to the user some insight to the information delivered. He calls them *experiential environments*. These environments can provide insight, by immersing the user to the data, allowing him to explore, experience and interact with it. In other words, they are used to bring the user into the information space available and assign him an active role in the information retrieval process, where relevancy of retrieved items is constantly evaluated and compared to nearby, related sources until the user gathers a list of data items to satisfy his information needs.

Access requirements also evolved in accordance to content and retrieval techniques evolution. It started with physical collections and local access, where the user needed to visit an information organization like a library to gain access to data items during office hours, and evolved with digital libraries and the World Wide Web to anytime access to remote systems, where the user can benefit from round-the-clock access services to digital and digitized information content. During the last years, we see one more evolution step; anywhere access to information content, i.e., the user is equipped with a mobile terminal which wirelessly communicates with computer networks and the Internet to access information content on demand, whether in an office, a teaching class, a park, or while traveling. Users are not concerned about the location of the data source as long as its quality and credibility is assured. They are interested in the result of data assimilation. Laptops, Personal Digital Assistants (PDAs), smartphones and Ultra Mobile Personal Computers (UMPCs) are some typical examples of popular mobile computing terminals to gain access to electronic information services. Using these devices, users can access traditional searching tools like library catalogs (On line Public Access Catalogs, OPACs) and indexes, as well as powerful search engines like Google and more sophisticated information retrieval tools, such as recommendation wizards, often used in large scale electronic stores.

In large public and academic libraries, such as the Library of Congress, the New York Public Library, and the Harvard Libraries, the two collections (physical and digital) are kept separately and as a result users can seek for information either by searching in the electronic catalogs from a PC or by walking to and browsing through the stacks. To avoid moving between the two spaces and overcome the discontinuity of searching in two different areas these spaces need to be brought close. The recent advancements in handheld computing devices like PDAs enable them with high resolution, colorful graphic displays, and wireless communication features. For instance, the device can wirelessly connect to a local computer network and the library's electronic services, and due to its inherent mobility its user can walk into the physical information space with an open window to the digital space, right on his palm. This allows for a uniform seeking procedure that integrates physical and electronic information collections into one, namely a hybrid information space that resembles the vision of Dr. R. Jain about experiential environments, where exploration and not querying is the predominant seeking interaction mode.

In this work we present the design procedure of creating a new library service that supports library patrons in seeking information within hybrid spaces using handheld devices, such as PDAs and smartphones. We also describe the evaluation phase of the design cycle, which aims at assessing the user satisfaction for the new service and present the derived results. Section II discusses related work in mobile computing for information services. The design procedure and service functionalities supported are presented in Section III. The evaluation method is described in Section IV and the results are presented in Section V. Section VI concludes this article with a discussion on the findings, limitations, and a brief description of future work.

II. Related Work

The potential raised by mobile devices in providing anywhere/anytime access to reference material and storing information locally, was quickly acknowledged by field practitioners, especially in healthcare environments [9]. In the beginning of the current decade several Health Sciences libraries, such as Libraries at Virginia Commonwealth Universities (VCU), were among the first to explore the PDA supporting services for medical doctors and paramedic personnel [10] [11] [12]. These mostly involved PDAs which were used for accessing reference content stored locally, such as the ePocrates clinical drug database and medical records, dictionaries and textbooks as well as writing and beaming prescribing aids.

Soon after these paradigms, devices where equipped with increased memory capacity, more efficient batteries, and higher resolution screens, making them all-around, valuable assistants for information advising. As a result there was a need for faster data exchange protocols, either wireline (Universal Serial Bus, USB) or wireless (Bluetooth and WiFi). Wireless communication features were a key factor in the usage and adoption of these portable-computing devices from a wide audience, since their users could also access and retrieve content not only locally stored on their device but also located in remote information management systems, such as digital libraries.

Buchanan, Jones, and Marsden [13] present an evaluation study on the usage of PDAs to access a remote Greenstone-based Digital Library. Their study focus mainly on the presentation issues occurring when searching and delivering content in small screen devices. However, no focus has been given to the usage of the PDAs in conventional libraries.

SmartLibrary [14] was a PDA-driven project started at Oulu University in Finland, where handheld devices were used to enable map-based guidance for book finding. A small search interface was used to submit a query to the library's OPAC and get a list of books that matched the searching criteria. Upon selection of a record from the list, the user could see its metadata and a small image of the library's floorplan, indicating the position of the book. Jones et al. [15] at Cornell University studied several application scenarios of wireless mobile devices in a library setting; these included query submitting to the OPAC from anyplace within the library, collaborative searching by leaving notes, sending emails, and communicating in real-time with group members while browsing the stacks in the library. In addition, the device could be used to capture some data from the books (e.g., by scanning or photographing part of it) and then moving the data to a laptop or desktop computer. However, neither in Oulu nor at Cornell universities access was provided to an information management system with structured and semi-structured data of electronic form, such as a digital library.

A closer approach to the usage of mobile computing devices to enrich information from the physical space with unstructured information (social tags and annotations) is the MoTag system [16], which uses PDAs to access G-Portal. G-Portal is a digital library of geospatial and geo-referenced resources that holds also social tags concerning the accessibility of public buildings and other similar structures. During their visit in a certain place, PDA users can search the G-Portal for any tags left by previous visitors, submit a photo of a location, create new tags, and also add comments and time-stamps. Similar examples come from the tourist industry and the museums. Many researchers have studied the use of PDAs in the context of city and museum guides [17] [18] for navigation and brief personalized information presentation [19] [20]. In these systems handhelds are used to display a floor plan of the current area. The map indicates nearby objects or exhibits which are available for the user to interact with in order to retrieve short descriptions about the objects and navigate in the area.

Most of the current research efforts focus on the development of applications that either facilitate mobile searching in the digital space or use the handhelds to provide navigation instructions in the physical space. Even though many libraries keep a wealth of recorded knowledge in both physical and digital form, to the authors' knowledge no studies have been made to assess the impact of a new mobile service that supports library patrons in seeking information in hybrid collections.

III. Developing the Prototype

This paper presents the design, implementation and evaluation of a service that uses handheld computing devices capable of accessing the Web to support students in searching and browsing large information environments, such as an academic library that holds data records in both physical and electronic form. In this section we describe the first stages of the design and implementation procedure, shown in Figure 1.

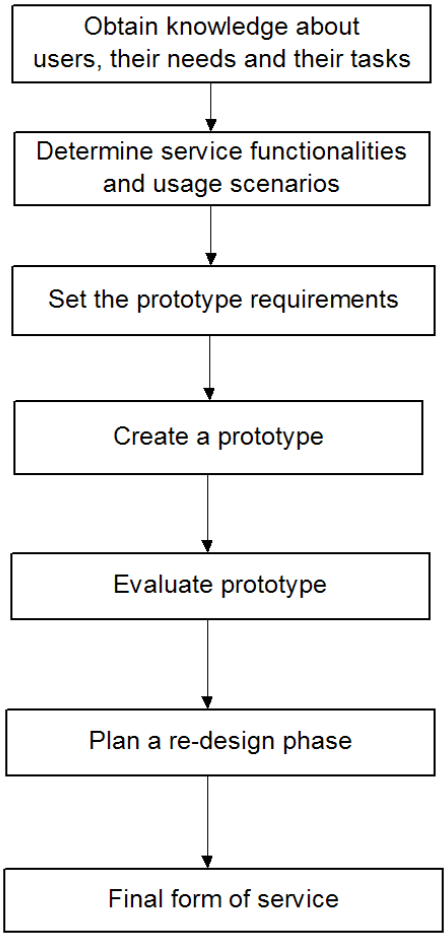


Fig. 1: Flowgram of the methodology adopted

A. The Methodology Outline

We started by conducting a focus group with 9 experts in the fields of computer science and librarianship to gain some insight of the typical library users, their needs, and seeking strategies. With this information at hand, the next step was to define the functionalities to be supported by the new service and create the necessary tools for these functionalities. Once the service functionalities were determined, we could create some typical usage scenarios to be used in our evaluation study.

The insight and data obtained from the focus group was used to inform the prototype design procedure, by providing guidelines on the interface design, set the requirements, and describe the service goals. We also had to extend the current infrastructure at the Library of Panteion University (Athens, Greece) to support wireless communication with the mobile terminals.

The next step was the design of the evaluation phase. We decided to follow a multi-strategy research, a term borrowed

from Layder [21], i.e., combine qualitative and quantitative research. More information on this kind of strategy can be found in [22]. During this phase we had to determine the research goals, determine the evaluation criteria and metrics to be used, construct a research model that visualizes the relationships among criteria, choose and create appropriate data collection instruments, such as interview guides and questionnaires, conduct a user study to collect the necessary data, analyze it, and finally report findings of our research.

These results would be used to start a redesign cycle, where prototype, scenarios, and evaluation procedure would be properly modified to improve the new service and gain some better insight to the extent that students accept it and embrace it.

B. Determining the Service Functionalities

The goal of this work was to create a new service that supports library patrons with handheld computing devices while seeking in both physical and digital information collections, in a unified way, by means of a user-centric approach, i.e., keeping the user active during the whole seeking procedure. Therefore, first we had to understand how students use the library's resources (catalogs, indexes, classification scheme, etc.) and identify any patterns in their seeking strategies and information needs. In addition, we had to explore the technical aspects rising from transferring the user-system interaction in a mobile context.

We chose to conduct a focus group in order to explore in depth these issues, both in terms of service functionalities and technical opportunities and constraints. We were particularly interested in how participants in the focus group discussion respond to each other's views and build up a view out of the interaction that takes place during the focus group discussion.

Blackburn and Stokes [23] found that discussion in groups of more than eight participants were difficult to manage. Therefore, our group consisted of 1 moderator, 2 observers, 5 librarian experts, and 4 experts from computer science. Four librarian experts were working in large academic libraries and one in a hospital library, and had a clear understanding of their library patrons and interactions with the library's resources. Two of the computer experts were working in ICT companies and the other two were working in universities in Athens, Greece. All participants joined the focus group after prior invitation. In general, invitations to group members were sent to people from the two fields that know each other, in order to explore collective understanding or shared meanings held within a work group, such as library employees.

Prior to the discussion, the participants were informed on the topics and procedure, as well as our intention to record the discussion in audio/video format. The participants received a list of scenarios making use of the functionalities for a proposed mobile service. These scenarios were derived from the literature and the participants were asked to provide their viewpoints for similar applications in a library. During the meeting they were encouraged to express their views on the proposed service functionalities so that we could study not only what they say but also how they say it, and how meaning is collectively constructed. The moderator would bring attention to specific points that are of potential interest to the focus group goals that they are not picked-up, and refocus the participants' attention to the topics of interest, in case the discussion goes off. The video recordings and the observer notes would later enable us to study and analyze the discussion.

Compared to individual interviews, focus group discussions many times appear to be less efficient due to several limitations. The following are some typical examples of these limitations; the extent to which it is appropriate to control the interaction between participants in order to have an in-depth discussion with multiple viewpoints and stay focused on the specific topics of interest; group effects such as dealing with reticent speakers and those who dominate the discussion. Asch's experiments [24] revealed that an emerging group view might mean that a perfectly legitimate perspective held by a minority of speakers may be suppressed. Therefore the moderator had to control the discussion and make clear that other peoples' viewpoints are definitely required.

To ensure that at the end of discussion we would have a clear viewpoint of each speaker on the topics and the proposed service functionalities discussed, we asked them to fill a short questionnaire used to express their attitude on a 5 -point Likert scale towards adopting (or not) the proposed information seeking aids.

To transcribe the focus group discussion we had to analyze data captured from two video cameras, a microphone, the observers' notes, and the questionnaires. We created a table in which each row represented a discussion topic, each column represented a speaker, and each cell included the corresponding time-stamped user comments, any observer notes, and the questionnaire score. These procedures allowed us to easily summarize the discussion, compare participants' viewpoints, and study the procedure of forming a group view.

To select the functionalities to be implemented for the new service we set an acceptance threshold, proposed by Nielsen [25], based on the emphasis given during the discussion; for a functionality to be selected it should (a) have an average score over 4, (b) at least 7 participants (80%) should have given it the top rates (4 or 5), and (c) no more than 1 participant (10%) should have given it the lowest rate (1).

The average values (AVGs) and the standard deviations (SDs) of the functionalities that survived the selection criteria were: (a) the wireless access to the OPAC and the e-resources of the library (AVG= 4.63, SD= 0.52), (b) the use of a map indicating a book's location in the stacks (AVG= 4.25, SD= 0.70), (c) the ability to communicate with the mobile device directly with other on-line users or send a short message/email to be received later (AVG= 4.13, SD= 0.99), (d) the ability to download/ save/disseminate electronic files retrieved, such as journal articles and lecture notes (AVG= 4.38, SD= 0.52), and (e) the ability of taking some quick notes either written or verbal (AVG= 4.0, SD= 0.75).

Service functionalities that did not survive the selection

criteria were mostly due to two reasons; privacy violation concerns and reduced utility value. These functionalities were: on-line user tracking to provide navigation instructions within a library setting (AVG= 3.5, SD= 1.20); storing user navigation routes to extract information about subject areas of interest and other preferences (AVG= 2.63, SD= 1.06); creating a patron profile to keep personal information (AVG= 3.50, SD= 1.20) that would be stored in the library's servers and updated from the student-system-content interaction in order to create content for personalized information services (e.g., alerts, notifications, recommendations, interfaces, etc.); wireless printing (AVG= 3.75, SD= 0.89) so that the students could immediately send a note or article for printing from anyplace within the library setting; route recommendation to collect books of interest (AVG= 3.00, SD= 1.20), which was shown not to be of particular interest to the students due to familiarity with the small size of academic libraries and the small number of books usually borrowed.

In addition, automatic metadata retrieval of books by detecting a Radio Frequency Identification (RFID) tag placed in the book (similar to scanning a barcode) was considered an attractive feature for the service (AVG= 4.13, SD= 0.64), especially when trying to detect and retrieve certain book or content within a stack. However, this feature was not implemented during the first design cycle due to extra equipment cost and time constraints.

With the service functionalities and corresponding tools determined, we could proceed to scenario descriptions, design requirements and guidelines, and prototype implementation. In a typical usage scenario a student uses the device to submit a query to the mobile OPAC. From the results list she sees the desired book and other related print works. Using the stylus she taps on the desired book to retrieve its metadata and sees that there are a few copies available on the shelf and a map indicating the location of the book in the stacks. While walking to the stacks to locate the book she activates her instant messaging (IM) account. Having found the book, she takes a quick note for the other related books of the author and sends a short question to the on-line librarian asking to inform her on due dates for previously borrowed books. Without needing to head for the computer room, she now searches the library's electronic resources for relevant entries. The results list shows a couple of records that seem relevant. She decides to download an article and send an email with its metadata, including a download link, to a colleague.

C. Designing and Implementing the Prototype

The focus group discussion revealed the need to design for users with diverse experiences, skills and knowledge concerning the information technology and collection usage, and cater for both novice and experienced users. In addition, we had to make the user-device interaction simple so that users could keep interacting with the physical environment and collections as much as possible, and spend their time effectively towards fulfilling their seeking goals. In other words, we had to keep their mental effort workload at low levels, so that they could keep touch with both information domains while seeking, despite frequent interactions.

Yet, the biggest challenge for the design phase was to create an interactive information service that would provide information on the spot, in the desired level, with flexible search options, via a device with constrained computing and interaction resources (e.g., processing power, screen size, and lack of keyboard). Whenever possible, interaction with physical objects, e.g., via metadata codes as information containers or pointers to other resources (Barcodes, Quick Response (QR) codes, RFID tags) should be exploited in order to speed-up the seeking process. That way, we can avoid unnecessary steps in searching, and further enhance experiential seeking and integration of the two information spaces (physical and digital) [26].

Regarding access to the service, it was decided that it would be implemented using a client-server architecture to reduce computing demands on resources to the mobile device, and that service should be web-based so that it can be accessed by any computing device capable of web browsing. Furthermore, we had to install a wireless computing network (WiFi Local Area Network) to make the service available from any place within the library setting.

In addition, the architecture should be modular to be easily upgradeable, i.e., each component in the architecture should be easily removed in the future and replaced by an improved version of it with better performance characteristics. The new service should also be designed for at least comparative usability to the currently available seeking service, i.e., catalog searching from a desktop terminal.

To create the new service, we consulted some of the most representative interface guidelines available; Shneiderman's and Plaisant's "Golden Rules for Interface Design" [27] and the "Ten Usability Heuristics" by Nielsen [28] apply to handheld design, since they are independent of specific technologies and device form factors. In addition, we considered basic design principles from "Apple Human Interface Guidelines" [29] and "Gnome Human Interface Guidelines" [30]. These guidelines typically include principles such as design for a variety of people profiles, using meaningful metaphors between application service and real world working cases, keep the application interfaces consistent, keep the user informed during processing and idle times, keep the interaction simple and pleasant, put the user in control of the interaction with the system, cater for simple and intuitive interaction, design well-defined dialogues, provide simple and unambiguous navigation, forgive the user when making mistakes, provide adequate help and examples for complex tasks, and provide feedback and communication on users' actions.

In addition, we studied three of the most comprehensive interface guidelines available for mobile devices; the PalmOS User Interface Guidelines by PalmSource Inc (now owned by Access Systems Americas, Inc.) [31]; Windows Mobile 6.0 - Design Guidelines by Microsoft Corp. [32]; and iPhone Human Interface Guidelines by Apple Corp. [33].

Palm presents some basic design principles and guidelines

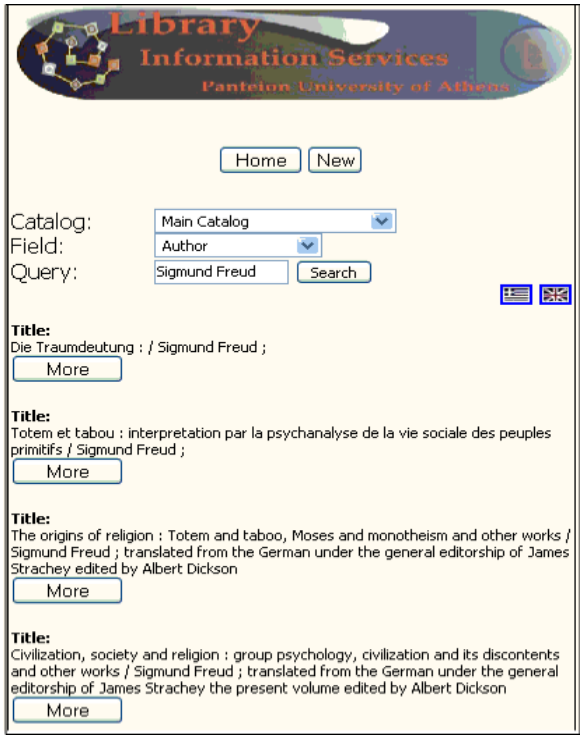


Fig. 2: The search interface for the mobile device

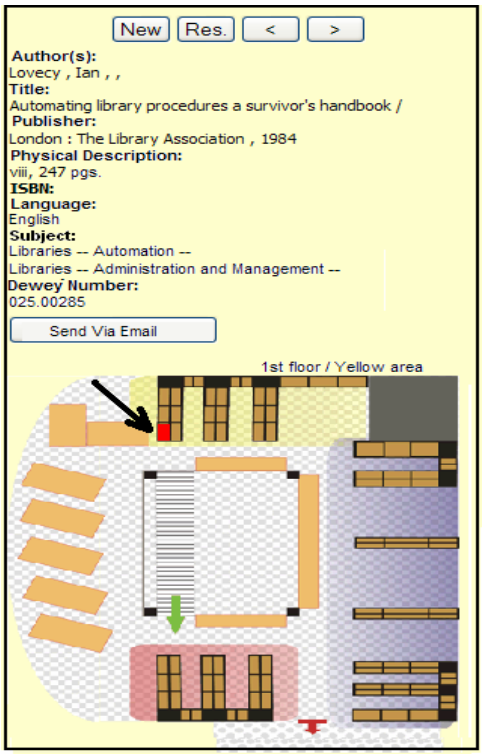


Fig. 3: Metadata for a record in the library's catalog (OPAC)

for the design process and concludes with recommendations and descriptions for getting feedback from the user and giving output back to him. Microsoft instead of giving some basic design principles, provides some application specific guidelines regarding home screen, web site design for mobile devices, navigation, screen rotation, soft-keys and menu operation, usability and interfaces for user-device interaction (e.g., screen layout and text input). Apple starts by covering the fundamental human interface design principles, describes how to apply them in designs for mobile applications, and moves on to description of the various views and controls that are available to the designer, along with guidance on using them effectively. Some of these guidelines are summarized here; design with pocket size in mind, i.e., limit data entry, hide unnecessary menus, do not use toolbars, provide only options that are usually needed to save screen space; keep interaction fast and simple by increasing speed and minimizing required steps to issue a command, and optimize frequent tasks; provide seamless connection with desktop computers since handhelds are used to extend desktop capabilities with the mobility feature rather than replace them; whenever possible choose "low-absorbing" interaction techniques and reduce short-term memory load to prevent user from loosing contact with the two information domains as a result of dealing with interaction issues; design for short, frequently interrupted tasks since the users will be moving in the library, thus constantly changing their working environment; ensure easy and permanent access to all library resources and areas where the user is expected to move; cater for effective and usable content delivery; design for familiarity with existing tools, i.e., use desktop-similar interfaces to benefit from usage familiarity.

To implement the prototype system we first had to create a mobile version of the search interface for the library's sources, suitable for small-screen devices. Two snapshots of the interface are shown in Figure 2 and Figure 3. The service is available from Panteion Library's website [34]. That interface would be used to build a searching query by submitting the searching term(s), the searching field (author or title), and the searching collection (either library's book catalog or the electronic resources). Upon a query submission a results list is sent back to the handheld and the user can tap on the record of interest to see its metadata. Each record from the physical collection is associated with a map indicating the corresponding item's location. For records in the electronic sources of the library, the user has the option of sending its metadata to an email account. These will typically include a downloading Uniform Resource Locator (URL). However, users are also allowed to download any available full-text material on the mobile devices. Furthermore, the instant messaging tool allows for short dialogs with other on-line users and library staff.

IV. Evaluation Design

Having developed the prototype, the next phase as shown in Figure 1 was its evaluation. The prototype described above was our first attempt to implement the new service and therefore a descriptive evaluation design (also called observational design) was adapted, i.e., an approach that would collect data of diverse nature (unstructured data from qualitative research

and structured from quantitative research). This is sometimes called a multi-strategy approach as mentioned earlier. The advantage of this approach is that it produces rich information content and insight from the data collected, which could not be reached by choosing either a qualitative or a quantitative approach alone. Many researchers argue that this triangulation can provide confidence in findings deriving from the study [35] [36], whereas not all researchers agree that a multi-strategy approach is always desirable or feasible [37] [38]. On the other hand, the amount of combined research in the social sciences has been increasing since 1980 [22].

Our evaluation objective was to study the usage of a new service in a library setting containing information in both physical and electronic format, and study the students' satisfaction and intention to use it. Particularly, we were interested in whether the mobility offered would be a valuable feature to the users while seeking in diverse information domains, and identify which factors and service capabilities mostly affected their interaction with the mobile device.

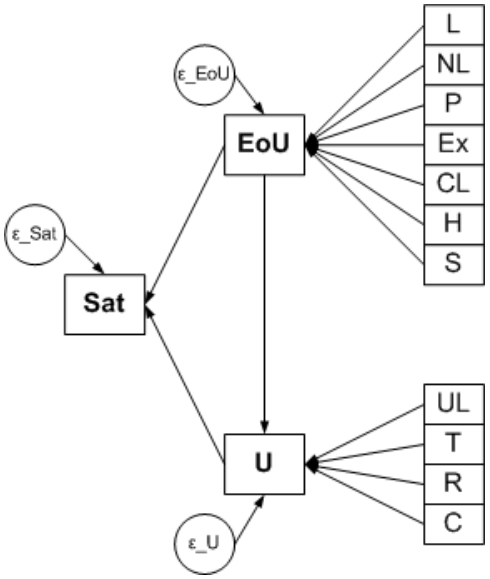


Fig. 4: Criteria and path relations of the evaluation model

A. Evaluation Criteria and Model

Several researchers agree that usefulness and usability are the most significant concepts for the user-centered evaluation of information services [39] [40]. Therefore, in a user-centered model that evaluates the impact of the new service to its users we need to examine the users' *Satisfaction (Sat), Usefulness (U),* and *Ease of Use (EoU)* towards that service and the effects between them. Usefulness is defined as "the degree to which a person finds that using a particular system or service will enhance his/her job performance". Ease of Use is defined as "the degree to which a person finds that using a particular system will be free of effort". We hypothesize satisfaction to be expressed in terms of usefulness and ease of use and positively related to both of them, i.e., the bigger the usefulness of the service, the bigger the user satisfaction. On the other hand, *Usefulness* and *Ease of Use* are complex constructs and therefore they can be broken down to simpler indicators that are easier to measure.

To assess the usefulness of the new service in finding and collecting the records of interest in hybrid collections, we use the following indicators (criteria): *utilitarian value (UL)*, which refers to the value the new service has in supporting the users to achieve their goals [41]; *time (T)*, which refers to the time earned from the usage of the service; *relevance (R)*, which refers to the relations among retrieved records from diverse information domains; *perceived completeness (C)*, which refers to the user's information needs obtained by the usage of the new service [42] [43]. Time saving and utilitarian value were addressed towards specific actions that were described in the scenario of Section III-B, such as information searching, file download and storage, communication, etc.

On the other side, ease of use is assessed by taking into account several interface attributes [41] [25], such as *learnability (L)*, easiness of transition between seeking tasks when following a *non-linear (NL)* seeking strategy, *organization and presentation (P)* of information delivered, *easiness of executing* and completing (Ex) the various seeking tasks, *clarity (CL)* and understandability of interaction with the service, on-line *help (H)* adequacy, and easiness in getting remote *support (S)* while seeking.

Table I summarizes the criteria referred in the bibliography and adopted in the present study whereas Figure 4 shows the relationships among them along with the corresponding error (ε) estimators. Relationships in the model described above can be moderated by various user-related factors; for example, users of different background, e.g., different level of computing experience, may perceive differently the ease of use of the new service. In addition, many usability studies have shown that user interfaces have a strong impact on Ease of Use. The current study describes our first evaluation approach in exploring users' attitudes towards the new service, and therefore its nature is rather exploratory than confirmative. For this reason, we do not pose and validate any testing hypotheses.

TABLE I: Evaluation criteria

Construct	Criteria
Ease of Use	a) Learnability, b) Task transition, c) Information presentation & layout d) Ease of task execution, e) Clarity, f) Help g) Remote support
Usefulness	a) Utilitarian value, b) Time saving, c) Relevance, d) Completeness

B. Experimental Setup

The next phase was to setup an experiment to collect data from users of the service. To recruit users we planned to have the evaluation conducted at an academic library (Panteion University, Greece), where students and academic stuff would

be typical users of the library. Since this was the first evaluation of the service we were mostly interested in collecting data and insight from a wide variety of users rather than a particular group (e.g., freshmen). Therefore all typical library patrons were eligible for participation in our data collection experiment. To increase the number of participants and reduce the experiment costs in terms of time and human resources, we asked for the contribution of teachers in 3 academic classes at Ionian and Panteion universities. Graduate and post-graduate students were motivated by their teachers to participate in the evaluation. They were encouraged to use the new service in order to collect bibliography records for their semester projects, resulting in 77 participants.

Prior to the experiment, participants were invited into a 30 minute briefing session where they were informed about the goal and the procedure of the experiment and also had a hands-on experience with the PDAs. According to the procedure, each student would borrow the PDA from the library's help-desk and would also be given a usage scenario similar to that described in Section III-B. Students were allowed to change the order of tasks described in the scenario, but they had to complete all the tasks. While participants were interacting with the mobile device and the service interfaces, their sessions were recorded (screen-captured) and transmitted in real-time to a remote PC. Two observers were also present to watch the interaction and guide the users through the procedure. The remote recording technique produced valuable content for the qualitative analysis of the next phase, in a way that is less intrusive to the experiment subjects. Upon completion of the tasks students either participated in an in-person interview or were asked to fill a questionnaire, describing some of their profile characteristics such as age, academic level and computing experience, as well as their experience from the interaction with the new service. Ten students were randomly chosen to be interviewed, resulting in 10 in-person interviews and 67 questionnaires (http://dlib.ionio.gr/hls/texts/evals/ev1/qsts).

V. Analysis and Results

In the following paragraphs we describe the analysis procedure and findings occurring from both quantitative and qualitative data collected from the evaluation phase.

A. Collecting and Analyzing Qualitative Data

The interview method was chosen in order to accumulate the users' comments aiming at a straight and representative qualitative evaluation of the validity of our research assumptions. The 10 interviews conducted where strictly personal in order to avoid any effect between the participants.

Semi-structured interviews were conducted based on both open-ended and closed questions to compensate for the drawbacks of each form. Two were the reasons for following the specific type of interview; the researchers intended to assure the aggregation of a minimum set of data and, on the other hand, to give the opportunity to the users to express their opinion freely without the interviewer losing control of the discussion.

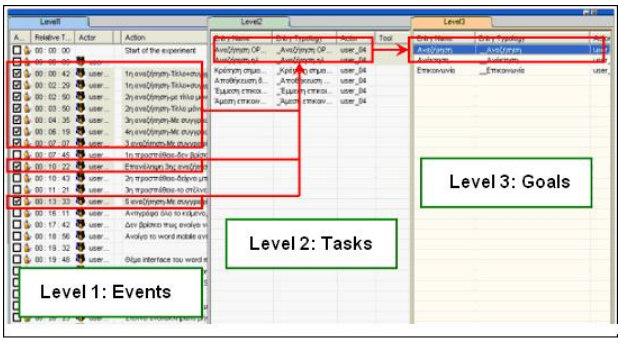


Fig. 5: Activity Lens: hierarchical view of multilevel analysis

Each interview lasted nearly 20 minutes and was recorded on video with the acceptance of the interviewees. Their opinions on usefulness, usability, and satisfaction for the new service were assessed using a 5 -point Likert scale (ranging from strongly disagree to strongly agree) against certain statements that the interviewer posed.

A remarkable amount of data was collected by the end of the interviews requiring an in depth analysis. For that purpose we used ActivityLens (AL) software [44], which is a tool especially designed to support ethnographic research studies and facilitates the analysis of data from multiple sources. These data can be audio/video recordings, log files, images, and text files (even hand-taken observation notes). AL permits integration and synchronization of the heterogeneous collected data. The AL environment was developed by the Human−Computer Interaction Research Group (HCI Group), of the Electrical & Computer Engineering Department, at University of Patras. AL was chosen among equivalent software, e.g., NVivo, Observer, and Transana because of its strength as a tool for qualitative analysis, its ease of use, the structural composition of data, and the full support that the creative team committed to offer.

ActivityLens supports hierarchical task analysis and therefore the recorded interaction of each user was classified in three levels of abstraction. We inserted coarse features of the collected data on the first level and extracted more qualitative results using the filters that the software offers on higher levels. At the lowest level (Operations or Events Level) the events for a task completion were annotated, according to a set of typologies; for instance an event could be annotated as "successful search to OPAC", "unsuccessful copy of metadata to the notepad", etc. In the middle level (Action or Tasks Level) the events of a user were grouped into tasks, according to the usage scenario, e.g., searching, typing, chatting, emailing, etc., with new typologies used to annotate the successful completeness (or not) of a task. Finally, at the third level (Activity or Goals Level) tasks were outlined as goals (seeking, communicating, etc.), while new typologies denote if the user achieved a goal. The hierarchical view of multilevel analysis is shown in Figure 5.

Regarding the users' profile it was found that more than half of the interviewees were visiting the library frequently

and were familiar with its physical space. In addition, 9 of them also use its electronic resources and 6 use web search engines in a daily basis. However, only 3 had previously used a mobile computing device, such as a PDA for an extended period of time.

The interviews revealed that users were very enthusiastic for the option of quickly searching in both information spaces from anyplace within the library, right when the need arises. This finding indicates the usefulness of the new service, which is also verified by the interviewees' Likert rates.

With most usability scores for the service interfaces above 4 (in scales ranging from 1 to 5), we conclude that the users found it easy to interact with the device. This was due to design resemblance of the mobile interfaces to their desktop counterparts. The most useful functionality was the capability of emailing the retrieved information (AVG= 4.44, SD= 0.73), followed by information storage in mobile disks, such as a memory card (AVG= 4.33, SD= 0.71). One of our participants emphasised on the usefulness of these functionalities by saying *"it is really important for me to store and send via mail anything that interests me the exact time that I find it, as simple as the pressing of a button"*.

Similarly the average usefulness of searching in the OPAC using the device was 4.33 (SD= 0.71), while it was found very convenient and easy to use it anytime/anywhere (AVG= 4.67, SD= 0.50), in contrast to the library's terminals. The usefulness of searching e-resources with a handheld was mostly affected by the relevance of the retrieved information (AVG= 4.11, SD= 1.05). This functionality also received a high usability score (AVG= 4.33, SD= 1.12). It is worth noting that those participants who quickly retrieved information records highly relevant to their interests, tended to perceive the functionality of searching in the electronic resources as the most useful. In addition, participants asked for the option of simultaneously choosing both collections (OPAC and e-resources) as searching targets.

Concerning the ability of taking quick notes, only three of the interviewees found it useful (AVG= 3.56, SD= 1.42) and reported that it was easy to copy/paste metadata (AVG= 4.22, SD= 0.97). The majority of our sample is used in keeping hand-written notes. In fact one of them mentioned: *"I always carry my own notebook and I am used in working that way. I can hardly change this habit even though I overcame all the usability problems that I came accross"*. On the other hand, 30% of our sample underlined the usefulness of taking notes on a handheld device because they considered it as a time saving procedure due to copy and paste commands available.

The video recordings revealed difficulties in text input using the virtual keyboard, which was a totally new experience for seven of the participants and gathered 45% of the complaints regarding the user-device interaction. The small screen size and stylus followed at 14% each, and the remaining 23% regarded navigation instructions, presentation style, device dimensions, and interfaces. However participants stated that after some training period these difficulties would not be strong enough to obscure the usefulness of the new service.

Some tasks gathered negative comments regarding their utilitarian value. At the bottom of the rank participants placed the use of the navigation map into the physical area (AVG= 2.33, SD= 1.66). Moreover the usability of this functionality was characterized indifferent (AVG= 3.0, SD= 1.22) and that was partly due to its reduced usefulness. Participants considered the service useful for new visitors, like freshmen, and for larger buildings. Surprisingly, one of the interviewees justified the lack of usefulness of the navigation map by saying *"It is boring to look for a book in the library. I prefer to ask directly the librarian instead of wandering through the stacks"*.

Furthermore, participants would prefer more vivid identification patterns, such as the existence of an indicator of the user's position in real-time. For similar reasons participants rated low the usefulness of the synchronous communication with the reference librarian (AVG= 3.22, SD= 1.39). Concerning the information completeness, they would like to have an indication of the number of hits in the search results as well as a relevance indicator next to each record.

Overall, the majority of the participants (nine out of ten) declared satisfied with the new service and they described it as innovative, interesting, and interactive. Anywhere/anytime access to the library's content and services is time saving and enables the users to easily swap the seeking target collection in an iterative fashion until they are satisfied with the resulting list. The meaning and the usefulness of the new service is reflected on the statement of one participant: *"The handheld device allowed me to implement a combined search to the hybrid information space, namely I can search and retrieve both books and electronic sources at the same time achieving more complete results"*.

Regarding the interaction with the mobile devices, users do not find the device's constraining resources (screen-size, lack of keyboard, low memory, etc.) to be a good reason to reject the new service. All of the interviewees intended to reuse the new service and recommend it to a friend or colleague.

B. Quantitative Analysis

All data for the quantitative analysis came from questionnaires that the participants used, in order to extract information about their profile and assess the evaluation criteria to be used. We used open-ended and multiple choice questions to collect these data. Questions regarding their user profile were coded using nominal variables (such as gender). Table II shows the frequency of their responses for their profile. We see that the majority of the participants in our study were female, active information searchers through the Internet channels and e-resources, and holding a bachelor degree.

Likert scales were chosen to express the extent to which participants agree (or disagree) to several statements, related to the evaluation criteria, such as *"It is easy for me to learn how to use the mobile device"*. Satisfaction (Sat) and perceived completeness (C) where assessed in 10-point Likert scales so that subjects could easier assess their attitudes towards these criteria. The rest of the criteria were assessed on 7-point scales. To proceed to the analysis stage, participants' responses

TABLE II: Frequency table showing the users' profile

Category	n	Percentage (%)
Male	18	26.8
Female	48	71.6
E-source usage	49	73.1
BSc level	53	79.1
MSc level	9	13.4
PhD level	5	7.5

TABLE III: Dominant predictors of user satisfaction

	B	SE B	β
Constant	0.35	10.95	
Time earned	5.08	1.28	.44*
Learnability	5.65	2.10	.31**
Completeness	0.20	0.80	.27***

Note: R^2= .57, *p≤.001, **p≤.01, ***p≤.05

were assigned to ordinal variables, i.e., besides recording an attitude/belief towards a statement we also recorded the order in which these attitudes occur. Like ordinal variables, interval (or scale) variables are used when the intervals between data points are equal for the whole measurement scale, so that there is a meaningful interpretation of the differences between data points. However, Jöreskog and Sörbom [45] suggest that an interval variable should be used only when data can be measured on at least a 15-point scale.

The statistical analysis is sensitive to missing data and there are several approaches available in statistical packages to handle such a situation [46]. For our analysis, missing values in the questionnaires were excluded pairwise, which means that if a person had a missing value for a particular variable, then his/her data were excluded from calculations involving only this variable.

For the current study we chose to use multiple regression analysis as a method of describing the relations and effects among the recorded variables, rather than predicting an outcome from recorded indicators. Having our model specified in Section IV-A the next step was to check whether the model can be identified. *Model identification* refers to deciding whether a set of unique parameter estimates can be computed for the regression equations. In other words, whether the number of parameters to be estimated equals the number of available equations. This occurs when the number of distinct values in the variance-covariance matrix (Table IV) of the indicators recorded, equals or exceeds the parameters to be estimated. Multiple regression models are always considered just-identified [47], i.e., all of the model parameters (beta weights or path coefficients) can be uniquely determined because there is enough information available in the variance-covariance matrix.

With our evaluation model identified, we can proceed to *model estimation*, that is compute the sample regression weights for the independent predictor variables. For these calculations we used the SPSS statistical package [48]. Stepwise regression analysis was performed twice, with the criteria presented in Table I used as independent variables, and Ease of Use and Usefulness as predicted outcomes. This analysis reveals which set of predictors is most important in explaining the variance in the predicted outcome, thus paying particular attention to them during development stages.

In the construct of *Usefulness* two independent variables were found to be the dominant predictors; *time earned* (t(57)= 5.190, p≤ .001) and *utilitarian value* of accessing the seeking service from anyplace within the library (t(57)= 2.267, p≤ .05). These criteria were found to account for 53% of variance in usefulness ($R^2 = .532$, $F = 32.346$, p≤ .05) and highlighted the fact that the participants perceive usefulness as the efficient access to resources from anywhere, in a ubiquitous fashion. As in the qualitative study, the navigation map aid and the assistance for locating books had no significant effect in usefulness and satisfaction, probably due to familiarity with the library environment and collections.

In the construct of *Ease of Use* five variables were found to account for 59% of its variance (R^2= .596, $F = 12.385$, p≤ .001): information *presentation* (t(42)= 2.826, p=≤ .01), *clarity* (t(42)= 2.763, p≤ .01), *easiness to execute communication tasks* (t(42)= 3.216, p≤ .01), *easiness to execute moving/storage tasks* (t(42)= -2.518, p≤ .05), and *help* (t(42)= 2.756, p= .01). The results of this analysis demonstrate the importance of interface characteristics and help functionalities in the perceived ease of use, as well as the easiness of executing crucial tasks that handhelds support.

Stepwise regression was performed to define which criteria from both categories are significant predictors of Satisfaction. In general the Ease of Use criteria account for 50.4% (R^2= .504, $F = 15.221$, p= .01), while the Usefulness criteria account for 48% (R^2= .483, $F = 26.679$, p= .01) in Satisfaction variance.

As shown in Table III three criteria were found to account for 57% in Satisfaction variance (R^2= .572, $F = 19,178$, p= .01). These are *time earned*, *learnability*, and *completeness* of the retrieved content. Figure 6 shows how this regression model of Satisfaction matches the recorded values for Satisfaction from the use of the new service. The three indicators reveal why participants are willing to use mobile computing devices in the library; it helps them in retrieving information content from multiple and diverse sources in a way that is quick and easy to learn.

The significance of the test statistics in our latest regression model does not mean by itself that there is a strong effect (relationship) between predictor variables and the recorded Satisfaction (recall that standardized β values indicate the relative importance of indicators in predicting an outcome). The importance of the chosen predictors is obtained by the effect size (ES), which is an objective and standardized measure of the magnitude of the recorded effect. The effect size is computed as ES= $R^2 - [p/(N-1)]$, where R^2= .572, p= 3 predictor variables and N= 64 observations (we used 64

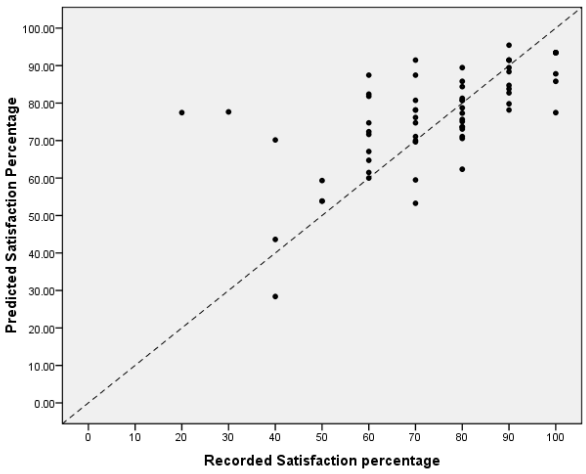


Fig. 6: Predicted satisfaction versus recorded satisfaction

measurements instead of 67 due to missing values). Therefore, $ES = .572 - [3/(64-1)] = .524$, which indicates a large effect size according to Cohen [49]. Miles and Shevlin [50] produced several graphs that illustrate the sample size needed to achieve different levels of predictive power, for different effect sizes, as the number of indicators (predictors) increases. From one of these diagrams we can see that for a large effect to be recorded and 3 predictors to be used, the required sample size should be approximately 40 participants. Therefore, with a sample of 64 subjects we can have confidence on the reliability of our regression model.

VI. Discussion and conclusion

This paper presents the design and evaluation procedure of a new library service that supports its users with handheld computing devices while seeking in hybrid information spaces. We described the stages of identifying the users' needs and service functionalities to implement, as well as the evaluation design, criteria, model, and methodology to assess their satisfaction for the new service.

We used a triangulation evaluation approach to assess the impact of the new service in the users' natural environment, i.e., a library setting. This approach included data from semi-structured interviews with the users, observation notes, video recordings from their interaction, and questionnaires administered to the users. This multi-strategy approach allowed us to gain better insights about the effect of the new service on the users and their needs. Besides studying how the service was used we were able to see and understand why was it used that way, i.e., study and analyze the non-linear seeking behavior of our experiment participants. In addition, qualitative findings helped us in better explaining relations among measurements and effects captured by the quantitative recordings. For instance besides describing the relation among measurement variables we could also explain and verify causality among them, which was used in our evaluation model specification. In general, the qualitative and quantitative findings provide valuable information, such as the fact that users envisage the proliferation of digital libraries by means of mobile devices, in order to raise space barriers, speed-up the seeking process and better experience the current information landscape that surrounds them. Furthermore users believe that the unified search for physical and electronic resources is an important feature with the interfaces kept as simple as possible, which was in agreement with the recommendation of experts.

The insight gained from this study provides valuable knowledge and data to proceed to inform decisions about the next phases of the service development. However, there are certain limitations in generalizing our findings to a bigger population. Our data come from a small sample size (N= 77), with participants from only 3 academic departments, i.e., social and humanitarian sciences. We can clearly see from Table II that their frequency distributions are far from uniform. In addition, all data recorded in the questionnaire are subjective measurements, i.e., they describe the users' attitudes towards various aspects of our evaluation criteria. Subjective measurements are always subject to bigger error levels and therefore reduce the explanatory power of our findings. To overcome this limitation we need to further explore the relationships and effects among recorded variables and use advanced statistical analysis techniques, such as exploratory and confirmatory factor analysis, which requires a much bigger sample of participants [51]. We also note that our model's fit index, to the data recorded, is R^2= .572, i.e., our model accounts for 57% of the variance in recorded satisfaction. In other words, more than 40% of variance is unexplained, indicating that we need to extent our evaluation model to include other factors that can have an effect in recorded satisfaction.

Therefore, based on the evaluation findings future work includes the extension of the prototype features in order to simultaneously submit queries in multiple sources and the resolution of interface problems that impede interaction. In addition, with the insight gained from the current study about the users' interaction with the service we plan to further continue our analysis with an experimental evaluation design, i.e., a design in which participants will be divided into two groups; test and control. This method enables us to investigate and better understand any effects and factors that significantly affect the users' efficiency while using a mobile device to seek in hybrid collections and therefore the new service's acceptance and usage.

Acknowledgment

This work is part of the PENED'03/791 project (http://dlib.ionio.gr/hls/), funded by the Reinforcement Programme of Human Research Manpower, under the Operational Programme "Competitiveness (2000-2006)". The authors wish to acknowledge Giannis Tsakonas and personnel at the Panteion Library Administration Office for their organization and technical support in conducting this study.

REFERENCES

[1] S. Veronikis, D. Gavrilis, K. Zoutsou, and C. Papatheodorou, "Using Handhelds to Search in Physical and Digital Information Spaces", in *Proc. 2nd Int. Conf. Mobile and Ubiquitous Computing, Systems, Services, and Technologies (UBICOMM)*, Valencia, Spain, pp. 225-230, 2008.

[2] K. G. Coffman and A. M. Odlyzko, "The Size and Growth Rate of the Internet", Center for Discrete Mathematics & Theoretical Computer Science, AT&T Labs, Austin, TX, Tech. Rep. 99-11, Oct. 1998.

[3] *World Internet Usage Statistics News and World Population Stats* [Online]. Available: http://www.internetworldstats.com/stats.htm (2009, May 12)

[4] *The Perseus Digital Library Project* [Online]. Available: http://www.perseus.tufts.edu/hopper/ (2009, May 12).

[5] *Project Gutenberg* [Online]. Available: http://www.gutenberg.org/wiki/Main_Page (2009, May 12).

[6] *ibiblio* [Online]. Available: http://www.ibiblio.org/ (2009, May 12).

[7] E. F. Codd Associates (1998) *Providing OLAP to User Analysts: An IT Mandate* White paper [Online]. Available: http://www.uniriotec.br/tanaka/SAIN/providing_olap_to_user_analysts.pdf

[8] R. Jain, "Unified access to universal knowledge: next generation search experience", unpublished white paper [Online], Available: http://ngs.ics.uci.edu, May 2009.

[9] M. Peterson, "Library service delivery via hand-held computers - the right information at the point of care", *J. Health Information and Libraries*, 21(1), pp. 52-56, 2004.

[10] K. Sommers, J. Hesler, and J. Bostick, "Little guys make a big splash: PDA projects at Virginia Commonwealth University", in *Proc. 29th Annu. ACM SIGUCCS Conf. on User Services*, Portland, Oregon, pp. 190-193, 2001.

[11] J. P. Shipman and A. C. Morton, "The new black bag: PDAs, health care and library services", *J. of Reference Services Review*, 29(3), pp. 229-238, 2001.

[12] P. Scollin, J. Callahan, A. Mehta, and E. Garcia, "The PDA as a Reference Tool: The Libraries Role in Enhancing Nursing Education", *J. Computer Informatics Nursing*, 24(4), pp. 208-213, 2006.

[13] G. Buchanan, M. Jones, and G. Marsden, "Exploring Small Screen Digital Library Access with the Greenstone Digital Library", in *Proc. 6th European Conf. Research and Advanced Technology for Digital Libraries*, pp. 583-596, 2002.

[14] M. Aittola, T. Ryhänen, and T. Ojala, "Smart Library: Location-Aware Mobile Library Service", in *Proc. 5th Symposium on Human Computer Interaction with Mobile Devices and Services*, Udine, Italy, pp.411-416, 2003.

[15] M. L. Jones, R. H. Rieger, P. Treadwell, and G. K. Gay, "Live from the stacks: user feedback on mobile computers and wireless tools for library patrons", in *Proc. 5th ACM Conf. Digital Libraries*, pp. 95-102, 2000.

[16] D. H. Goh, L. L. Sepoetro, M. Qi, R. Ramakrishnan, Y. L. Theng, F. Puspitasari, and E. P. Lim, "Mobile tagging and accessibility information sharing using a geospatial digital library", in *Proc. 10th Int. Conf. Asian Digital Libraries*, Hanoi, Vietnam, pp. 287-296. , 2007.

[17] J. Baus and C. Kray, "A Survey of Mobile Guides", in *MGUIDES '03 Workshop on Mobile Guides at Mobile HCI*, Udine, Italy, 2003.

[18] F. Garzotto, P. Paolini, M. Speroni, B. Proll, W. Retschitzegger, and W. Schwinger, "Ubiquitous Access to Cultural Tourism Portals", in *Proc. Database and Expert Systems Applications, 15th Int. Workshop*, Washington, DC, pp. 67-72, 2004.

[19] R. Oppermann and M. Specht, "A Context-Sensitive Nomadic Exhibition Guide", in *Proc. Handheld and Ubiquitous Computing: 2nd Int. HUC Symp.*, pp. 127-142, 2000.

[20] M. Fleck, M. Frid, T. Kindberg, E. O'Brien-Strain, R. Rajani, and M. Spasojevic, "From informing to remembering: ubiquitous systems in interactive museums", *IEEE Pervasive Computing*, 1(2), pp. 13-21, 2002.

[21] D. Layder, *New Strategies in Social Research*, Cambridge, UK: Polity Press, 1992.

[22] A. Bryman and E. Bell, *Business Research Methods*, 2nd ed. Oxford University Press, UK: 2006.

[23] R. Blackburn and D. Stokes, "Breaking Down the Barriers: Using Focus Groups to Research Small Medium-Sized Enterprises", in *J. International Small Business*, 19(1), pp. 44-67, 2000.

[24] S. E. Asch, "Effects of Group Pressure upon the Modification and Distortion of Judgements", in H. Guetzkow, Ed., *Groups, Leadership and Men*, Pittsburgh, Carnegie Press, 1951.

[25] J. Nielsen, *Usability Engineering*, Boston: Morgan Kaufmann, 1993.

[26] R. Blum, K. Khakzar, and W. Winzerling, "Mobile Design Guidelines in the Context of Retail Sales Support", in K. Asai, Ed., *Human Computer Interaction: New Developments*, Vienna, Austria: InTech Education and Publishing, 2008.

[27] B. Shneiderman and C. Plaisant, *Designing the User Interface Strategies for Effective Human-Computer Interaction,* 4th ed. Amsterdam: Addison-Wesley, 2004.

[28] J. Nielsen, "Heuristic Evaluation", in J. Nielsen and R. L. Mack, Eds. *Usability Inspection Methods*, New York, NY: John Wiley & Sons, pp. 25-62, 1994.

[29] *Apple Human Interface Guidelines*, Apple Computer Inc. [Online]. Available: http://developer.apple.com/documentation/UserExperience/ (2009, January 12).

[30] *Gnome Human Interface Guidelines* [Online]. Available: http://library.gnome.org/devel/hig-book/2.24/ (2009, January 12).

[31] *Palm OS User Interface Guidelines*, PalmSource Inc., 2003, [Online]. Available: http://www.accessdevnet.com/docs/uiguidelines.pdf (2009, January 12).

[32] *Microsoft Corp. Design Guidelines, Windows Mobile Version 6.0* [Online]. Available: http://msdn.microsoft.com/en-us/library/bb158602.aspx (2009, January 12).

[33] *iPhone Human Interface Guidelines User Experience*, Apple Computer Inc, 2008 [Online]. Available: http://developer.apple.com/iphone/library/documentation (2009, January 12).

[34] *Mobile OPAC* [Online]. Available: http://library.panteion.gr/mobile/opac.php?sid=&lang=en (2009, May 12).

[35] E. J. Webb, D. T. Campbell, R. D. Schwartz, and L. Sechrest, *Unobtrusive Measures: Noncreative Measures in the Social Sciences*, Chicago: Rand McNally, 1966.

[36] S. Zamanou and S. R. Glaser, "Moving toward Participation and Involvement", in *J. Group and Organization Management*, 19(4), pp. 475-502, 1994.

[37] Y. S. Lincoln and E. G. Guba, *Naturalistic Inquiry*, Newbury Park, CA: Sage Publications, 1985.

[38] T. A. Schwandt, "Solutions to the paradigm conflict: Coping with uncertainty", in *J. Contemporary Ethnography*, 17(4), pp. 379-407, 1989.

[39] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C. P. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Sølvberg, "Evaluation of Digital Libraries", in *Int. J. Digital Libraries*, 8(1), pp. 21-38, 2007.

[40] G. Tsakonas and C. Papatheodorou, "Exploring usefulness and usability in the evaluation of open access digital libraries", in *J. Information Processing & Management*, 44(3), pp. 1234-1250, 2007.

[41] J. Jeng, "What is usability in the context of digital library and how it can be measured?", in *J. Information Technology and Libraries*, 24(2), pp. 47-56, 2005.

[42] G. Tyburski. *Criteria for Quality in Information*, Ballard Spahr Andrews & Ingersoll, LLP [Internet]. Available: http://www.virtualchase.com/quality/criteria_print.html (2009, May 12).

[43] R. Fidel and M. Green, "The many faces of accessibility: engineers perceptions of information sources", in *J. Information Processing & Management*, 40(3), pp. 563-581, 2004.

[44] A. Stoica, G. Fiotakis, D. Raptis, I. Papadimitriou, V. Komis, and N. Avouris, "Field evaluation of collaborative mobile applications", in J. Lumsden, Ed., *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, Hershey, PA: Idea Group Publishers, 2007.

[45] K. Jöreskog and D. Sörbom, *PRELIS2: User's reference guide*, Lincolnwood, IL: Scientific Software International, 1996.

[46] A. Field, *Discovering Statistics Using SPSS*, 2nd ed. London, UK: Sage Publications Ltd., 2005.

[47] R. E. Schumacker and R. G. Lomax, *A Beginner's Guide to Structural Equation Modeling*, 2nd ed., Philadelphia, PA: Lawrence Erlbaum Associates, 2004.

[48] SPSS, Inc. SPSS Statistical Analysis Software. Available: http://www.spss.com/statistics/

[49] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Inc., 1988.

[50] J. Miles and M. Shevlin *Applying Regression and Correlation: A Guide for Students and Researchers*, London, UK: Sage Publications, 2001.

[51] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, 4th ed. Boston: Allyn & Bacon, 2001.

TABLE IV: The variance-covariance matrix for the recorded evaluation criteria

	Sat	U	EoU	L	NL	P	Ex	CL	H	S	UL	T	R	C
Sat	315.33													
U	7.70	0.65												
EoU	8.57	0.36	1.08											
L	6.94	0.33	0.30	0.94										
NL	2.67	0.10	0.00	0.48	1.27									
P	5.33	0.22	0.31	0.26	0.13	0.33								
Ex	8.34	0.47	0.34	0.79	0.63	0.34	1.27							
CL	4.08	0.13	0.26	0.16	0.10	0.11	0.13	0.42						
H	9.66	0.51	0.50	0.60	0.28	0.26	0.68	0.17	1.07					
S	1.62	0.04	0.40	0.07	0.51	0.24	0.17	0.01	0.03	2.12				
UL	4.90	0.39	0.17	0.30	0.25	0.13	0.43	0.08	0.44	0.13	0.81			
T	12.31	0.80	0.48	0.66	0.25	0.34	0.93	0.12	0.75	0.16	0.71	2.04		
R	6.70	0.21	0.39	0.29	0.08	0.20	0.34	0.19	0.42	0.12	0.24	0.45	0.75	
C	223.98	2.69	7.83	6.12	8.44	3.63	6.96	5.50	9.71	2.26	3.75	4.71	9.11	523.78

Performance of Spectral Amplitude Warp based WDFTC in a Noisy Phoneme and Word Recognition Tasks

R. Muralishankar
PES Centre for Intelligent Systems
Dept. of Telecommunication Engineering,
PES Institute of Technology, Bangalore, India.
muralishankar@pes.edu

H. N. Shankar
PES Centre for Intelligent Systems
Dept. of Telecommunication Engineering,
PES Institute of Technology, Bangalore, India.
hnshankar@pes.edu

Abstract

In this paper, we investigate the noise robustness of three features, namely, the warped discrete Fourier transform cepstrum (WDFTC, [1]), perceptual minimum variance distortionless response (PMVDR) and Mel-frequency cepstral coefficients (MFCC). We generate WDFTC and PMVDR features by all-pass based warping; we use spectral warping for MFCC. PMVDR and WDFTC use warped-LP and warped discrete Fourier transforms, respectively. We employ WDFTC, PMVDR and MFCC features in continuous noisy monophone and word recognition tasks using the TIMIT corpus. We also test these features on gender-specific monophone and word recognition tasks. Further, we employ spectral amplitude warping (SAW) in WDFTC feature extraction (WDFTC_SAW) and demonstrate enhanced robustness of this feature. We observe that SAW does not improve robustness for the MFCC and PMVDR features. Finally, we report the recognition performance and discuss many interesting properties of these features. Our study shows that the PMVDR and WDFTC_SAW achieve recognition performance superior to the MFCC and WDFTC in noisy conditions.

Index Terms:Robustness, Speech recognition, Warped Discrete Fourier Transform, Cepstrum, WDFTC, PMVDR, Spectral Amplitude Warping, WDFTC_SAW.

1. Introduction

Contemporary automatic speech recognition (ASR) systems perform satisfactorily when the test and training conditions are close. However, ASR performance degrades rapidly in various conditions such as background acoustical noise, stressed speech (*e.g.*, lombard), channel conditions, and speaker variability [2, 3]. With additive acoustical noise the problem is as pragmatic as it is challenging. Interestingly, humans do a far better job than ASR systems in noise, thus pointing to scope for further improvement [2]. Moreover, rapid proliferation of mobile phones concomitant with the large number of interactive voice applications being developed around them continues to present increasingly varied and complex acoustical backgrounds to the ASR systems [4].

In general, there are three approaches towards addressing the problem of noise robustness in ASR. The first incorporates a speech enhancement unit as a part of the feature extraction process, thereby presenting clean features to the ASR unit. Missing features approach [5], vector Taylor series approach [6] and spectral subtraction techniques [7] are instances in point. The second approach involves compensating the trained ASR models using techniques such as Parallel Model Combination (PMC) and dynamic Hidden Markov Model (HMM) variance compensation [8]. The third aims to develop new feature analysis methods which are relatively robust to distortion such as relative spectra (RASTA) [9] or cepstral mean subtraction (CMS). It is within the domain of robust features that in a companion paper we developed and introduced in the warped discrete cosine transform cepstrum (WDCTC) [10]. There, we benchmarked the new feature against the popular Mel-frequency cepstral coefficients (MFCC) in terms of its statistical properties and performance in simple recognition tasks [11]. A new feature representation called the Perceptual-MVDR (PMVDR) [12] has been proposed by Yapanel et al. They compute cepstral coefficients from the speech signal. Warping is incorporated directly into the DFT power spectrum. A variant of this feature, proposed in [13], uses warped-LP coefficients in generating warped-MVDR spectrum. The building block for all these features is the warping adopted to transforms or to the LP model.

In this paper, we propose a variant of MFCC, the warped discrete Fourier transform cepstrum (WDFTC); spectral warping is achieved using the warped discrete Fourier transform (WDFT). We then employ spectral amplitude warping (SAW) in addition to the frequency warping to generate WDFTC, i.e., WDFTC_SAW. We perform a comparative analysis of the features derived from the warping with the MFCC. The four features that we examine in our com-

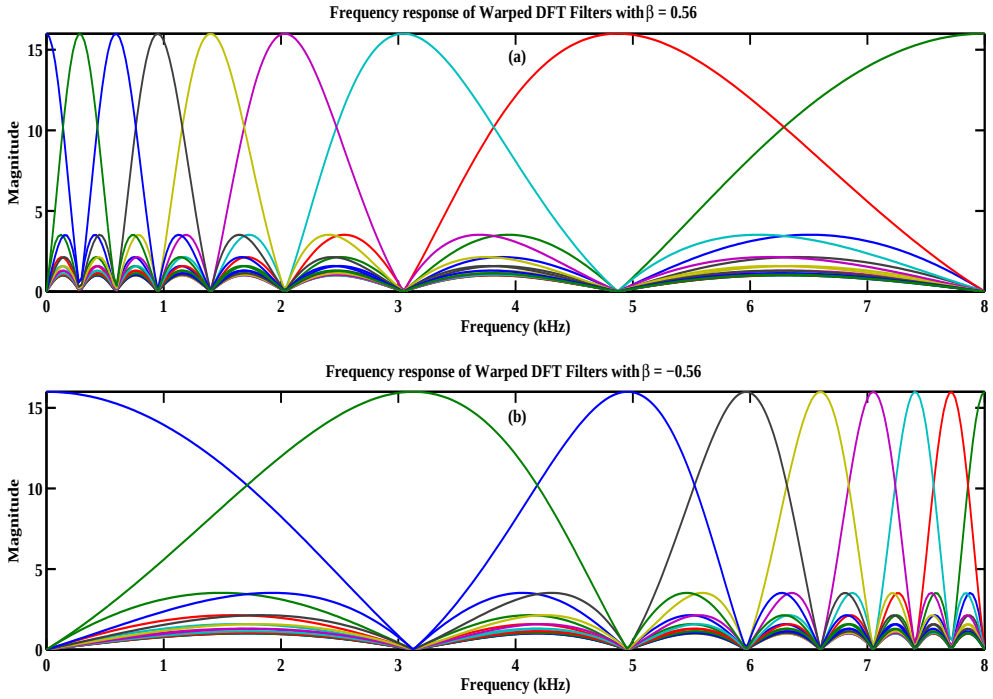


Figure 1. 16-Band Warped Filter bank shown for $f_s/2$. (a) $\beta = 0.56$ (b) $\beta = -0.56$

parative analysis are WDFTC_SAW, WDFTC, PMVDR and MFCC. We focus on studying their noise robustness properties. Particularly, we benchmark WDFTC_SAW, WDFTC and PMVDR against the MFCC in monophone and word recognition using the TIMIT corpus without using language models in monophone recognition and postprocessing of features like CMS. Thus the recognition performance is due to the features alone. We test the features in six different noise conditions – babble, car, fan, factory, tank and F-16 cockpit noise – at signal-to-noise ratio (SNR) from 0 dB through 20 dB. We report simulations reflecting phoneme recognition rates and recognition accuracies for monophone recognition.

In the second part of this work, we use Sphinx-III recognizer for word recognition task using TIMIT corpus. We report word error rate (WER) of the four features under clean and babble noise conditions. Finally, we highlight several interesting observations on noise robustness properties of warp-based features.

2. Warped-DFT Cepstrum

In this section, we briefly review WDFTC with a dyad of purposes in mind. First, to serve a didactic cause and second, to facilitate unfolding of the notations used in the sequel.

Let the N-point DFT of the input vector $[x(0), x(1), ..., x(N-1)]^T$ be given by $\{X(0), X(1), ..., X(N-1)\}$, where the frequency samples of the $z-$transform of the sequence evaluated at uniformly-spaced points $z = e^{\frac{j2\pi k}{N}}$, $0 \leqslant k \leqslant N-1$, on the unit circle are

$$X(k) = X(z)|_{z=e^{\frac{j2\pi k}{N}}} = \sum_{n=0}^{N-1} x(n)\, e^{\frac{j2\pi kn}{N}} \qquad (1)$$

for $k = 0, 1, ..., N-1$. For spectral analysis applications, DFT provides a fixed frequency resolution given by $2\pi/N$ over $[0, 2\pi]$. WDFT proposed in [14] is the most general form of DFT that can be employed to evaluate the frequency samples arbitrarily at distinct points in the z-plane. If z_k, $0 \leqslant k \leqslant N-1$, denote distinct frequency points in the z-plane, the N-point WDFT of the length-N sequence is given by

$$X_{WDFT}(k) = X(z_k) = \sum_{n=0}^{N-1} x(n) {z_k}^{-n}, 0 \leqslant k \leqslant N-1. \qquad (2)$$

Now, incorporating a nonlinear frequency resolution closely following the psychoacoustic Bark scale, yields enhanced representation for speech. Thus we warp DFT by an all-pass transformation $z^{-1} = A(z)$:

$$A(z) = \frac{-\beta + z^{-1}}{1 - \beta z^{-1}}, \qquad (3)$$

where β controls warping. It may be useful to note that Smith and Abel [15] have shown that for $\beta = 0.56$ warping closely resembles psychoacoustic Bark scale for sampling at $16kHz$. We also use $\beta = 0.56$ in computing the perceptually motivated speech spectrum. WDFT filters for length-16 is in Fig-

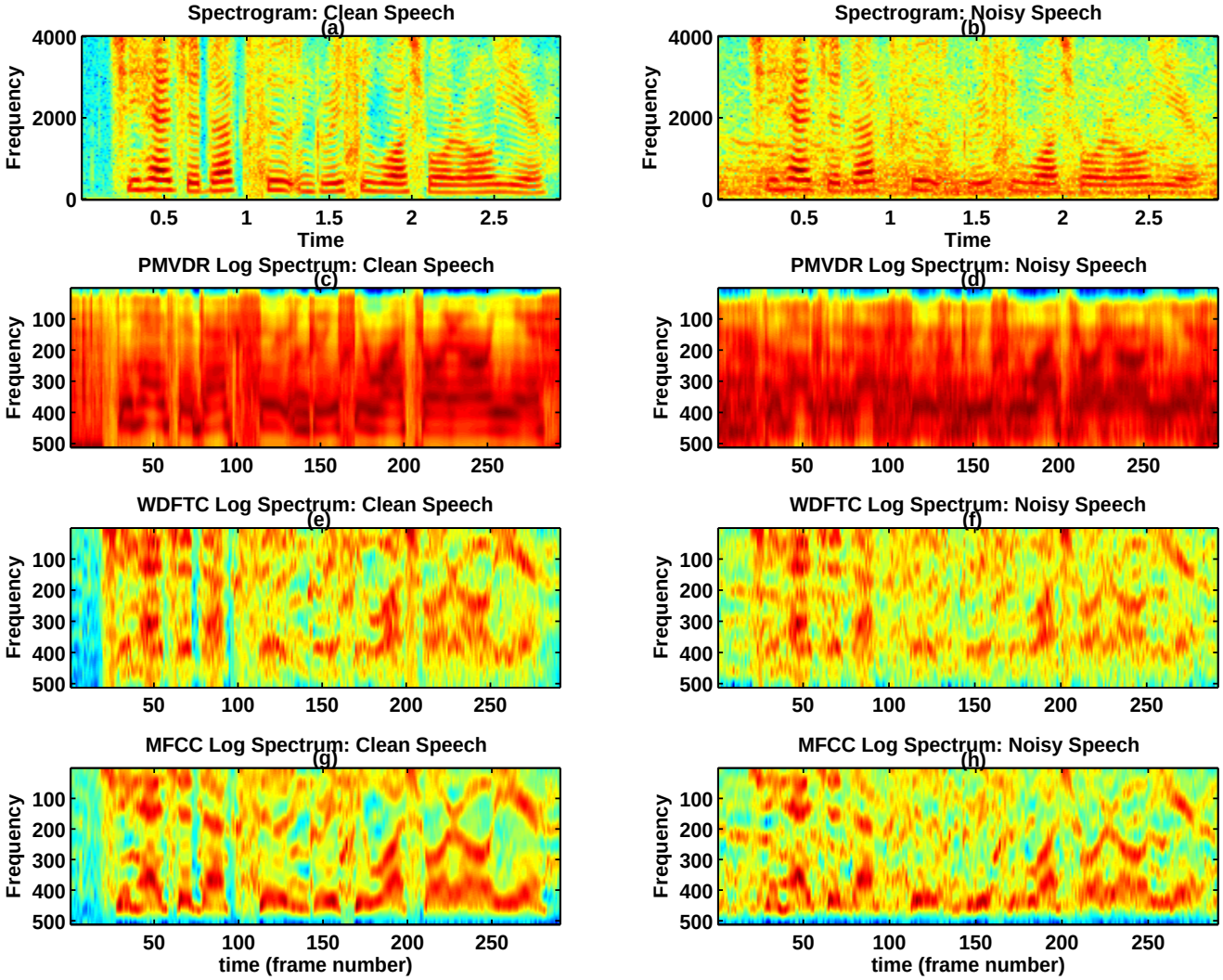


Figure 2. Illustrating the impact of 5 dB babble noise on PMVDR, WDFTC and MFCC Log spectra: (a) Spectrogram of clean speech, (b) Spectrogram of noise corrupted speech (a), (c) PMVDR log spectrum of clean speech (a), (d) PMVDR log spectrum of noise corrupted speech (a), (e) WDFTC log spectrum of clean speech (a), and (f) WDFTC log spectrum of noise corrupted speech (a, (g) MFCC log spectrum of clean speech (a), and (h) MFCC log spectrum of noise corrupted speech (a).

ure 1. Details of the implementation of WDFT are in [14]. The WDFTC algorithm is outlined in Algorithm 1.

3 MVDR Spectral Envelope Estimation

MVDR spectral estimation has been explored for speech parameterization [16, 17, 18]. Here, we present only the computational algorithm and general properties of MVDR and perceptual MVDR (PMVDR). In the MVDR spectrum estimation method, the power spectral density at ω_l is determined by filtering the signal by a distortionless FIR filter, $h(n)$, designed to minimize the output power while constraining the filter gain to unity at the frequency of interest, ω_l. This provides a lower bias with a smaller filter length. The parametric

Algorithm 1 Algorithm to compute the WDFTC

1: Obtain an N-point WDFT, $X_{WDFT}(k)$, $0 \leq k \leq N-1$ for a finite duration, real sequence $x(n)$, $0 \leq n \leq N-1$.
2: Compute $\zeta(k)$ of the WDFT coefficients:

$$\zeta(k) = |X_{WDFT}(k)|, \tag{4}$$

where $|.|$ evaluates the absolute value.
3: Compute the WDFTC $\widehat{x}(n)$ as

$$\widehat{x}(n) = (IDCT(\ln(\zeta(k))). \tag{5}$$

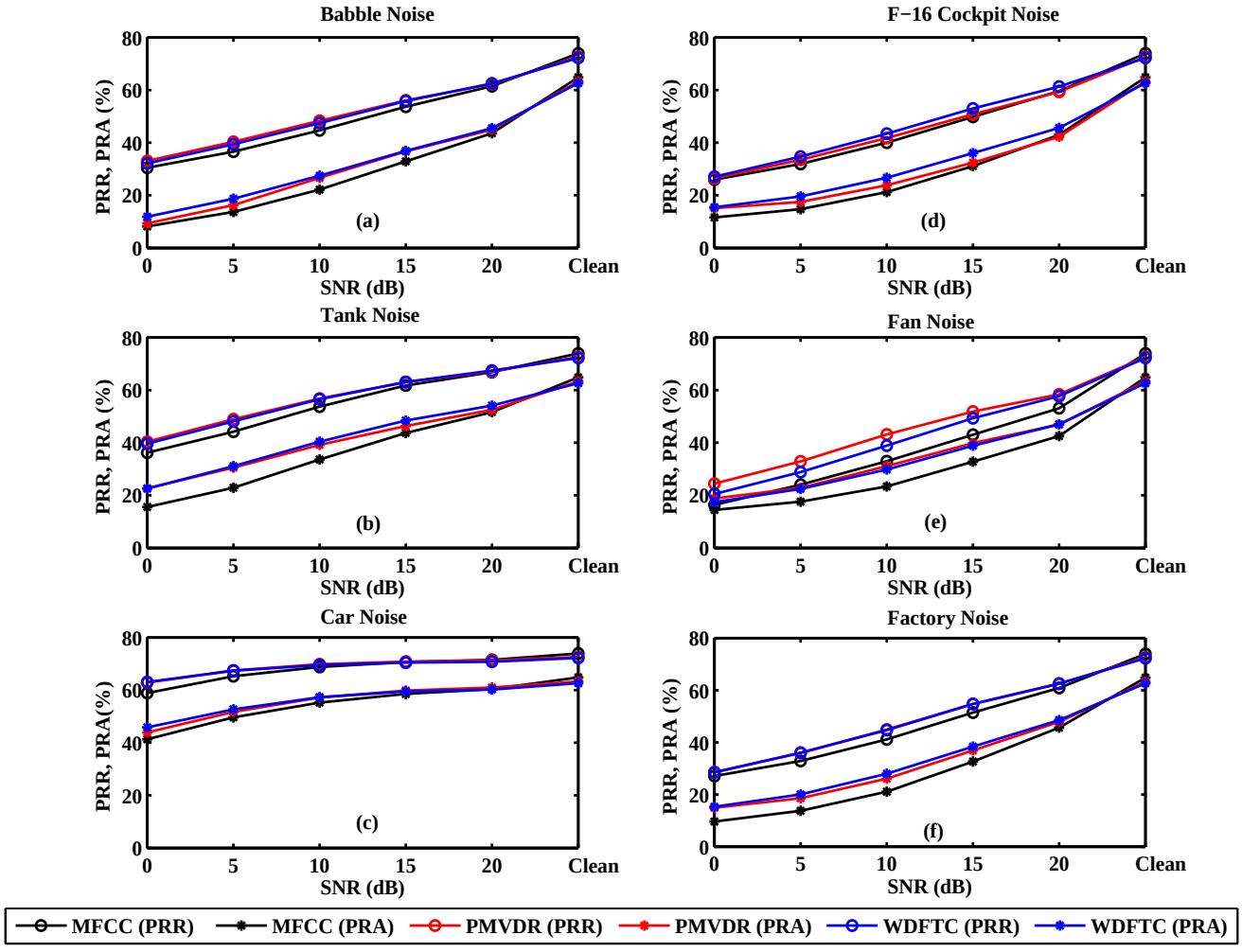


Figure 3. Illustrating the impact of different noises at various SNRs on MFCC, WDFTC and PMVDR phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on the entire TIMIT corpus using diagonal (DC) covariance.

form of the Mth order MVDR spectrum is given by

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2}. \quad (6)$$

The MVDR coefficients, $\mu(k)$, are obtained from a non-iterative computation using the LP coefficients a_k and prediction error variance, P_e.

$$\mu(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} L a_i a_{i+k}^*, & k = 0, ..., M \\ \mu^*(-k), & k = -M, ..., -1 \end{cases} \quad (7)$$

where $L = (M + 1 - k - 2i)$. We compute the MVDR envelope using LP coefficients of order M and the prediction error power, ϵ_M, as

$$S_{MVDR}(e^{j\omega}) = \frac{\epsilon_M}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}}. \quad (8)$$

4. PMVDR Feature Extraction

MVDR spectrum exhibits useful properties such as low variance, low distortion and good spectral envelope matching across a wide range of pitch frequencies. Therefore it is widely considered as a robust speech parameterization technique in speech recognition. MVDR has been used in spectral estimation [17] and in envelope estimation [18]. A natural extension to the MVDR scheme is the incorporation of the perceptually motivated mel frequency into the otherwise linear frequency scale. In [18], perceptual information was incor-

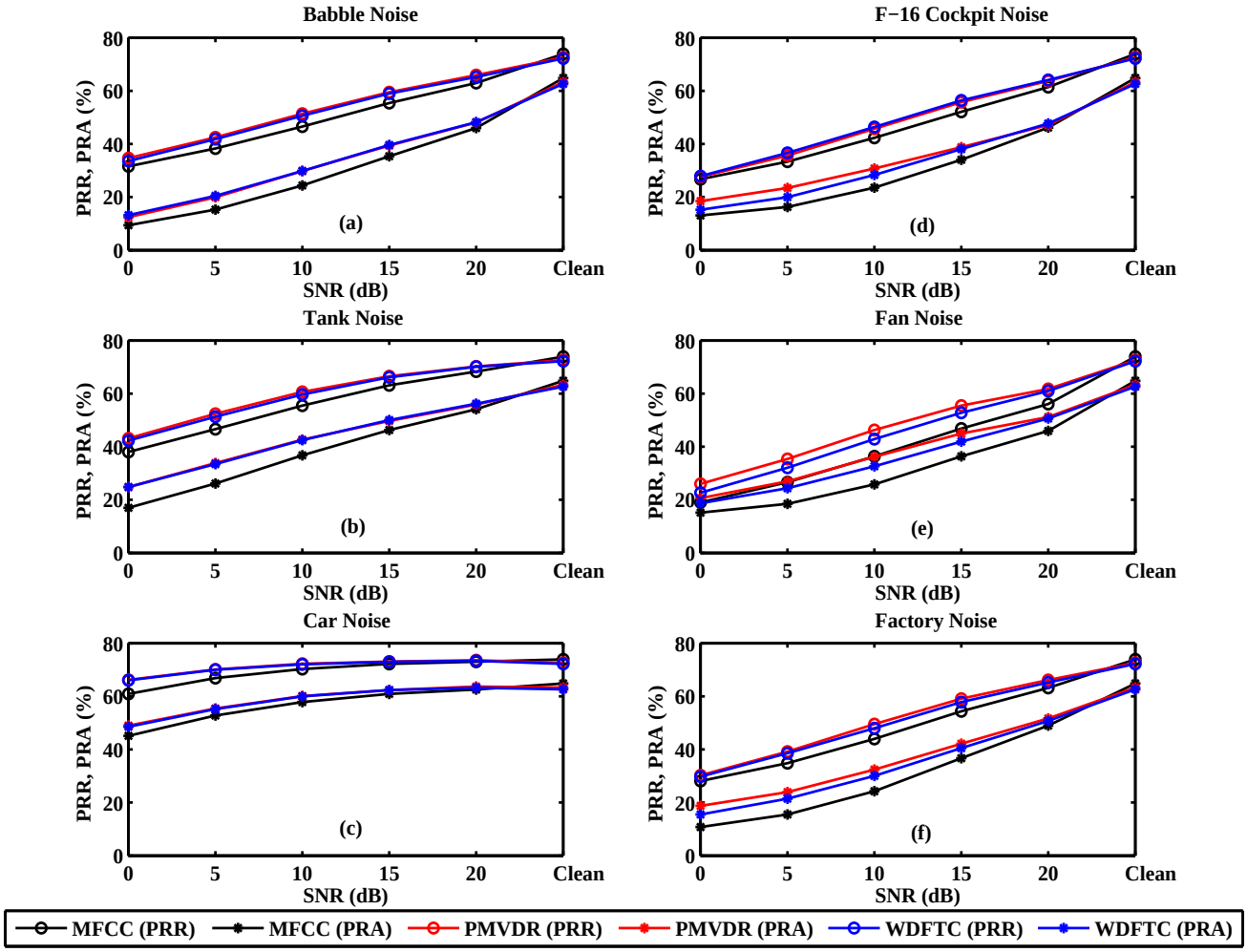


Figure 4. Illustrating the impact of different noises at various SNRs on MFCC, WDFTC and PMVDR phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from a male speaker in the TIMIT corpus using diagonal (DC) covariance.

porated directly into spectral estimation by using mel-scaled filter banks. It was easily seen that the filter bank structure is only a rough approximation to the perceptual scale since it samples the perceptual spectrum at the center frequencies of the filter bank. Furthermore, the filter bank is less effective in completely removing the harmonic excitation information from the spectrum. Alternatively, the use of warping techniques is also popular in contemporary literature, e.g., incorporating warping directly into the DFT power spectrum [12], or the use of warped-LP coefficients in generating the warped-MVDR spectrum [13]. Presently, we generate the PMVDR features using the warped-LP coefficients.

5. Spectral Amplitude Warping (SAW)

Features discussed so far employ frequency warping with improved resolution in the lower frequency band of the power spectrum. SNR in this band is often higher than at high frequencies. The result of nonuniform resolution provided by the frequency warping in turn helps in improved phoneme recognition performance specifically under noisy conditions. It is well known that the acoustic models generated by clean speech performs poorly under mismatched conditions and in a simulation the same models trained on mismatched conditions perform superior to the former case. Generating such an acoustic model is possible only for a few simulated mismatched conditions. It is useful here to incorporate the general effect of noise (such as reduced peak to valley differ-

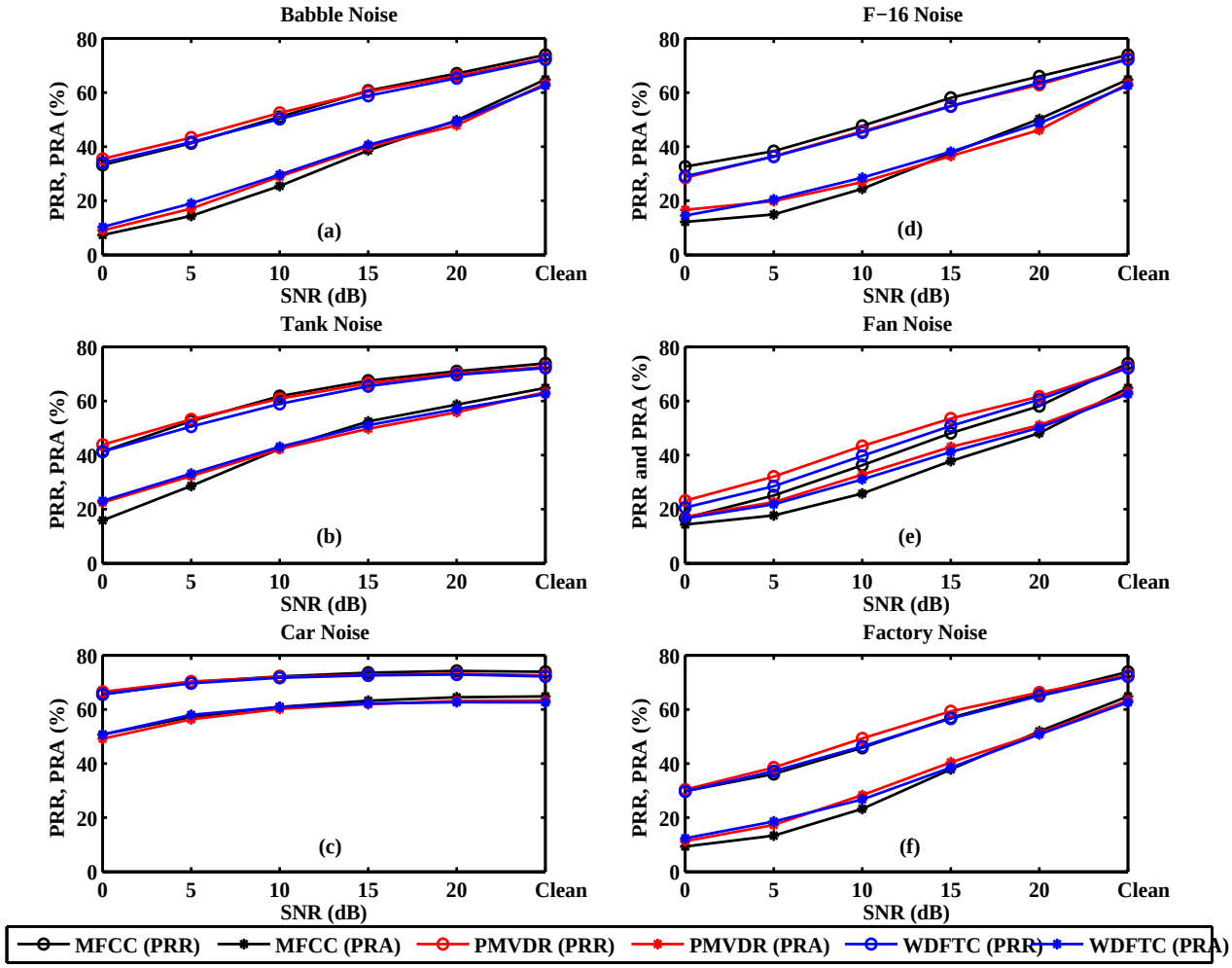


Figure 5. Illustrating the impact of different noises at various SNRs on MFCC, WDFTC and PMVDR phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from female speaker in the TIMIT corpus using diagonal (DC) covariance.

ences and change in the formant positions) on speech spectrum used in front-end feature extraction.

In this paper, we model reduction in peak-to-valley difference using Spectral amplitude warping (SAW). SAW has been employed to shape coding noise in speech and audio coders [19] in pre- and post-processing blocks to provide a non-linear transformations to the signal short-time spectrum before and after encoding. Adopting SAW improves the noise shaping capability of an existing coder without modifying the coder itself. It is reported [19] that the output quality of G.722 wideband speech coder operating at 48 kbps is close to the same coder operating at 64 kbps.

Consider

$$X_w(k) = f_{nl}(X(k)), \tag{9}$$

where $X(k)$ and $X_w(k)$ are the DFT spectra of signals $x(n)$ and $x_w(n)$ respectively, and $f_{nl}(\cdot)$ is nonlinear. In [19],

$$X_w(k) = X(k)\frac{|X(k)|^{\alpha(k)}}{|X(k)|}, \tag{10}$$

where $\alpha(k)$. There, $\alpha(k) = 0.5\ \forall k$. This reduces the dynamic range of $|X_w(k)|$. The attenuation is relatively higher for larger $|X(k)|$. In a nut shell, the effects of this transformation are (a) attenuation of the formants; and (b) attenuation of the harmonics. The disparity between peak and valleys is reduced both at the formant and harmonic levels. We use this transformation to generate front-end features from the transformed spectra so as to distort the speech spectra equivalent to the effect of noise on it. Consequently acoustic models generated from these spectra perform better than those from clean speech spectra. In our monophone and con-

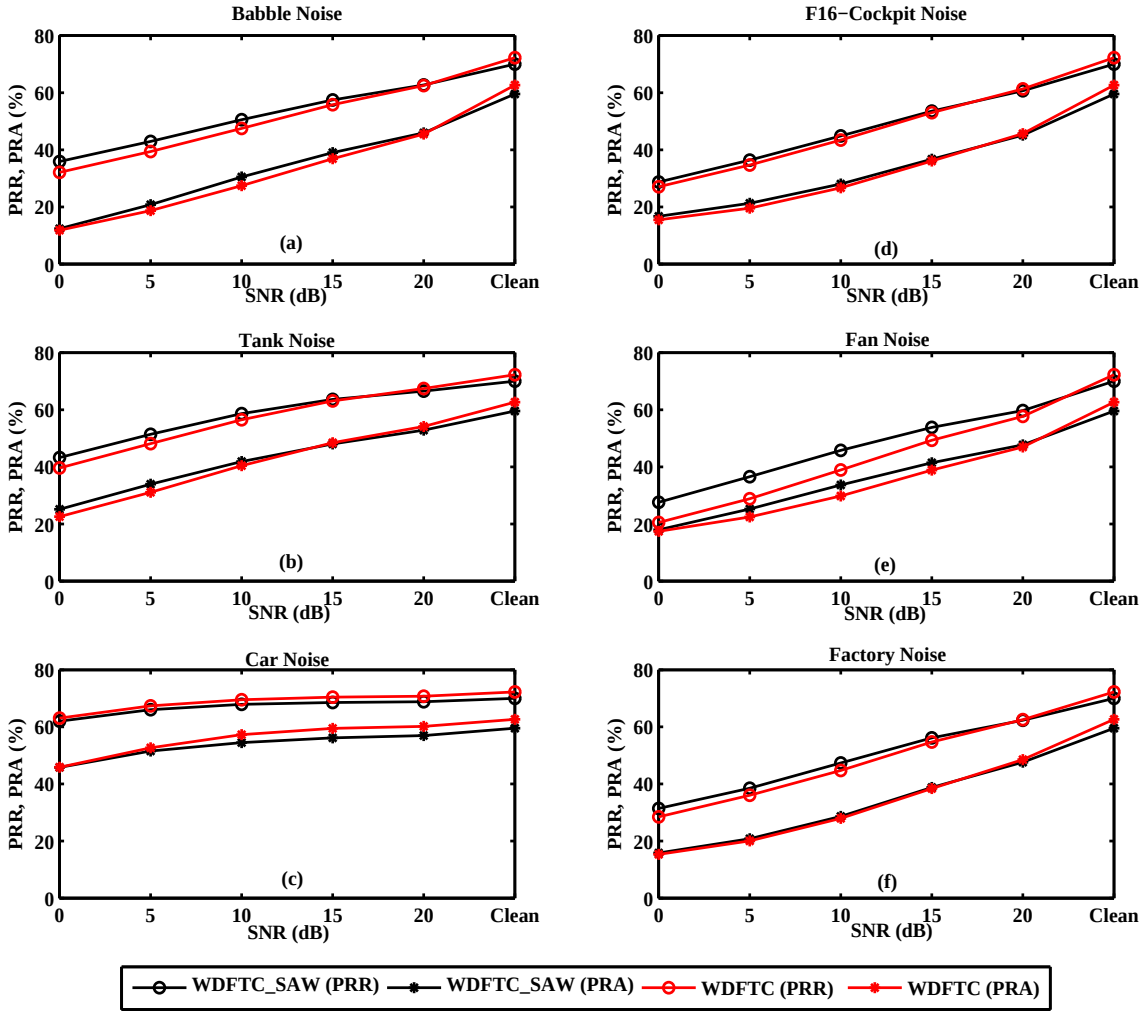


Figure 6. Illustrating the impact of different noises at various SNRs on WDFTC and WDFTC_SAW phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on the entire TIMIT corpus using diagonal (DC) covariance.

tinuous speech recognition we adopt this transformation with $\alpha(k) = 0.5$.

6. The Monophone Recognition Setup

We use GMM-HMM where the front-end features are the WDFTC or PMVDR or MFCC. We train and test the HMM recognizer using the HTK Toolkit [20], and the entire TIMIT corpus of 6300 sentences recorded from 630 speakers. We train phonetic HMMs using speech from the TIMIT train set with 462 speakers and test on the TIMIT test set with 168 speakers. It is common that the 61 TIMIT phonemes are mapped to a reduced set of 39 phonemes after training and testing, and the results are reported on this reduced set [21]. We train the HMMs using diagonal covariances (DCs). We use a 3-state HMM model for each of the 39 phonemes and for each state, a mixture splitting procedure under the DC setup [20]. The mixture splitting procedure starts with one mixture per state and it goes up to 8 mixtures in four steps with a re-estimation algorithm in each step. Finally, we present monophone performance under a DC setup for clean and noisy cases.

All the speech files are sampled at 16 kHz and pre-emphasized with $1 - 0.97z^{-1}$. They are then Hamming windowed. Speech signal is analyzed every 10 ms with a frame width of 25 ms. We generate 13-dimensional WDFTC, PMVDR and MFCC features from each speech frame including the zeroth coefficient. The WDFTC and MFCC are generated using the Algorithm 1 and a mel-scale triangular filter bank with 24 filter bank channels respectively. We append a

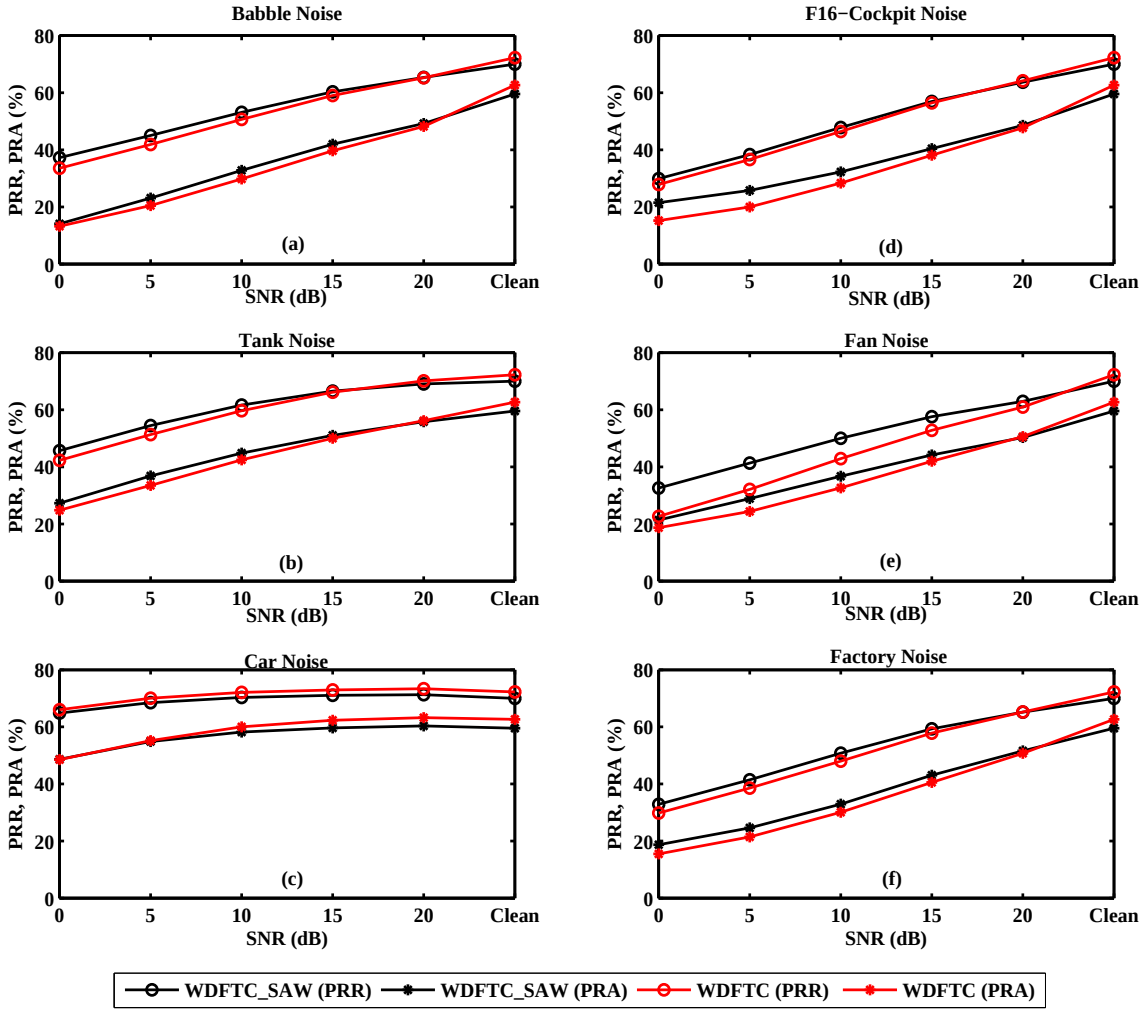


Figure 7. Illustrating the impact of different noises at various SNRs on WDFTC and WDFTC_SAW phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from a male speaker in the TIMIT corpus using diagonal (DC) covariance.

26-dimensional delta and delta-delta cepstral features to 13-dimensional WDFTC and MFCC. We then employ the procedure in [22] to obtain delta and delta-delta for MFCC. We may employ dynamic spectral parameters [23] to compute delta and delta-delta for PMVDR and WDFTC.

7. The Word Recognition Problem

We employ the Sphinx-III speech recognizer [24]and the TIMIT speech database to evaluate the WDFTC_SAW front-end parameters uunlike traditional feature like MFCC, WDFTC and PMVDR. The utterances of the words in the transcriptions of the database are generated by the lexicon provided with the database. We use a phoneme set of 43 symbols including silence.

7.1 Generating Acoustic Models

For each of the above described features we train one acoustic model in two stages – encoding and decoding.

7.1.1 Encoding

Using the transcriptions in the database, the acoustic models are force-aligned against the transcription of the training data; thus pronunciations of the words with multiple phonetic representations are extracted. New acoustic models are synthesized by tieing the existing models following the same procedure. As before, the feature vectors are 13-D. We pre-emphasize the signal in the time domain with a factor of 0.97. The resulting speech waveforms are segmented into 25 ms frames with a step of 10 ms. The first and second deriva-

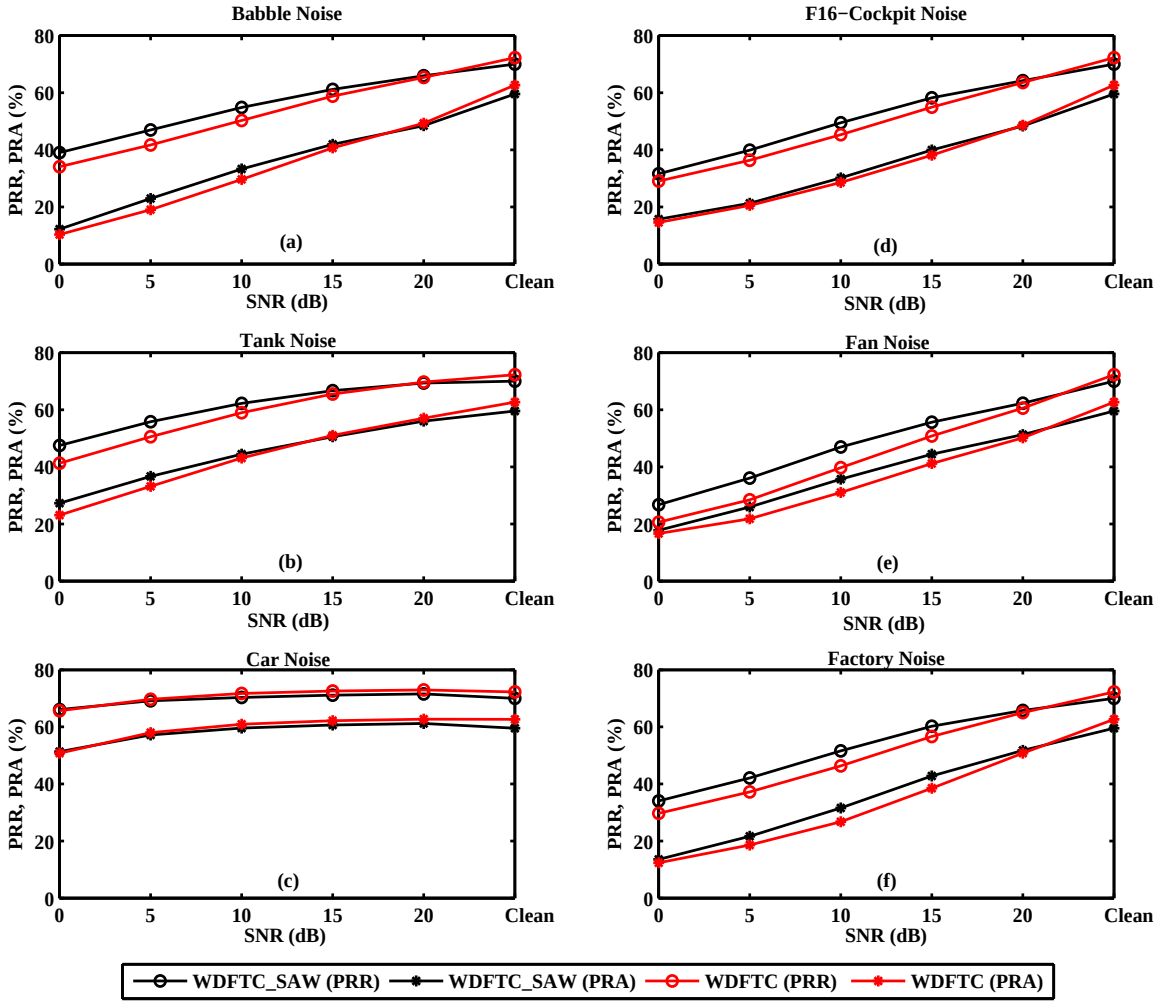


Figure 8. Illustrating the impact of different noises at various SNRs on WDFTC and WDFTC_SAW phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from female speaker in the TIMIT corpus using diagonal (DC) covariance.

tive coefficients are appended to the static features in 13-D, finally resulting in feature vectors in 39-D.

Using the 39-D feature vectors so extracted we build context independent (CI) phone models for each of the 42 monophones on a 3-state Bakis-topology HMMs [25] with a non-emitting terminating state. CI phone models are trained by the Baum-Welch algorithm. Next, context-dependent (CD) untied triphone models are trained for every triphone that occurs at least 8 times in the training data. The CI model parameters initialize the parameters of the CD models. The CD models are now trained as above through the Baum-Welch algorithm. Decision trees determining similar HMM states of all untied models are built in order to be merge the common states or senones. In all, 1000 senones are trained. Decision trees were pruned to restrict the number of leaves to within a prespecified number of tied states. Every state of all the HMMs was modelled by a mixture of 16 Gaussians.

7.1.2 Decoding

We employ the Sphinx-III decoder. Throughout the recognition process the most probable sequence of words is considered as the recognized one. This result depends on two factors, namely, the acoustic score that the HMM models provide and the probability of the existence of the sequence of words called language weight. We use a 3-gram language model [26]. The training corpus of the language model included all the transcriptions of the database.

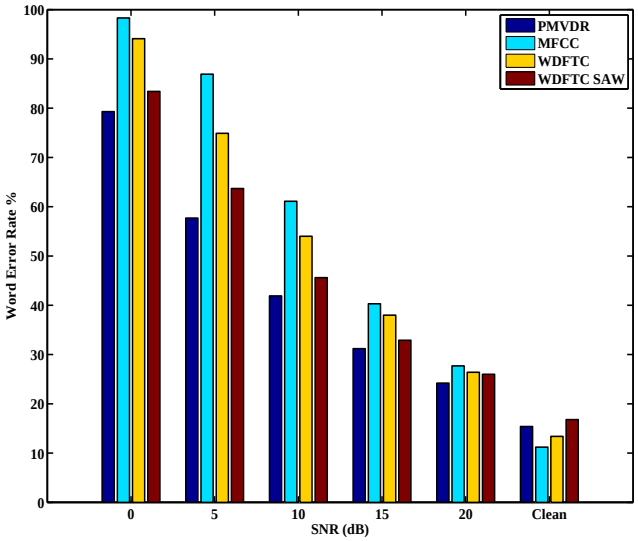


Figure 9. Word Error Rate (WER) of MFCC, PMVDR, WDFTC and WDFTC_SAW features under noise-free and Babble noise conditions

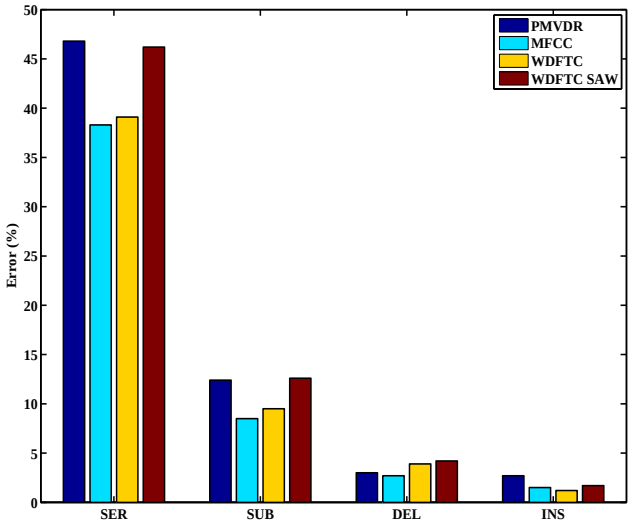


Figure 10. Sentence Error Rate (SER) and Errors in Word Recognition Performance

Table 1. Monophone Performance of MFCC, PMVDR, WDFTC and WDFTC_SAW on gender independent and specific speech samples from the clean TIMIT Corpus

Gender	All		Male		Female	
Feature	PRR	PRA	PRR	PRA	PRR	PRA
MFCC	73.9	64.8	75.8	65.3	76.2	67.8
PMVDR	72.6	63.3	75.0	65.3	74.6	65.6
WDFTC	72.2	62.6	74.9	65.9	74.3	65.3
WDFTC_SAW	70.0	59.5	72.5	62.8	72.5	62.9

8. Results and Discussions

Figure 2 shows the spectrograms of a TIMIT sentence (*She had your dark suit in greasy wash water all year*) and spectrograms of cepstral features from the PMVDR, WDFTC and MFCC in Fig. 2(a),(c),(e) and (g) respectively. Correspondingly, their noisy versions with 5 dB 'babble' noise are in Fig. 2(b),(d),(f) and (h) respectively. Robustness of WDFTC and PMVDR *vis-a-vis* MFCC is evident. Further, PRR and PRA for the PMVDR, WDFTC and MFCC features on the clean TIMIT corpus are outlined in Table 1 for gender independent and specific samples from the entire dataset. It can be seen from the tables that the MFCC performs slightly better than both the PMVDR and WDFTC with a 1-2% margin on the PRR and PRA for the entire dataset.

It may be noted from Figs. 3, 4 and 5 that the car, fan and tank noises are narrowband and stationary relative to the factory, F-16 cockpit and babble noises. From Fig. 3, it is observed that while the PMVDR and WDFTC exhibit a consistent degradation with falling SNR, their PRAs and PRRs performances are better than the MFCC. Further, it can be seen that the performance of PMVDR is marginally better than that of WDFTC and appreciably better than that of MFCC. The sensitivity of the acoustic models trained on MFCC in noisy backgrounds is clearly obviated by this observation. In comparison the PMVDR and WDFTC acoustic models seem to degrade gracefully with falling SNRs. The DC models of WDFTC and PMVDR are closer to each other in performance. MFCC exhibit poor performance under DC condition for most of the noise cases and SNRs. In the case of WDFTC and PMVDR, the DC models seem to give a better PRA for narrowband and broadband noise types, respectively. This is especially true for the low values of SNR. We proceed to test MFCC, PMVDR and WDFTC on gender-specific samples from the TIMIT database. We generate and test gender-specific cases by employing DC models; the phoneme recognition performance is shown in Figs. 4 and 5. We observe that PMVDR and WDFTC outperform MFCC in all noisy conditions except with F-16 cockpit noise.

We adopt WDFTC_SAW as a front-end feature for monophone and word recognition. Figure 6 presents the comparative performance of the WDFTC and WDFTC_SAW features. We observe that WDFTC_SAW outperforms WDFTC, particularly, under low SNR conditions except the case for car noise. Substantial improvement in PRR and PRA performance can also be seen for fan noise. Results from the gender-specific tests is shown in Figs. 7 and 8. Even in these tests SAW has improved the overall performance of WDFTC and the performance difference with the PMVDR is minimal. However, this improved performance is at the cost of marginal fall in noise-free condition.

In word recognition, we rank list word recognition performance of these features with clean and babble noise for SNR from 0 to 20 dB, vide Fig.9. It may be observed that the MFCC has lower WER compared to other features under clean conditions. However, PMVDR achieves lower WER under babble noise conditions for SNRs from 0 to 20 dB. The performance with WDFTC_SAW feature is better than WDFTC and MFCC. The performance difference between WDFTC_SAW and PMVDR is minimal.

Figure 10 shows sentence error rate (SER) and word recognition errors such as Substitutions (SUB), deletions (DEL) and Insertions (INS) from these four features under noise-free conditions. We can observe that MFCC has lowest SER, SUB, DEL and INS errors. PMVDR has heighest SER, substitution and insertion errors. Finally, adopting SAW to generate MFCC and PMVDR does not yield good results in both monophone and word recognition tasks.

In general, it can be easily concluded on the basis of all the results presented above that WDFTC_SAW, WDFTC and PMVDR outperform MFCC in different noise types and SNRs. This may be attributed to an all-pass based spectral warping in PMVDR, WDFTC and WDFTC_SAW feature generation. It has been shown [27] in addition that warping reduces the dynamic range of features and we believe that it plays an important role in noise robustness of the features. This is not very surprising as it is well known that the low-frequency spectrum of speech contains more energy than the high-frequency spectrum, and therefore it is more robust to additive noise. The high frequency spectrum in contrast is more susceptible to distortion by additive noise. Naturally, we achieve noise robustness by adopting a warping scheme which boosts at low frequency and bucks at high frequency. In other words, robustness is achieved by retaining reliable spectral bands in speech spectrum and eliminating bands where SNR is poor. The non-unitary nature of the WDFT which amplifies the low frequency spectrum and attenuates the high frequency part of the spectra also helps in highlighting the signal and downplaying the impact of the additive noise. It may be useful to note that the better performance of the PMVDR, WDFTC and WDFTC_SAW are attributed to their noise robustness and low feature variance.

9. Conclusion

In this paper, we presented exhaustive results illustrating the noise robust properties of the WDFTC, WDFTC_SAW and PMVDR features which have been benchmarked with MFCC. We introduced SAW to enhance robustness of WDFTC and demonatrate enhanced performance in noisy conditions. MFCC and PMVDR however failed to match the recognition results of these features alone. It was also shown that unlike MFCC, the WDFTC_SAW, WDFTC and PMVDR models degrade gracefully with falling SNR. It is important to note that we have not employed postprocessing of features like cepstral mean subtraction and variance normalization. Our experiments essentially shed light on some very useful properties of MFCC, WDFTC, WDFTC_SAW and PMVDR features that may be useful in improving the performance of current ASR systems in noise. Therein lies the take-home message.

References

[1] R. Muralishankar and D. O'Shaughnessy, "A comparative analysis of noise robust speech features extracted from all-pass based warping with mfcc in a noisy phoneme recognition," in *Proc. ICDT 2008*, Bucharest, Romania, July 2008, pp. 180–185.

[2] Douglas O'Shaughnessy, "Interacting with computers with voice: automatic speech recognition and synthesis," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1272–1305, Sept. 2003.

[3] John H.L. Hansen and M.A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. on Speech & Audio Proc.*, vol. 3, no. 5, pp. 415–421, Sep. 1995.

[4] I Varga, S Aalburg, B Andrassy, S Astrov, J.G. Bauer, C Beaugeant, C Geissler, and H Hoge, "ASR in mobile phones - an industrial approach," *IEEE Trans. on Speech & Audio Proc.*, vol. 10, no. 8, pp. 562–569, Nov. 2002.

[5] Bhiksha Raj and R. M. Stern, "Missing feature approaches in speech recognition," *IEEE Signal Proc. Mag.*, vol. 22, no. 5, pp. 101–116, Sept. 2005.

[6] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *ICASSP*, May 1996, pp. 733–736.

[7] Guillaume Lathoud, Mathew Magimai.-Doss, Bertrand Mesot, and Herve Bourland, "Unsupervised spectral subtraction for noise-robust ASR," in *ASRU*, Dec. 2005, pp. 343–348.

[8] Li Deng, Jasha Droppo, and Alex Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainity computed from a parametric model of speech distortion," *IEEE Trans. on Speech & Audio Proc.*, vol. 13, no. 3, pp. 412–421, May 2005.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech & Audio Proc.*, vol. 2, pp. 578–589, Oct. 1994.

[10] R. Muralishankar, A. Sangwan, and D. O'Shaughnessy, "Warped Discrete Cosine Transform Cepstrum: A new feature for speech processing," in *Proc. EUSIPCO*, Sep. 2005.

[11] A. Sangwan, R. Muralishankar, and D. O'Shaughnessy, "Performance analysis of the Warped Discrete Cosine Transform Cepstrum with MFCC using different classifiers," *MLSP*, pp. 99–104, Sept. 2005.

[12] U. H. Yapanel and John. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

[13] M. Wolfel, John. McDonough, and A. Waibel, "Warping and scaling of the minimum variance distortionless response," in *ASRU*, 2003, pp. 387–392.

[14] S. Bagchi and S. K. Mitra, *Nonuniform Discrete Fourier Transform and its Signal Processing Applications*, Norwell, MA: Kluwer, 1999.

[15] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. on Speech & Audio Proc.*, vol. 7, pp. 697–708, June 1999.

[16] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. on Speech & Audio Proc.*, vol. 8, no. 3, pp. 221–239, May 2000.

[17] S. Dharanipragada and B. D. Rao, "MVDR-based feature extraction for robust speech recognition," in *ICASSP*, May 2001, vol. 1, pp. 309–312.

[18] U. H. Yapanel and S. Dharanipragada, "Perceptual MVDR-based cepstral coefficients(PMCCs)for noise robust speech recognition," in *ICASSP*, May 2003, vol. 1, pp. 644–647.

[19] R. Lefebvre and C. Laflamme, "Spectral amplitude warping SAW for noise spectrum shaping in audio coding," in *Proc. ICASSP*, Apr. 1997, vol. 1, pp. 335–338.

[20] S. J. Young, *HTK Version 3.3: Reference Manual and User Manual*, Cambridge Univ. Engg. Dept.-Speech Grp., 2005.

[21] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoust. Speech & Signal Proc.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[22] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoust., Speech, Sig. Proc.*, vol. 34, no. 2, pp. 52–59, Feb. 1986.

[23] S. M. Ahadi, H. Sheikhzadeh, R. L. Brennan, G. H. Freeman, and E. Chou, "On the use of dynamic spectral parameters in speech recognition," in *Proc. ISSPIT*, Dec. 2003, pp. 757–760.

[24] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Trans. on Acoust. Speech & Signal Proc.*, vol. 38, no. 1, pp. 35–45, Jan. 1990.

[25] R. Bakis, "Continuous speech word recognition via centi-second acoustic states," in *Proc. of ASA Meeting (Washington, DC)*, Apr. 1976.

[26] "The CMU-Cambridge statistical language modelling toolkit, v2," Available: http://www.speech.cs.cmu.edu/SLM/toolkit_documentation.html.

[27] R. Muralishankar, A. Sangwan, and D. O'Shaughnessy, "Theoretical complex cepstrum of DCT and warped DCT filters," *Proc. IEEE Signal Proc. Ltrs.*, vol. 14, no. 5, pp. 367–370, May 2007.

Fairness index in single and double star Networks

Marc GILG
University of
Haute-Alsace,
34 rue du Grillenbreit,
F-68000 COLMAR,
France
marc.gilg@uha.fr

Abderrahim MAKHLOUF
University of Pierre and Marie Curie,
Doctoral School of Computer Science,
Telecom and Electronics,
4 Place Jussieu
F-75252 PARIS Cedex 05, France
abderrahim.makhlouf@etu.upmc.fr

Pascal LORENZ
University of
Haute-Alsace,
34 rue du Grillenbreit,
F-68000 COLMAR,
France
pascal.lorenz@uha.fr

Abstract

In wireless network, the communication works in half duplex mode and nodes can interfere together. In this context, fairness is not obvious. This paper will focus on fairness in the received packets by each node. Fairness is evaluated for static networks topologies called Single Star Network or Double Star Network. The fairness is quantified by its index. In this work, the evaluation of fairness index for double star network is given. Some value of this index are not possible for double star network topology. For example the index of one can only be possible if the double star network is simular to star network. Then the star networks are studied and some simulations are used to illustrate the way to get fairness in the network by controlling the flow rates.

1 Introduction

The performance of wireless 802.11 MAC protocol is generally evaluated with two parameters : collision probability and fairness [1],[13]. The fairness algorithm was widely studied by different research groups. Jain, Chiu and Hawe give us a mathematical definition in [6]. Their paper introduces the fairness index for any kind of resource sharing. This definition will be applied to the packet rates received by nodes.

TCP fairness was studied in [4], where the authors propose a distributed algorithm on neighbors to improve TCP fairness. In another paper [11] the authors propose a scheduling algorithm to get fairness in a multi-hop wireless network. Some papers are also based on the study of a distributed algorithm to maximalize throughput and fairness in Ad Hoc networks [2],[3] .

This paper is focused in fairness on node reception rates in an Ad-Hoc wireless network. A general introduction for fairness and fairness index is given. After that, a description of the network characteristic is done. For a theoretical analysis of the problem, the fairness index is evaluated for a basic network called Double Star Network and Star network. The existence of some value of fairness index is proved. We recall form [15] that exact fairness, such that fairness index is one, is not possible until Double Star Network degenerates in a Single Star Network. The fairness of Star network is studied and an algorithm is recalled from [14]. The simulation is done with NS2-2.33, will show that fairness can be accomplished by limiting the transmission rate of some nodes.

2 Network model and fairness

This work is done in the area of Ad-Hoc wireless network. An Ad-Hoc network is made of wireless nodes which establish wireless communications between themselves. In this context, there is no central infrastructure. This means that the nodes are equivalent. A node can be in two states at a given time, transmission or reception. The limitation of radio communications implies that the communications of a channel are limited in distance and they act in half duplex mode. These characteristics are described as follow :

2.1 Network model

We consider an Ad-Hoc network such as :

- The network is packet-switched
- The nodes are in half-duplex mode
- Only nodes in some distance can communicate
- The time is divided in time slots
- Packets are sent in time slots

2.2 Fairness index

In an Ad-Hoc network, the nodes have an equivalent role. Because the communications are in half-duplex mode, the position of the node in the network topology is dramatically related to its transmission rate. Some nodes with high degrees of connectivity in transmission will interfere with a high number of neighbors. This implies that the transmission and reception rates of the nodes are different.

In this paper, we try to give a fair access to each node. Fairness can be expressed as a mathematical formula given in [6]. The formula is based on a resource independent model and can be used to express fairness for any shared resource. It is also independent of network scalability. The resource will be applied to the reception rate of each node.

We recall the fairness index definition form [6] :

Definition 2.1. The fairness index of a shared resource x_i is given by

$$f(x_i) = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2} \quad (1)$$

We will apply this approach to the reception rate of a node. Here are some of the notations :

Notation. The following notations will be used in this work :

- Let x_i be the reception rate of node X_i
- Let $r_{j,i}$ be the reception rate of node X_i of the packets send by X_j
- Let D_i be the degree of node X_i
- Let S_j^i be the transmission rate of node X_j which is a neighbor of node X_i

Remark 1. Using the previous notations, we have a reception rate x_i of node X_i :

$$x_i = \sum_{j=1}^{D_i} r_{j,i}$$

Node X_j has transmission rate S_j. Therefore we have $r_{j,i} = S_j^i$. This gives us :

$$x_i = \sum_{j=1}^{D_i} S_j^i \quad (2)$$

3 Double Star Network

Compute a fairness condition on reception rates is not easy for some Ad-Hoc networks. From a theoretical approach, the star double network and star network are introduced. This network is simple enough to compute a relation on transmission rates to get fairness. It can represent a subgraph of an Ad-Hoc network composed of a central nodes and its neighbors.

3.1 Definition

Definition 3.1 (Double Star network)**.** The double star network $SN_{n,m}$ is composed of $n + m + 1$ nodes $\{X_0, X_1, \ldots, X_n, X_{n+1}, ..X_{n+m}\}$ where :

- $\{X_i, i \leq n\}$ are neighbors of the node X_0, $\{X_0, X_i, n+1 \leq i \leq n+m\}$ are neighbors of the node X_n
- and there is no connection between X_i, X_j for $i, j > 0$ only X_0 is the neighbor of X_n.

This is an example :

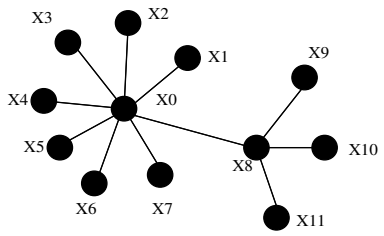


Figure 1. The $SN_{8,3}$ double star network

When the central nodes X_0 or X_n sends a packets, it will be received by their neighbors and this is not fair. Each set of packets respectively sent by X_i where $0 < i < n+1$ and or by X_j where $n < j < m+n+1$ will be seen respectively only once by X_0 or X_n.

3.2 Fairness index of the double star network

We can compute the fairness index for a double star network :

Lemma 3.1. *For a double star network $SN_{n,m}$, the fairness index α for the reception rate is given by the equation :*

$$\begin{aligned} &2\left(2 - \alpha(n+m+1)\right) X^2 - 2\alpha(n+m+1)Y^2 + \\ &\mathcal{Q}(S_0, S_n) = 0 \end{aligned} \quad (3)$$

where :

$$X = \sum_{i=1}^{n-1} S_i + \frac{(2n - \alpha(n+m+1))S_0}{2(2-\alpha(n+m+1))} + \frac{(2(m+1) - \alpha(n+m+1))S_n}{2(2-\alpha(n+m+1))}$$

$$Y = \sum_{i=n+1}^{n+m} S_i - \frac{\alpha(n+m+1)S_0 - \alpha(n+m+1)S_n}{2\alpha(n+m+1)}$$

$$\mathcal{Q}(S_0, S_n) = AS_0^2 + 2BS_0S_n + CS_n^2$$

is a quadratique form where

$$A = -\frac{(n-1)(n+m+1)\alpha((n+m+1)\alpha - n - 1)}{(n+m+1)\alpha - 2}$$

$$B = \frac{m(n-1)(n+m+1)\alpha}{\alpha(n+m+1) - 2}$$

$$C = -\frac{(n+m+1)^2\alpha^2(2m-3) - 2(n+m+1)\alpha(m^2-2)}{2(\alpha(n+m+1)-2)}$$

Proof. Let α be :

$$\alpha = f(x)$$

some basic computation gives the equation (3). □

The equation (3) might have no solutions. Let's see which value α will give no trivial solutions. Equation (10) is a quadratique relation in X and Y it has a constant term $\mathcal{Q}(S_0, S_n)$.

3.2.1 The coefficients of X^2 and Y^2

The coefficient of Y^2 is $-2\alpha(n+m+1)$ it is negative because α is a fairness index and therefore α is positive.

The coefficient of X^2 is $2(2-\alpha(n+m+1))$. Its sign depends on $2-\alpha(n+m+1)$.

The following lemma gives some results about the existence of solution of (3) :

Lemma 3.2. *For a double star network, the existence of fairness index α is submit to the following rules :*

- *If $\frac{2}{n+m+1} > \alpha$ then there exits no trivial solutions for equation (3).*
- *If $\alpha > \frac{2}{n+m+1}$ then there exists no trivial solutions for equation (3) if $\mathcal{Q}(S_0, S_n) > 0$.*

Proof. If $\frac{2}{n+m+1} > \alpha$ then the coefficient of X^2 and Y^2 have opposite sign. This implies the existence of solutions. If $\alpha > \frac{2}{n+m+1}$ then the coefficient of X^2 and Y^2 have same sign. To have no trivial solution, the quadratique form $\mathcal{Q}(S_0, S_n)$ has to be positive. □

The lemma 3.2 shows the importance of the sign of the quadratique form $\mathcal{Q}(S_0, S_n)$ if $\alpha > \frac{2}{n+m+1}$. In the next part, the sign of quadratique form will be studied.

3.2.2 The sign of the quadratique form $\mathcal{Q}(S_0, S_n)$

Recall that :

$$\mathcal{Q}(S_0, S_n) = AS_0^2 + 2BS_0S_n + CS_n^2 \tag{4}$$

In the lemma 3.2, the sign of the quadratique form $\mathcal{Q}(S_0, S_n)$ is important for $\alpha > \frac{2}{n+m+1}$. Let's suppose that $(n+m+1)\alpha - 2 > 0$, then the denominator of A, B and C are positive. Let's study their numerator :

- The sign of A is the sign of $n+1-(n+m+1)\alpha$.
- The sign of B is positive because $n > 1$.
- The sign of C is the sign of $2(m^2-2) - (n+m+1)\alpha(2m-3)$.

The next lemma will give conditions on n and m to have A, B and C positive.

Lemma 3.3. *If $\frac{2}{n+m+1} < \alpha$ and $m \geq 1$ then*

$$\frac{m^2-2}{2m-3} \geq 1$$

and C is positive if

$$\frac{2}{m+n+1} < \alpha \leq \frac{2(m^2-2)}{(n+m+1)(2m-3)}$$

Proof. If $\alpha > \frac{2}{n+m+1}\frac{m^2-2}{2m-3}$ then C is negative else C is positive. The sign of C is given by

$$2(m^2-2) - (n+m+1)\alpha(2m-3)$$

Let's prove that :

$$2(m^2-2) - (n+m+1)\alpha(2m-3) < 0$$

This implies that :

$$\frac{2}{n+m+1}\frac{m^2-2}{2m-3} < \alpha$$

□

Lemma 3.4. *A is positive if :*

$$\frac{2}{n+m+1} < \alpha \leq \frac{n+1}{n+m+1}$$

Proof. The sign of A is given by :

$$n+1-(n+m+1)\alpha$$

□

The next theorem will give a condition for the quadratique form $\mathcal{Q}(S_0, S_n)$ to be positive.

theorem 3.1. *If* $1 < n \leq m$ *and*

$$\frac{2}{n+m+1} < \alpha \leq \frac{n+1}{n+m+1}$$

then the quadratique form $\mathcal{Q}(S_0, S_n)$ *is positive.*

Proof. B is positive, and because

$$\frac{2}{n+m+1} < \alpha \leq \frac{n+1}{n+m+1}$$

A is positive according to lemma 3.4. Recall the condition for C to be positive given by lemma 3.3 :

$$\frac{2}{m+n+1} < \alpha \leq \frac{2(m^2-2)}{(n+m+1)(2m-3)}$$

Let's prove that :

$$\frac{n+1}{m+n+1} \leq \frac{2(m^2-2)}{(n+m+1)(2m-3)}$$

if $1 \leq m$.

$n+m+1$ is positive, this implies to prove :

$$n+1 \leq \frac{2(m^2-2)}{2m-3}$$

Because $n \leq m$ we have $n+1 \leq m+1$ Let's prove that

$$n+1 \leq m+1 \leq \frac{2(m^2-2)}{2m-3}$$

We suppose that $1 < m$ then $2m-3$ is positive and the condition is equivalent to :

$$(m+1)(2m-3) \leq 2(m^2-2)$$

This is equivalent to

$$0 \leq 2(m^2-2) - (m+1)(2m-3) = m-1$$

. This is true because $1 < m$. □

The theorem 3.1 and lemma 3.2 implie the following corollary :

Corollary 3.1. *If* $1 < n \leq m$*, then in a double star network it is possible to get a fairness index* α *such that*

$$\alpha \leq \frac{n+1}{n+m+1}$$

If $n > m > 1$ then we set $Y_0 = X_n, Y_1 = X_{n+1}, \ldots, Y_{p-1} = X_{n+m}$, $Y_p = X_0, Y_{p+1} = X_1, \ldots Y_{p+q} = X_{n-1}$ this gives a $SN_{p,q}$ double star network with $p = m+1$ and $q = n-1$. Notice that $p < q$ and we prove the following corollary :

Corollary 3.2. *If* $1 < m < n$ *then in a double star network it is possible to get a fairness index* α *such that*

$$\alpha \leq \frac{m+2}{n+m+1}$$

3.2.3 Maximal fairness index for double star network.

The corollaries 3.1 and 3.2 give a upper bound for the fairness index of a double star network. This is shown in the next lemma :

Lemma 3.5. *Let* α *be a fairness index of a double star network* $SN_{n,m}$ *which has no zero data rate reception.*

- *If* $1 < n \leq m$ *then* $\alpha \leq \frac{n+1}{n+m+1}$ *exists*
- *If* $1 < m < n$ *then* $\alpha \leq \frac{m+2}{n+m+1}$ *exists*

To validate the theoretical analysis, the next section will present some simulations.

3.3 Double star network simulations

NS2-2.33 is used for the next simulations. CBR over UDP traffic is used. The nodes $X_i, 1 \leq i \leq n-1$ and $X_j, n+1 \leq j \leq n+m$ have a CBR rate of 0.5Mb. The nodes X_0 and X_n have a CBR rate going for 0.1Mbps to 1.0Mbps with a step of 0.5Mbps. The Ad-Hoc routing protocol DSDV is active. The CBR traffic goes from X_0 to $X_i, 1 \leq i \leq n$, for X_n to $X_j, n+1 \leq l \leq n+m$ and backwards.

3.3.1 The $SN_{3,8}$ double star network.

In this case, $n = 3$, $m = 8$ and $n < m$. The lemma 3.5 shows that there can exist fairness index α such that

$$\alpha \leq \frac{1}{3}$$

We do simulation for 1000 time slots, and we get the following results show in this figure 3.3.1 :

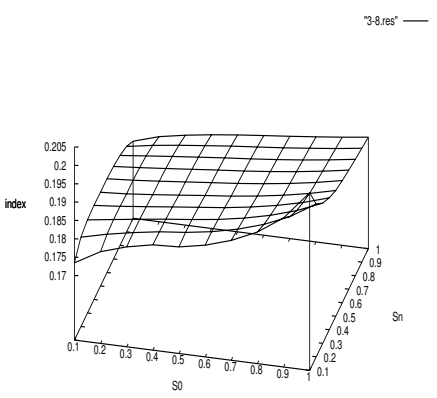


Figure 2. $SN_{3,8}$ **fairness index**

We notice that the maximal fairness index is given for $S_0 = 1.0Mbps$ and $S_n = 0.1Mbs$. Its value is 0.2008 witch is in the range given by lemma 3.5.

3.3.2 The $SN_{8,3}$ double star network.

In this case, $n = 8$, $m = 3$ and $n > m$. The lemma 3.5 shows that there can exist fairness index α such that

$$\alpha \leq \frac{5}{12}$$

We do simulation for 1000 time slots, and we get the following results show in this figure 3.3.2 :

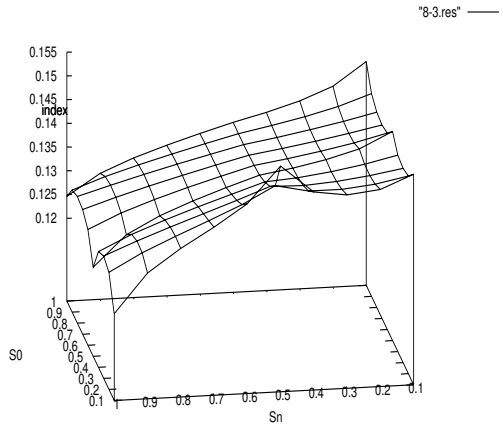


Figure 3. $SN_{8,3}$ **fairness index**

We notice that the maximal fairness index is given for $S_0 = 0.5Mbps$ and $S_n = 0.5Mbs$. Its value is 0.1501 witch is lower than $\frac{1}{6} = \frac{2}{n+m+1}$.

If α will be greater than $\frac{1}{6}$, then the equation 3 will be the equation of an ellipse with variable X and Y. It will be a very lucky to have X, Y on this ellipse. If the fairness index is lower than $\frac{1}{6}$ then equation 3 is easier to solve.

We have given results of existence of fairness index if $\alpha \leq \frac{n+1}{n+m+1}$ or if $\alpha \leq \frac{m+2}{n+m+1}$, but what happens if the fairness index is greater as this values, say $\alpha = 1$. The next section will give an answer in this case.

3.4 Fairness for double star network.

A network is fair if the fairness index is 1. This was studied in the conference [15], and we recall the main results in this section.

To be fair in a double star network, the fairness index of the network must be 1. This implies the following lemma :

Lemma 3.6. *A double star network $SN_{n,m}$ is fair if and only if :*

$$n\left(a - X - Y\right)^2 + m\left(a + Y\right)^2 + m(n-1)a^2 = 0 \quad (5)$$

where

- $a = S_0 - S_n$
- $X = \sum_{i=1}^{n-1} S_i$
- $Y = \sum_{i=n+1}^{n+m} S_i$

Proof. If the network is fair, the fairness index is

$$f(x) = 1$$

this gives :

$$\left(\sum_{i=1}^{n} S_i + nS_0 + \sum_{i=n+1}^{n+m} S_i + S_0 + mS_n\right)^2 = (n+m+1)\times$$
$$\left(\left(\sum_{i=1}^{n} S_i\right)^2 + (n+1)S_0^2 + \left(\sum_{i=n+1}^{n+m} S_i\right)^2 + mS_n^2\right)$$

By direct computation, we get the condition (5). □

Remark 2. In equation (5), we can observe that all terms are positive or null. This implies that each term has to be zero for the relation to be validate. We have to discuss about the value of n and m.

3.4.1 Case $n \neq 0, m = 0$

If $m = 0$ and $n \neq 0$, the relation (5) becomes :

$$n(a - X)^2 = 0 \quad (6)$$

This implies that $a = X$. This is the result given for a star network in the paper [14].

3.4.2 Case $n = 0, m \neq 0$

If $n = 0$ and $m \neq 0$ the relation (5) becomes :

$$\left(a + Y\right)^2 - a^2 = 0 \quad (7)$$

This implies that $Y = 0$ or $Y = -2 * a$. Because $n = 0$ we have $a = 0$ and then $Y = 0$. In this configuration, no node is transmitting.

3.4.3 Case $n = 1, m \neq 0$

If $n = 1$, then $X = 0$, and the relation (5) becomes:

$$(a - Y)^2 + m\,(a + Y)^2 = 0 \tag{8}$$

Because $m \neq 0$, this implies that :

$$\begin{cases} a - Y = 0 \\ a + Y = 0 \end{cases}$$

In this case, we get $Y = 0$ and $a = 0$. Then only X_0 and X_1 are transmitting with the same rate S_0.

3.4.4 Case $n = 1, m = 0$

If $n = 1$, then $X = 0$, and the relation (5) becomes:

$$(a - Y)^2 = 0 \tag{9}$$

This implies that $Y = a$, and then we get :

$$S_0 = \sum_{i=1}^{m+1} S_i$$

This is the result for the star network SN_{m+1}.

3.4.5 Case $n \neq 0, n \neq 1, m \neq 0$

In this case, the equation (5) has no coefficient equal to zero. Then every term should be equal to zero :

$$\begin{cases} a - X - Y = 0 \\ a + Y = 0 \\ a = 0 \end{cases}$$

The only solution is $a = 0, X = 0, Y = 0$ there is no packet transmitted.

The next theorem was proved :

theorem 3.2. *The fairness of packet transmitted in a $SN_{n,m}$ double star network is given by:*

- $S_0 = \sum_{i=1}^{n} S_i$, *if* $< m = 0$ *and* $n > 0$ *this is a* SN_n *star network,*
- $S_1 = \sum_{i=2}^{m+1} S_i$ *if* $n = 1$ *and* $m > 0$, *this is a* SN_m *star network,*

If $m > 0$ and $n > 1$, then transmitting a packet broke the fairness condition.

This theorem shows that exact fairness can't exist in a non degenerate double-star network. In the next section, we recall the results for [14] whitch gives the results in the degenerate case called star network.

4 Star Network

Definition 4.1 (Star network)**.** The star network SN_n is composed of $n + 1$ nodes $\{X_0, X_1, \ldots, X_n\}$ where $\{X_1, \ldots, X_n\}$ are neighbors of the node X_0, and there is no connexion between X_i, X_j for $i, j > 0$.

The next graph shows a star network :

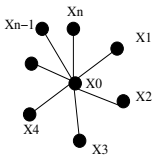


Figure 4. A star network

In this network, the central node X_0 has a degree of n, and each of its neighbors has a degree of 1. This is a very unfair communication channel repartition. When the node X_0 sends a packet, it will be received by the n neighbors. This packet will be seen n times in the network. On the other hand, each packet send by $X_i, i \geq 1$ will be seen only once by X_0. Intuitively, we see that to get fairness, the transmission rate of X_0 must balance the transmission rates of all other nodes. This will be shown in the next section.

4.1 Fairness of the star network

We can compute the fairness conditions for a star network :

Lemma 4.1. *For a star network SN_n, the fairness for the reception rates hold only and only if :*

$$S_0 - \sum_{i=1}^{n} S_i^0 = 0 \tag{10}$$

Proof. Using the expression (2) of x_i, the reception rate of node X_i, we have :

$$f(x) = \frac{\left(\sum_{i=0}^{n} \sum_{j=1}^{D_i} S_j^i\right)^2}{(n+1)\sum_{i=0}^{n}\left(\sum_{j=1}^{D_i} S_j^i\right)^2}$$

If the network is fair, we must have $f(x) = 1$. By a direct computation, we get :

$$S_0 - \sum_{i=1}^{n} S_i^0 = 0$$

□

The relation gives a direct condition on transmission rates to achieve fairness. It is much simpler to compare transmission rates than to evaluate the fairness index. The fairness index is based on a continuous function for transmission rates. This implies that if the difference (10) is close to zero, then the fairness index is close to 1. To get a fair star network, the condition (10) must to move closer. This remark lets us introduce a fairness algorithm which will try to minimize the difference (10) to achieve fairness.

Remark 3. The star network is fair only and only if the transmission rate S_0 is the sum of the transmission rates of all the neighbors of X_0.

4.2 Fairness algorithm

An Ad-Hoc network is seen from a node Y_i as a star network SN_d where d is the degree of the node Y_i. Following (3), we can imagine that the node Y_i adjusts its transmission rate such that it corresponds to the sum of the reception rates. It gets from its neighbors. This gives us the following algorithm running on each node Y_i and using a parameter s given by the administrator of the network :

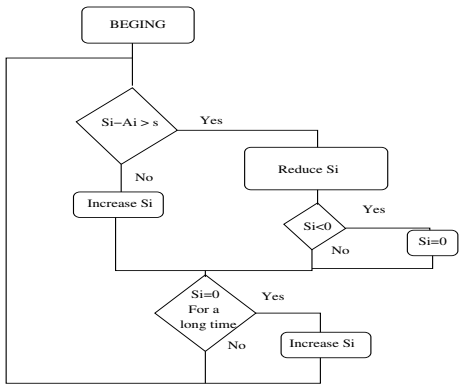


Figure 5. Algorithm

1. Computes the sum of the neighbors transmission rates A_i
2. Compares the sum A_i to the node Y_i transmission rate S_i.
3. If $S_i - A_i > s$ then reduce S_i, if S_j becomes negative, then set it to 0.
4. If $S_i - A_i < s$ then increase S_i if it is possible.
5. If $S_j = 0$ for a long time, then increase it.
6. Go to step 1

This algorithm acts only on the transmission rates. It tries to adjust the difference $S_i - A_i$ to be close to zero. To do this, it needs to have some control over S_i. For a star network, the theoretical approach shows that minimizing the difference $S_i - A_i$ will increase fairness. But it can also be used on any Ad-Hoc networks.

The parameter s controls the sensibility of the algorithm to the standard access algorithm. If s is null, the algorithm will try to always get an exact fairness. This is not realistic, and this can decrease the performance of the network. If s is too high, then the algorithm will have no influence on fairness.

4.3 Simulations

We use the network simulator ns2 to do the simulations. The DSDV routing protocol is used. First, we will do the simulation with the original ns2. After that, we modify ns2 to simulate our algorithm. The transmission rate will be computed on TCP packets sent by each nodes. We use FTP agent to simulate traffic.

4.3.1 The star network SN_6

We will now use SN_6 star network for simulation where two FTP connections (up and down) are established between X_0 and $X_i, i > 0$. The following graph shows the fairness index in a function of time :

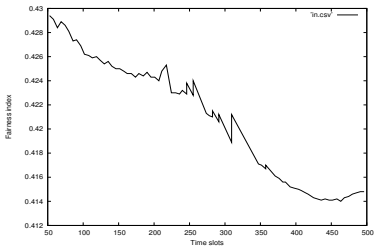


Figure 6. Fairness of a SN_6 star network

We can also compute the difference (10) in function of time slots :

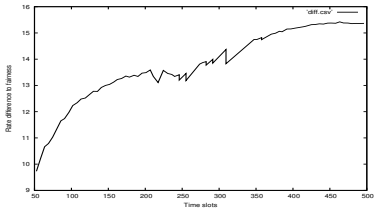


Figure 7. Rate difference to fairness

We notice that the difference (10) increases in time, which is coherent with the fact that the fairness index is decreasing. The aim of our algorithm is to keep the difference (10) close to zero.

We apply our algorithm to this network with $s = 500$. To reduce S_j the algorithm changes the rate from 1Mb to 0.5Mb. When the delay becomes too long, the algorithm reset the node in the standard rate.

This gives us the following graphs :

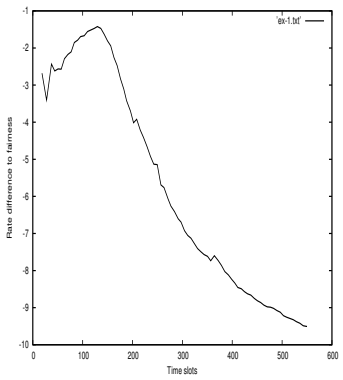


Figure 8. Fairness of a SN_6 star network with modified access algorithm

Remark that the fairness index goes to 0.43 which is better than the simulation done by the default algorithm of ns2.

We can also compute the difference (10) in function of time slots. This is shown in the next graph, see how the algorithm acts on to minimize the difference.

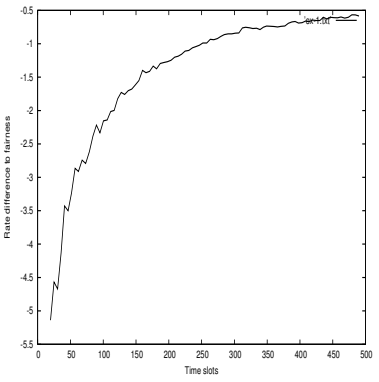


Figure 9. Rate difference to fairness

We notice that the difference rate (10) goes from 15 to -10 which is closer to zero. This improves the fairness index from 0.41 to 0.43. The fairness index is still far from 1, but we can expect that it will be better if the simulation time goes to infinity as described in the following figure.

4.3.2 The star network SN_8

We will now use the SN_8 star network for simulation where two FTP connections (up and down) are established between X_0 and $X_i, i > 0$. The following graph shows the fairness index in a function of time :

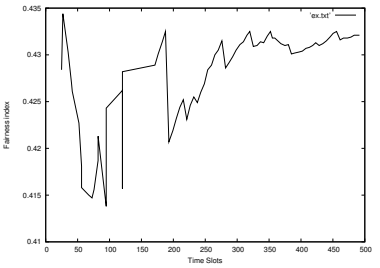


Figure 10. Fairness of a SN_8 star network

We can also compute the difference (10) in function of time slots :

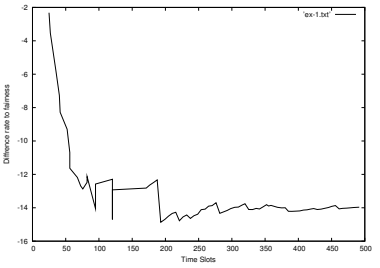


Figure 11. Rate difference to fairness

We can notice that the fairness index is around 0.432 at time slot 500 and the difference (10) is around -14. This confirms that for the star network SN_8, the behavior is not fair. Now we will try to see what is happening with our algorithm. The fairness index is shown in the following graph :

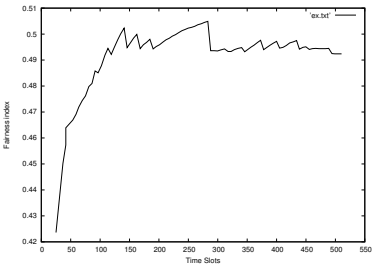


Figure 12. Fairness of a SN_8 star network

This shows that the fairness index is better than the ns2 standard case. At 500, the fairness index is higher than 0.49

and is still increasing. Therefore the algorithm gives better results. Let's take a look at the difference of rate to fairness given by (10) :

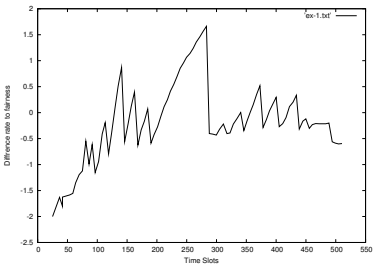


Figure 13. Difference of rate to fairness of a SN_8 star network

This graph shows that the algorithm is working well. The difference is less that -0.6 at time slot 500. The algorithm seems to be efficient.

4.4 A no star network

In this example, the algorithm is applied to no star network. The topology of the network is given by the graph :

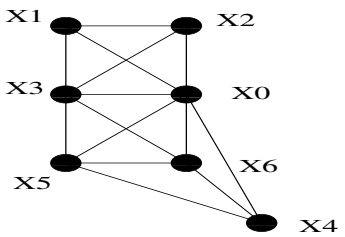


Figure 14. A no star network

There is an FTP traffic simulated for each node to its neighbors. When we applied the standard ns2 simulator, this gave for the fairness index the following result:

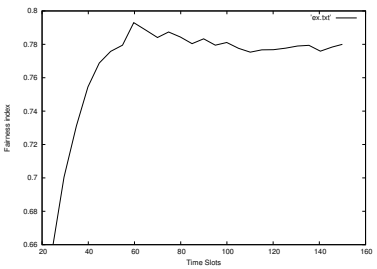


Figure 15. Fairness of the no star network

We can see that the node X_0 is connected to every other node in the network. This node can play the same role as the central node for a star network. The rate difference (10) can be evaluated for this node. This gives the next graph :

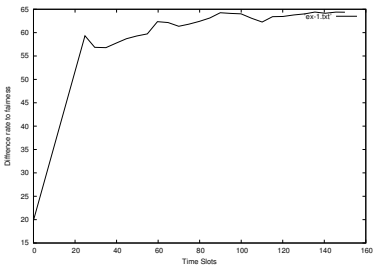


Figure 16. Difference rate of the no star network

The difference is increasing according to the fairness index of the network. This let us supposes that the fairness index is related to the rate difference of X_0 to its neighbors. The algorithm for star networks can be apply to reduce the rate difference (10). When the algorithm is used, the rate difference (10) reacts as following graph :

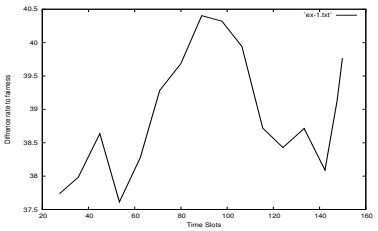


Figure 17. Difference rate of the no star network with our algorithm

The rate difference goes down from 65 to less than 40. The fairness index is shown in the next graph :

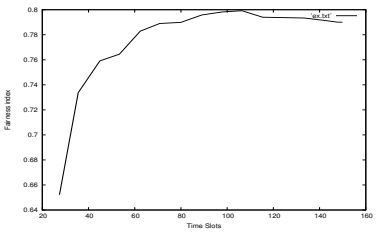


Figure 18. Fairness index of the no star network with algorithm running

Thus the fairness index tends to be closer to 0.8.

In this example, we applied our algorithm to a no star network. The algorithm improves the fairness index.

5 Conclusion

In this work, the study is focused on the fairness of the reception rate. After some generalities, a double star network and a single star network are introduced. This network enables us to compute the fairness index. For double star network, we give some upper bound on fairness index to guaranty their existence. We prove that fairness can only exist if double star network degenerates in star network. Fairness is studied in star networks. Then we elaborate an algorithm to get fairness. The algorithm needs only to know the reception rate and the transmission rate of the node and it can be used on every Ad-Hoc network. But in this case the influence on the fairness index is not developed. Never less and example are shown where the algorithm improves fairness.

The simulation shows that the fairness index is improved for a star network if we apply our algorithm. But the fairness index doesn't seem to react very efficiently. We can expect better results if the simulation time goes to infinity.

In some further work, we propose to apply our algorithm to more complex networks to approach a general Ad-Hoc network. A first step is to compute the maximal fairness index for some topology and then we expect to modify our algorithm to reach this maximal fairness index for a more general Ad-Hoc network.

References

[1] Can Emre Koksal, Hisham Kassab, Hari Balakrishnan, *An Analysis of Shor-Term Fairness in Wireless Media Access Protocols*, ACM SIGMETRICS 2000, Santa Clara, CA, June 2000, p. 118-119.

[2] Marc Gilg, Pascal Lorenz *A Totally Distributed and Adjustable Scheduling Algorithm in Wireless Ad Hoc Networks* International Conference on Networking and Services, ICNS'2005, October 23-28, 2005, Papeete, Tahiti, French Polynesia.

[3] Marc Gilg, Pascal Lorenz *An Adjustable Scheduling Algorithm in Wireless Ad Hoc Networks*, 3rd European Conference on Universal Multiservice Networks, October 25-27, 2004, Porto, Portugal, LNCS 3262, pp 216-226.

[4] Kaixin Xu, Mario Gerla, Lantao Qi, Yantai Shu *Enhancing TCP fairness in Ad Hoc Wireless Networks using neighbourhood RED* MobiCom'O3, September 14-19,2003, San Diego, USA, ACM 0-89791-88-6/97/05

[5] Thomas Kunz , Hao Zhang *Transport Layer Fairness and Congestion Control in Multihop Wireless Networks* Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2007), October 2007

[6] R. Jain, D. Chiu, and W. Hawe, *A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems*, DEC Research Report TR-301, September 1984.

[7] Lachlan L.H. Andrew , Stephen V. Hanly , Rami G. Mukhtar *Active Queue Management for Fair Resource Allocation in Wireless Networks* IEEE Transactions on Mobile Computing, February 2008 pp. 231-246

[8] Nitin Vaidya , Anurag Dugar , Seema Gupta , Paramvir Bahl *Distributed Fair Scheduling in a Wireless LAN* IEEE Transactions on Mobile Computing, November 2005 pp. 616-629

[9] Mohammad Mahfuzul Islam , Manzur Murshed *Min-Max Fairness Scheme for Resource Allocation in Cellular Multimedia Networks* International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II, April 2005 pp. 265-270

[10] *Fair Scheduling over multiple servers with flow-dependent server rate* S.R. Mohanty , L.N. Bhuyan Proceedings. 2006 31st IEEE Conference on Local Computer Networks, November 2006 pp. 73-80

[11] A. Sagora, D.J. Vergados, D. D. Vergados, *On Per-Flow Fairness and Scheduling in Wireless Multi-hop Networks* ICC 2008, IEEE, Workshop proceedings

[12] Sunwoong Choi , Kihong Park , Chong-kwon Kim *Performance Impact of Interlayer Dependence in Infrastructure WLANs* IEEE Transactions on Mobile Computing, July 2006 pp. 829-845

[13] Yu Yan Ming, Joe Timoney, Linda Doyle and Donal O'Mahony *Evaluation of Channel Fairness Models for Ad-Hoc networks* Proceedings of the First Joint IEI/IEE Symposium on Telecommunications Systems Research, Dublin, November 27th, 2001

[14] Marc Gilg, Abderrahim Makhlouf, Pascal Lorenz *Fairness in a Static Wireless Network* International Conference on Services and Networks Communications, ICSNC08, October 25-30, 2008, Sliema, Malta.

[15] A. MAKHLOUF, M. GILG, P. LORENZ *Fairness in Double Star Ad Hoc Network* The Fifth International Conference on Networking and Services ICNS 2009, IEEE, 20-25 april, Valencia

Assurance-driven design in Problem Oriented Engineering*

Jon G. Hall Lucia Rapanotti
Department of Computing
The Open University, UK
{J.G.Hall,L.Rapanotti}@open.ac.uk

Abstract

The design of assurance cases is hampered by the posit-and-prove approach to software and systems engineering; it has been observed that, traditionally, a product is produced and then evidence from the development is looked for to build an assurance case. Although post-hoc assured development is possible, it often results in errors being uncovered late—leading to costly redevelopment—or to systems being over-engineered—which also escalates cost. As a consequence, there has been a recent move towards the proactive design of the assurance case. Assurance-driven design sees assurance as a driving force in design. Assurance-driven design is suggestive of how the design process should be shaped for assurance. It is not, however, a prescriptive method; rather it allows an organisation to assess their assurance needs according to their developmental needs, including their attitude to risk, and to adapt their processes accordingly.

We have situated the work within Problem Oriented Engineering, a design framework inspired by Gentzen-style systems, with its root in requirement and software engineering. In the paper we present the main elements of the approach and report on its application in real-world projects.

Keywords: Dependability, Software Engineering, Assurance Case, Problem Oriented Engineering, Engineering Design

1 Introduction

By engineering design (shortly, design), we refer to the creative, iterative and often open-ended endevour of conceiving and developing products, systems and processes (adapted from [2]).

Engineering design by necessity includes the identification and clarification of requirements, the understanding and structuring of the context into which the engineered system will be deployed, the detailing of a design for a solution that can ensure satisfaction of the requirements in context, and the construction of arguments to assure the validating stake-holders that the solution will provide the functionality and qualities that are needed. The last of these is the concern of this paper.

Typically, for software at least, even though evidence is gathered during development the collation, documentation and quality injection of the assurance argument follows construction; perhaps this is because software development is currently sufficiently difficult without having to serve the needs of two masters: code *and* assurance. If software and assurance argument could be developed together, then developmental risk could be managed better–development errors that weaken an assurance argument could be found earlier in the process—as could developmental cost—by removing the compensating tendency to over-engineer.

Assurance-driven design (ADD), introduced in [1], does not make development any simpler; rather, it makes the building of an assurance argument a driver for development. Accepting this, however, ADD can guide the developer: by providing a more specific focus on those parts of a system that *require* assurance; by providing early feedback on design decisions; by capturing coverage of the design space; and, last but not least, by delivering an assurance argument alongside the product.

Our work on assurance-driven design is situated within Problem Oriented Engineering (POE), our framework for engineering design (instantiated for software in [3, 4]). The techniques we propose have no particular dependence on a software development context; indeed, our main example combines software and educational materials design and it is the assurance of their combined qualities that will drive our development.

The paper is structured as follows. Section 2 provides the briefest introduction to POE. In Section 3 we develop assurance-driven design, and in Section 4 illustrate its use through its application to a real-world problem. Section 5

*An expanded version of [1]

relates our work to that of others, and Section 6 reflects on what has been achieved and concludes the paper.

2 Problem Oriented Engineering

Problem Oriented Engineering is a framework for engineering design, similar in intent to Gentzen's Natural Deduction [5], presented as a sequent calculus. As such, POE supports rather than guides its user as to the particular sequence of design steps that will be used; the user choosing the sequence of steps that they deem most appropriate to the context of application. The basis of POE is the *problem* for representing *design problems* requiring designed solutions. *Problem transformations* transform problems into others in ways that preserve solutions (in a sense that will become clear). When we have managed to transform a problem to axioms[1] we have solved the problem, and we will have a designed solution for our efforts. A comprehensive presentation of POE is beyond he scope of this paper and can be found in [3, 4].

2.1 Problem

A problem has three descriptive elements: that of an existing real-world problem context, W; that of a requirement, R; and that of a solution, S. We write the problem with elements W, S and R as $W, S \vdash R$. What is known of a problem element is captured in its description; descriptions can be written in any appropriate language: examples include natural language, Alloy ([6]), and machine language. Solving a problem is finding S that satisfies R in the context of W.

Figure 1 gives an example of engineering design problem (shortly problem), described in a Problem-Frame-like notation ([7]). The problem (from a real world case study [8, 9]) is that of defining a controller to release decoy flare from a military aircraft: essentially decoy flares provide defence against incoming missile attack. The context includes a Pilot, a Defence system and some other existing hardware, represented in the figure as named undecorated rectangles. The solution to be designed is Decoy Controller, represented as a named decorated rectangle. The arc annotations are shared phenomena: for instance, the Pilot can send an ok command to the Decoy Controller. The solution needs to satisfy the Safe decoy control requirement, represented as a named dotted ellipse, for the safe release of decoys. Formally, in POE, this problem is represented as:

$$\text{Defence System}^{con}, \text{Dispenser Unit}^{out}_{fire,sel}, \text{Aircraft Status System}^{air}, \text{Pilot}^{ok}, \text{Decoy Controller}^{fire,sel}_{con,out,air,ok} \vdash \text{SDC}^{fire,sel}_{con,out,air,ok}$$

but we use both notations interchangeably.

[1] An *axiomatic problem* is a problem whose adequate, i.e., fit-for-purpose, solution is already known.

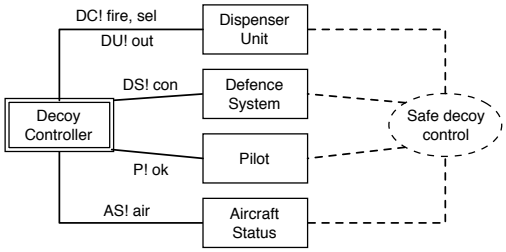


Figure 1. The Decoy Controller Problem

2.2 Problem transformations and justification obligations

Problem transformations capture discrete steps in the problem solving process. Many classes of transformation are recognised in POE, reflecting a variety of engineering practices reported in the literature or observed elsewhere. Problem transformations relate a problem and a justification to (a set of) problems. Problem transformations conform to the following general pattern (whose notation is based on that of [5]). Suppose we have *conclusion* problem $P : W, S \vdash R$, *premise* problems $P_i : W_i, S_i \vdash R_i$, $i = 1, ..., n$, $(n \geq 0)$ and *justification* J, then we will write:

$$\frac{P_1 : W_1, S_1 \vdash R_1 \quad ... \quad P_n : W_n, S_n \vdash R_n}{P : W, S \vdash R}\ \begin{matrix}[\text{NAME}] \\ \langle\langle J \rangle\rangle\end{matrix}$$

to mean that, derived from an application of the NAME problem transformation schema (discussed below):

> S is a solution of $W, S \vdash R$ with *adequacy argument* $(AA_1 \wedge ... \wedge AA_n) \wedge J$ whenever $S_1, ..., S_n$ are solutions of $W_1, S_1 \vdash R_1, ..., W_n, S_n \vdash R_n$, with adequacy arguments $AA_1, ..., AA_n$, respectively.

Engineering design under POE proceeds in a step-wise manner with the application of problem transformation schemata, examples of which appear below: the initial problem forms the root of a *development tree* with transformations applied to extend the tree upwards towards its leaves. A problem is solved for a stake-holder S if the development tree is complete, and the adequacy argument constructed for that tree convinces S that the solution is adequate. For technical reasons[2], we write

$$\overline{P}$$

to indicate that problem $P = W, S \vdash R$ is solved. As P will be fully detailed in determining the solution — to the satisfaction of stake-holders — the indication that P is solved is without justification. For the technical details, see [4].

A partial development tree is shown in Figure 2.

[2] Simply, that we may indicate an axiom in a Gentzen system thus.

$$\dfrac{\dfrac{\overline{P_3 : W_3, S_3 \vdash R_3} \qquad P_4 : W_4, S_4 \vdash R_4}{P_2 : W_2, S_2 \vdash R_2}\ {\scriptstyle [N_2]}\ {\scriptstyle \langle\langle J_2 \rangle\rangle}}{P_1 : W_1, S_2 \vdash R_1}\ {\scriptstyle [N_1]}\ {\scriptstyle \langle\langle J_1 \rangle\rangle}$$

Figure 2. A POE partial development tree

The figure contains four nodes, one for each of the problems P_1, P_2, P_3 and P_4. The problem transformation that gave the problem solver P_1 is justified by J_1, whereas the branching to problems P_2 and P_3 is justified by J_2. From the tree, we see that P_3 is solved. P_4 remains unsolved, so that the adequacy argument for the tree is incomplete; from the definition above, the incomplete adequacy argument is:

$$J_2 \wedge J_1$$

2.2.1 "Have we done enough?"

At any point in a development we can ask if we have done enough, i.e., if we were to declare our development complete would we be able to satisfy the validating stake-holders? This question is most obviously asked of the complete development, in which case an affirmative answer convinces all stake-holders that we have an adequate solution to the whole problem.

As previously mentioned, a completed development in POE is represented as a complete development tree, i.e., a tree in which no problems exist to be solved. The development is successful if the adequacy argument, AA, satisfies the stake-holders of the adequacy of the solution. For any stake-holder S, then, we have done enough if

$$\mathsf{AA} \text{ convinces } \mathsf{S}.$$

Consider again the form of the adequacy argument given a partial tree, such as that in Figure 2. Suppose that S is a stake-holder for problem P_1. Should P_1 be solvable, we would wish to find justification J_3 and solved problem P_5, say, such that

$$J_3 \wedge J_2 \wedge J_1 \text{ convinces } \mathsf{S} \qquad \text{and} \qquad \dfrac{\overline{P_5}\ {\scriptstyle [N_3]}\ {\scriptstyle \langle\langle J_3 \rangle\rangle}}{P_4}$$

If we were free to choose P_5 without any reference to the requirements of the argument that establishes it as fit-for-purpose (that formed when J_3 is added to the adequacy argument) it would be unlikely to result in something that could be justified. Of the techniques mentioned in the introduction to this paper, the 'posit' of 'posit and prove' is moving towards this 'free' choice; moreover, over-engineering a solution simply allows the engineer a freer choice.

As we begin to balance the choice of P_5 and J_3, we move towards the position of assurance-driven design, in which the requirements for justification motivate the design. The techniques we introduce in this paper allow us to structure the development so that this balance can occur. Primary amongst them is the construction of projections from the overall development tree into, what might be called, 'stake-holder spaces' in which validation takes place.

2.2.2 A formal backwater: the weakest pre-justification

Although it is — currently — only of theoretical interest, by inspection, there is a best such justification that, given an incomplete development tree, completes the adequacy argument so as to just satisfy the stake-holder, S. By analogy to other formal systems, we term this the *weakest pre-justification*, J_{wpj}, such that, if IAA is the current (incomplete) adequacy argument for that tree, then for any K

$$(\mathsf{K} \wedge \mathsf{IAA} \text{ convinces } \mathsf{S}) \Rightarrow (\mathsf{K} \Rightarrow J_{wpj})$$

3 Assurance-driven design

A metaphor for engineering design under POE is that one grows a forest of trees. Each tree in the forest grows from a root problem through problem transformations that generate problems like branches; with happy resonance, the tree's *stake*-holders guide the growth of the tree. Some trees, those that have root problems that are validatably solvable for its stake-holders will grow until they end with solved problem leaves.

There are many reasons why the forest has many trees: described elsewhere [10], but only of note in this paper, is the preservation of a record of unsuccessful design steps, i.e., design steps that are not validatable for the current stake-holders, which cause a development to backtrack to a point where a different approach can be taken. The backtracked sub-trees are kept as record of unsuccessful development strategies[3].

For this paper, we note simply that development trees grow through the developer's careful choice of effective design steps. To produce an effective design step, the developer must consider both the problem(s) that the step will produce towards solution *and* what is the justification obligation that will satisfy the *validating stake-holders*. With the discharged justification obligations forming the basis of the adequacy argument, the result of a sequence of effective design steps is a solution *together with its assurance*

[3] Backtracked trees are not 'deadwood'; rather they stand as proof of design space exploration, with their structure being reusable for, for instance, other stake-holders' problems. Unsolved problems that remain in backtracked trees do not affect the completed status of a development.

argument[4]. We have observed the interplay of design steps and their justification under POE (for instance, [11]), and have developed a simple, composable process pattern—the POE Process Pattern—that guides their effective interleaving. The structuring of the problem solving activity through the POE process pattern is the basis of assurance-driven design.

We note that a problem transformation schema is applied to a conclusion problem, and that the development tree is extended up by the application. There is no necessity for any premise problem to be determined before the justification is added. Indeed, one could see the problem solver saying "It is fashionable to have a fan oven in stainless steel", and then searching for an oven that fits the bill[5].

It is determining the needs for the justification, rather than for the premise problem(s), that motivates us to introduce assurance-driven design: assurance-driven design determines the justification first, and then looks for the corresponding premise problem.

3.1 The POE process pattern

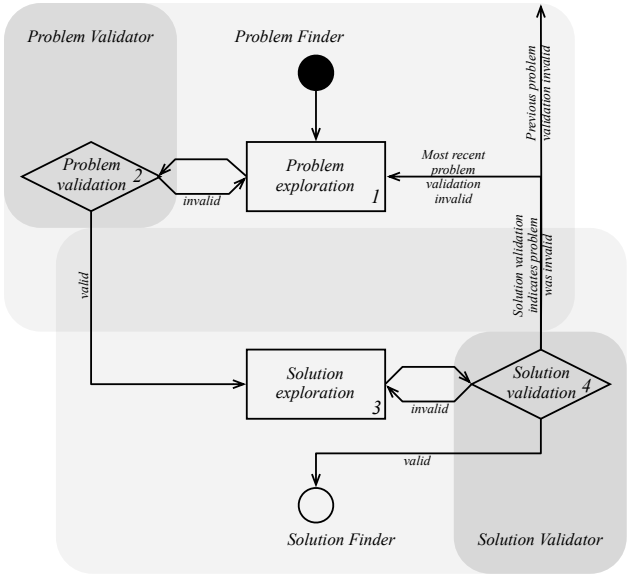


Figure 3. The POE Process Pattern for assurance-driven design

The POE process pattern shown in Figure 3 is described in a variant of the UML activity diagram notation [12]: rectangles are resource consuming activities; diamonds indicate choice points; the flow of information is indicated by arrows; the scope of the various roles is indicated by shading, overlapping shading indicating potential need for communication between roles. Referring to the numbers in the figure: first explore the problem better to understand it (*1*), checking that understanding through problem validation (*2*), iterating problem exploration as necessary; then explore the solution better to understand its design (*3*), checking that understanding through solution validation (*4*), iterating solution exploration as necessary.

The role of a *problem finder* during problem exploration is to explore their understanding of the problem (or part thereof), perhaps with the help of others. The goal of problem exploration is to produce descriptions of the problem that will satisfy the problem-validators(s) at problem validation. Similarly, the role of *solution finder* during solution exploration is to explore their understanding the solution (or part thereof) to the problem, again perhaps with the help of others. The goal of solution exploration is to produce descriptions of the solution that will satisfy the solution validator(s) at solution validation.

The role of a *problem validator* is to validate a candidate problem description. There are many familiar examples of problem validator. These include, but are not limited to:

- the customer or client — those that pay for a product or service;
- the regulator — those whose remit is the certification of safety of a safety of a safety-critical product, for instance;
- the business analyst — whose role is to determine whether the problem lies within the development organisation's business expertise envelope;
- the end-user — those who will use the product or service when commissioned.

It is a problem validator's role to answer the question "Is this (partial) problem description valid?" Depending on a problem validator's answer, the Problem Finder will need to re-explore the problem (when the answer is "No!"), or task the Solution Finder to find a (partial) solution (when the answer is "Yes!").

The role of the *solution validator(s)* is to validate a (candidate or partial) solution description, such as a candidate architecture (a partial solution) or choice of component (something of complete functionality). Although present in every commercial development, the roles of solution validator may be less familiar to the reader. They include, but are not limited to:

- a technical analyst — whose role is to determine whether a proffered solution is within the development organisation's technology expertise envelope;

[4]If there are no validating stake-holders for a development, the justification obligations can be ignored.

[5]Of course, we could have written such a statement as part of the requirement, but that would have been the stake-holder's statement, not the problem solver's.

- an oracle — who determines, for instance, which of a number of features should be included in the next release;

- a unit, module, or system tester; a project manager— who needs to timebox particular activities.

It is the solution validator's role to answer the question "Is this (candidate or partial) solution description valid?" Depending on their response, the problem solver may need to re-explore the solution (when the answer is "No!"), move back to exploring this or a previous problem (when the answer is "No, but it throws new light on the problem!"), or moving on to the next problem stage (when the answer is "Yes!").

The potential for looping in the POE process pattern concerns unsuccessful attempts to validate, and is indicated by arrows labelled *invalid* in the figure. Those leading back to exploration activities, of which there are two, continue their respective exploration activities in the obvious way. The other two invalid arrows lead from a failed solution validation to restart a problem exploration when the indication is that it was wrong. Examples of this latter form of failure are well known in the literature. For instance, Don Firesmith, in an upcoming book [13], talks about the need for architecture *re*-engineering in the light of inadequately specified quality requirements [part of an earlier problem exploration]:

> [...] it is often not until late in the project that the stakeholders recognize that the achieved levels of quality are inadequate. By then [...] the architecture has been essentially completed [solution exploration], and the system design and implementation has been based on the inadequate architecture.

In this way, recognising late that inadequately specified quality requirements (as discovered through problem exploration and validated at problem validation) have not been met can be very difficult and expensive to fix; leading to revisiting a long past problem, that of re-establishing the architecturally significant quality requirements[6].

Although we do not consider developmental risk explicitly in this paper, we note that feedback within the process has an impact on resources: an unsuccessful validation indicates that some previous exploration was invalid, to a greater or lesser extent. Moreover, some proportion of the development resource that will have expended during and subsequent to that exploration — the impact of the failed validation — will have been lost[7].

[6]Firesmith cites Boeing's selection of the Pratt and Whitneys PW2037 Engine for the Boeing 757 [14] as an instance of this problem.

[7]Work on risk management in POE is in preparation at the time of writing.

After successful problem validation, handover between the problem and solution finders occurs. In problem and solution finder are the same person, this raises no issues. Otherwise, it is possible to consider the solution finder as a problem validator, so that they receive a description of the problem that they have validated as the basis of their solution exploration. Symmetry dictates that the problem finder should have a role in solution finding too.

3.1.1 Building potent design processes

Although the POE process pattern provides a structure for problem solving, in its raw form, a problem will only be solved (i.e., the end state in Figure 3 is reached) when, after iteration, a validated problem is provided with a validated solution. This 'bang-bang' approach is suitable for simple problems, but is unlikely to form the basis of any realistically complex problem encountered in software engineering.

To add the necessary complexity, the POE process pattern combines with itself in three basic ways; in combination, it is again a process that can be combined. The three ways it can be combined are in sequence, in parallel and in a fractal-like manner, as suggested in Figure 4, and as described in the sequel.

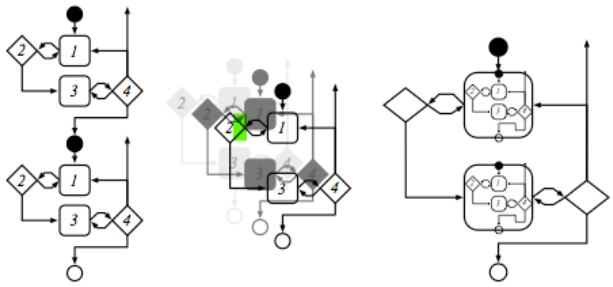


Figure 4. (a) Sequential, (b) Parallel and (c) Fractal-like combination

Sequential Design By identifying the end of one complete problem solving cycle with the start of another (see Figure 4(a)), we move a partially solved problem to the next phase: using the validated solution to explore the problem further. In [4], we show how a partial solution in the form of an architecture can lead to more detailed problem exploration: in that paper, we use the Model-View-Controller architecture to structure the solution of a problem, simplifying the problem to one of defining first the Model, then the View and finally the Controller.

In sequence, the POE process pattern models (more or less traditional) design processes in which architectures are

used as structure in the solution space according to architecturally significant requirement and qualities, and according to developmental requirements.

Parallel Design By identifying many instances of the POE process pattern through the start state, many problem solvers can solve problems in parallel. Architectures that admit such concurrent problem solving, and that might be discovered in a sequential prelude to such a process, are evident in many areas. One of timely relevance, given their current popularity, is open source projects, such as the GNU Classpath project whose goal is to provide

> 'a 100% free, clean room implementation of the standard class libraries for compilers and runtime environments for the java programming language.'

Concurrent development may place demands on the resources shared throughout the concurrent design. For instance, during problem and solution validation should access to the various stake-holders be co-ordinated, or should individual problem and solution finders be allowed access to them as and when necessary?

Communications between those involved in parallel development is an issue on the GNU Classpath project, and it is not surprising that explicit guidance exists to i) partition work through a task list and a mailing list, ii) contact the central maintainer of the project when the developer wishes to make certain non-trivial additions to the project, iii) global announcements whenever important bugs are fixed or when 'nice new functionality' is introduced.

Fractal-like Design Fractal structures are self-similar in the sense that the whole structure resembles the parts it is made of [15]. Another way to look at it is that the whole is generated from simple building blocks, with complexity emerging through recursion of the simple generators. By analogy, problem solving under the POE process pattern is structurally simple and admits recursive application in that problem solving activity can occur in the Problem Exploration and Solution Exploration parts of the POE process pattern. In the next section, we show how this leads to our notion of assurance-driven design.

3.1.2 The 'fractal' nature of validation

Given that problem and solution exploration can both be instances of the POE process pattern, let us consider the problems and solutions they work with.

As Problem Exploration leads to Problem Validation, it is 'complete' when we have delivered a problem description that satisfies the problem validator. That is, Problem Exploration is complete when we have found a

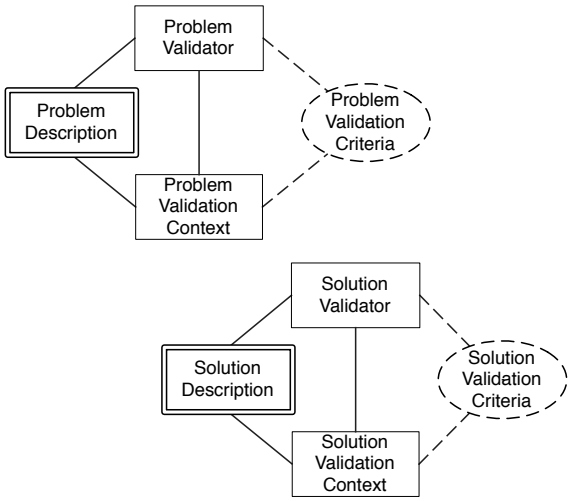


Figure 5. (a) Problem exploration as a problem validation problem, and (b) Solution exploration as a solution validation problem.

Problem Description that solves the following *problem validation problem*, illustrated in Figure 5(a):

Problem Validation Context, Problem Validator,
Problem Description ⊢ Problem Validation Requirements

The Problem Validation Context (PVC) is a description of the context in which the validation of the problem will be undertaken, and will need to be found as part of the fractal problem exploration phase of the outer problem exploration, as will the Problem Validation Requirements (PVR), i.e., the requirements that will need to be met for the problem to be validated. Note that the Problem Validator (PV) is an explicit domain in the context.

Symmetrically, solution exploration can be seen as complete when we have found a Solution Description that, when considered in the Solution Validation Context (SVC), satisfies the Solution Validation Requirements (SVR) of the Solution Validator (SV). As a POE problem, this is the *solution validation problem*, illustrated in Figure 5(b):

Solution Validation Context, Solution Validator,
Solution Description ⊢ Solution Validation Requirements

Although the fractal-like nature makes an easy clarity somewhat difficult, the view we have just presented fits well with practice. Indeed, discussions that lead to an agreed (i.e., validated) collection of use-cases [16] can be seen as a technique for producing a problem description that satisfies the problem validation problem. Moreover, discussions that lead to an agreed collection of acceptance tests can be seen as a solution description that satisfies the solution validation

problem. Requirements engineering, consisting of elicitation, analysis, specification can be seen as a technique for partial problem exploration; pattern oriented analysis and design is a technique for partial solution exploration.

In terms of Section 2.2, each validation is a projection of a whole development tree's adequacy argument into the stake-holder space determined by the validation context and validation requirements and, for a properly engineered solution, considering each of the adequacy problems is important.

4 Assurance-driven design in practice

The companion paper, [1], presented the assurance-driven design of a safety-critical subsystem of an aircraft. In this paper, we present a very different problem, that of the assurance-driven design of a research programme for The Open University. Whereas the aircraft example involved just a single stake-holder — the regulator for the system — this paper's example involved over 50 stake-holders as problem and solution validators. The project manager for the programme is the second author. For more information about that project, and a discussion of how POE was adopted in practice, please refer to [17, 18].

4.1 Notation

Because of the needs of the problem, we have augmented the traditional Gentzen-style notation to support better the separation of the problem and solution explorations, and to link validation problems to the justification of which they form a part. Figure 6 illustrates the differences. In the figure, we see the traditional transformations involving the problems labelled 'design problem' that will be familiar from Section 2.2. The triangular structures that extend the horizontal bar indicates the collection of validation problems associated with the step: by convention, when written on the right they are problem validation problem, when on the left they are solution validation problem[8].

4.2 Example

The Computing Department at The Open University is in the process of developing a new part-time MPhil programme to be delivered at a distance, supported by a blend of synchronous and asynchronous internet and web technologies — the *eMPhil*. The eMPhil is innovative in many ways in its adoption and use of emergent technology, like Second Life and Moodle, to support the core processes of the programme (the interested reader is directed to [19] for details).

[8]Because of the separation of problem and solution validations, never will problem and solution validations need to appear in the same step.

The eMPhil project team was faced with a complex socio-technical problem, that of the adoption and development of appropriate software systems and the definition of new processes and practices, of the design and delivery of induction and training activities for staff and students, and the institution of a framework for quality assurance, monitoring and continuous process improvement. The project also found itself with many stake-holder groups, those who would play problem validators, such as the Head of the Department of Computing, and solution validators, including Head of the Research Degrees Committee and Pro-Vice Chancellor for Research and Enterprise. The difficulties of managing the design and validation of the programme partially motivated the development of and application of the techniques described in this paper.

4.2.1 The problem

The eMPhil was required to meet a number of objectives:

- for the Head of the Department of Computing: to enhance and develop the department's provision to its graduate community; to increase the overall amount of research supervision that takes place within the department;
- for at-a-distance students: to make available technology for their support; to provide as a forum for that student community; to allow those unable to commit time for a PhD a research degree to study for;
- for academics wishing to promote research in their area: to create cohorts of research students on specific research themes and projects;
- for the Head of the Research Degrees Committee and Pro-Vice Chancellor Research and Enterprise: to support the development of research skills; to comply with university policy on research student induction and training; to comply with national standards [20].

The eMPhil core project team — the problem and solution finders — was composed of four academics, with the second author as project leader. The POE process pattern was used as described below to shape the project, with the techniques described earlier in the paper used to drive and manage its development and risks, as well as to identify the eMPhil project's needs for resource and communication.

4.2.2 The process

Figure 6 illustrates the early design steps taken by the development team towards a solution to the problem. The first transformation (bottom of the figure) achieves a first characterization of the problem context and requirement (from

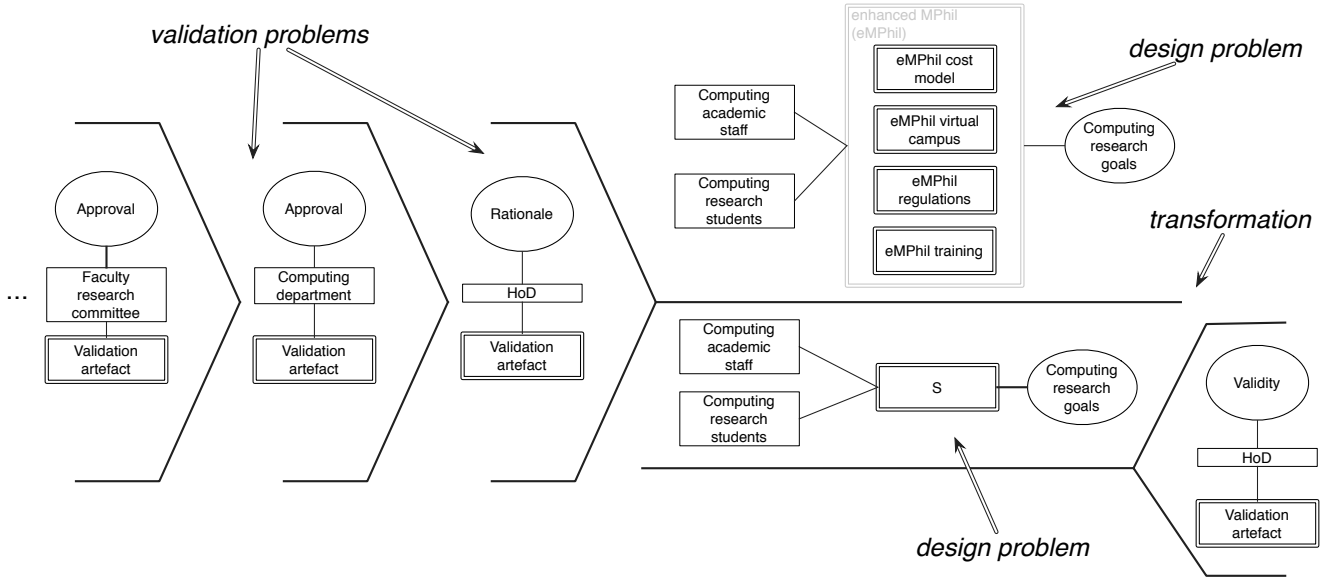


Figure 6. Early design steps

an empty conclusion problem—the start of all POE Design explorations), with a problem validator identified as the Head of the Computing Department (HoD). The HoD set the strategic goals which constituted the initial requirement description, and led to the inclusion of Computing academic staff and research students as a first approximation for the problem context. This initial problem exploration was coupled with the solution of the associated validation problem, consisting in making sure that the HoD's strategic intent was understood correctly by the problem solver.

The next transformation (top of Figure 6) captures an early solution exploration activity, in which a candidate solution architecture is starting to emerge, that of a new research degree, an MPhil, to be delivered in part-time mode at a distance. Note the validation problems to the left. The initial buy-in for the new degree was sought from the HoD, as the person in charge of releasing resources for the project, and with whom the rationale for the proposed solution was discussed. Approval from the HoD then triggered a comprehensive approval process throughout the organisation, reflecting its power structure (each validation problem concerns stake-holders at different management levels). Downside risks at these point were very high, with assurance taking precedence and greatly influencing the design.

Subsequent design alternated between further problem and solution explorations, and related validation, as illustrated in Figure 8, which provides a snapshot of the design tree after the first 10 months' development from the developer team perspective, assuming both problem and solution finder roles. Transformations labelled A and B at the bottom of the figure correspond to the early steps we have just described. From the initial solution architecture, a number of sub-problems were then identified (transformation C) each addressing complementary aspects of the solution, such as the design of its technological infrastructure, a related cost model, a programme of user induction and training, a system of monitoring and evaluation, etc. Each sub-problem was then taken forward through further transformations and related stake-holder validation, with the design problems at the top representing either solved sub-problems or open problems in the process of being addressed.

Note that some of the steps introduce sub-problems, which lead to branches in the design tree. This happens when a number of solution components have been identified, each to be designed, together with their architectural relations and mutual constraints. The POE transformation which generates them, called *solution expansion*, generates appropriate sub-problems for each to-be-designed component, based on such architectural knowledge. Each sub-problem then becomes the root of a (sub-) design tree.

4.2.3 Fractal Problem Validation

As introduced in Section 3, validation problems are problems too, and so their solution can be arrived at through a problem solving process, and hence (should they have solution) solvable in our POE framework, i.e., they should be treated as any other problem, with problem finder exploring the problem, obtaining problem validation, and so on. In the augmented notation, the obvious place for the validation problem development is in the extension to the horizontal bar; see Figure 7. However, such diagrams quickly become unwieldy, and a more pragmatic approach was nec-

essary in which the validation problem development was placed in a separate file, with indicators (again, Figure 7, on the left) for the state of each validation problem and hyperlinks used for easy access to the embedded validation problem development. It became apparent that the indicators formed a useful proxy for developmental risk associated with an unsolved validation problem, that risk being associated with the progress made in the solution of the main problem as opposed to the validation problem. We used a simple semaphore system for the risk indicators. Given the lack of tool support, this was deemed a simple, but useful tool from a project management perspective; of course, a more accurate estimation of risk would have required more sophisticated tools. Figure 7 gives an intuition of the meaning of the risk indicator: to the right is the equivalent fully expanded validation problem.

4.3 Early evaluation

The experience on the project so far has been very encouraging, and has clearly indicated that the conceptual tools offered by POE, including assurance-driven design were able to cater for all relevant aspects of the project. Design forests provide a powerful summary of the development, with all critical decision points clearly exposed, and all sub-problems (solved and unresolved) and their relation clearly identified. The risk indicators, despite their lack of sophistication, were considered very useful in signposting critical parts of the development. The notation was also considered an effective communication tool: its relative simplicity and abstraction allowed even non technical stake-holders, like senior managers in academic and academic-related units, to grasp the essence of the project with very little explanation required. The inclusion of validation problems within the development tree, with the explicit acknowledgement of all relevant stake-holders, was also considered a valuable tool to gauge the criticality of each design step, as well as to focus attention on the aspects of the problem of significance to each stake-holder. For instance, the high criticality of initial approval process is evidenced by the large set of validation problems in the early stages of development, in which the validation effort largely outweighed the effort to produce an initial outline for the solution, but greatly reduced the risk of the programme not to be deemed viable by management later on.

5 Related Work

Work on assurance cases is found in the area of dependability, from which two main structured notations for expressing safety cases have emerged. One is the goal-structuring notation (GSN) [21], a graphical argumentation notation which allows the representation of individual elements of a safety argument and their relations. Elements include: goals (used to represent requirements and claims about the system), context (used to represent the rationale for the approach and the context in which goals are stated), solutions (used to give evidence of goal satisfaction) and strategies (the approach used to identify sub-goals). The other, is Adelard's Claim-Argument-Evidence (ASCAD) approach [22], which is based on Toulmins work on argumentation and includes: claims (same as Toulmin's claims), evidence (same as Toulmin's grounds) and argument (combination of Toulmin's warrant and backing). More recently, Habli and Kelly [23] have also suggested ways in which product and process evidence could be combined in GSN assurance cases. One of the difficulties of these approaches is that they were not conceived to provide an integrated approach to safety development and, by and large, use artifacts and processes which may parallel but not integrate with software development. Instead, a main aim of our work is to allow for the efficient co-design of both software and assurance case based on artefacts and processes which are common to both. Some very recent work by Strunk and Knight [24] proposes Assurance Based Development (ABD) in which a safety-critical system and its assurance case are developed in parallel through the combined used of Problem Frames [7] and GSN. Although this work shares some of our goals, it is still rather preliminary for a meaningful comparison with POE.

A more mature process model, which shares something with POE, is the CHOAS model and lifecycle of Raccoon [25]. In this model fractal invocations of problem solving processes are combined to provide a rich model of software development, which is then used as the basis for a critical review of software engineering, of its processes and its practices. Raccoon's review leads to the conclusion that neither separately nor together do top-down or bottom-up developments tell the whole story; hence, a 'middle out theory' is proposed, based on the work that developers do to link high level project issues to, essentially, code structures. It is an attractive theory, and we wish to explore the ways in which fractal invocation in POE and assurance-driven design satisfy the criteria laid down for it.

6 Discussion and conclusion

The POE notion of problem suggests a separation of context, requirement and solution, with explicit descriptions of what is given, what is required and what is designed. This improves the traceability of artefacts and their relation, as well as exposing the assumptions upon which they are based to scrutiny and validation. That all descriptions are generated through problem transformation forces the inclusion of an explicit justification that such assumptions are realistic

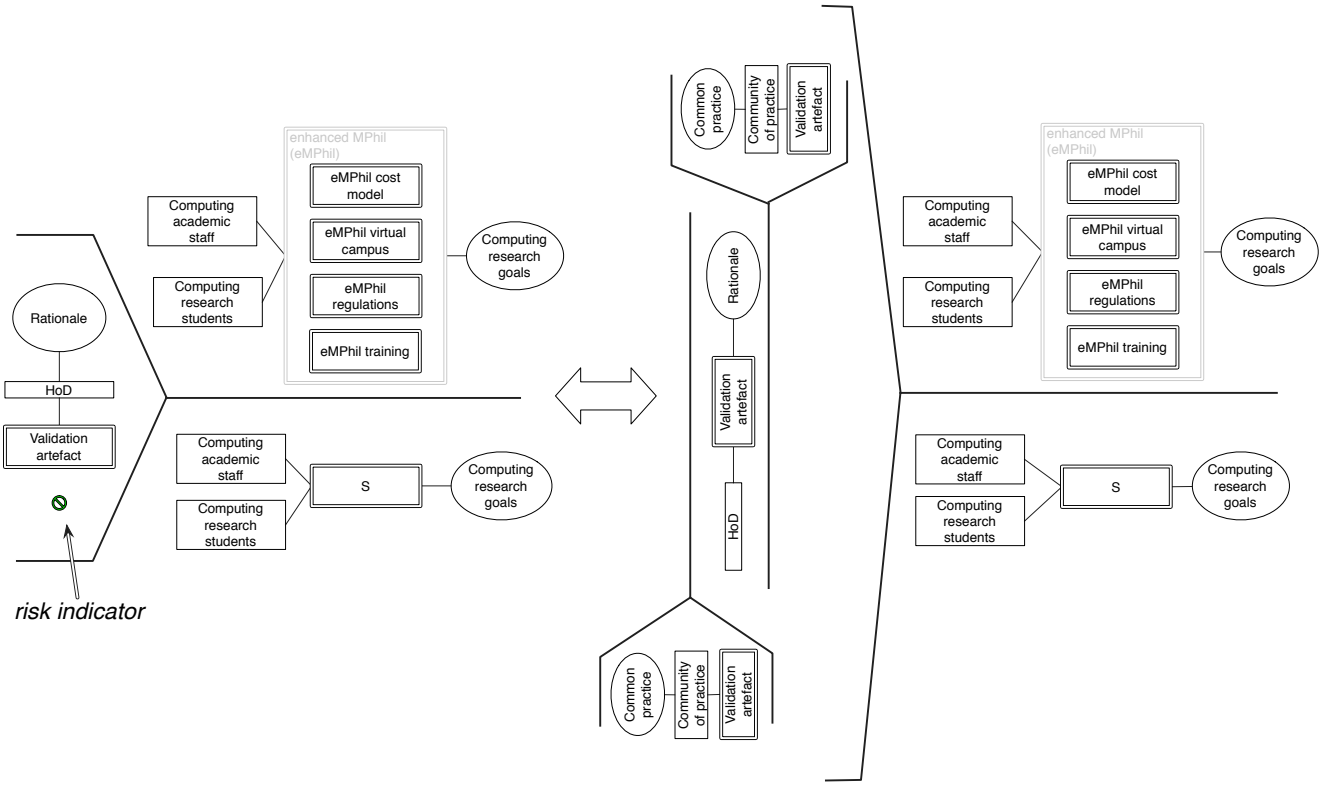


Figure 7. Risk indicator and its 'fractal' validation problem

and reasonable. In particular, requirements are justified as valid, are fully traceable with respect to the designed system (and *vice versa*), and evidence of their satisfaction is provided by the adequacy argument of a completed POE development tree.

We have shown (a) how (partial) problem and solution validation can be used to manage developmental risk and (b) how an assurance arguments can be constructed alongside the development of a product. Developmental risks arise from tentative transformation which are not completely justified: in such cases concerns can be stated as suspended justification obligations to be discharged later on in the process. This adds the flexibility of trying out solutions, while still retaining the rigour of development and clearly identifying points where backtracking may occur.

Although other approaches provide a focus on an assurance argument, the possibility of having the assurance argument *drive* development is an option that appears unique to ADD and POE.

Finally, POE defines a clear formal structure in which the various elements of evidence fit, that is whether they are associated with the distinguished parts of a development problem or the justifications of the transformation applied to solve it. This provides a fundamental clarification of the type of evidence provided and reasoning applied. Moreover, that the form of justification is not prescribed under POE signifies that all required forms of reasoning can be accommodated, from deductive to judgemental, within a single development.

Acknowledgments

We acknowledge the financial support of IBM, under the Eclipse Innovation Grants, and of SE Validation Limited. Our thanks go to Derek Mannering at General Dynamics UK, Lucy Hunt at Getronics plc, Jens Jorgensen and Simon Tjell of Aarhus University, Andrés Silva of the University of Madrid, Colin Brain of SE Validation Ltd, Anthony Finkelstein at UCL (who suggested the discussion of "Have we done enough?"), and John Knight at UVA. L. B. S. Raccoon has read all of our work and made truly insightful comments. Finally, thanks go to our many colleagues in the Computing Department at The Open University, particularly Michael Jackson.

References

[1] Jon G. Hall and Lucia Rapanotti. Assurance-driven design. In *Proceedings of the Third International Conference on Software Engineering Advances (ICSEA 2008)*. Published

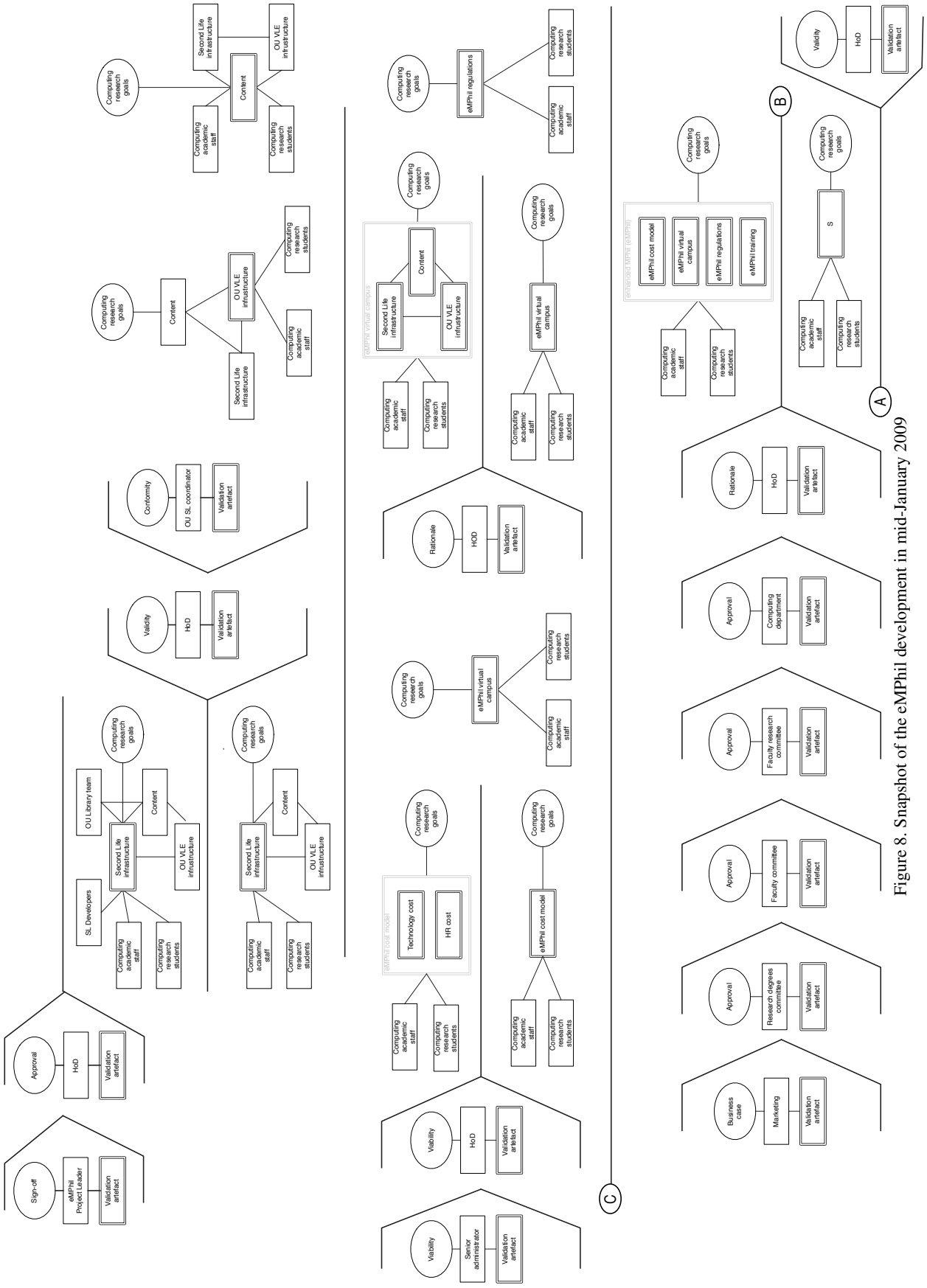


Figure 8. Snapshot of the eMPhil development in mid-January 2009

by the IEEE Computer Society, 2008. Also available as Open University Computing Department Technical Report #2007/15.

[2] Engineering Council of South Africa Standards and Procedures System Definition of Terms to Support the ECSA Standards and Procedures System.

[3] Jon G. Hall, Lucia Rapanotti, and Michael Jackson. Problem oriented software engineering: A design-theoretic framework for software engineering. In *Proceedings of 5th IEEE International Conference on Software Engineering and Formal Methods*, pages 15–24. IEEE Computer Society Press, 2007. doi:10.1109/SEFM.2007.29.

[4] Jon G. Hall, Lucia Rapanotti, and Michael Jackson. Problem-oriented software engineering: solving the package router control problem. *IEEE Trans. Software Eng.*, 2008. doi:10.1109/TSE.2007.70769.

[5] M. E. Szabo, editor. *Gentzen, G.: The Collected Papers of Gerhard Gentzen*. Amsterdam, Netherlands: North-Holland, 1969.

[6] Daniel Jackson. *Software Abstractions: Logic, Language, and Analysis*. MIT Press, Cambridge, MA, 2006.

[7] Michael A. Jackson. *Problem Frames: Analyzing and Structuring Software Development Problems*. Addison-Wesley Publishing Company, 1st edition, 2001.

[8] Derek Mannering, Jon G. Hall, and Lucia Rapanotti. Towards normal design for safety-critical systems. In M. B. Dwyer and A. Lopes, editors, *Proceedings of ETAPS Fundamental Approaches to Software Engineering (FASE) '07*, volume 4422 of *Lecture Notes in Computer Science*, pages 398–411. Springer Verlag Berlin Heidelberg, 2007.

[9] Jon G. Hall, Derek Mannering, and Lucia Rapanotti. Arguing safety with problem oriented software engineering. In *10th IEEE International Symposium on High Assurance System Engineering (HASE)*, Dallas, Texas, 2007.

[10] Jon G. Hall and Lucia Rapanotti. The discipline of natural design. In *Proceedings of the Design Research Society Conference 2008*. Design Research Society, 2008.

[11] Derek Mannering, Jon G. Hall, and Lucia Rapanotti. Safety process improvement with POSE & Alloy. In Francesca Saglietti and Norbert Oster, editors, *Computer Safety, Reliability and Security (SAFECOMP 2007)*, volume 4680 of *Lecture Notes in Computer Science*, pages 252–257, Nuremberg, Germany, September 2007. Springer-Verlag.

[12] OMG. Unified Modeling Language (UML), version 2.0. http://www.omg.org/technology/documents/formal/uml.htm. Last checked: May 2009.

[13] Donald Firesmith. *The Method Framework for Engineering System Architectures*. CRC Press, 2008.

[14] James P. Womack and Daniel T. Jones. *Lean Thinking – Banish Waste and Create Wealth in Your Corporation*. Simon and Schuster, 1996.

[15] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley-Blackwell, 2nd edition, 2003.

[16] Alistair Cockburn. *Writing Effective Use Cases*. Addison-Wesley, 2001.

[17] Lucia Rapanotti and Jon G. Hall. Designing an online part-time master of philosophy with problem oriented engineering. In *Proceedings of the Fourth International Conference on Internet and Web Applications and Services*, Venice, Italy, May 24-28 2009. IEEE Press.

[18] Lucia Rapanotti and Jon G. Hall. Problem oriented engineering in action: experience from the frontline of postgraduate education. Technical Report TR2008/16, The Open University, 2008.

[19] Lucia Rapanotti, Leonor M. Barroca, Maria Vargas-Vera, and Shailey Minocha. deepthink: a second life campus for part-time research students at a distance. Technical Report TR2009/1, The Open University, 2009.

[20] UK GRAD, Joint Skills Statement of Skills Training Requirements. http://www.grad.ac.uk/jss/ Last checked: May 2009.

[21] Tim Kelly. A systematic approach to safety case management. In *Proceedings SAE 2004 World Congress*, Detroit, US, 2004.

[22] R. Bloomfield, P. Bishop, C. Jones, and P. Froome. *ASCAD - Adelard Safety Case Development Manual*, 1998.

[23] I. Habli and T. Kelly. Achieving integrated process and product safety arguments. In *Proceedings of 15th Safety Critical Systems Symposium (SSS'07)*. Springer, 2007.

[24] Elisabeth A. Strunk and John C. Knight. The essential synthesis of problem frames and assurance cases. *Expert Systems*, 25(1):9–27, 2008.

[25] L. B. S. Raccoon. The Chaos model and the Chaos cycle. *SIGSOFT Softw. Eng. Notes*, 20(1):55–66, 1995.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
issn: 1942-2679

International Journal On Advances in Internet Technology
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

International Journal On Advances in Life Sciences
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

International Journal On Advances in Networks and Services
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

International Journal On Advances in Security
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

International Journal On Advances in Software
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
issn: 1942-2628

International Journal On Advances in Systems and Measurements
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

International Journal On Advances in Telecommunications
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601