Revista

Telecomunicações

Volume 11 Número 01 Maio de 2008 ISSN 1516-2338

Editorial

Tutoriais

Artigos Científicos

Engenharia Aplicada

Inatel
Instituto Nacional de Telecomunicações

Revista Científica/Periódica publicada pelo Inatel - Instituto Nacional de Telecomunicações

Diretor: Wander Wilson Chaves
Vice-diretor: Carlos Roberto dos Santos

Expediente

Assessoria de Comunicação & Marketing - ASCOM
e-mail: ascom@inatel.br

Diagramação
ASCOM
Setor de Editoração Eletrônica

Tiragem: 1.500 exemplares

Instituto Nacional de Telecomunicações
Av. João de Camargo, 510
Caixa Postal: 05
Santa Rita do Sapucaí - MG - BRASIL
CEP 37540-000
Tel: (35) 3471.9200 Fax: (35) 3471.9314
e-mail: inatel@inatel.br
http://www.inatel.br

EDITORIAL

Nas últimas duas décadas temos vivenciado um ritmo alucinante de transformação nas tecnologias, notadamente na área de telecomunicações. Dentre as possíveis causas para este fenômeno podemos citar a privatização do setor, o advento da Internet e a massificação da telefonia celular. É de se supor que a introdução da TV digital em nosso país ajude a intensificar ainda mais esta corrida tecnológica.

Como atores principais neste processo, penso que devemos nos sentir orgulhosos de nossas conquistas, mas ao mesmo tempo apreensivos sobre as consequências desta corrida desenfreada: mudanças sociais dramáticas vêm sendo observadas, desde a necessidade cada vez mais premente de atualização por parte dos profissionais da área, passando por novas formas de ensino, até a mudança de comportamento dos cidadãos comuns que usam estas tecnologias.

Embora nosso papel enquanto engenheiros, professores ou pesquisadores não seja especificamente pensar nestas questões, penso que uma reflexão sobre o tema é interessante, talvez apenas para dar um sentido maior ao que fazemos no nosso dia a dia.

Neste número temos dois tutoriais, um sobre a modulação MSK e outro sobre coompressão fractal de imagens. Na seção de artigos científicos, temos um trabalho sobre tráfego de redes. Finalizando este número, temos três artigos de engenharia aplicada: o primeiro aplicado a redes óticas, o segundo sobre o mercado de tecnológico, e o último sobre antenas rádio base. Este último é uma reedição do artigo originalmente publicado na edição de julho de 2007. Houve um engano de nossa parte, e publicamos a primeira versão do artigo, e não a versão final. Desta forma, para não prejudicar o autor, resolvemos publicar nesta edição a versão correta de seu trabalho.

Saudações,

Carlos Alberto Ynoguti
Editor

Contributions to the Understanding of the MSK Modulation

Dayan Adionel Guimarães

***Abstract*—This tutorial deals with key aspects of the MSK (Minimum Shift Keying) modulation, aiming at unveiling some of its hidden concepts. Signal generation and demodulation are analyzed in detail. Common questions concerning the study of the MSK modulation are addressed and answered, e.g. the similarities and differences among MSK, Sunde's FSK (Frequency Shift Keying) and SQPSK (Staggered Quaternary Phase-Shift Keying) or OQPSK (Offset QPSK); the relation among the modulating data stream, its differentially-decoded version, the frequency shifts and the phase shifts of the modulated signal, and the MSK signal-space representation.**

Index Terms—MSK, FSK, SQPSK and OQPSK modulations.

***Resumo*—Este tutorial trata de aspectos chave sobre a modulação MSK, objetivando revelar alguns dos seus conceitos muitas vezes não revelados explicitamente. A geração e a demodulação do sinal MSK são analisadas em detalhe. Ao longo do trabalho procura-se responder a algumas questões intrigantes relacionadas com, por exemplo, as similaridades e diferenças entre as modulações MSK, FSK (de Sunde) e SQPSK ou OQPSK e a relação entre a seqüência moduladora, sua versão decodificada diferencialmente, os desvios de freqüência e de fase do sinal modulado e a representação do sinal MSK no espaço euclidiano.**

Palavras chave— Modulações MSK, FSK, SQPSK e OQPSK.

I. INTRODUCTION

The Minimum Shift Keying (MSK) modulation, also known as "fast FSK" [1], was first considered during the early 60s and 70s [2]-[4], and its characteristics have gained the attention of the scientific community during the subsequent decades.

MSK modulation has features such as constant envelope, compact spectrum and good error performance, which are all desirable in many digital communication systems. Its utilization goes from the Global System for Mobile Communication (GSM), in which a Gaussian-filtered MSK (GMSK) modulation is employed, to micro-satellite communications, positioning and navigation systems, hybrid optical/wireless communication systems, deep space communications and, more recently, to the Blue Ray disc technology [5], only to mention a few examples.

Like many recently rediscovered technologies developed several years, or even decades ago, the MSK modulation seems to be one more idea whose time has come.

Manuscript received on December 9, 2006.

D. A. Guimarães (dayan@inatel.br) is with INATEL - Instituto Nacional de Telecomunicações. Av. João de Camargo, 510 - Santa Rita do Sapucaí - MG - Brazil - 37540-000.

Although covered in many papers and good books on Digital Communications, some of the concepts of this modulation are hidden or difficult to understand, representing opportunities for alternative approaches, like the one adopted in this tutorial. This approach is intended to help everyone who wants to have an understanding about the MSK modulation, especially the practicing engineers and the first-level graduate students in Telecommunications. It addresses some key questions about the MSK modulation, such as:

1 – To which extent the MSK modulation can be regarded as a special case of the conventional Sunde's [6] [7, p. 381] FSK (Frequency Shift Keying) modulation?
2 – To which extent the MSK modulation can be detected in the same way as the Sunde's FSK modulation?
3 – To which extent the MSK modulation can be regarded as a special case of the SQPSK or OQPSK (Staggered or Offset QPSK) modulation?
4 – To which extent the frequency and phase shifts of an MSK signal are related to the modulating data sequence?
5 – To which extent the phase shifts of an MSK signal can be related to the phase transition diagram on its signal-space representation?

The remaining of this work is organized as follows: Section II addresses some fundamental concepts about the signal-space representation, the complex representation of signals and systems, and the minimum separation between tones in an orthogonal FSK signaling. Section III is devoted to the analysis of the signal construction from the signal-space expansion and the complex representation approaches. The MSK spectral content, receiver structure and system performance are also analyzed in Section III. Further attributes and uses of the MSK modulation are summarized in Section IV, and Section V addresses the answers to the questions highlighted above, concluding the work.

II. BASIC CONCEPTS

In this section the reader are invited to revisit some fundamental concepts about signal-space representation and complex representation of signals and systems. Although applicable to the study of digital communications in general, these two concepts are essential for the study at hand, and will give us insight on different forms of MSK signal generation and detection. Additionally, the minimum tone separation for coherent detection of orthogonal FSK is analyzed, aiming at

justifying the term *minimum* in the name of the MSK modulation.

A. Signal-space representation

The signal-space representation is constructed on the basis of linear combination theory, and it is very analogous to the vector algebra theory. Let us define an N-dimensional Euclidian space spanned by N orthogonal axes. Let us also define a set of orthogonal vectors $\{\boldsymbol{\phi}_j\}$, $j = 1, 2, \ldots, N$, normalized in the sense that they have unit length. These vectors are said to be *orthonormal* and to form an *orthornormal basis*.

Any vector $\mathbf{v}_i$, $i = 1, 2, \ldots, M$ in the Euclidian space can be generated through the linear combination

$$\mathbf{v}_i = \sum_{j=1}^{N} v_{ij}\boldsymbol{\phi}_j \tag{1}$$

where the coefficients v_{ij} correspond to the projection of the i-th vector on the j-th base vector. Their values can be determined by the dot product (or inner product) between $\mathbf{v}_i$ and $\boldsymbol{\phi}_j$, that is

$$v_{ij} = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\phi}_j \tag{2}$$

where the superscript T denotes matrix transposition, $\mathbf{v}_i = [v_{i1}\ v_{i2}\ \ldots\ v_{iN}]^{\mathrm{T}}$ and $\boldsymbol{\phi}_j$ is also an N-dimensional vector with a 1 in the j-th position and zeros otherwise, that is $\boldsymbol{\phi}_j = [0\ 1\ 0\ \ldots\ 0]^{\mathrm{T}}$ for $j = 2$ as an example.

Figure 1 illustrates these concepts for a two-dimensional ($N = 2$) Euclidian space and for two vectors ($M = 2$). The axes were labeled in a way to resemble the orthonormal base-vectors.

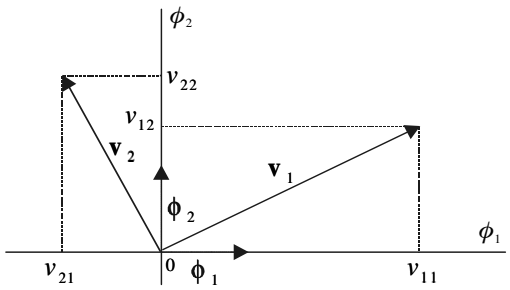


Figure 1. Vector-space representation for $M = 2$ and $N = 2$.

In a similar way, one can use the Euclidian space to represent coefficients that, in a linear combination, will give rise to signals instead of vectors. Then we have the signals

$$s_i(t) = \sum_{j=1}^{N} s_{ij}\phi_j(t)\ ,\ i = 1, 2, \ldots, M \tag{3}$$

where, now, the set $\{\phi_j(t)\}$ comprises N orthonormal *base-functions*, one function being orthogonal to each other and having unit energy, that is:

$$\int_0^T \phi_i(t)\phi_j(t)dt = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{4}$$

The set of functions $\{\phi_j(t)\}$ are also said to be *orthonormal* and to form an *orthornormal basis*.

Through Figure 1 it can be seen that the value of a coefficient is proportional to a measure of the orthogonality between the analyzed vector and the corresponding base-vector: the greater the orthogonality, the lesser the value of the coefficient. By analogy to the vector algebra, we can determine the values of the coefficients in (3) through a measure of orthogonality between the analyzed waveform and the corresponding base-function, which leads intuitively to

$$s_{ij} = \int_0^T s_i(t)\phi_j(t)dt,\ \begin{cases} i = 1, 2, \ldots, M \\ j = 1, 2, \ldots, N \end{cases} \tag{5}$$

In fact (5) has a formal mathematical justification, which can be obtained by operating generically with (3) and (4):

$$\begin{aligned} \int_0^T x(t)\,y(t)dt &= \int_0^T \sum_{j=1}^{N} x_j\phi_j(t) \sum_{k=1}^{N} y_k\phi_k(t)\,dt \\ &= \sum_{j=1}^{N}\sum_{k=1}^{N} x_j y_k \int_0^T \phi_j(t)\phi_k(t)dt \\ &= \sum_{j=1}^{N} x_j y_j = \mathbf{x}^{\mathrm{T}}\mathbf{y} \end{aligned} \tag{6}$$

Expression (6) states that the correlation in time domain has the inner product as its equivalent in the vector domain.

We are now ready to define the signal-space representation: since knowing the set of coefficients and base-functions is as good as to know the waveform signals themselves, we can also represent signals in a Euclidian space. In this representation we use points instead of vectors, to avoid polluting unnecessarily the graph. This kind of plot is also called *signal constellation*. Figure 2 shows a two-dimensional signal-space used to represent the signals $s_1(t)$ and $s_2(t)$ through the corresponding *signal vectors* $\mathbf{s}_1$ and $\mathbf{s}_2$.

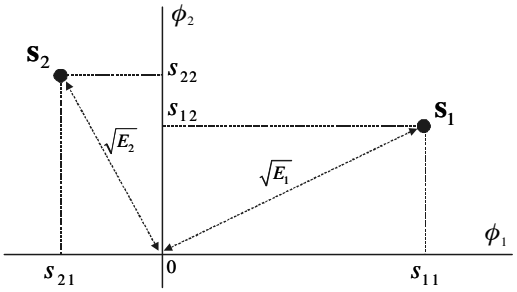


Figure 2. Signal-space representation for $M = 2$ and $N = 2$.

As can be noticed from Figure 2, the norm of a signal vector, that is, the length of this vector can be determined in the light of equation (6) by:

$$\sqrt{s_{i1}^2 + s_{i2}^2} = \sqrt{\mathbf{s}_i^T\mathbf{s}_i} = \sqrt{\int_0^T s_i^2(t)dt} = \sqrt{E_i} \tag{7}$$

Generally speaking, the distance from any signal vector to the origin of the coordinates is equal to the square root of the corresponding signal energy:

$$E_i = \int_0^T s_i^2(t)dt = \mathbf{s}_i^T \mathbf{s}_i = \sum_{j=1}^N s_{ij}^2 = \|\mathbf{s}_i\|^2 \quad (8)$$

As a complementary result, the squared Euclidian distance between two signal vectors is obtained through

$$d_{ik}^2 = \|\mathbf{s}_i - \mathbf{s}_k\|^2 = \sum_{j=1}^N \left(s_{ij} - s_{kj}\right)^2 = \int_0^T \left[s_i(t) - s_k(t)\right]^2 dt \quad (9)$$

The concepts just described will be used later for the understanding of a particular form for the MSK signal generation and detection.

B. Complex representation of signals and systems

We start by reviewing the concept of Hilbert transform. Following [7] and [8], let $g(t)$ be a signal with Fourier transform $G(f)$. The Hilbert transform of $g(t)$ and the corresponding inverse transform are defined respectively by

$$\hat{g}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(\tau)}{t-\tau} d\tau = \int_{-\infty}^{\infty} g(\tau) \frac{1}{\pi(t-\tau)} d\tau \quad (10)$$

and

$$g(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\hat{g}(\tau)}{t-\tau} d\tau \quad (11)$$

In (10) we can identify that the Hilbert transform of $g(t)$ is the convolution between $g(t)$ and the function $1/\pi t$.

By recalling that a convolution in the time domain corresponds to a multiplication in the frequency domain, and by using the Fourier transform pair

$$\frac{1}{\pi t} \rightleftharpoons -j\,\mathrm{sgn}(f)\,, \quad (12)$$

where sgn(f) is the *sign* function or *signum* function defined by

$$\mathrm{sgn}(f) = \begin{cases} 1, & f > 0 \\ 0, & f = 0 \\ -1, & f < 0 \end{cases}, \quad (13)$$

then we can write:

$$\hat{G}(f) = -j\,\mathrm{sgn}(f) G(f) \quad (14)$$

Analyzing (14) we can see that the Hilbert transform of $g(t)$ corresponds to a phase shift of –90º for the positive frequencies of $G(f)$ and +90º for the its negative frequencies.

Let us now make use of another definition: the *analytic signal* or *pre-envelope* of $g(t)$:

$$g_+(t) = g(t) + j\hat{g}(t) \quad (15)$$

from where, using (14) and the definition of the *signum* function given in (13), we can obtain

$$G_+(f) = G(f) + \mathrm{sgn}(f)G(f) = \begin{cases} 2G(f), & f > 0 \\ G(0), & f = 0 \\ 0, & f < 0 \end{cases} \quad (16)$$

Now, consider a band-pass signal $g(t)$ whose bandwidth is essentially confined in $2W$ Hz and is small compared to the its carrier frequency f_c. According to (16), the analytic spectrum $G_+(f)$ is centered about f_c and contains only positive frequency components. Then, using the frequency-shifting property of the Fourier transform we can write:

$$g_+(t) = \tilde{g}(t)\exp(j2\pi f_c t) \quad (17)$$

where $\tilde{g}(t)$ is called the *complex envelope* of the signal $g(t)$ and it is clearly a low-pass signal.

Since $g_+(t)$ is a band-pass signal, we can determine the low-pass signal $\tilde{g}(t)$ through a frequency translation of $g_+(t)$ back to about $f = 0$. Using again the frequency-shifting property of the Fourier transform we can write

$$\begin{aligned} \tilde{g}(t) &= g_+(t)\exp(-j2\pi f_c t) \\ &= \left[g(t) + j\hat{g}(t)\right]\exp(-j2\pi f_c t) \end{aligned} \quad (18)$$

or, equivalently,

$$g(t) + j\hat{g}(t) = \tilde{g}(t)\exp(j2\pi f_c t) \quad (19)$$

Since the signal $g(t)$ is the real part of the left side of the expression above, we can obtain a very useful representation:

$$g(t) = \mathrm{Re}\left[\tilde{g}(t)\exp(j2\pi f_c t)\right] \quad (20)$$

Generally speaking, $\tilde{g}(t)$ can be a complex quantity, which can be expressed in the Cartesian form by:

$$\tilde{g}(t) = g_I(t) + jg_Q(t) \quad (21)$$

where the subscripts I and Q stand for *in-phase* and *quadrature*. Then, by substituting (21) in (20) we have, after some simplifications:

$$g(t) = g_I(t)\cos(2\pi f_c t) - g_Q(t)\sin(2\pi f_c t) \quad (22)$$

Both $g_I(t)$ and $g_Q(t)$ are low-pass signals and are called the in-phase component and the quadrature component of the signal $g(t)$, respectively. This is why we call $\tilde{g}(t)$ the *equivalent low-pass* version of the band-pass signal $g(t)$. This result will be used later on in this tutorial to describe a particular form for the MSK signal generation and detection.

Rewriting expression (21) in its polar form we have:

$$\tilde{g}(t) = a(t)\exp[j\theta(t)]\,, \quad (23)$$

from where, using (20), we can obtain

$$\begin{aligned} g(t) &= \mathrm{Re}\left[\tilde{g}(t)\exp(j2\pi f_c t)\right] \\ &= \mathrm{Re}\left\{a(t)\exp[j\theta(t)]\exp(j2\pi f_c t)\right\} \\ &= a(t)\cos\left[2\pi f_c t + \theta(t)\right] \end{aligned} \tag{24}$$

In (24), $a(t) = |\tilde{g}(t)|$ is the envelope of the band-pass signal $g(t)$, or the amplitude modulated component of $g(t)$, and $\theta(t)$ is its phase, or the phase-modulated component of $g(t)$. This result will also be used later as a means for understanding the MSK signal generation.

Taking the Fourier transform of $g(t)$ we know to obtain its frequency content. If $g(t)$ is a voltage signal, then the magnitude of its Fourier transform will result in a, say, "voltage spectral density". Then, using (20) we get:

$$G(f) = \Im\{g(t)\} = \int_{-\infty}^{\infty}\left\{\mathrm{Re}\left[\tilde{g}(t)\,\mathrm{e}^{j2\pi f_c t}\right]\right\}\mathrm{e}^{-j2\pi f t}\,dt \tag{25}$$

Using the identity $\mathrm{Re}[C] = ½[C + C^*]$ in (25), and applying the Fourier transform properties $x^*(t) \leftrightharpoons X^*(-f)$ and $x(t)\exp(j2\pi f_c t) \leftrightharpoons X(f - f_c)$, we obtain:

$$\begin{aligned} G(f) &= \frac{1}{2}\int_{-\infty}^{\infty}\left[\tilde{g}(t)\,\mathrm{e}^{j2\pi f_c t} + \tilde{g}^*(t)\,\mathrm{e}^{-j2\pi f_c t}\right]\mathrm{e}^{-j2\pi f t}\,dt \\ &= \frac{1}{2}\left[\tilde{G}\left(f - f_c\right) + \tilde{G}^*\left(-f - f_c\right)\right] \end{aligned} \tag{26}$$

If $g(t)$ is a sample function of an stationary random process $G(t)$, it has infinity energy and, hence, its Fourier transform does not exist. In this case the spectral content of $G(t)$ is given by its *power spectral density* (PSD), which is obtained from the Fourier transform of the auto-correlation function $R_G(\tau)$ of the random process, as follows [8, p. 67]:

$$S(f) = \int_{-\infty}^{\infty} R_G(\tau)\exp(-j2\pi f\tau)\,d\tau \tag{27}$$

The PSD for a stationary random process can also be estimated through [7, p. 51]:

$$S(f) = \lim_{\Pi\to\infty}\frac{1}{\Pi}E\left[\left|G_\Pi(f)\right|^2\right] \tag{28}$$

where $G_\Pi(f)$ is the Fourier transform obtained from the sample process $g_\Pi(t)$, which is $g(t)$ truncated from $-\Pi/2$ to $\Pi/2$. The function $|G_\Pi(f)|^2$ is called the *energy spectral density* of the energy signal $g_\Pi(t)$. If the signal is deterministic, (28) can also be used, without the expectation operation [11, p. 31].

However, if the Fourier transform $G(f)$ exists and is exact, according to which was stated before equation (25) $S(f)$ can be simply determined by the squared-modulus of $G(f)$, that is,

$$S(f) = \left|G(f)\right|^2 = \frac{1}{4}\left[\left|\tilde{G}\left(f - f_c\right) + \tilde{G}^*\left(-f - f_c\right)\right|^2\right] \tag{29}$$

Using a simplified notation, and the fact that $|C|^2 = CC^*$, we can rewrite (31) as follows:

$$\begin{aligned} S(f) &= \frac{1}{4}\left[\left|X(f) + X^*(-f)\right|^2\right] \\ &= \frac{1}{4}\left\{\left[X(f) + X^*(-f)\right]\left[X^*(f) + X(-f)\right]\right\} \\ &= \frac{1}{4}\begin{bmatrix} X(f)X^*(f) + X^*(-f)X(-f) \\ +X(f)X(-f) + X^*(-f)X^*(f) \end{bmatrix} \\ &= \frac{1}{4}\begin{bmatrix} |X(f)|^2 + |X(-f)|^2 \\ +X(f)X(-f) + X^*(-f)X^*(f) \end{bmatrix} \end{aligned} \tag{30}$$

By recognizing that $X(f)$ and $X(-f)$ are band-limited, band-pass signals, the products $X(f)X(-f)$ and $X^*(f)X^*(-f)$ in (30) vanish to zero. Going back to the normal notation, we get:

$$\begin{aligned} S(f) &= \frac{1}{4}\left[\left|\tilde{G}\left(f - f_c\right)\right|^2 + \left|\tilde{G}\left(-f - f_c\right)\right|^2\right] \\ &= \frac{1}{4}\left[S_B\left(f - f_c\right) + S_B\left(-f - f_c\right)\right] \end{aligned} \tag{31}$$

Equation (31) states that we can easily obtain the power spectral density $S(f)$ of a band-pass signal by translating the power spectral density $S_B(f)$ of the low-pass equivalent, and its mirror image, to the frequencies f_c and $-f_c$, respectively, and multiplying the result by ¼.

C. Minimum frequency separation for coherent detection

It may be somewhat obvious for some readers that MSK is a form of orthogonal frequency shift keying modulation, but our aim in this subsection is to give reasons for the term *minimum* in the Minimum Shift Keying nomenclature.

To be coherently orthogonal in the signaling interval T, two cosine functions with different frequencies must satisfy

$$\int_0^T \cos\left(2\pi f_1 t\right)\cos\left(2\pi f_2 t\right)dt = 0 \tag{32}$$

Using the identity $\cos\alpha\cdot\cos\beta = ½[\cos(\alpha - \beta) + \cos(\alpha + \beta)]$ in the expression above we obtain:

$$\tfrac{1}{2}\int_0^T \cos\left[2\pi\left(f_1 - f_2\right)t\right]dt + \tfrac{1}{2}\int_0^T \cos\left[2\pi\left(f_1 + f_2\right)t\right]dt = 0 \tag{33}$$

from where, after some manipulations, we get:

$$\frac{\sin\left[2\pi\left(f_1 - f_2\right)T\right]}{4\pi\left(f_1 - f_2\right)} + \frac{\sin\left[2\pi\left(f_1 + f_2\right)T\right]}{4\pi\left(f_1 + f_2\right)} = 0 \tag{34}$$

Since for practical purposes the sum $f_1 + f_2 >> 1$, the second term in the left-hand side of (34) is approximately zero, which results in

$$\begin{aligned} &\sin\left[2\pi\left(f_1 - f_2\right)T\right] = 0 \\ &\Rightarrow \left(f_1 - f_2\right) = \frac{k}{2T},\ k \text{ inteiro} \end{aligned} \tag{35}$$

Then, the minimum frequency separation between tones for an orthogonal FSK with coherent detection is

$$(f_1 - f_2) = \frac{1}{2T}, \quad (36)$$

which justifies the name Minimum Shift Keying for the MSK modulation.

III. MSK Signal Generation And Detection

In this section, the MSK signal generation and detection are analyzed in detail, based first on a complex representation approach, and then, based on a signal-space representation approach. The MSK power spectral density is also considered. However, first we introduce the basics about the MSK and the conventional binary FSK modulation. At the end of the section these modulations are revisited, aiming at establishing their similarities and differences from the design of the transmitter and the receiver perspective.

The receiver structures and performances considered in this section assume that the system operates on an Additive White Gaussian Noise (AWGN) channel.

A. MSK and conventional binary FSK

A continuous-phase, frequency-shift keying (CPFSK) signal can be described as a phase-modulated signal using (24), as shown by:

$$s(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left[2\pi f_c t + \theta(t)\right] \quad (37)$$

where E_b is the average energy per bit and T_b is the bit duration.

The time derivative of the phase evolution $\theta(t)$ in (37) gives rise to the CPFSK instantaneous angular frequency shift. Then, in a given bit interval $\theta(t)$ increases or decreases linearly, depending on the desired transmitted tone, as described by:

$$\theta(t) = \theta(0) \pm \frac{\pi h}{T_b} t, \; 0 \le t \le T_b \quad (38)$$

where $\theta(0)$ accounts for the accumulated phase history until instant $t = 0$ and h is a measure of the frequency deviation. If $h = 1$ we have the conventional form of binary FSK modulation, also know as Sunde's FSK [6] [7, p. 381], in which the tone separation is obtained from (38) as $1/T_b$ Hz.

Generalizing (38), at any time instant the phase evolution can be determined by

$$\theta(t) = \theta(0) + \frac{\pi h}{T_b} \int_0^t b(t)dt \quad (39)$$

where $b(t) \in \{\pm 1\}$ is the waveform related to the information sequence, such that a -1 represents a bit 0 and a $+1$ represents a bit 1.

The modulated signal described by (37) and (39) can be generated by means of a continuous-phase VCO (voltage controlled oscillator) having $b(t)$ as its input, and configured with center frequency f_c Hz and gain $h/(2T_b)$ Hz/volt.

Example 1 - Suppose we want to transmit the information sequence [1 0 1 0 0 0 1 1]. Following (39), with $h = 1$, we shall have the phase evolution illustrated in Figure 3. Also in Figure 3 are plotted the waveform $b(t)$ and the resultant FSK modulated signal $s(t)$ for $f_c = 2/T_b$ Hz. The resultant tones are then at frequencies $f_1 = 5/(2T_b)$ Hz and $f_2 = 3/(2T_b)$ Hz.

A careful look at Figure 3 shows that phase transitions from one bit to the next lead to the same value, using modulo 2π algebra (a phase transition of $+\pi$ is equal to a phase transition of $-\pi$, modulo 2π). Then, the receiver is not able to explore any phase information in the conventional Sunde's FSK modulation.

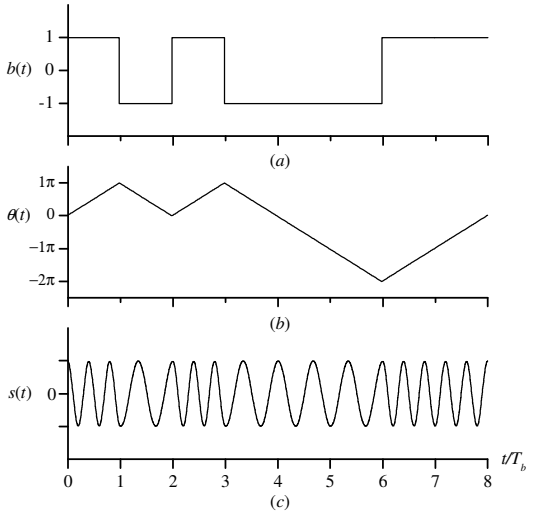


Figure 3. Information sequence (*a*), phase evolution (*b*) and modulated signal (*c*) for the Sunde's FSK modulation.

Now, let us make $h = ½$ in (39). In this case we have the minimum tone separation of $1/(2T_b)$ Hz, and, through (37), we shall generate an MSK signal.

Example 2 - Suppose again that we want to transmit the information sequence [1 0 1 0 0 0 1 1]. According to (39), with $h = ½$, we shall have the phase evolution shown in Figure 4. The waveform $b(t)$ and the resultant MSK modulated signal $s(t)$ for $f_c = 1/T_b$ Hz are also plotted. The resultant tones are at frequencies $f_1 = 5/(4T_b)$ Hz and $f_2 = 3/(4T_b)$ Hz.

As can be noticed from Figure 4, phase transitions from a bit to the next one lead to different values, modulo 2π. Then, it is possible to explore some phase information with the MSK modulation. This is indeed the motivation for the use of MSK: the receiver can explore phase transitions in order to benefit from this additional information to improve performance.

B. MSK signal generation and detection from the complex representation approach

The generation of $s(t)$ through (37) and (39), though straightforward from the implementation point of view, brings no or little insight on how the receiver can be constructed in order to explore the phase information in the modulated signal. Then we are forced to obtain alternative mathematical models for representing $s(t)$.

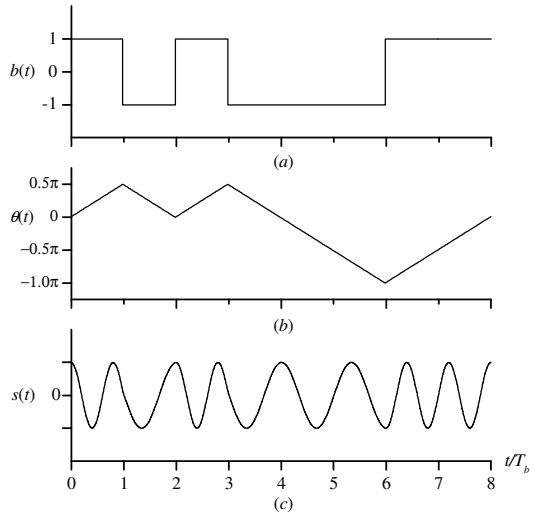


Figure 4. Information sequence (*a*), phase evolution (*b*) and modulated signal (*c*) for the MSK modulation.

To simplify matters, consider initially only the first bit interval. Using $h = ½$ in (38) and the identity $\cos(\alpha \pm \beta) = \cos\alpha\cdot\cos\beta \mp \sin\alpha\cdot\sin\beta$, we can rewrite (37) as follows:

$$\begin{aligned} s(t) &= \sqrt{\frac{2E_b}{T_b}}\cos\left[\theta(0) \pm \frac{\pi}{2T_b}t\right]\cos\left(2\pi f_c t\right) \\ &\quad -\sqrt{\frac{2E_b}{T_b}}\sin\left[\theta(0) \pm \frac{\pi}{2T_b}t\right]\sin\left(2\pi f_c t\right)\ ,\ 0 \le t \le T_b \\ &= s_I(t)\cos\left(2\pi f_c t\right) - s_Q(t)\sin\left(2\pi f_c t\right) \end{aligned} \tag{40}$$

Making use of (22) and applying again the identity $\cos(\alpha \pm \beta) = \cos\alpha\cdot\cos\beta \mp \sin\alpha\cdot\sin\beta$ to the in-phase component of $s(t)$, and, without loss of generality, assuming $\theta(0) = 0$, we get

$$\begin{aligned} s_I(t) &= \sqrt{\frac{2E_b}{T_b}}\cos\left[\pm\frac{\pi}{2T_b}t\right] \\ &= \pm\sqrt{\frac{2E_b}{T_b}}\cos\left(\frac{\pi}{2T_b}t\right) \end{aligned},\ -T_b \le t \le T_b \tag{41}$$

Since $\theta(0) = 0$, before $t = 0$ the phase evolution was a positive or negative slope going towards zero, depending on the previous bit. Then, the result in (41) is an increasing cosine function from $–T_b$ to 0. Thus, $s_I(t)$ can be interpreted as a half-cycle cosine function from the whole interval $(–T_b, T_b]$.

Similarly, the quadrature component of $s(t)$ can be written as follows:

$$s_Q(t) = \pm\sqrt{\frac{2E_b}{T_b}}\sin\left(\frac{\pi}{2T_b}t\right),\ 0 \le t \le 2T_b \tag{42}$$

where we have made use of $\theta(0) = 0$ and of the identity $\sin(\alpha \pm \beta) = \sin\alpha\cdot\cos\beta \pm \cos\alpha\cdot\sin\beta$. We have also made use of the relation $\cos[\theta(0)] = \cos[\theta(T_b) \mp \pi/2] = \pm\sin[\theta(T_b)] = \pm 1$.

Since $\theta(T_b) = \pm\pi/2$, depending on the information bit during the interval $(0, T_b]$, we shall have $\sin[\theta(t)]$ going towards zero during the interval T_b to $2T_b$, regardless the information bit during this interval. Thus, $s_Q(t)$ can be viewed as a half-cycle sine function from the whole interval $(0, 2T_b]$, the polarity of which depending on the information bit during the interval $[0, T_b)$.

Using the results (41) and (42) in (40), we obtain:

$$\begin{aligned} s(t) &= \pm\sqrt{\frac{2E_b}{T_b}}\cos\left(\frac{\pi}{2T_b}t\right)\cos\left(2\pi f_c t\right) \\ &\quad \mp\sqrt{\frac{2E_b}{T_b}}\sin\left(\frac{\pi}{2T_b}t\right)\sin\left(2\pi f_c t\right) \end{aligned} \tag{43}$$

where the polarity of both terms in a given bit interval are not necessarily the same.

Following [4, p. 18], we can rewrite (43) as:

$$\begin{aligned} s(t) &= \sqrt{\frac{2}{T_b}}a_I(t)\cos\left(\frac{\pi}{2T_b}t\right)\cos\left(2\pi f_c t\right) \\ &\quad -\sqrt{\frac{2}{T_b}}a_Q(t)\sin\left(\frac{\pi}{2T_b}t\right)\sin\left(2\pi f_c t\right) \end{aligned} \tag{44}$$

where we have defined $a_I(t)$ and $a_Q(t)$ as random sequences of rectangular pulses with amplitudes $\pm\sqrt{E_b}$ and duration $2T_b$ seconds. These sequences are associated to the polarities of the half-cycle cosine and sine functions as follows: if $a_I(t)$ is positive, $s_I(t)$ follows the function $\cos\{[\pi/(2T_b)]t\}$; if $a_I(t)$ is negative, $s_I(t)$ corresponds to $–\cos\{[\pi/(2T_b)]t\}$. The same happens with $s_Q(t)$: if $a_Q(t)$ is positive, $s_Q(t)$ follows the function $\sin\{[\pi/(2T_b)]t\}$; if $a_Q(t)$ is negative, $s_Q(t)$ corresponds to $–\sin\{[\pi/(2T_b)]t\}$.

From the above discussion we can conclude that, depending on the information bit to be transmitted, the in-phase and quadrature components of $s(t)$ can change their polarities each $2T_b$ seconds, and that the half-cycle cosine and sine functions are offset from each other by T_b seconds. However, we are not still able to easily obtain the information sequence responsible for generating a given sequence of polarities. This would demand us to come back to the general analysis presented in Section III-A, specifically to equation (37), thus making difficult the visualization of the implementation issues for the MSK modem.

Then, for the time being we assume a given sequence of pulses for $s_I(t)$ and $s_Q(t)$, and later we determine the information sequence based on the analysis of this assumption. A general rule will arise from this analysis.

Example 3 – In a 8-bit interval, let $s_I(t)$ and $s_Q(t)$ assume the sequence of half-cycle cosine and sine pulses shown in Figure 5. For reference, in this figure the functions $\cos\{[\pi/(2T_b)]t\}$ and $–\sin\{[\pi/(2T_b)]t\}$ are also plotted, in dashed lines, and are given the polarities of the waveforms $a_I(t)$ and $a_Q(t)$. Combining the waveforms in Figure 5 according to (40), we get the results in Figure 6. In this figure the waveforms $s_I(t)$ and $s_Q(t)$ are also plotted, in dashed lines. The carrier

frequency in this example is $f_c = 1/T_b$ Hz. The resultant tones are then at frequencies $f_1 = 5/(4T_b)$ Hz and $f_2 = 3/(4T_b)$ Hz.

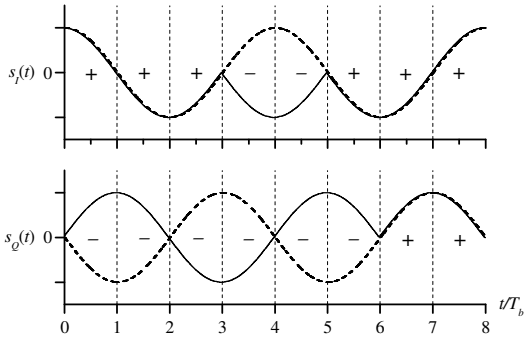


Figure 5. Generation of the MSK signal: base-band in-phase and quadrature components.

Observing the modulated signal $s(t)$ in Figure 6 we can notice that, if a bit 1 is associated to the tone of greater frequency, the corresponding modulating sequence should be **d** = [1 1 1 0 0 1 0 0]. Let us now define a new sequence **i** in which the exclusive-or (XOR) operation between a given bit and the previous one results in a bit of the sequence **d**. This new sequence is **i** = [1 0 1 0 0 0 1 1 1]. Sequence **d** can be seen as a *differentially decoded* version of **i**. Additionally, suppose that the sequence **i** is parallelized to form the sequences of odd and even symbols of duration $2T_b$, $\mathbf{i}_o$ = [1 1 0 1 1] and $\mathbf{i}_e$ = [0 0 0 1], respectively. Now, suppose that each symbol of these sequences is converted to $\pm\sqrt{E_b}$. The great achievement here is that these new parallel sequences, if they are off-set to each other T_b seconds, are exactly the waveforms $a_I(t)$ and $a_Q(t)$.

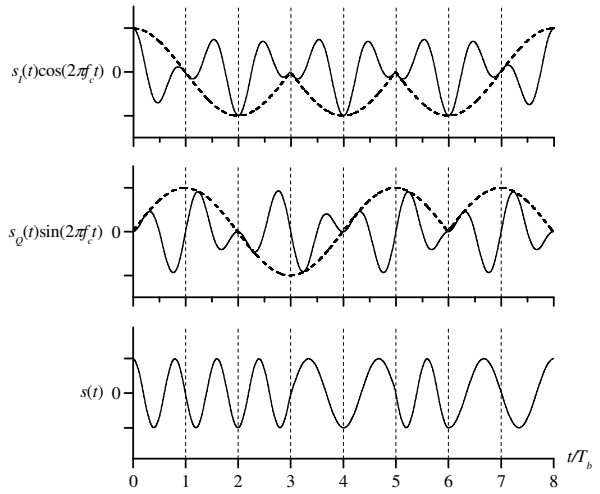


Figure 6. Generation of the MSK signal: modulated in-phase and quadrature components and the resultant MSK signal.

We then conclude that the MSK signal can be generated through (44), where the waveforms $a_I(t)$ and $a_Q(t)$ are the serial-to-parallel (S/P) converted version of the information sequence, with bit 1 converted to $+\sqrt{E_b}$ and bit 0 converted to $-\sqrt{E_b}$. Additionally, the sequence $a_Q(t)$ has to be offset T_b seconds from $a_I(t)$, before they multiply the corresponding remaining terms in (44). Figure 7 illustrates the structure of the MSK modulator constructed according to complex representation approach just described.

The MSK signal just analyzed can also be generated by means of a VCO configured with center frequency f_c Hz and gain $1/(4T_b)$ Hz/volt. However, since the frequency shifts in the modulated signal do not directly correspond to the information sequence, the input of the VCO must be the differentially decoded version of this information sequence, converted to $\{\pm 1\}$.

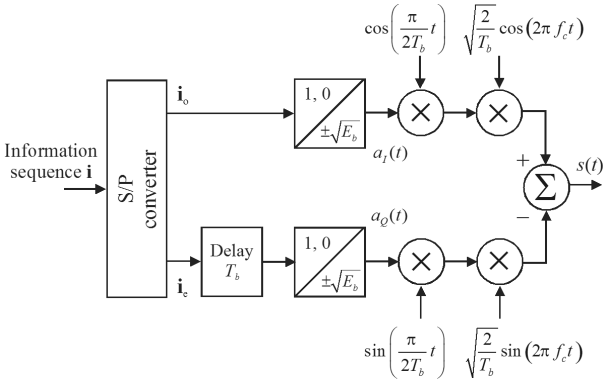


Figure 7. MSK modulator constructed according to the complex representation approach.

Some authors claim that the MSK modulation is a special form of OQPSK (or SQPSK) modulation where the pulse shaping are half-cycle cosine and sine functions, instead of the rectangular shaping functions used in OQPSK. However, in despite of being true, this statement must be carefully interpreted. From (43) we can see that, in fact, the shapes of the pulses that modulate the quadrature carriers are half-cycle cosine and sine functions. Nevertheless, they are not a simple reshaping of the waveforms $a_I(t)$ and $a_Q(t)$. Before modulating the quadrature carriers, $a_I(t)$ and $a_Q(t)$ are modified by the polarities of the waveforms $\cos\{[\pi/(2T_b)]t\}$ and $\sin\{[\pi/(2T_b)]t\}$ in each $2T_b$ interval. But we can make a small modification in the above structure to implement the MSK modulation in the same way we implement an OQPSK modulator, the unique difference being the shape of the pulses that modulate the quadrature carriers. We just have to use the modulus $|\cos\{[\pi/(2T_b)]t\}|$ and $|\sin\{[\pi/(2T_b)]t\}|$ in (44). The resultant structure is shown in Figure 8. In this figure we have used additional simplifications to make the MSK modulator structure closer to a more practical one: the quadrature carriers were generated from a single oscillator and the pulse-shaping functions were implemented via low-pass filters with identical impulse responses given by

$$h(t) = \begin{cases} \sin\left(\dfrac{\pi}{2T_b}t\right), & 0 \le t \le 2T_b \\ 0, \text{ otherwise} \end{cases} \tag{45}$$

The inputs to these filters are PAM (Pulse Amplitude Modulation) sequences having very short durations (approximating unit impulses) and amplitudes of $+\sqrt{E_b}$.

We can recall that at the beginning of Section III-B we have made the assumption that $\theta(0) = 0$. This assumption was adopted only to facilitate the mathematical description of the MSK modulation. In fact, from an implementation perspective, any initial phase is allowed for the quadrature carriers. However, regardless of this initial phase, the designer must only guarantee the correct phase alignment among the quadrature carriers, the pulse shaping functions and the sequences $a_I(t)$ and $a_Q(t)$.

In the light of the similarities between the MSK and OQPSK modulations, we are now able to understand possible structures for the MSK demodulator. We know that a conventional QPSK modulator can be interpreted as two BPSK (Binary Phase-Shift Keying) modulators, each of them making use of one of the two quadrature carriers. Then, the QPSK demodulator can be implemented as two independent BPSK demodulators. The decisions made by each of these demodulators are parallel-to-serial (P/S) converted to form the estimate of the transmitted bit sequence. The OQPSK demodulator follows the same rule, with the difference that one of the estimated parallel sequences is offset T_b seconds from the other. Then, before P/S conversion these sequences must be aligned in time.

If we use $|\cos\{[\pi/(2T_b)]t\}|$ and $|\sin\{[\pi/(2T_b)]t\}|$ in the modulator of Figure 7 or adopt a more practical solution, as shown in Figure 8, we must use $|\cos\{[\pi/(2T_b)]t\}|$ and $|\sin\{[\pi/(2T_b)]t\}|$ in the demodulator in Figure 9. In this case the correspondence between the MSK and the OQPSK demodulators exists, the unique difference being the shape of the pulses that multiply the quadrature carriers. If we do not apply the modulus operation at the modulator we have its complex representation approach realization and, in this case, we just have to use the original $\cos\{[\pi/(2T_b)]t\}$ and $\sin\{[\pi/(2T_b)]t\}$ in the demodulator of Figure 9.

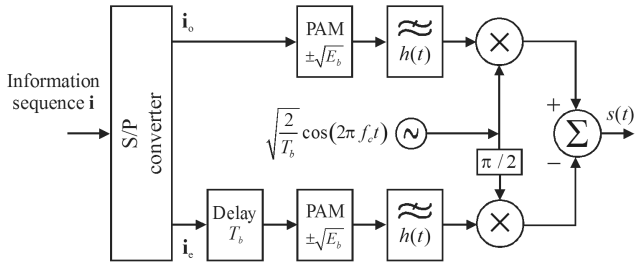


Figure 8. A more practical MSK modulator constructed according to the equivalence with the OQPSK modulation.

The received signal in Figure 9 is coherently correlated, in one arm of the receiver, with the result of the multiplication between the in-phase carrier and the shaping function $\cos\{[\pi/(2T_b)]t\}$. In the other arm, the received signal is correlated with the result of the multiplication between the quadrature carrier and the shaping function $-\sin\{[\pi/(2T_b)]t\}$. These correlations are made in a $2T_b$ seconds interval, reflecting the duration of the half-cycle cosine and sine functions, and are time-aligned with these functions. The estimated sequences $\hat{\mathbf{i}}_e$ and $\hat{\mathbf{i}}_o$ are then time-aligned and P/S converted to form the estimate of the transmitted sequence, $\hat{\mathbf{i}}$.

If, for some reason, it is necessary to represent a bit 1 in the sequence **d** by the tone of lower frequency, the only thing we have to do is to invert the minus signal in the summation block in Figure 7 or Figure 8, and invert the minus signal in the bottom multiplier block in Figure 9.

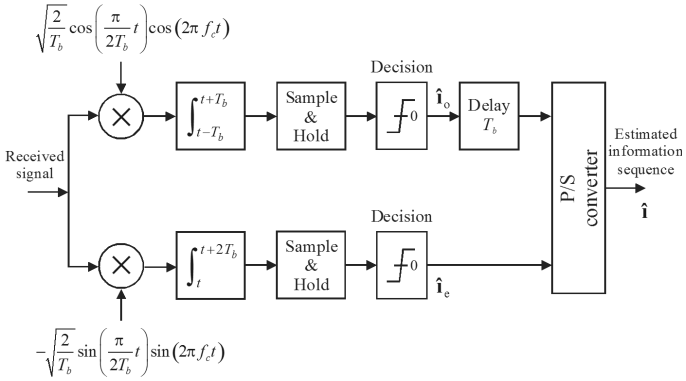


Figure 9. MSK demodulator constructed according to the complex representation approach.

C. MSK signal generation and detection from the signal-space representation approach

We are now able to determine the orthonormal base-functions responsible for generating the MSK signal. Recalling that we are talking about a binary orthogonal signaling, the base-functions can be directly obtained from (44) as follows:

$$\phi_1(t) = \sqrt{\frac{2}{T_b}} \cos\left(\frac{\pi}{2T_b} t\right) \cos\left(2\pi f_c t\right) \tag{46}$$

$$\phi_2(t) = \sqrt{\frac{2}{T_b}} \sin\left(\frac{\pi}{2T_b} t\right) \sin\left(2\pi f_c t\right) \tag{47}$$

These base-functions, differently from what is stated in [7, p. 390], are defined for any interval, not only from 0 to T_b.

Comparing (44) with (46) and (47) we readily see that the MSK signal vectors are determined by the amplitudes of the waveforms $a_I(t)$ and $a_Q(t)$ defined in (44), in each bit interval:

$$\mathbf{s}_i = \begin{bmatrix} s_{i1} \\ s_{i2} \end{bmatrix} = \begin{bmatrix} \pm\sqrt{E_b} \\ \pm\sqrt{E_b} \end{bmatrix}, \; i = 1,\, 2,\, 3,\, 4 \tag{48}$$

Then, as shown in Figure 10, the signal-space diagram for the MSK modulation comprises four signal vectors, despite of MSK be a binary modulation. The mapping between these vectors and the information bits is determined via the differentially decoded version of the information bits. The following example is meant to clarify these statements.

Example 4 – Let the sequence of signal vector polarities be [+ −], [+ −], [+ −], [− −], [− −], [+ −], [+ +] and [+ +], generated on a bit-by-bit basis. In this sequence, the polarities on the left refer to s_{i1}, and those on the right refer to s_{i2}. These

polarities are the same as those considered in Figure 5 and, as we already know from Example 3, they are associated to the information sequence **i** = [1 0 1 0 0 0 1 1 1] and to its differentially decoded version **d** = [1 1 1 0 0 1 0 0]. From this example it is possible to draw the mapping between the signal vectors and the differentially decoded version of the information bits, as shown in Table I.

TABLE I – MAPPING BETWEEN THE MSK SIGNAL VECTORS AND THE DIFFERENTIALLY DECODED VERSION OF THE INFORMATION BITS

i	Bits	Signal vector coordinates	
		s_{i1}	s_{i2}
1	0	$+\sqrt{E_b}$	$+\sqrt{E_b}$
2	1	$+\sqrt{E_b}$	$-\sqrt{E_b}$
3	0	$-\sqrt{E_b}$	$-\sqrt{E_b}$
4	1	$-\sqrt{E_b}$	$+\sqrt{E_b}$

Since MSK is a continuous phase modulation, no abrupt phase transition occurs when a symbol changes. The circumference in Figure 10 illustrates this smooth phase transitions between any pair of symbols. They can be observed in a *x*-*y* plot, with $s_I(t)$ applied to the *x*-axis and $s_Q(t)$ applied to the *y*-axis (see Figure 5).

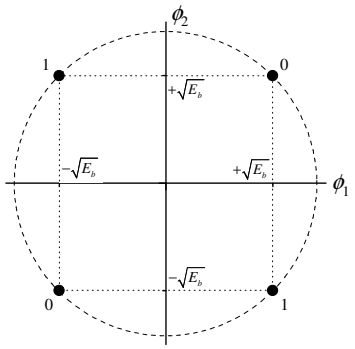


Figure 10. MSK constellation.

Observing (46) and (47) we see that the base-functions $\phi_1(t)$ and $\phi_2(t)$ correspond to the modulation of the quadrature carriers by the waveforms $\cos\{[\pi/(2T_b)]t\}$ and $\sin\{[\pi/(2T_b)]t\}$, respectively. Comparing (46) and (47) with (44), we see that the base-function $\phi_1(t)$ are multiplied by $a_I(t)$, the base-function $\phi_2(t)$ is multiplied by $a_Q(t)$, and the results are added to form the MSK signal $s(t)$. Figure 11 illustrates the generation of the MSK signal form this signal-space representation approach. The signal vector polarities associated to the waveforms $a_I(t)$ and $a_Q(t)$ are the same as those used in Example 4.

As we did with the complex representation approach, now we shall construct the modulator structure based on the signal-space representation. As a matter of fact, if we group together the two upper mixers and group together the two lower mixers in Figure 7 this job is already done. But we shall manipulate the base-function expressions to get an alternative structure. First, let us expand $\phi_1(t)$ using the identity $\cos\alpha\cdot\cos\beta = \frac{1}{2}[\cos(\alpha-\beta)+\cos(\alpha+\beta)]$:

$$\begin{aligned}\phi_1(t) &= \sqrt{\frac{2}{T_b}}\cos\left(\frac{\pi}{2T_b}t\right)\cos\left(2\pi f_c t\right)\\ &= \sqrt{\frac{1}{2T_b}}\cos\left(2\pi f_2 t\right)+\sqrt{\frac{1}{2T_b}}\cos\left(2\pi f_1 t\right)\end{aligned} \tag{49}$$

Now, let us expand $\phi_2(t)$ using the identity $\sin\alpha\cdot\sin\beta = \frac{1}{2}[\cos(\alpha-\beta)-\cos(\alpha+\beta)]$:

$$\begin{aligned}\phi_2(t) &= \sqrt{\frac{2}{T_b}}\sin\left(\frac{\pi}{2T_b}t\right)\sin\left(2\pi f_c t\right)\\ &= \sqrt{\frac{1}{2T_b}}\cos\left(2\pi f_2 t\right)-\sqrt{\frac{1}{2T_b}}\cos\left(2\pi f_1 t\right)\end{aligned} \tag{50}$$

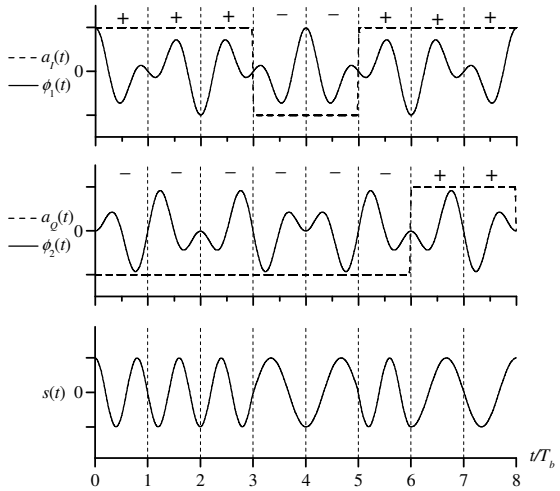


Figure 11. Generation of the MSK signal: base-functions, coefficients and the resultant MSK signal.

Figure 12 shows the MSK modulator constructed according to the interpretation of expressions (49) and (50). The two cosine functions are multiplied to generate the tones with frequencies f_1 and f_2, according to (49). Each of these tones is selected through the band-pass filters shown in this figure, and the results are combined according to (49) and (50) to generate the base-functions. Finally, these base-functions are multiplied by the corresponding waveforms $a_I(t)$ and $a_Q(t)$ and the results are added-up to form the MSK signal. The approach at hand can also consider the demodulator shown in Figure 9, where we readily identify the use of the base-functions $\phi_1(t)$ and $\phi_2(t)$ feeding the correlators.

We can see that, operating in different ways with the mathematical model of the MSK signal, it is possible to construct different, but equivalent structures. More structures would be possible if an alternative mathematical model were adopted. These comments are also valid to the construction of the MSK demodulator. In [11, pp. 299-307] the reader can find several forms for the implementation of an MSK modem, along with different approaches on its construction.

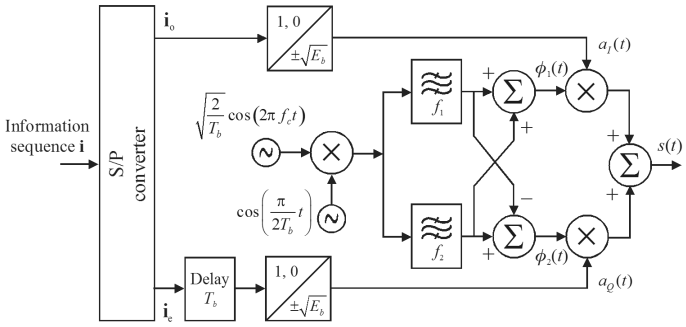


Figure 12. MSK modulator constructed according to the signal-space representation approach.

D. Bit error probability for the MSK modulation

We can see through Figure 8 and equation (44) that the modulator transmits two independent sequences using two quadrature carriers, and through Figure 9 we can see that the demodulator detects these sequences independently. Consequently, we can state that the modulator can be interpreted as formed by two independent BPSK-like modulators and that the demodulator can be interpreted as formed by two independent BPSK-like demodulators. The difference to the conventional BPSK modulator and demodulator is the presence of half-cycle sine and cosine pulse-shaping functions. The energy per symbol for each of these two component BPSK modulators is easily found to be

$$\begin{aligned}\xi &= \frac{2E_b}{T_b}\int_0^{2T_b}\cos^2\left(\frac{\pi}{2T_b}t\right)\cos^2\left(2\pi f_c t\right)dt \\ &= \frac{2E_b}{T_b}\int_0^{2T_b}\sin^2\left(\frac{\pi}{2T_b}t\right)\sin^2\left(2\pi f_c t\right)dt = E_b\end{aligned} \tag{51}$$

where, for simplification purposes, we have adopted the carrier frequency f_c as an integer multiple of $1/(2T_b)$. The energy per MSK symbol is the sum of the symbol energies in the quadrature modulated carriers, that is $E = 2E_b$, a value that can also be obtained from the constellation in Figure 10.

Confusions may arise here: the duration of one bit is of course T_b seconds, and we must make the bit decisions in a bit-by-bit basis. But the phase information at the MSK receiver is explored in $2T_b$ seconds intervals, so that the effective energy collected by this receiver corresponds to observations made during intervals of $2T_b$ seconds.

From the above discussion we can conclude that the bit error probability for the MSK modulation on the AWGN channel, considering equally-likely bits, can be determined by the average of the bit error probabilities for the two component BPSK detectors [8, p. 271], which results in:

$$P_b = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right), \tag{52}$$

where N_0 is the AWGN power spectral density and $\operatorname{erfc}(u)$ is the complimentary error function of the argument. This result shows that the performance of the MSK modulation is the same as the performance of the BPSK and QPSK modulations, and is 3 dB more energy-efficient than the conventional BFSK with coherent detection [7, p. 418].

E. MSK signal generation and detection from a conventional Sunde's FSK approach

Suppose now that we aim at generating an MSK signal using the conventional FSK approach, but with the minimum tone separation $(f_1 - f_2) = 1/(2T_b)$ Hz. The modulator would appear like in Figure 13. This form of FSK signal generation guarantees phase continuity only if the tone separation is a multiple of $1/T_b$ and the carrier frequency is a multiple of $1/(2T_b)$. Then, the modulated signal in Figure 13 will show phase discontinuities, which does not correspond to an MSK signal. MSK and binary FSK signals are the same if they are generated according to (37) and (39), using $h = ½$.

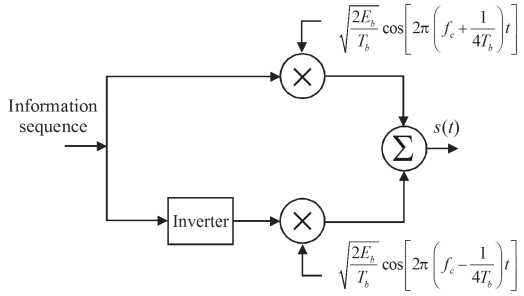


Figure 13. A try for generating an MSK signal from the conventional binary FSK implementation approach.

Now, following [9], suppose that we want to detect an MSK signal using a conventional coherent FSK demodulator. We would be tempted to think that it is just necessary to correlate the received signal with base-functions formed by the cosine tones with frequencies f_1 and f_2, during T_b seconds intervals, and that the decision would be made in favor of the greatest correlator output. However, the phase continuity and phase dependency imposed by the MSK signal construction do not permit the use of the above approach. This is illustrated in Figure 14, were we have plotted an MSK signal and the cosine base-functions with frequencies f_1 and f_2 separated by $1/(2T_b)$ Hz. Observe that, in several intervals, there are no phase coherence between the modulated signal and the base-functions with the same frequency, a behavior that would lead to detection errors.

Let us elaborate a little bit more on this issue. From Figure 14 we can see that when no phase coherence occurs, the MSK signal is at 180º out of phase from the corresponding base-function. Then, by comparing the magnitudes of the correlators outputs we are still able to make correct decisions. But we cannot forget that, unless the MSK signal is generated directly from the realization of (37) and (39) with $h = ½$, the estimated bits would correspond to a differentially decoded version of the information bits. To get the estimates of the information bits we have to apply the inverse operation on the estimated bits through the exclusive OR (XOR) between a given bit and the previous XOR result (see Example 3 and the corresponding comments). However, this operation can lead to the opposite decisions, since a differentially decoded 1 can result from the information sequence 01 or 10, and a differentially decoded 0 can result from the information

sequence 00 or 11. Inserting a differential coder at the transmitter input and a differential decoder at the receiver output easily solves this ambiguity problem.

Finally, we shall have the transmitter and receiver structures shown in Figure 15.

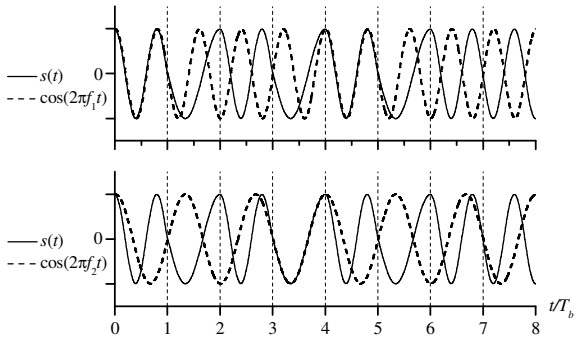


Figure 14. MSK signal $s(t)$, the $\cos(2\pi f_1 t)$ and $\cos(2\pi f_2 t)$.

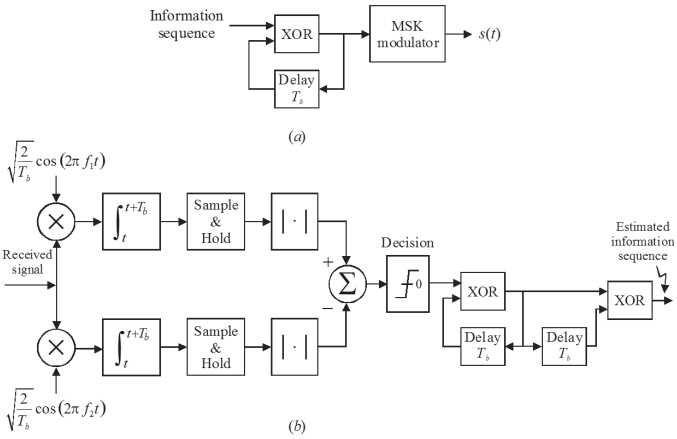


Figure 15. MSK modulator with conventional FSK detection: modified MSK transmitter (*a*) and detection via a modified coherent binary FSK receiver (*b*).

Since the receiver in Figure 15 is not exploring any phase information, we expect a worse performance as compared to the one provided by the appropriate MSK receiver. Furthermore, although the channel noise is Gaussian, the noise in the decision variable is not. Then, the analytical process for obtaining an expression for the bit error probability P_e for the receiver under investigation is quite involved and is beyond the scope of this work. Nevertheless, a numerical calculation of P_e was made and a simulation of the system in Figure 15 was carried out. Both results agreed and showed that the performance lies in between a coherently detected and a non-coherently detected binary FSK, as shown in Figure 16, and is approximately 3.05 dB worse than the P_e obtained with the MSK receiver. This is an attractive result, since the P_e curves for the coherent and the non-coherent FSK differs asymptotically in about 1 dB [7 p. 418], and we are using a transmitted signal that has the most compact spectrum among the coherent and orthogonal CPFSK modulations [9].

Using a more practical and simplified approach, the MSK modulator in Figure 15-*a* can be replaced by a VCO, eliminating the need for the three differential circuits used by the complete system. This alternative was also simulated and the BER was the same as the one obtained with the simulation of the complete system depicted by Figure 15.

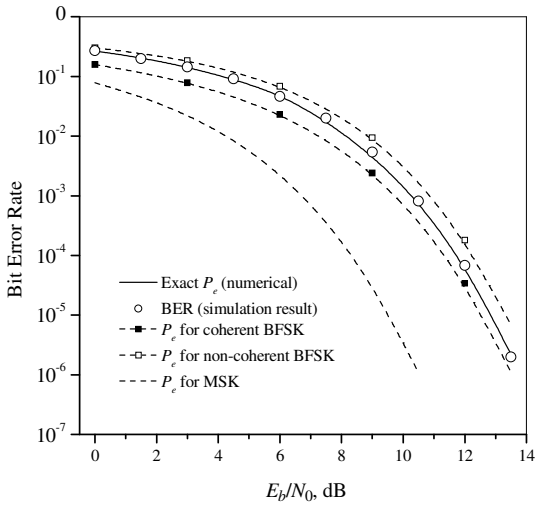


Figure 16. Performance results for MSK, coherent and non-coherent BFSK and for the system depicted in Figure 15. The channel is AWGN [9].

F. Power spectral density of the MSK signal

We saw in Section II that in order to obtain the PSD of a modulated signal, we can determine the PSD of its complex envelope representation and, using (31), convert the result to the desired PSD. According to (22), the MSK signal can be written as:

$$s(t) = s_I(t)\cos\left(2\pi f_c t\right) - s_Q(t)\sin\left(2\pi f_c t\right) \tag{53}$$

from where the complex envelope given by (21) is

$$\tilde{s}(t) = s_I(t) + j s_Q(t) \tag{54}$$

For the MSK modulation, the low-pass in-phase and quadrature components in (54) are random waveforms in which the pulses with duration $2T_b$ can assume positive or negative values according to:

$$\begin{aligned} s_I(t) &= \sum_k I_k\, p\left(t - 2kT_b\right),\ -\infty \le k \le \infty \\ s_Q(t) &= \sum_k Q_k\, p\left(t - 2kT_b\right),\ -\infty \le k \le \infty \end{aligned} \tag{55}$$

where $p(t)$ is the shaping pulse with half-cycle sine format:

$$p(t) = \sqrt{\frac{2E_b}{T_b}}\sin\left(\frac{\pi}{2T_b}t\right),\ 0 \le t \le 2T_b\,, \tag{56}$$

and $\{I_k\}$ and $\{Q_k\}$ are random antipodal sequences $\in \{\pm 1\}$ associated to the odd and even information bits, respectively (see Example 3) or, equivalently, associated to the waveforms $a_I(t)$ and $a_Q(t)$ in (44).

It is a well-known result that the power spectral density of a random antipodal sequence can be determined by dividing the *energy spectral density* (ESD) of the shaping pulse by the pulse duration [7, p. 48] [8, p. 207]. By recalling that the ESD of a pulse is the squared-modulus of its Fourier transform, then

the PSD of $s_I(t)$, which is equal to the PSD of $s_Q(t)$, can be easily determined. Furthermore, we know that the in-phase and quadrature components of the MSK signal are independent to each other. Then, the PSD of (54) can be obtained through

$$S_B(f) = 2\frac{|P(f)|^2}{2T_b}, \tag{57}$$

and the PSD of the MSK signal can be finally obtained using the above result in (31).

Following the procedure just described, the PSD of the base-band MSK signal in (54) can be obtained from [8, p. 214] and is given by

$$S_B(f) = \frac{32E_b}{\pi^2}\left(\frac{\cos 2\pi fT_b}{1-16f^2T_b^2}\right)^2 \tag{58}$$

Equation (58) is plotted in Figure 17, along with the base-band PSD of the QPSK modulation, for comparison purposes. To draw this figure, both MSK and QPSK signals were set to the same average power.

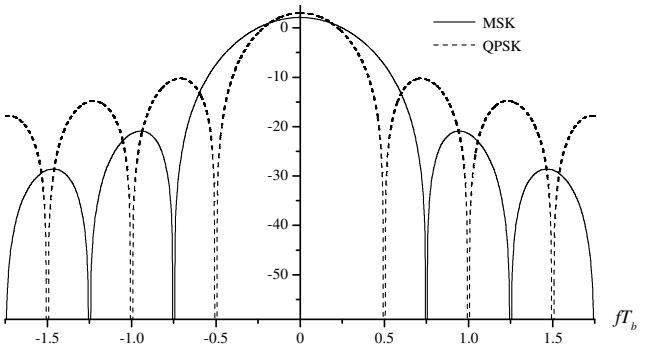


Figure 17. Normalized base-band PSD, in dBm/Hz, for the MSK and the QPSK modulations with the same average power.

It can be seen from Figure 17 that, although the main lobe of the MSK spectrum is wider that the main lobe of the QPSK one, the PSD of the MSK decreases faster with frequency. For QPSK, approximately 90% of the modulated signal power is concentrated in the main lobe. For MSK, this quantity increases to approximately 99%. This is a desired attribute of the MSK modulation, which makes it attractive due to easy filtering and, consequently, low adjacent channel interference.

Detailed and more complete considerations about the power spectral characteristics of continuous-phase modulated signals can be found in [8, pp. 209-219].

IV. Further Attributes and Uses of the MSK

In this section we summarize some MSK-related topics concerning additional attributes and applications of this modulation. We start by revisiting the application of the MSK in the recently-developed Blue-Ray technology [5], and as the base for implementing the GMSK modulation used, for instance, in the GSM standard [7, pp. 396-400]. In the case of the GSM standard, a Gaussian-filtered version of the information sequence is applied to an MSK modulator, resulting in the GMSK signal. This is done to increase the spectral efficiency of the MSK modulation, with the penalty of a possibly small reduction in performance due to inter-symbol interference introduced by the Gaussian filtering process.

As mentioned at the beginning of this paper, the MSK modulation is also attractive because of its constant envelope, a characteristic that can be observed in all FSK-type modulations. Although *M*-PSK modulations also have constant envelopes, this is valid only if no filtering is applied to the signal. When the modulated signal is filtered before going through some non-linear distortion, such as non-linear amplification, out-of-band and in-band spurious can be generated due to envelope fluctuations that occur during abrupt phase transitions. Non-constant envelopes can also show high peak-to-average power ratios (PAPR), making it difficult the project of high dynamic range and power-efficient non-linear amplifiers. The MSK modulation, even after filtering, has low PAPR, becoming attractive in these cases.

The MSK modulation can also be viewed as a special form of coded-modulation scheme in which the phase continuity restrictions introduce some sort of redundancy and, consequently, error correction capabilities. This attribute is explored in detail in [10].

In [12], J. K. Omura, et. al apply the MSK modulation to achieve code-division multiple access (CDMA) capability in a spread spectrum system.

Finally, although MSK is usually associated to the binary case, that is, $M = 2$, its concepts are generalized to the *M*-ary case in [13] and [14]. A multi-amplitude, continuous-phase modulation approach is considered in [8, pp. 200-203], where the signal amplitude is allowed to vary, while the phase trajectory is constrained to be continuous. Generalized MSK is also considered in [15].

V. Conclusions

We are now armed with enough concepts to give possible answers (*A*) to the questions (*Q*) listed at the end of Section I:

Q: To which extent the MSK modulation can be regarded as a special case of the Sunde's FSK modulation? *A*: We saw that MSK is in fact a special form of FSK with the minimum tone separation for orthogonality and coherent detection. However, the MSK signal construction gives to the receiver the ability to explore phase information for performance improvement, which does not happen with the conventional FSK modulation. As we saw in Section III-E, the conventional binary FSK signal with minimum tone separation does not correspond to an MSK signal and does not exhibit phase continuity for all bit transitions.

Q: To which extent the MSK modulation can be detected as the conventional Sunde's FSK modulation? *A*: From the analysis in Section III-E we conclude that an MSK signal can be detected as a conventional binary FSK, but it is necessary to make modifications at the transmitter and at the receiver, according to the block diagram shown in Figure 15. Since this modified receiver explores no phase information, the performance will not be the same as that provided by the appropriate MSK receiver.

Q: To which extent the MSK modulation can be regarded as a special case of the SQPSK or OQPSK (Staggered or

Offset QPSK) modulation? *A*: The MSK modulation is indeed a special form of OQPSK (or SQPSK) modulation, where the pulse shaping are half-cycle cosine and sine functions instead of the rectangular shaping functions used in OQPSK. But this is not a direct interpretation of the MSK signal construction. To shown perfect equivalence with the OQPSK modulation, the MSK transmitter must be implemented according to Figure 8. The receiver structure is kept unchanged, according to the block diagram shown in Figure 9.

Q: To which extent the frequency and phase shifts of an MSK signal are related to the modulating data sequence? *A*: If the modulated signal is generated through the realization of (37) and (39), using $h = ½$, then there will be a direct correspondence, that is, bit 0 will be represented by the tone with frequency, say, f_2 (or vice-versa), and bit 1 will be represented by the tone with frequency f_1 (or vice-versa). However, by generating the MSK signal through the other ways shown is this tutorial, the frequency shifts will correspond to a differentially decoded version of the modulating data sequence.

Q: To which extent the phase shifts of an MSK signal can be related to the phase transition diagram on its signal-space representation? *A*: The MSK signal is constructed in a way that, besides phase continuity, it exhibits phase transitions that helps the receiver improve the detection performance. This is done because phase transitions from one bit to the next lead to different values, modulo 2π (see Figure 4). A bit one increases the phase in $\pi/2$ radians and a bit 0 decreases the phase in $\pi/2$ radians. If these bits are or are not the information bits, it depends on how the MSK signal is generated: directly via (37) or indirectly (see former question and answer). Concerning the phase shifts of an MSK signal, they cannot be directly mapped on the signal-space symbol transitions. Two reasons support this conclusion: firstly, since a given signal-space diagram can represent a base-band or a band-pass signaling, it is not always able to represent phase transitions of a modulated signal, though it can happen with some modulations, such as *M*-PSK and *M*-QAM. Secondly, discrete points in a signal space cannot represent continuous-phase signals, because the phase of the carrier is time-variant [8, pp. 199-200]. As an example, two consecutive ones correspond to the same coordinates in Figure 10, but we know that the carrier phase changes $+\pi/2$ radians from its preceding value, in a continuous way. A solution to this is to have a three-dimensional diagram with axes $s_I(t)$, $s_Q(t)$ and t, in which the phase trajectory can be recorded [8, pp. 194-195]. Figure 18 illustrates this representation.

REFERENCES

[1] R. de Buda, "Coherent demodulation of frequency-shift keying with low deviation ratio", *IEEE Trans. on Communications*, vol. COM-20, no. 3, pp. 429-436, June 1972.

[2] M. L. Doelz and E. H. Heald, "Minimum shift data communication system", *United States Patent* 2,917,417, March 28, 1961.

[3] S. A. Groaemeyer and A.L. McBride, "MSK and offset QPSK modulation", *IEEE Trans. on Communications*, August 1976.

[4] S. Pasupathy, "Minimum Shift Keying: A Spectrally Efficient Modulation", *IEEE Communications Magazine*, vol. 17, no. 4, pp. 14-22, July 1979.

[5] *Blu-Ray Disc Recordable Format – Part 1: Physical Specifications*. Available at http://www.blu-raydisc.com/assets/downloadablefile/BD-R_Physical_3rd_edition_0602f1-13322.pdf (last access: August, 06, 2007).

[6] E. D. Sunde, "Ideal binary pulse transmission by AM and FM", *Bell Systems Technical Journal*, vol. 38, pp. 1357-1426, Nov. 1959.

[7] S. Haykin, *Communication Systems*, 4th Edition - John Wiley and Sons, Inc.: New York, USA, 2001.

[8] J. G. Proakis, *Digital Communications* – 3rd Edition, McGraw Hill, Inc.: USA, 1995.

[9] D. A. Guimarães, A Simple FFSK Modulator and its Coherent Demodulator, *IEICE Trans. Fundamentals*. Vol. E91-A, No. 3, pp. 909-910, March 2008.

[10] H. Leib, S. Pasupathy, "Error Control Properties of Minimum Shift Keying", *IEEE Communications Magazine*, vol.31 No.1, pp. 52-61, January 1993.

[11] S. Benedetto, and E. Biglieri, *Principles of Digital Transmission With Wireless Applications*. Kluwer Academic and Plenum Publishers: New York, 1999.

[12] J. K. Omura et. al., "MSK spread-spectrum receiver which allows CDMA operations", *United States Patent 5,963,585*, October 5, 1999.

[13] M. K. Simon, "A generalization of minimum shift keying (MSK) type signaling based upon input data symbol pulse shaping", *IEEE Trans. on Communications*, vol. COM-24, pp. 845-856, August 1976.

[14] I. Korn, "Generalized MSK", *IEEE Trans. on Information Theory*, vol. IT-26, no. 2, pp. 234–238, March 1980.

[15] R. Sadr and J. K. Omura, "Generalized minimum shift-keying modulation techniques", *IEEE Trans. on Communications*, Volume 36, Issue 1, pp. 32-40, Jan 1988.

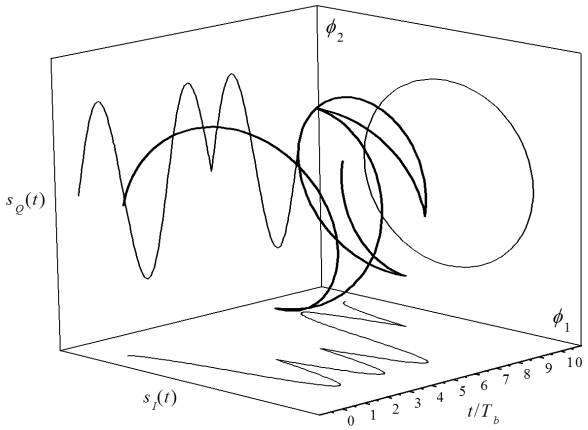


Figure 18. Phase trajectory of an MSK signal. The projections of this trajectory on all planes are also shown.

Dayan Adionel Guimarães was Born in Carrancas, MG, Brazil, on March 01, 1969. He holds the titles: Electronics Technician (ETE "FMC", 1987), Electrical Engineer (Inatel, 1994), Specialist in Data Communication Engineering (Inatel, 2003) and in Human Resources Management (FAI, 1996), Master in Electrical Engineering (Unicamp, 1998) and Doctor in Electrical Engineering (Unicamp, 2003).

From 1988 to 1993 he developed equipment for Industrial Instrumentation and Control, and was also the Manufacturing and Product Engineering Supervisor at *SENSE Sensores e Instrumentos*. Since January 1995 he is Professor at Inatel where, for eight years, he was responsible for the structure that supports practical teaching activities for the Electrical Engineering undergraduate course. His research includes the general aspects on Digital and Mobile Communications, specifically Multi-Carrier CDMA systems, and coding for fading channels, specifically Block Turbo Codes.

Dr. Dayan is member of the *Telecomunicações* magazine's Editorial Board, member of the Inatel's Master Degree Counseling Board and of the IEICE (*Institute of Electronics, Information and Communication Engineers*), Japan.

Codificação Fractal de Imagens

Ana Lúcia Mendes Cruz
UNICAMP
analucia@decom.fee.unicamp.br

Fernando Silvestre da Silva
UNICAMP
silva@decom.fee.unicamp.br

Yuzo Iano
UNICAMP
yuzo@decom.fee.unicamp.br

Roger Fredy Larico Chavez
UNICAMP
rlarico@decom.fee.unicamp.br

***Abstract*— This paper presents fractal bases and its potentiality when applied to image coding as a new ally on the establishment of new coding standards, especially when merged to the new techniques presented in literature. The basic topics about fractal compression, specially the PIFS coders (*Partitioned Iterated Function Systems*) are included, illustrative and mathematically. A really complete approach concerning the traditional literature and state of art literature follows about the topic including hybrid coders, such as a quick comparison involving fractal, wavelets, hybrids techniques and JPEG2000.**

***Index Terms*— Image Coding, Fractals, coding time, wavelets.**

***Resumo*— Este artigo tem por objetivo apresentar as bases e potencialidades da codificação fractal de imagens como aliada no estabelecimento de novos padrões de compressão, especialmente quando combinados a técnicas recentemente apresentadas na literatura. Os princípios básicos da compressão fractal, em especial os codificadores PIFS (*Sistema de Funções Iterativas Particionadas*), são apresentados, ilustrativa e matematicamente. Uma abordagem bastante completa acerca da bibliografia tradicional e estado da arte contendo inclusive algoritmos híbridos é apresenta da sobre o tema, culminando com uma breve comparação sobre sistemas fractais, wavelets, híbridos e JPEG2000.**

***Palavras chave*— Codificação de Imagens, fractais, velocidade de codificação, wavelets.**

I. INTRODUÇÃO

Imagens digitais exigem uma parcela cada vez maior do mundo da informação. Basta observar os avanços contínuos na tecnologia de impressoras, celulares e câmeras digitais na última década, por exemplo. Novas técnicas não somente transformaram as aplicações e serviços existentes, como a distribuição de vídeo para entretenimento residencial, mas também geraram novas indústrias e serviços tais como vídeo-conferência, distribuição por satélite direto às residências, gravação de vídeo digital, serviços vídeo *on demand*, HDTV (*High Definition Television*), vídeo em dispositivos móveis, *streaming* de vídeo, entre outros [1] [2].

Embora os métodos de compressão de imagens sejam de uso extensivo hoje, a demanda pelo aumento da capacidade de armazenamento e de velocidade de transmissão exige pesquisas contínuas em busca de novos métodos ou na melhoria dos já existentes. Muitas áreas da ciência, em especial a engenharia e matemática, têm se dedicado a essa questão. Novas tecnologias, a exemplo das técnicas fractais, se tornam aliadas potenciais no estabelecimento de novos padrões de compressão de imagens.

A. Codificação Fractal de Imagens

A codificação fractal de imagens consiste em representar os blocos da imagem através de coeficientes de transformações contrativas, explorando o conceito de auto-similaridade. Assim, nesse tipo de codificação, ao invés de armazenar/transmitir os blocos da imagem como uma coleção de *pixels*, somente são enviados/armazenados os coeficientes dessas transformações. Esse princípio de funcionamento permite obter altas taxas de compressão mantendo ótima qualidade visual.

Entre as vantagens da compressão fractal estão a rápida decodificação, transmissão progressiva, boa compressão e o fato de praticamente não necessitar de codificadores entrópicos [1][3][4]. Por se tratar de uma técnica ainda pouco explorada, se apresenta como um vasto campo para a pesquisa, dado o seu alto potencial para a compressão de imagens. Um dos aspectos que encorajam a codificação fractal é que ela é completamente paralelizável. Hardware paralelo especializado pode, portanto, reduzir significativamente o tempo de codificação.

Apesar das características vantajosas da compressão fractal, o tempo exaustivo de processamento é a principal barreira na implementação de sistemas práticos que façam uso dessa tecnologia [5][6]. Diversas tentativas recentes têm sido feitas na tentativa de minimizar esta questão, abordando desde vetores de características até redes neurais. Muita pesquisa ainda se faz necessária para destravar o potencial fractal latente para a compressão de imagens. Sem dúvida, essas pesquisas devem residir na questão de promover a aceleração

· Manuscrito recebido em 24 de março de 2007; revisado em 07 de junho de 2007.

A. L. M. Cruz (analucia@decom.fee.unicamp.br), F. S. Silva (silva@decom.fee.unicamp.br), Y. Iano (yuzo@decom.fee.unicamp.br), e R.Chavez (rlarico@decom.fee.unicamp.br) pertencem ao Departamento de Comunicações – DECOM da Faculdade de Engenharia Elétrica e de Computação da UNICAMP .

da etapa de compressão dessa nova tecnologia.

II. Sistema de Funções Iterativas (IFS): bases da compressão fractal de imagens

O termo fractal foi primeiramente introduzido por Benoit Mandelbrot em 1983 [7]. A propriedade-chave que caracteriza os fractais e os diferencia das demais técnicas é a auto-similaridade, isto é, os fractais apresentam a mesma complexidade de detalhamento independente da escala em que são observados. Barnsley and Sloan [8] foram os primeiros a reconhecer o potencial da aplicação da teoria fractal ao problema da compressão de imagens e patentearam sua idéia em 1990 e 1991 [9][10]. Era o chamado IFS (*Iterated Funcion Systems* ou Sistema de Funções Iterativas).

Em 92, Jacquin [11] introduziu o conceito de particionamento por blocos, criando a técnica conhecida como PIFS (*Partitioned Iterated Function System* ou Sistema de Funções Iterativas Particionadas). O PIFS é o sistema que melhor se adequa às imagens reais, sendo efetivamente utilizado nos codificadores fractais atuais [5]. Contudo, a compreensão do IFS é essencial para se entender o modo como a compressão fractal trabalha [12].

Uma analogia simples para explicar esses princípios básicos de funcionamento da idéia fractal foi feita por Fisher [3]: essa analogia consiste num tipo especial de fotocopiadora que reduz a imagem a ser copiada pela metade e a reproduz três vezes na cópia de saída (reduzidas e deslocadas), como mostrado na Fig. 1.

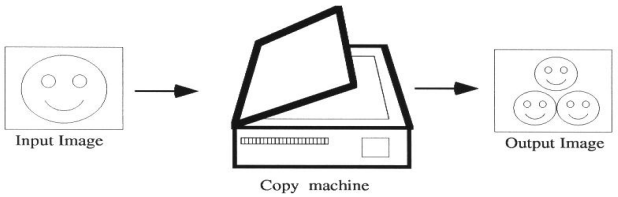


Fig. 1 – Exemplo de Fotocopiadora de Fisher. [3]

A idéia básica consiste em passar essa cópia de saída novamente pela copiadora, num processo recursivo. Após várias iterações desse processo, percebe-se na Fig. 2 que mesmo com diferentes imagens de entrada, todas as cópias de saída parecem convergir para a mesma imagem final, mostrada na Fig. 2c. Essa imagem é o chamado "atrator" para essa máquina fotocopiadora especifica. Logo, a imagem inicial não afeta o atrator final; de fato somente a posição e orientação das cópias (transformações afins) determinam como a imagem final se parecerá.

Uma vez que a fotocopiadora reduz a imagem de entrada, qualquer imagem inicial será reduzida a um ponto na medida em que aumentam as iterações. As transformações possuem, portanto, a limitação de serem "contrativas", ou seja, a distância entre dois pontos quaisquer da imagem de saída (após transformação) precisa ser menor do que a distância entre os pontos correspondentes na imagem de entrada [3].

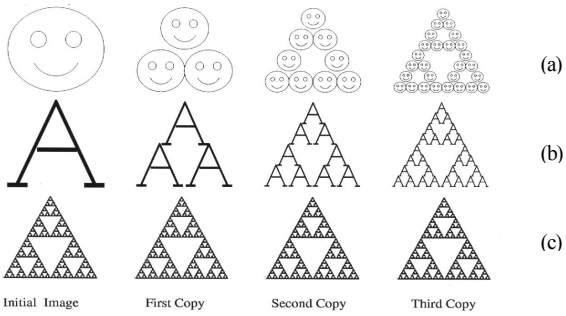


Fig. 2 – Primeiras 3 cópias geradas pela fotocopiadora para 3 imagens diferentes. [3]

Na prática, transformações na forma (1) permitem gerar uma grande variedade de interessantes atratores. Essas transformações afins polinomiais de primeira ordem podem torcer, deslizar, rotacionar, escalar ou deslocar a imagem de entrada, dependendo dos valores dos coeficientes.

$$\omega_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix} \quad (1)$$

Assim, matematicamente, um sistema de funções iterativas (IFS) em R^2 consiste de uma coleção de transformações contrativas $\{\omega_i : R^2 \rightarrow R^2 \mid i=1,...n\}$ que mapeiam o plano R^2 para si mesmo em cada iteração. Essa coleção de transformações define o mapa [3]:

$$W(\cdot) = \bigcup_{i=1}^{n} \omega_i(\cdot) \quad (2)$$

Foi Hutchinson [13] quem provou que se algumas condições forem satisfeitas e se esse mapa W for contrativo, a recursão converge para o atrator. Logo, como o mapa W (ou os ω_i) determina completamente uma única imagem [3], é possível determinar o atrator, que é a imagem que deseja-se recuperar.

Assim, dada uma imagem inicial A_0, aplicando uma vez o conjunto de transformações W, gera-se $A_1=W(A_0)$; a segunda vez gerando $A_2=W(A_1)=W(W(A_0))\equiv W^{o2}(A_0)$ e assim por diante. Dessa forma, o atrator é o conjunto limite:

$$A_f = \lim_{n\to\infty} \mathrm{W}^{\mathrm{on}}(\mathrm{A}_0) \quad (3)$$

que independe da escolha de A_0.

Na Fig. 3 é apresentado um exemplo prático da aplicação desta teoria. Note que partiremos de uma mesma imagem e aplicaremos a ela três mapas contrativos diferentes. O primeiro mapa consiste em aplicar três transformações que simplesmente subamostram e deslocam a imagem. O segundo consiste em subamostragem com duas rotações mais inversão e duas subamostragens com dois deslocamentos. O terceiro mapa abrange rotações, subamostragens, contrações e deslocamentos. Note que dependendo do mapa aplicado na recursão, é gerado um atrator diferente. Note também que o atrator independe da imagem inicial. Trocar a imagem inicial por outra e aplicar as mesmas transformações recursivamente, fará o sistema convergir para o mesmo atrator.

Torna-se prioritário frisar que cada atrator é formado por copias reduzidas e transformadas de si mesmo, implicando na auto-similaridade. O atrator, ou seja, a imagem que deseja-se

recuperar é formado por auto-similaridades de si próprio. Este conceito é fundamental para se compreender os sistemas de codificação fractal utilizados hoje.

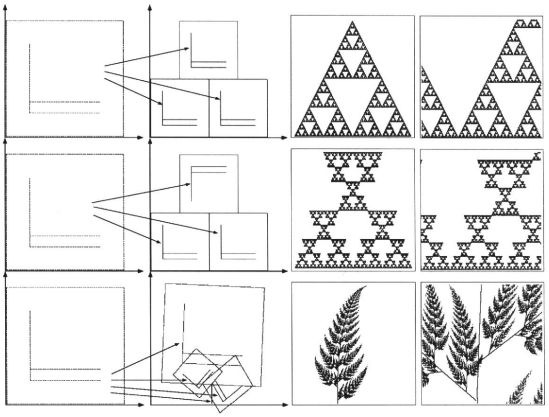


Fig. 3 – Transformações Afins, Atratores e Ampliação no Atrator. [3]

Finalmente, Barnsley [14] foi o primeiro a perceber que armazenar os coeficientes das transformações afins ao invés de armazenar ou transmitir a imagem como uma coleção de *pixels*, poderia gerar significativa compressão. Uma vez armazenados os coeficientes, para reconstruir a imagem basta aplicar recursivamente as transformações por eles determinadas em uma imagem qualquer.

Abordagens matemáticas mais detalhadas podem ser encontradas em [1][3][4][15][16], mas basicamente os dois principais teoremas nos quais esta fundamentada a teoria fractal apresentada são o Teorema do Mapeamento Contrativo e o Teorema da Colagem. As principais definições e propriedades acerca do Sistema de Funções Iterativas são apresentados a seguir.

A. Espaço em que os fractais existem e a Contratividade: abordagem matemática

Lema A (Contratividade): "Seja (X, d) um espaço métrico completo, onde d é a métrica associada que mede a distância entre os elementos do espaço X. Então, o espaço dos fractais ($\mathcal{H}(\boldsymbol{X})$, h) é o espaço dos subconjuntos compactos e não-vazios do espaço X, sendo também um espaço métrico completo [16]. Seja também $\{\omega_n : X \to X$, $n=1, 2, ..., N\}$ um conjunto de mapeamentos contrativos neste espaço, e portanto, em ($\mathcal{H}(\boldsymbol{X})$, h) (onde h é a métrica de *Hausdorff*) [16]. O fator de contratividade de ω_n é dado por s_n para cada n. Definindo $W : \mathcal{H}(\boldsymbol{X}) \to \mathcal{H}(\boldsymbol{X})$ por:

$$W(A) = \omega_1(A) \bigcup \omega_2(A) \bigcup \ldots \bigcup \omega_N(A) \tag{4}$$

$$W(A) = \bigcup_{n=1}^{N} \omega_n(A), \text{ para cada } A \in \mathcal{H}(\mathbf{X}). \tag{5}$$

então W é um mapeamento contrativo com fator de contratividade $s = max \{s_n , n=1, 2, ..., N \}$".

Através do lema A, os sistemas de funções iterativas são definidos por [16]:

Definição 1: "Um Sistema de Funções Iterativas – IFS – consiste em um espaço métrico completo ($\boldsymbol{X}$,d) e um conjunto finito de mapeamentos contrativos $\omega_n : \boldsymbol{X} \to \boldsymbol{X}$, com os respectivos fatores de contratividade s_n, para $n=1, 2, ..., N$. A notação para a IFS é $\{(\boldsymbol{X}, d)$; ω_n, $n=1, 2, ..., N\}$ e seu fator de contratividade é $s = max \{s_n , n=1, 2, ..., N \}$".

Teorema A: "Seja $\{(\boldsymbol{X}, d)$; ω_n, $n=1, 2, ..., N\}$ um sistema de funções iterativas com fator de contratividade s. Então a transformação $W : \mathcal{H}(\boldsymbol{X}) \to \mathcal{H}(\boldsymbol{X})$ definida por:

$$W(A) = \bigcup_{n=1}^{N} \omega_n(A) \tag{6}$$

para todo $A \in \mathcal{H}(\boldsymbol{X})$, é um mapeamento contrativo no espaço métrico completo ($\mathcal{H}(\boldsymbol{X})$, h) com fator de contratividade s, ou seja:

$$h(W(A), W(B)) \le s \cdot h(A, B), \tag{7}$$

para todo $A,B \in \mathcal{H}(\boldsymbol{X})$. A existência e unicidade do ponto fixo, $A_f \in \mathcal{H}(\boldsymbol{X})$, chamado de *atrator* da IFS, é dado pela aplicação do teorema do ponto fixo de mapas contrativos, que leva a:

$$A_f = \lim_{n \to \infty} W^{on}(A), \quad \forall A \in \mathcal{H}(\boldsymbol{X}), \tag{8}$$

de forma que $A_f = W(A_f)$.

A distância entre um dado $A \in \mathcal{H}(\boldsymbol{X})$ e o atrator A_f obedece à inequação

(TEOREMA DA COLAGEM):

$$h(A, A_f) \le \frac{1}{1-s} h(A, W(A))" \tag{9}$$

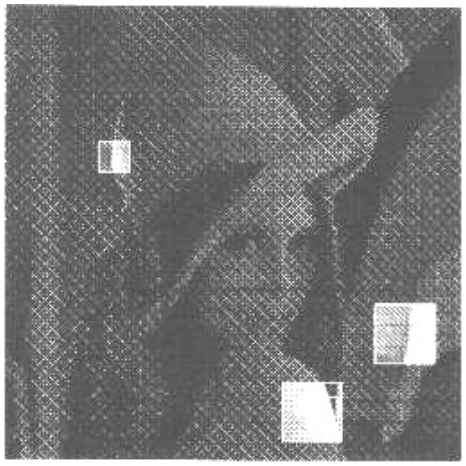
Esse teorema resume as características que possibilitam o modelamento de fractais como sendo subconjuntos compactos de um espaço métrico completo que pode ser completamente especificado através de um sistema de equações (mapas contrativos).

III. SISTEMA DE FUNÇÕES ITERATIVAS PARTICIONADAS (PIFS): CODIFICADORES FRACTAIS ATUAIS

Uma imagem real, em geral, não apresenta a auto-similaridade apresentada, dita "auto-similaridade global", ou seja, a imagem não parece conter transformações afins de si mesma. Ao invés disso, algumas de suas áreas apresentam similaridades entre si (similar por partes). Na Fig. 4 podem ser vistas similaridades entre as regiões de diferentes tamanhos destacadas no chapéu e espelho e entre regiões destacadas no ombro [3]. Para codificar essas imagens reais é necessária a aplicação do PIFS.

A distinção entre esse tipo de similaridade e a dos fractais apresentados na Fig. 3 é que, naquele caso, a imagem era formada por cópias transformadas de si mesma *inteira* e aqui a imagem é formada por cópias propriamente transformadas de *partes* de si mesma. Dessa forma, as partes transformadas geralmente não se encaixam perfeitamente. Por conseqüência, a imagem codificada pelo PIFS não será uma cópia idêntica da imagem original, mas sim uma aproximação [1].

(a)



(b)

Fig. 4 – (a) Imagem Lena Original (256 x 256 pixels); (b) Exemplos de similaridades nessa imagem. [3]

A. Codificação PIFS

Assim, a codificação PIFS se resume a procurar, para cada bloco da imagem que se deseja codificar, pelo bloco de tamanho maior que seja auto-similar. O bloco que se deseja codificar é chamado de *range-block*. Os blocos de tamanho maior que serão comparados com cada *range-block* são chamados de *domain-blocks*, que conjuntamente compõem o que se chama de *domain pool*, na verdade um espaço de procura que pode ou não conter informação parcialmente redundante (*overlap* entre *domain-blocks*). Assim, a cada *domain-block*, são aplicadas subamostragem, filtragem de média e transformações-afins ω_i (estas últimas compostas por uma parte geométrica e ajustes de contraste e brilho). Os resultados de todas essas transformações são comparados com o *range-block* que está sendo codificado [1].

A escolha pelo melhor *matching* (casamento) é feita minimizando uma medida de distorção. Ao encontrar o par *domain-range* somente são enviados os coeficientes da transformação afim que transforma o *domain-block* D_i no *range-block* R_i e o endereço do *domain-block* que foi escolhido como seu equivalente.

A necessidade de se encontrar o melhor *matching* surge uma vez que, em geral, a intenção de determinar a região D_i da imagem que após a transformação gere exatamente R_i não pode ser alcançada. Isso ocorre porque uma imagem real dificilmente é composta somente por partes que, após a transformação, se encaixem perfeitamente em outra parte da mesma imagem.

Assim, o que sempre é possível, é tentar encontrar uma outra imagem *f'* tal que $f' = f_\infty$ e que a métrica *d(f, f')* seja pequena o suficiente, ou seja, procura-se o mapa *W* tal que o seu ponto fixo *f'* seja próximo (ou se pareça com) a imagem *f*. A métrica *d(f, f')* pode ser definida de várias formas, contudo, a métrica utilizada na quase totalidade dos trabalhos em codificação fractal é a métrica *rms*, mostrada na equação (10), escolha que facilitará a determinação dos valores dos fatores de contraste s_i e brilho o_i, conforme será apresentado à frente.

$$d_{rms}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y} \rangle} \tag{10}$$

Onde o operador <·,·> simboliza o produto interno padrão.

O processo de codificação, então, se resume a encontrar as regiões D_i e os mapas ω_i tais que ao aplicar ω_i em D_i obtém-se algo mais próximo possível da região R_i.

Para formalizar matematicamente a codificação de imagens com PIFS é necessário primeiramente definir o modelo matemático que será usado para a imagem natural:

Definição 2: "Seja $I = [0;1] \subset \mathfrak{R}$ o intervalo unitário na reta real e, conseqüentemente, I^2 o quadrado unitário no plano euclidiano. Uma imagem natural pode ser vista como o gráfico de uma função $f \in \boldsymbol{F}$ tal que $\boldsymbol{F} = \{ f \mid f : I^2 \rightarrow \mathfrak{R} \}$".

Deve-se notar que a função *f* mapeia o quadrado unitário para o intervalo unitário, mas considera-se que a imagem de *f* seja $\mathfrak{R}$ para que as somas e subtrações de imagens fiquem bem definidas. Embora o *range* e o *domain* da função sejam, respectivamente, R_i x *I* e D_i x *I*, é usual referenciá-los apenas como *range-block* R_i e *domain-block* D_i.

Suponha que se deseja codificar por PIFS uma imagem *f*. Ou seja, deseja-se encontrar a coleção de mapas $\omega_1, \omega_2, ..., \omega_n$ com:

$$\omega_i \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \\ o_i \end{bmatrix}, \; i=1, 2, ..., n \tag{11}$$

onde $\{a_i, b_i, c_i, d_i, e_i, f_i\}$ define a transformação geométrica; s_i determina o contraste e o_i o brilho. Todos esses componentes compõem a transformação afim, mas por vezes, a transformação geométrica é referenciada separadamente, uma vez que é transmitida separadamente do brilho e contraste no *bitstream* do sinal codificado. A transformação geométrica é enviada em 3 bits. Tipicamente o brilho é enviado em 7 bits, e o contraste em 5 bits [3].

$W(f)$ é definido por:

$$W(f) = \bigcup_{i=1}^{N} \omega_i \left(f \cap \left(D_i \times I \right) \right) \tag{12}$$

de modo que $f = f_\infty$ seja o atrator de *W*. Em outras palavras, o

PIFS consiste na aplicação recursiva de um mapa W composto por uma coleção de transformações contrativas ω_i, com a diferença de que agora a imagem será processada por blocos, ou seja, cada transformação será aplicada somente sobre o *domain-block*, ao invés de ser aplicado na imagem toda. A união de todas as partes transformadas irá formar a nova imagem após a iteração.

Usando a notação $\omega_i(f)$ para indicar $\omega_i(f \cap (D_i \times I))$, a equação do ponto fixo pode ser expressa como:

$$f_\infty = W(f_\infty) = \omega_1(f_\infty) \cup \omega_2(f_\infty) \cup \ldots \omega_n(f_\infty). \quad (13)$$

O objetivo, na prática, é encontrar $f' = f_\infty$ tal que o erro $d_{sup}(f, f')$ seja pequeno o suficiente, conforme já colocado. Nesse caso:

$$f \approx f' = W(f') \approx W(f) = \omega_1(f) \cup \omega_2(f) \cup \ldots \omega_n(f) \quad (14)$$

Então, é suficiente aproximar os blocos-R_i dos blocos-D_i transformados minimizando a medida de distorção:

$$\min_{\omega_i} \{d\ (f \cap (R_i \times I), \omega_i(f))\}, \quad i = 1, \ldots, n \quad (15)$$

Quando se utiliza partição *quadtree* para processamentos da imagem, o formato de D_i e R_i é sempre quadrado e o tamanho do *domain-block* é, em geral, o dobro do tamanho do *range-block*. Assim, a escolha da parte geométrica do mapa é limitada a apenas uma redução de tamanho por um fator de 2 seguida de uma dentre as 8 opções: 4 rotações (0º, 90º, 180º e 270º) e 4 inversões (vertical, horizontal, diagonal principal, diagonal secundária). Assim, a minimização da equação (15) fica bastante simplificada, bastando encontrar os valores s_i e o_i que minimizam a distância em cada uma das 8 transformações geométricas possíveis.

B. Cálculo dos Valores de s_i e o_i

Sejam $\boldsymbol{r}_i$ o *range-block* transformado em vetor; $\boldsymbol{d}_i$ o *domain-block* transformado em vetor (após subamostragem e transformação geométrica); $\boldsymbol{1}$ o vetor de uns"; e *NxN* as dimensões do *range-block*. Para calcular os valores de s_i e o_i ótimos para cada candidato a *domain-block-equivalente* em uma dada transformação geométrica utiliza-se a expressão:

$$\min_{s_i\, o_i}\left(d_{rms}\left(\boldsymbol{r}_i,\ s_i \cdot \boldsymbol{d}_i + o_i \cdot \boldsymbol{1}\right)\right) = \left(\sum_{j=1}^{N^2} \left| r_{ij} - \left(s_i \cdot d_{ij} + o_i\right)\right|^2\right)^{1/2} \quad (16)$$

Para resolver essa equação utilizar-se-á uma a métrica *rms* mostrada na equação (10). Assim:

$$d^2_{rms}\left(\boldsymbol{r}_i,\ s_i \cdot \boldsymbol{d}_i + o_i \cdot \boldsymbol{1}\right) = \left\langle \boldsymbol{r}_i - \left(s_i \cdot \boldsymbol{d}_i + o_i \cdot \boldsymbol{1}\right), \boldsymbol{r}_i - \left(s_i \cdot \boldsymbol{d}_i + o_i \cdot \boldsymbol{1}\right)\right\rangle =$$
$$= s_i^2 \langle \boldsymbol{d}_i, \boldsymbol{d}_i \rangle + 2 o_i s_i \langle \boldsymbol{d}_i, \boldsymbol{1} \rangle + o_i^2 \langle \boldsymbol{1}, \boldsymbol{1} \rangle - 2 s_i \langle \boldsymbol{d}_i, \boldsymbol{r}_i \rangle - 2 o_i \langle \boldsymbol{1}, \boldsymbol{r}_i \rangle + \langle \boldsymbol{r}_i, \boldsymbol{r}_i \rangle \quad (17)$$

Diferenciando com relação a s_i e a o_i e igualando a zero obtém-se os valores:

$$s_i = \frac{\langle \boldsymbol{d}_i, \boldsymbol{1} \rangle \langle \boldsymbol{1}, \boldsymbol{r}_i \rangle - \langle \boldsymbol{1}, \boldsymbol{1} \rangle \langle \boldsymbol{d}_i, \boldsymbol{r}_i \rangle}{\langle \boldsymbol{d}_i, \boldsymbol{1} \rangle^2 - \langle \boldsymbol{1}, \boldsymbol{1} \rangle \langle \boldsymbol{d}_i, \boldsymbol{d}_i \rangle} \quad e \quad o_i = \frac{\langle \boldsymbol{d}_i, \boldsymbol{1} \rangle \langle \boldsymbol{d}_i, \boldsymbol{r}_i \rangle - \langle \boldsymbol{1}, \boldsymbol{r}_i \rangle \langle \boldsymbol{d}_i, \boldsymbol{d}_i \rangle}{\langle \boldsymbol{d}_i, \boldsymbol{1} \rangle^2 - \langle \boldsymbol{1}, \boldsymbol{1} \rangle \langle \boldsymbol{d}_i, \boldsymbol{d}_i \rangle} \quad (18)$$

Substituindo pela definição do produto interno:

$$s_i = \frac{\sum \boldsymbol{d}_i \cdot \sum \boldsymbol{r}_i - N^2 \cdot \sum \boldsymbol{d}_i \boldsymbol{r}_i}{\left(\sum \boldsymbol{d}_i\right)^2 - N^2 \cdot \sum \boldsymbol{d}_i^2} \quad e \quad o_i = \frac{\sum \boldsymbol{d}_i \cdot \sum \boldsymbol{d}_i \boldsymbol{r}_i - \sum \boldsymbol{r}_i \cdot \sum \boldsymbol{d}_i^2}{\left(\sum \boldsymbol{d}_i\right)^2 - N^2 \cdot \sum \boldsymbol{d}_i^2} \quad (19)$$

onde:

$\sum \boldsymbol{r}_i$ = soma de todos os elementos do vetor $\boldsymbol{r}_i$;

$\sum \boldsymbol{d}_i$ = soma de todos os elementos do vetor $\boldsymbol{d}_i$;

$\sum \boldsymbol{r}_i \boldsymbol{d}_i$ = soma do produto elemento a elemento (produto interno) dos vetores $\boldsymbol{d}_i$ e $\boldsymbol{r}_i$;

$\sum \boldsymbol{d}_i^2$ = soma do quadrado de todos os elementos do vetor $\boldsymbol{d}_i$;

N^2 = número de elementos de cada vetor.

Note que os valores de s_i e o_i calculados por estas equações não são limitados em termos de amplitude e precisarão ser quantizados. Para calcular o erro de aproximação usam-se os valores quantizados, pois serão estes os valores enviados ao decodificador. Outro fator de restrição é o máximo valor de contraste permitido s_{max} (que é um parâmetro do codificador). Caso o valor calculado ultrapasse s_{max}, o valor s_{max} é usado para o cálculo do erro [3].

C. Decodificação PIFS

Uma vez transmitidos os dados, cada iteração do processo de *decodificação* de imagens naturais em escala de cinza através do PIFS é realizado da seguinte forma: uma máscara seleciona uma parte de uma imagem no decodificador (*domain-block*), na qual aplica-se uma subamostragem e uma transformação-afim ω_i (já transmitida). A seguir, esse *domain-block*, já subamostrado e transformado, é copiado para uma região específica desta mesma imagem (*range-block*). Esse processo é repetido para todas as partes da imagem. O conjunto final de todos os *range-blocks* formará a imagem já decodificada após esta iteração [1].

Assim, para realizar a codificação e decodificação através do PIFS é necessária primeiramente a definição de quatro aspectos básicos [16]:

- O número de cópias transformadas que irão compor a imagem de saída, ou seja, o número de *range-blocks*. Em geral essa informação é transmitida de forma implícita, como no caso da partição *quadtree*;
- A máscara que selecionará qual a parte afetada (*domain-block*) em cada transformação.
- O ajuste de brilho e contraste para cada *domain-block*;
- Os demais coeficientes das transformações afins (além de contraste e brilho) associados a cada *domain-block*.

Dada uma imagem *f* em escala de cinza, uma iteração do processo de decodificação utilizando uma partição com *N*

partes (*N range-blocks*) pode ser descrito como a aplicação do mapeamento *W* da seguinte forma [16]:

$$W(f) = \omega_1(f) \cup \omega_2(f) \cup ..... \cup \omega_N(f)$$

onde ω_i é aplicada *somente* sobre a respectiva região D_i. É importante salientar que associado a cada ω_i tem-se uma região D_i na imagem de entrada e uma região R_i na imagem de saída.

Como *W(f)* representa uma imagem, deve-se salientar que $\bigcup R_i = I^2$ e que $R_i \bigcap R_j = \varnothing$, se $i \neq j$, ou seja, R_i é uma *partição* da imagem de saída (I^2). Dessa forma, pode-se dizer que o processo iterativo é tal que a saída da primeira iteração f_1 é dada por $f_1=W(f_0)$, a da segunda $f_2=W(f_1)=W(W(f_0))=W^{o2}(f_0)$ e assim por diante.

A Fig. 5 ilustra um exemplo do processo de decodificação. Parte-se de uma imagem qualquer, e aplica-se recursivamente o mapa. O resultado de algumas iterações pode ser observado. Em geral, após a quarta ou quinta iteração, a seqüência já converge.

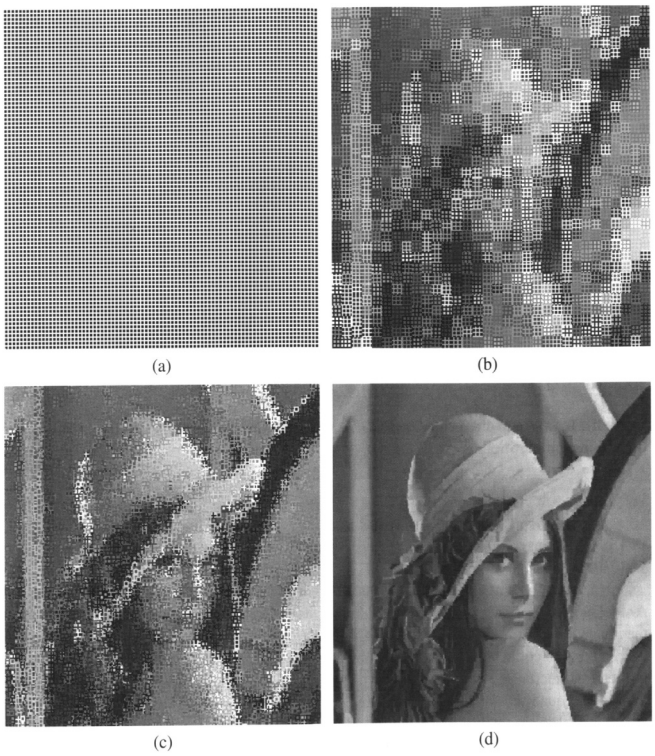


Fig.5 – Decodificação fractal da imagem Lena: a) imagem inicial arbitrária; b) após a 1ª iteração; c) após a 2ª iteração; d) após a 10ª iteração. [16]

IV. CODIFICAÇÃO FRACTAL ACELERADA

A codificação fractal apresenta ótima qualidade de imagens, mesmo a baixas taxas de bits e a decodificação fractal é extremamente rápida. Entretanto, na codificação, cada *range-block NxN* é comparado a *todos* os *domain-blocks 2Nx2N* da *domain pool* dentro de uma imagem *MxM*. Como conseqüência, o algoritmo de *matching* possui complexidade computacional $O[M^4]$, o que exige um tempo exaustivo de processamento. Esse peso computacional é o principal motivo de hesitação da adoção de aplicações fractais, sendo similar ao peso computacional do processo de estimação de movimento de sistemas de codificação de vídeo como, por exemplo, o MPEG2, onde o processo de procura pelo melhor *matching* é responsável em média por 40% do tempo de codificação.

Basicamente, as tentativas de acelerar a codificação fractal consistem em modificar os seguintes aspectos: a composição da *domain pool*, o tipo de procura usado no *block matching*, ou a representação/quantização dos parâmetros transformados.

Diferentes tentativas de reduzir o tempo de compressão fractal têm sido tentadas. As mais tradicionais dizem respeito à restrição da *Domain Pool* via pré-classificação dos blocos. Por exemplo, em [17] Jacquin classificou os blocos usando informações de borda e, em [3], Fisher pré-classificou os blocos segundo seus valores de média e variância. Fisher [3] desenvolveu uma classificação de domínios com o objetivo de reduzir o número de comparações de *matching*. O ponto principal desse esquema é que, primeiramente, todos os *domain-blocks* são classificados. Durante a codificação, cada *range-block* é classificado e somente será comparado com os *domain-blocks* de mesmo tipo que o seu (segundo um critério baseado em média e variância), o que reduz significativamente o número de comparações de *matching*.

Bogdan e Meadows [18] aplicaram uma rede auto-organizadora de Kohonen ao problema da codificação fractal, requerendo treinamento do codificador sobre a imagem a ser codificada. Esses autores reduziram o número de comparações *range-domain*, mas não a complexidade da comparação em si. McGregor *et. al.* [19], por outro lado, trabalharam esses dois problemas usando uma procura do tipo K-D-tree nos *domain-blocks* combinada à extração de um pequeno número de características do bloco-imagem.

Um passo importante na questão da aceleração fractal foi dado em [20], e foi a inspiração do codificador de Cardinal de [21]. Em [20] Saupe reduziu a questão da procura pelo melhor matching à procura pelo vizinho mais próximo dentro de um espaço métrico conveniente. Nesta técnica, cada bloco é associado a um vetor de características tal que minimizar o erro de colagem entre o *range-block* e o *domain-block* equivale a minimizar a distância entre os vetores de característica correspondentes. A procura pelo melhor *matching* então consiste em simplesmente encontrar o vetor mais próximo no espaço vetorial. Deve-se colocar, entretanto, que o método de Saupe encontra o melhor matching e só depois calcula os valores de *si* e *oi*, o que pode ferir a restrição de contratividade sobre *si*. O método de Saupe pode ser feito de forma mais rápida se for usada uma estrutura K-D-*tree*.

Em [21], Cardinal reduziu o processo de procura pelas transformações afins de cada *range-block* ao problema geométrico clássico de busca pelos vizinhos mais próximos num espaço euclidiano. Baseando-se na partição geométrica do espaço de características dos *range-blocks*, Cardinal conseguiu boa aceleração na compressão sem perda de

qualidade. Idéias similares são apresentadas em [22][23]. Também pode ser utilizada a quantização vetorial como em [24][25]. Em [4] um método ligeiramente menos sofisticado que o de Cardinal [21] é apresentado, mas segue a mesma filosofia. Cardinal argumenta que tais técnicas tendem a superar as aproximações clássicas que têm por base a redução da *Domain Pool*.

Em 2002, Tong e Wong [26] apresentaram uma procura pelo vizinho mais próximo baseada na projeção ortogonal e pré-quantização dos parâmetros fractais transformados. Tong e Pi também desenvolveram um codificador de procura adaptativa que exclui um grande número de *domain-blocks* não qualificados [27] e Wang e Hsieh [28] elaboraram um método que explora a correlação entre *range-blocks* com o mesmo propósito. Bani Eqbal [29] propôs também um método de procura em árvore para a *Domain Pool*.

Em 2002, Chi *et al.* [30] apresentaram um novo modelo de métrica baseado na integral *fuzzy* de Sugeno combinada com uma partição *quadtree* obtendo melhor qualidade visual e redução no tempo de compressão.

Alguns algoritmos eficientes [31][32][33] foram também apresentados para aliviar a carga computacional mantendo a mesma qualidade de imagem do *Full Search*. Entretanto, pré-processamento é necessário para esses algoritmos.

Wen *et.al.* também em 2003 [34] utilizaram média e variância como método de classificação combinado com técnicas de redução nas transformações. Também em 2003, Siu *et. al.* [35] propuseram um esquema baseado numa condição de exclusão única e numa predição de contraste zero. A condição de exclusão única evita um grande número de comparações de matching, enquanto a predição de contraste zero é capaz de determinar se o fator contraste para um domain-block é zero ou não, calculando também a diferença entre o *range-block* e o *domain-block* transformado de maneira eficiente e exata.

Huang *et. al.* [36] propuseram um algoritmo também baseado em variância introduzindo dois parâmetros capazes de controlar a velocidade de codificação e a qualidade e em [37] é proposto um método de compressão fractal de imagens usando um vetor de características especial com o objetivo de classificar os *domain-blocks* da imagem num algoritmo bastante conveniente para a implementação em hardware.

Ao lado dessas tentativas de acelerar a codificação fractal pura, alguns codificadores híbridos têm também sido desenvolvidos. A relação entre fractal e codificadores baseados em transformada foi investigada em [8]-[42]. Davis [40][43][44] apresentou uma aproximação que usa elementos de ambos os tipos de compressão: fractal e *wavelet*. Em [41] argumenta-se que o mapeamento contrativo fractal poderia ser considerado como uma operação de predição do domínio *wavelet* e em [42] os autores fizeram o matching *domain-range* no domínio *wavelet* para obter uma compressão mais alta.

Li e Kuo [45] usaram o mapeamento contrativo fractal para predizer os coeficientes *wavelets* entre escalas e então codificaram o resíduo de predição com um codificador de plano de bits. Esse procedimento é diferente de [39] e [40] e de outros codificadores fractais convencionais, onde a imagem é codificada inteira por predição fractal.

A idéia por trás da maioria dos codificadores híbridos fractais-wavelet é aplicar a DWT (*Discrete Wavelet Transform* – Transformada Discreta Wavelet) à imagem e em seguida usar métodos fractais no domínio *wavelet* [5]. Sabe-se que a concentração de energia *wavelet* está localizada primariamente no canto superior esquerdo dos valores dos filtros passa-baixas, o que torna a subbanda de aproximação extremamente favorável à aplicação de técnicas fractais. Tal característica foi explorada em [46], no qual é proposto um codificador fractal acelerado que aplica a classificação de domínios de Fisher à subbanda passa-baixas da imagem transformada via wavelet e também um codificador SPIHT (*Set Partitioning in Hierarchical Trees*) modificado nos coeficientes remanescentes.

O número bastante restrito dessas publicações sobre Codificadores Híbridos Fractal-Wavelet e o fato de seus autores não explicitarem todos os parâmetros utilizados (dada a abrangência das duas técnicas) torna difícil a comparação via implementação desses codificadores, uma vez que existe quase uma impossibilidade de reproduzir seus resultados. Em contrapartida, os codificadores híbridos se tornam um vasto campo para a pesquisa.

V. Desempenho Comparativo entre Algoritmos

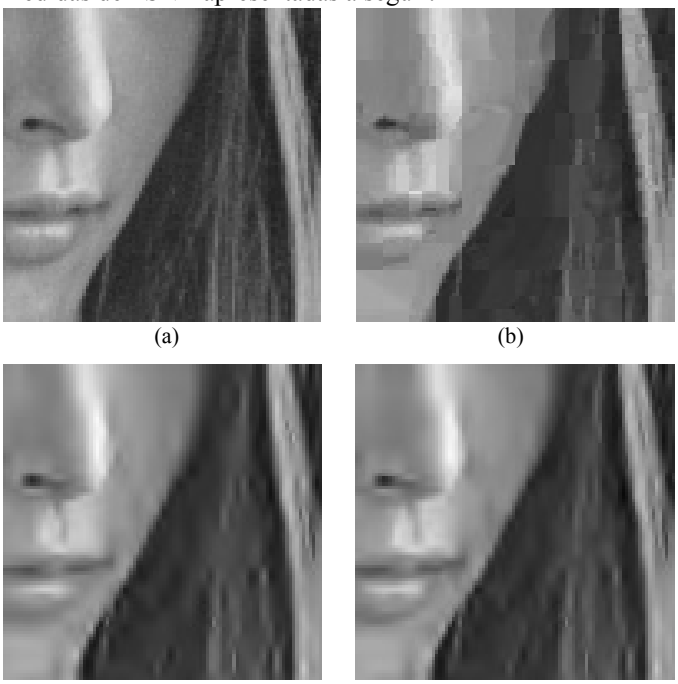
A título de ilustração, a seguir é realizada uma breve comparação entre algoritmos fractais puros e outras técnicas, incluindo técnicas híbridas (no caso, exemplificando técnicas mistas fractais e wavelets). Os codificadores fractais puros, wavelet puro, híbrido fractal-wavelet e JPEG2000 foram comparados em termos de qualidade visual, PSNR e tempo de codificação. O primeiro, o codificador fractal puro acelerado de Fisher [3], será referenciado como "QPIFS", enquanto a técnica wavelet pura será referenciada como "SPIHT". O codificador híbrido utilizado como referência ilustrativa foi extraído de [46], aonde se encontram maiores detalhes acerca das simulações.

Uma melhora significativa na qualidade subjetiva é freqüentemente observada em codificadores híbridos fractais-wavelets, evitando-se artefatos de "blurring" e blocagem. Observando a Fig. 6 (ampliação de Lena 512x512), pode-se notar que nas imagens reconstruídas pelo QPIFS artefatos de blocagem são altamente visíveis devido ao particionamento fractal por blocos.

Por outro lado, o esquema SPIHT causa evidente "blurring" em regiões como o cabelo e a boca de Lena, como conseqüência do *threshold* de quantização. O codificador híbrido usado no experimento apresenta melhoria significativa na qualidade visual da imagem reconstruída, combinando as vantagens de ambas as técnicas, evitando, desse modo, os

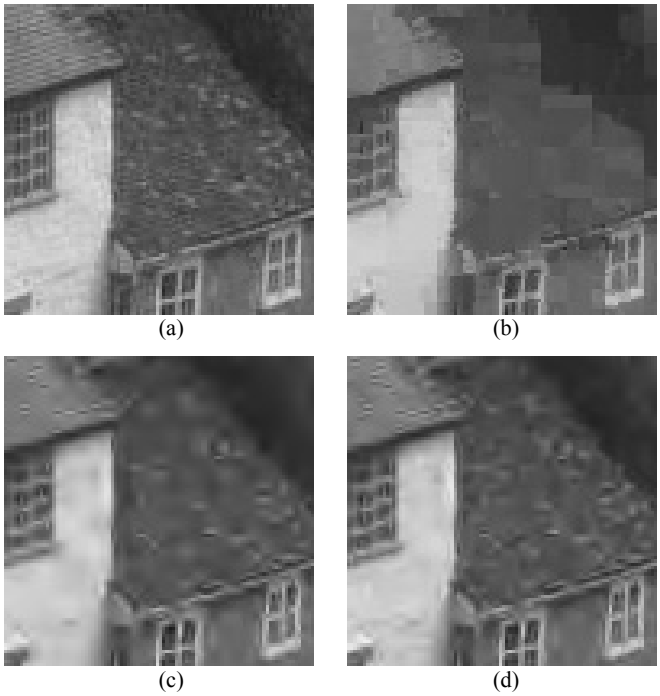
artefatos de compressão que cada técnica apresenta individualmente.

As mesmas conclusões podem ser observadas na Fig. 7, ampliação da imagem Goldhill 512x512. Resultados similares foram obtidos com o conjunto de imagens padrão em tons de cinza disponíveis no Waterloo Bragzone website[*1] a várias taxas de compressão, resultados estes corroborados nas medidas de PSNR apresentadas a seguir.



(a) (b) (c) (d)

Fig. 6 – Ampliações de Lena codificada a 0,32 bpp. (a)Original, (b)QPIFS, (c) SPIHT, (d) Híbrido



(a) (b) (c) (d)

Fig. 7 – Ampliações de Goldhill codificada a 0,32 bpp. (a)Original, (b)QPIFS, (c) SPIHT, (d) Híbrido

([*1]) http://links.uwaterloo.ca/bragzone

Com relação ao tempo de codificação, o codificador híbrido em [46] apresenta desempenho entre os esquemas SPIHT e QPIFS em todos os casos. O citado codificador híbrido apresentou uma média de 1,7 vezes o tempo de codificação do SPIHT, e o QPIFS apresentou uma média de 17 vezes o tempo de codificação do método híbrido. Dessa forma, o codificador híbrido ilustrado apresentou uma média de 94% de redução no tempo de codificação fractal puro, tempo este que pode ser consideravelmente melhorado caso sejam combinadas técnicas aceleradoras mais recentes.

O codificador JPEG2000 utilizado[*2] nas simulações, por se tratar de um software fechado, sem acesso ao código-fonte, impede comparações reais quanto ao tempo de processamento com os demais métodos. Contudo, comparações quanto às medidas de PSNR encontram-se na Fig. 8.

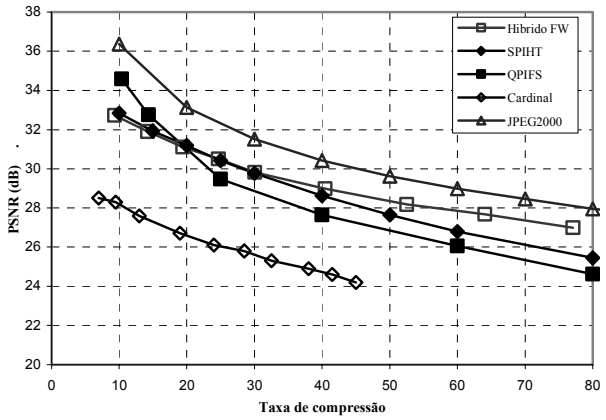


Fig. 8 – Curvas de taxa-distorção

Naturalmente o desempenho do JPEG2000 em PSNR e tempo de codificação é superior a todos os demais, por ser, mais do que um padrão já consolidado e exaustivamente testado, o principal padrão de codificação atual para imagens estáticas, fazendo uso de técnicas de compressão baseadas em tecnologia *wavelet* discreta em estado da arte e em codificação aritmética. Nenhum dos demais algoritmos utilizados na comparação contaram com codificação aritmética ou entrópica.

VI. CONCLUSÕES

É fundamental perceber que a combinação de técnicas fractais com outras técnicas, como as transformadas wavelets no exemplo apresentado, promovem um salto em desempenho na codificação fractal, tanto com relação à qualidade visual, como em medidas objetivas de PSNR e tempo de processamento, aproximando seu desempenho de padrões mais avançados e complexos como o JPEG2000.

Métodos fractais, devido ao tempo exaustivo de codificação são tradicionalmente mais convenientes para aplicações de arquivamento, tais como enciclopédias digitais, onde uma imagem é codificada uma vez e decodificada várias vezes. Este artigo procura destacar que a combinação com novas técnicas, tais como wavelets, novos codificadores entrópicos,

([*2]) http://www.luratech.com

técnicas de lifting otimizadas, hardware dedicado otimizado ou mesmo novos algoritmos de procura rápida podem permitir a extensão de técnicas fractais para tempo real.

VII. AGRADECIMENTOS

Este trabalho foi financiado pela FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo.

REFERÊNCIAS

[1] A.L. M. C.; Y. Iano – "Procedimentos Para Método Híbrido de Compressão de Imagens Digitais Utilizando Transformadas Wavelet e Codificação Fractal" – Dissertação de Doutorado, FEEC/UNICAMP, SP, Brasil, Maio, 2005.

[2] Special Issue on the H.264/AVC Video Coding Standard –*IEEE Transactions on CSVT*, vol.13, no.7, , Jul, 2003.

[3] Y. Fisher – "Fractal Image Compression, Theory and Application" – Ed. Springer-Verlag, New York Inc, USA, 1995.

[4] N. Lu – "Fractal Imaging" – Academic Press, San Diego, USA, 1997;

[5] S. T. Wealsted – "Fractal and Wavelet Image Compression Techniques" – Ed. SPIE Optical Engineering Press, Washington, USA, 1999.

[6] K. R. Rao; P. C. Yip – "The Transform and Data Compression Handbook" – Ed. CRC Press, LLC, USA, 2001.

[7] B. Mandelbrot – "The Fractal Geometry of Nature" – Ed. W. H. Freeman and Company, New York, USA, 1983.

[8] M. F. Barnsley; A. Sloan – "A better Way to Compress Images" – Byte, pp. 215-223, Jan, 1988.

[9] M. F. Barnsley; A. Sloan – "Method and Apparatus for Image Compression by Iterated Funcion System" – US Patent # 4.941.193, 1990.

[10] M. F. Barnsley; A. Sloan – "Method and Apparatus for Processing Digital Data" – US Patent # 5.065.447, 1991.

[11] A. Jacquin – "Image Coding Based on a Fractal Theory of Iterated Contractive Image Transformations" – IEEE Transactions on Image Processing, vol. 01, no. 01, pp. 18-30, 1992.

[12] M. Peruggia– "Discrete Iterated Function Systems" – Ed. A K Peters, Massachusetts, USA, 1993.

[13] J. E. Hutchinson – "Fractal and Self-Similarity" – Indiana University Mathematics Journal, Vol. 35, No. 5, 1981.

[14] M. F. Barnsley; A. Jacquin – "Applications of Recurrent Iterated Function Systems to Images" – SPIE Visual Communications and Image Processing , pp. 122-131, 1998.

[15] E. L. Lima – "Espaços métricos" - Ed. Projeto Euclides – Rio de Janeiro, Brasil, 1993.

[16] M. F. Barnsley– "Fractals Everywhere 2nd edition" – Ed. Morgan Kaufmann, Academic Press, San Diego, USA, 1993.

[17] A. Jacquin – "A Fractal Theory of Iterated Markov Operators with Applications to Digital Image Compression" – PhD. Thesis, Georgia Institute of Technology, USA, 1989.

[18] A. Bogdan; H. Meadows – "Kohonen Neural Network for Image Coding Based on Iteration Transformation Theory" – *Proceedings of SPIE*, vol. 1766, pp. 425-436, 1992.

[19] D. McGregor; R. J. Fryer; W. P. Cookshott; P. Murray – "Fast Fractal Transform Method for Data Compression" – *University of Strachclyde Research Report*, 94/156 [IKBS-17-94], 1994.

[20] D. Saupe – "Accelerating Fractal Image Compression by Multi-Dimensional Nearest Neighbor Search" –*Proceedings DCC'95 Data Compression Conference*, pp. 222-231, Mar 1995.

[21] J. Cardinal – "Fast Fractal Compression of Greyscale Images" – *IEEE Transactions on Image Processing*, vol. 10, no. 01, pp. 159-164, Jan., 2001.

[22] J. Kominek – "Algorithm for Fast Fractal Image Compression" – *Proceedings of SPIE Symp. Electronic Imaging: Science Technology*, vol. 2419, 1995.

[23] B. E. Wohlberg; G. de Jager – "Fast Image Domain Fractal Compression by DCT Domain Block Matching" – *Electron. Lett.*, vol. 31, pp. 869-870,1995.

[24] R. F. Sproull – "Refinements to Nearest Neighbor Searching in K-Dimensional Trees" – *Algorithmica*, vol. 6, pp. 579-589, 1991.

[25] I. Katsavounidis; C.-C. J. Kuo; Z. Zhang – "Fast Tree-Structured Nearest Neighbor Encoding for Vector Quantization" – *IEEE Transactions on Image Processing*, vol. 5, pp. 398-404, Fev, 1996.

[26] C. S. Tong; M. Wong – "Adaptative Approximate Nearest Neighbor Search for Fractal Image Compression" – IEEE Transactions on Image Processing, vol. 11, No.06, pp. 605-615, Jun, 2002.

[27] C. S. Tong; M. Pi – "Fast Fractal Image Encoding Based on Adaptative Search" – IEEE Transactions on Image Processing, vol. 10, No.09, pp. 1269-1277, Set, 2001.

[28] C. -C. Wang; C. –H. Hsieh – "An Efficient Fractal Image Coding Method Using Interblock Correlation Search" – IEEE Transactions on CSVT, vol. 11, No.02, pp. 257-261, Fev, 2001.

[29] B. B. Eqbal – "Enhancing the Speed of Fractal Image Compression" – *Opt. Eng*, vol. 34, No.06, pp. 1705-1710, Jun, 1995.

[30] J. Li; G. Chen; Z. Chi – "A Fuzzy Image Metric with Application on Fractal Coding" – *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 636-643, Jun, 2002.

[31] T. –K. Truong; J. -H, Jeng; I. S. Reed; P. C. Lee; A. Q. Li – "A Fast Encoding Algorithm for Fractal Image Compression Using DCT Inner Product" - *IEEE Transactions on Image Processing*, vol. 9, No.4, pp. 529-534, Abr, 2000.

[32] S. Lee; S. Ra – "An Analysis of Isometry Transforms in Frequency Domain for the Fast Fractal Coding" – *IEEE Signal Proc. Lett*, vol. 6, pp. 100-102, Maio, 1999.

[33] D. Saupe; H. Harteinstein – "Lossless Acceleration of Fractal Image Compression by Fast Convolution" – *Proc. IEEE Int. Conf. Image Processing*, vol. 1, pp. 185-188, Sep, 1996.

[34] Y. –G. Wu; M.-Z. Huang; Y.-L.Wen – "Fractal Image Compression with Variance and Mean" – *IEEE ICME*, vol. 1, pp. 353-356, 2003.

[35] C.-M. Lai; K.-M. Lam; W. –C. Siu– "A Fast Fractal Image Coding Based on Kick-Out and Zero Contrast Conditions" - *IEEE Transactions on Image Processing*, vol. 12, No.11, pp. 1398-1403, Nov, 2003.

[36] C. He; S. X. Yang; X. Huang– "Variance-Based Accelerating Scheme for Fractal Image Encoding" – *Eletronics Letters*, vol. 40, No.2, Jan, 2004.

[37] N. Rowshanbin; S. Samavi; S. Shirani – "Acceleration of Fractal Image Compression Using Characteristic Vector Classification" – *IEEE CCECE/CCGEI*, Ottawa, pp. 2057-2060, May, 2006.

[38] C. Caso; C. -C. J. Kuo – "New Results for Fractal/Wavelet Image Compression" – Proceedings of SPIE Visual Communications and Image Processing, vol. 2727, pp. 536-547, Mar, 1996.

[39] R. Rinaldo; G. Calvagno – "Image Coding by Block Prediction of Multiresolution Subimages" – IEEE Transactions on Image Processing, vol. 4, no. 07, pp. 909-920, Jul, 1995.

[40] G. M. Davis – "A Wavelet Based Analysis of Fractal Image Compression" – IEEE Transactions on Image Processing, vol. 7, no. 02, pp. 141-154, Fev, 1998.

[41] S. Asgari; T. Q. Nguyen; W. A. Sethares – "Wavelet-Based Fractal Transforms for Image Coding With no Search" – Proc IEEE International Conf. On Image Processing, ICIP97, 1997.

[42] D. Herbert; E. Soundarajan – "Fast Fractal Image compression With Triangulation Wavelets " – Proc. SPIE Conf. On Wavelets Applications in Signal and Image Processing VI, San Diego, USA, 1998.

[43] G. M. Davis – "Adaptative Self-Quantization of Wavelet Subtrees: A Wavelet-Based Theory of Fractal Image Compression " – Proc. SPIE Conf. On Wavelets Applications in Signal and Image Processing III, San Diego, USA, 1995.

[44] G. M. Davis – "Implicit Image Models for Fractal Image Compression " – Proc. SPIE Conf. On Wavelets Applications in Signal and Image Processing IV, Denver, USA, 1996.

[45] J. Li; C. –C. J. Kuo – "Image Compression With a Hybrid Wavelet-Fractal Coder" – IEEE Transactions on Image Processing, vol. 8, no. 06, pp. 868-874, Jun, 1999.

[46] Y. Iano, F. S. Silva, A. L.M. Cruz – "A Fast and Efficient Hybrid Fractal-Wavelet Image Coder" – IEEE Transactions on Image Processing, vol. 15, no. 1, pp. 98-105, Jan, 2006.

Ana Lúcia M. Cruz Nasceu em São Paulo, 1974. Concluiu o curso de Graduação em Engenharia Elétrica em 1998 e obteve o título de Mestrado em 2001 e de Doutorado em 2005, pela UNICAMP. Atualmente está no programa de Pós Doutoramento do Departamento de Comunicações da Faculdade de Engenharia Elétrica e Computação da Unicamp. Seus interesses incluem Processamento, compressão e codificação de Imagem e vídeo digitais, Televisão Digital e Comunicações por satélite.

Fernando S. da Silva Nasceu em São Paulo, 1973. Concluiu o curso de Graduação em Engenharia Elétrica em 1998 e obteve o título de Mestrado em 2001 e de Doutorado em 2005, pela UNICAMP. Atualmente está entrando no programa de pesquisador colaborador voluntário no Departamento de Comunicações da Faculdade de Engenharia Elétrica e Computação da Unicamp e leciona no Centro Universitário Salesiano de São Paulo (Unisal). Seus interesses incluem processamento, compressão e codificação de imagem e vídeo digitais, televisão digital e comunicações por satélite.

Yuzo Iano Nasceu em Lins, 1950. Concluiu o curso de Graduação em Engenharia Elétrica em 1972 e obteve o título de Mestrado em 1974 e de Doutorado em 1986, pela UNICAMP. Atualmente é Professor Adjunto no Departamento de Comunicações da Faculdade de Engenharia Elétrica e Computação da Unicamp e é responsável pelo Laboratório de Comunicações Visuais. Ele já desenvolveu projeto de processamento digital de sinais (áudio e imagem) em conjunto com o CPqD. Seus interesses incluem codificação de áudio e video, vídeo digital, transmissão digital de sinais, televisão digital e comunicações por satélite.

Roger F. L. Chavez concluiu o curso de Graduação em Engenharia de Sistemas na Universidade de San Agustín, Arequipa, Peru, em 2002. Recebeu seu título de mestre na Universidade Estadual de Campinas – UNICAMP – em 2007, especializado em reconhecimento de padrões. Atualmente, está no programa de doutorado no Laboratório de Comunicações Visuais do Depto. de Comunicações da Faculdade de Engenharia Elétrica e de Computação da UNICAMP. Seus interesses atuais concentram-se no desenvolvimento de um método eficiente para sincronismo de TV digital e códigos corretores de erro.

Esquema de Escalonamento de Fluxos de Dados Baseado nas Singularidades Locais do Tráfego Internet

Flávio Henrique Teles Vieira, Christian Jorge, & Lee Luan Ling

***Abstract*—The modern network traffic is composed of complex flows that present different statistical characteristics and quality of service requirements. This integration of flows motivates the introduction of new traffic congestion control and management schemes. In this work, we propose a new traffic flow scheduling scheme that incorporates the local singularity data of traffic processes. For this end, we initially verify the application of an algorithm that is based on the decay of wavelet coefficients in time windows to estimate the pointwise Hölder exponents. These exponents quantify the degree of traffic singularities. Next, we develop an adaptive prediction algorithm for the pointwise Hölder exponents of a traffic process. The simulations confirm that the proposed scheduling scheme provides lower data loss rate as well as higher link utilization than the GPS (Generalized Processor Sharing) scheme greatly used in network traffic routers.**

***Index Terms*— Network Traffic, Scheduling, Prediction, Multifractals, Hölder Exponent.**

***Resumo*—A integração de vários tipos de serviços nas redes de comunicações atuais traz consigo a necessidade de se introduzir novos esquemas de gerenciamento e controle de tráfego. Em alguns casos, os esquemas atuais podem ter sua eficiência reduzida devido ao comportamento repleto de diferentes singularidades locais dos fluxos de tráfego. Propomos neste artigo um esquema de escalonamento de fluxos de dados que utiliza informações da regularidade local de tráfego representada pelo expoente de Hölder pontual. Para tal, inicialmente verificamos a aplicação de um algoritmo de estimação dos expoentes de Hölder baseado no decaimento dos coeficientes wavelets em janelas de tempo. Em seguida, desenvolvemos um algoritmo adaptativo de predição dos expoentes de Hölder. Esse algoritmo de predição é incorporado na estratégia proposta de escalonamento de fluxos. As avaliações e simulações realizadas mostram que o esquema de escalonamento proposto provê uma menor perda de dados e uma utilização do enlace maior em comparação ao esquema de escalonamento GPS (Generalized Processor Sharing) convencional, implementado em muitos roteadores.**

***Palavras chave*—Tráfego de Redes, Escalonamento, Predição, Multifractais, Expoente de Hölder.**

Manuscrito recebido em 11 de Junho de 2007; revisado em 14 de Novembro de 2007.

F. H. T. Vieira (flavio@decom.fee.unicamp.br), C. Jorge (christian@fee.unicamp.br) e L. L. Ling (lee@decom.fee.unicamp.br) pertencem ao Departamento de Comunicações (DECOM) da FEEC (Faculdade de Engenharia Elétrica e de Computação) da UNICAMP. Av. Albert Einstein - 400 - 13083-852- Campinas - SP.

I. INTRODUÇÃO

Os serviços oferecidos pelas redes IP (Internet Protocol) atuais estão evoluindo do modelo de 'melhor-esforço' de transmissão da informação para um paradigma com várias classes de serviços, cada qual com seus requisitos de qualidade de serviço (QoS). Prover tais requisitos de qualidade de serviço usando a tecnologia IP é uma tarefa desafiadora para a engenharia de tráfego de redes.

Mecanismos preventivos são usados para alocar recursos com antecedência a fim de se evitar a ocorrência de congestionamentos. Nestes mecanismos, o algoritmo empregado de predição das características do tráfego desempenha um papel fundamental [2] [14]. Sabe-se que o desempenho de predição do tráfego depende de fatores tais como: quantidade de informações disponíveis do processo, a escala de tempo utilizada e o intervalo de predição. Um fator que tem grande influência no desempenho da predição é a própria natureza do tráfego [12] [14].

Na última década ocorreram mudanças significativas na compreensão do comportamento do tráfego de redes. Inicialmente houve a descoberta da propriedade de invariância à escala (*scaling*) do tráfego de pacotes [1]. Neste caso, modela-se o tráfego como um processo monofractal, cuja lei de escalas é determinada pelo parâmetro de Hurst. Entretanto em [6], verificou-se na verdade um comportamento multifractal do tráfego em escalas menores que algumas centenas de milissegundos. Esse comportamento se origina devido aos mecanismos do protocolo TCP/IP (Transmission Control Protocol) que fragmentam unidades de informação de uma camada de rede, em unidades menores na próxima camada [6].

Um processo multifractal é caracterizado por irregularidades (singularidades) locais mais acentuadas, com leis de escalas mais complexas do que se supunham os modelos monofractais. Neste contexto, costuma-se empregar o conceito de expoente de Hölder, oriundo da análise multifractal, para caracterizar a regularidade local do processo de tráfego, ou seja, o grau de rajadas presente nos dados [13].

A análise da regularidade local é importante para o gerenciamento e controle de congestionamento da rede por fornecer informações que podem tornar os esquemas de alocação de recursos mais eficientes. Com relação a isso, sabe-

se que um fluxo de tráfego de dados com alto grau de rajadas (grau alto de irregularidade) proporciona um menor aproveitamento dos recursos [1].

Assim sendo, este artigo contribui para o gerenciamento e controle preventivo do tráfego de redes através da proposta de uma nova disciplina de escalonamento de fluxos de dados em roteadores que leva em consideração as regularidades locais dos fluxos. Mostraremos que esta disciplina de escalonamento resulta em uma melhor distribuição da taxa de transmissão de um enlace para o escoamento dos fluxos presentes e uma menor perda de dados com relação à disciplina de escalonamento GPS (Generalized Processor Sharing).

O artigo está organizado da seguinte forma. Na seção II, caracterizamos a regularidade local de uma função utilizando a transformada wavelet. Na seção III, apresentamos um algoritmo para a estimação em janelas da regularidade local de fluxos de tráfego. A Seção IV é dedicada ao desenvolvimento de um novo algoritmo adaptativo de predição de séries temporais. Nesta mesma seção são analisados os erros de predição dos expoentes de Hölder. Na Seção V, propomos um esquema de escalonamento que incorpora a predição dos expoentes de Hölder pontuais como critério de alocação das taxas de transmissão de cada fluxo. Finalmente na Seção VI, apresentamos as conclusões obtidas.

II. Caracterização da Regularidade Local

A. Análise Wavelet

A transformada wavelet pode caracterizar o comportamento local e em escala de um sinal (processo) através de uma representação do mesmo no espaço e no domínio da freqüência. Conceitualmente, a transformada wavelet é um produto-convolução do sinal analisado com a wavelet-mãe ψ. Uma das ferramentas mais utilizadas da análise wavelet é o coeficiente wavelet. Os coeficientes wavelet $d_{j,k}$ são obtidos ajustando-se a wavelet-mãe ψ a uma determinada escala j e transladando-a até um ponto $2^j k$ do sinal, com $j,k \in Z$, ou seja [4]

$$d_{j,k} = 2^{-j} \int_{-\infty}^{\infty} f(x)\psi\left(2^{-j}x - k\right)dx \tag{1}$$

com $j,k \in Z$.

Pode-se dizer que com o aumento do valor de $d_{j,k}$ tem-se um aumento na variação do sinal no ponto $2^j k$ [6]. Esta propriedade é importante para a caracterização das singularidades de um sinal por meio do decaimento do valor absoluto dos coeficientes wavelet $d_{j,k}$ [9].

B. Expoente de Hölder Pontual

O expoente de Hölder pontual é capaz de descrever o grau de uma singularidade local, o que é interessante para a caracterização das rajadas de dados em redes de computadores. O expoente de Hölder pontual é definido da seguinte forma:

Definição 1 (Expoente de Hölder pontual): Seja α um número real estritamente positivo, K uma constante e $x_0 \in R$. A função $f : R \to R$ é $C^{\alpha}(x_0)$ se existir um polinômio P_n de grau $n < \alpha$ tal que

$$| f(x) - P_n(x - x_0) | \leq K | x - x_0 |^{\alpha} \tag{2}$$

O expoente de Hölder pontual α_p da função f em x_0 é definido como

$$\alpha_p(x_0) = \sup\{\alpha > 0 \mid f \in C^{\alpha}(x_0)\} \tag{3}$$

Note que o polinômio P_n pode ser encontrado mesmo se o desenvolvimento da função f em série de Taylor ao redor de x_0 não existir.

C. Espectro Multifractal

O espectro multifractal (ou espectro de singularidades) provê informações sobre quais singularidades ocorrem em um dado processo e quais singularidades predominam.

Definição 2 (Espectro multifractal): Seja f uma função: $[a,b] \to R$, a < b e seja $\alpha(x)$ o expoente de Hölder pontual de f em cada ponto $x \in [a,b]$. O espectro multifractal $D(\alpha)$ de f é definido como

$$D(\alpha) = d_H(\{x \mid \alpha(x) = \alpha\}) \tag{4}$$

em que d_H denota a dimensão de Hausdorff [5].

Uma maneira muito utilizada para a estimação do espectro multifractal de um sinal com suporte compacto consiste na aplicação da transformada de Legendre [5] [13]. Para isso, inicialmente estima-se a função de partição $S(q,j)$ do processo analisado. A função de partição $S(q,j)$ é calculada em termos dos coeficientes wavelets $d_{j,k}$ do sinal pela seguinte equação

$$S(q,j) = \sum_k | d_{j,k} |^q \tag{5}$$

Seja $\tau(q)$ a função estrutura definida como [5]

$$\tau(q) = \lim_{j \to -\infty} \frac{\log S(q,j)}{j \log 2} \tag{6}$$

Assim, o espectro multifractal $D(\alpha)$ do sinal analisado pode ser calculado como [13]

$$D(\alpha) = \tau^*(\alpha) \tag{7}$$

em que $\tau^*(\alpha)$ é a transformada inversa de Legendre da função estrutura $\tau(q)$, ou seja $\tau^*(\alpha) = \min_q(\alpha q - \tau(q))$.

A Figura 1 apresenta os expoentes de Hölder de um processo de tráfego Internet assim como o seu espectro multifractal de Legendre.

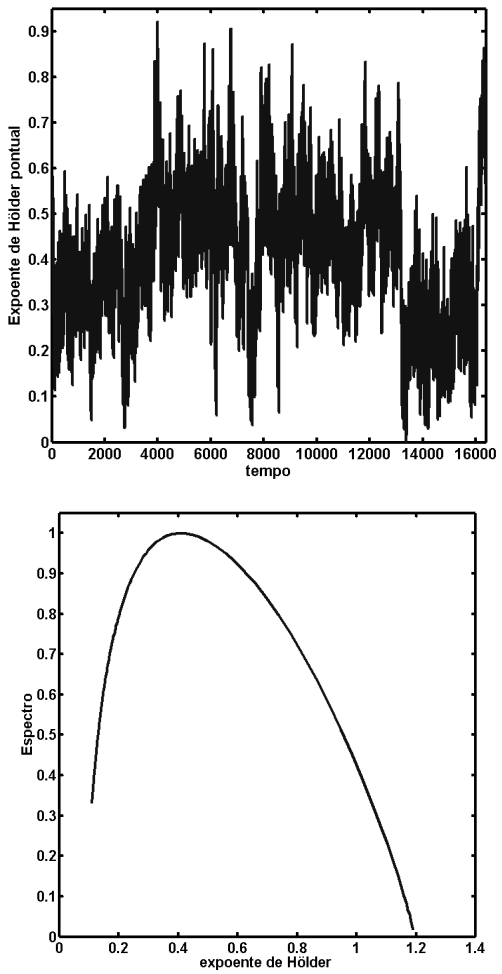


Fig. 1 - Acima: expoentes de Hölder pontuais referentes a um trecho da série de tráfego *lbl-pkt-5* na escala de tempo de 100 ms. Abaixo: espectro multifractal.

D. Tipos de Singularidades

Existem diferentes funções com um mesmo expoente de Hölder num determinado ponto, mas com comportamentos distintos ao redor desse mesmo ponto. Isso é devido ao tipo de singularidade que a função possui. Podemos destacar dois tipos [8] [19]:

Definição 3 (Singularidade não-oscilante): Seja $f_\alpha^{(-n)}$ a primitiva de *n*-ésima ordem da função f_α. Dizemos que f_α possui uma singularidade não-oscilante, com expoente de Hölder pontual α em x_0, se

$$\forall n \in N : f_\alpha^{(-n)} \in C^{\alpha+n}(x_0) \tag{8}$$

Desta forma, a função f_α deve ser regular o suficiente, para que dado um ponto x_0 com expoente de Hölder α, a derivada de f_α apresente expoente de Hölder α-1 em x_0 e a integral de f_α apresente expoente de Hölder α+1 no mesmo ponto. A função $f_\alpha(x) = |x - x_0|^\alpha$ representa o exemplo mais simples de singularidade não-oscilante em x_0.

Definição 4 (Singularidade oscilante): seja $g_{\alpha,\beta}^{(-n)}$ a primitiva de n-ésima ordem da função $g_{\alpha,\beta}$. Pode-se dizer que $g_{\alpha,\beta}$ possui uma singularidade oscilante em x_0, com expoente de Hölder pontual α e expoente de oscilação $\beta > 0$, se

$$\forall n \in N : g_{\alpha,\beta}^{(-n)} \in C^{\alpha+n(1+\beta)}(x_0) \tag{9}$$

Este tipo de singularidade ocorre em funções cujo expoente de Hölder pontual em x_0 aumenta mais que a unidade quando integradas. Isto se deve aos efeitos da suavização das oscilações presentes, causados pela operação de integração da função [9]. A função $g_{\alpha,\beta}(x) = |x - x_0|^\alpha \, sen\left(\frac{1}{|x - x_0|^\beta}\right)$ representa o exemplo mais simples de singularidade oscilante em x_0.

E. Decaimento dos Coeficientes Wavelets

O expoente de Hölder *h* de uma função no ponto x_0 pode ser aproximado usando $|d_{a,b}| \sim a^h$, com os coeficientes wavelets máximos absolutos $|d_{a,b}|_{b \in Z}$ imersos num cone do tipo $|x - x_0| \leq Ka$, onde *K* é uma constante [8][18]. Porém, esse tipo de análise só é condizente para singularidades não-oscilantes [8].

Para tráfego de redes, devido à alta irregularidade presente, devemos levar em consideração também as singularidades oscilantes que possuem expoente de oscilação $\beta > 0$. Nesse caso, os coeficientes wavelets máximos absolutos estão imersos num cone parabólico mais largo que o citado anteriormente, cuja equação é dada por $|x - x_0|^{1+\beta} \leq Ka$ [9].

Dessa forma, para uma caracterização correta do comportamento local de um processo, devemos levar em consideração não apenas a amplitude dos coeficientes wavelets, mas também sua localização no tempo. Assim, utiliza-se a seguinte desigualdade [3]:

$$|d_{a,b}| \leq K(a + |x - x_0|)^h \tag{10}$$

Com base nessa desigualdade, Seuret et al. [15] propuseram um algoritmo para estimação dos expoentes de Hölder pontuais tanto para singularidades oscilantes quanto não-oscilantes. Na próxima seção descreveremos como os fundamentos deste método podem ser aplicados para se realizar a estimação dos expoentes de Hölder em janelas de tempo.

III. ESTIMAÇÃO DA REGULARIDADE LOCAL

Nesta seção, apresentamos um método para estimar a regularidade local de um processo de tráfego.

Seja um sinal amostrado contendo 2^n amostras na escala de tempo *j* onde quanto maior o seu valor, maior a escala considerada.

A estimação do expoente de Hölder pontual para uma amostra k_0 pode ser feita considerando os seguintes passos [15]:

Passo 1) Construa, em um mesma figura, para cada $0 < j \le n$, a seguinte curva paramétrica (com parâmetro $k \le 2^n$):

$$x_j(k) = \log_2(2^j + |2^j k - x_0|) \quad (11)$$

$$y_j(k) = \log_2(|d_{j,k}|) \quad (12)$$

Passo 2) Encontre todas as retas *D*: $y = \alpha x + C$ que satisfaçam as duas restrições a seguir:

1. *D* está acima de todos os pontos $(x_j(k),y_j(k))$, ou seja:

$$\forall j, \forall k, \quad y_j(k) \ge \alpha x_j(k) + C \quad (13)$$

2. *D* toca uma das curvas paramétricas, ou seja, existe uma seqüência de pares (j_m, k_m) tal que

$$\lim_{m\to\infty} y_{j_m}(k_m) - (\alpha x_{j_m}(k_m) + C) = 0 \quad (14)$$

Passo 3) Calcule α_{max} o maior coeficiente angular encontrado em todas as retas *D* satisfazendo (13) e (14). O coeficiente α_{max} é o expoente de Hölder pontual do sinal para a amostra k_0.

Note que a estimação do expoente de Hölder pontual para o instante amostral k_0 se dá inicialmente construindo-se uma "nuvem" de pontos $(x_j(k),y_j(k))$. O expoente de Hölder para o instante de tempo k_0 corresponde ao coeficiente angular da reta que se encaixa tão precisamente quanto possível no topo dessa "nuvem" [15]. Com relação ao número de pontos na obtenção da nuvem de pontos, utiliza-se a restrição $3 \le x(j,k) \le \log_2(2^n) - 2$, sendo 2^n o número de amostras do processo. Isto permite que seja levado em conta os coeficientes wavelets máximos absolutos em uma extensão suficiente de expoentes de oscilação β e gerar uma quantidade de pontos $(x_j(k), y_j(k))$ tal que seja possível uma construção precisa da reta que toca o topo desta "nuvem". Para tal, utilizamos neste trabalho a função Morlet como wavelet-mãe [4].

A Figura 2 mostra o sinal $|x|^{0,7}\, sen\left(\frac{1}{|x|^{1,2}}\right)$ que apresenta uma singularidade oscilante, sua respectiva "nuvem de pontos" e a reta que toca o topo da nuvem.

A. Estimação do Expoente de Hölder Pontual em Janelas de Tempo

Nesta seção, propomos uma estratégia dinâmica para a estimação do expoente de Hölder pontual para um dado sinal. Esta estratégia se baseia na utilização de uma quantidade fixa de amostras consecutivas (janela de tempo) para a estimação do expoente de Hölder pontual relativo a cada amostra do sinal. Para cada amostra do processo estimamos o seu respectivo expoente de Hölder utilizando somente as amostras da janela ao invés de todas as amostras do processo. Dessa forma, utiliza-se uma menor quantidade de amostras (assim como uma menor quantidade de coeficientes wavelet) para a estimação do expoente em cada instante de tempo.

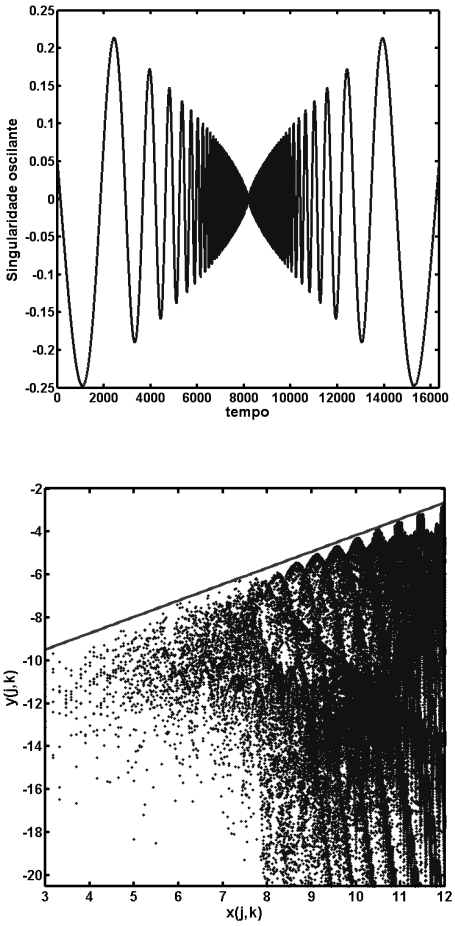


Fig.2. Acima: função $|x|^{0,7}\text{sen}(1/|x|^{1,2})$; Direita: "nuvem de pontos" associada ao ponto de singularidade oscilante da função. Abaixo: O coeficiente angular da reta é a estimativa do expoente de Hölder pontual.

A diminuição da quantidade de coeficientes wavelet utilizada em cada estimação possibilita um processamento mais rápido das informações, podendo ser usada em roteadores e *switches* como indicadores das condições de tráfego. Comparamos então os expoentes de Hölder pontuais estimados por meio de janelas de tempo com os expoentes de Hölder pontuais de referência. Definimos este último como sendo os expoentes de Hölder pontuais estimados utilizando-se todas as amostras disponíveis de um processo de tráfego (sem uso de janelas de tempo).

Utilizamos nas simulações processos que consistem das amostras de traços de tráfego Internet como: dec1-pkt, dec2-pkt e lbl5-pkt nas escalas de tempo de 100 e 200 ms. Estes traços contêm a quantidade de bytes transmitida em cada intervalo de tempo. Nas simulações, três tamanhos de janelas foram considerados: janela 12 (2^{12} = 4096 instantes de tempo), janela 13 (8192 instantes de tempo) e janela 14 (16384 instantes de tempo).

Como medida de desempenho da estimação dos expoentes de Hölder, utilizamos o conhecido erro quadrático médio normalizado (EQMN), definido como [10][16]:

$$EQMN = \frac{E[(h_{ref} - h_{jan})^2]}{\sigma^2{}_{h_{ref}}} \quad (15)$$

onde h_{jan} é o expoente de Hölder pontual estimado via janelamento, h_{ref} é o expoente de Hölder pontual usado como referência e $\sigma^2{}_{h_{ref}}$ é a variância dos expoentes de Hölder pontuais de referência.

TABELA I
ERRO QUADRÁTICO MÉDIO NORMALIZADO

	Janela 12	Janela 13	Janela 14
série dec-pkt-1	0,4899	0,2483	0,0426
série lbl5	0,5423	0,4332	0,2116

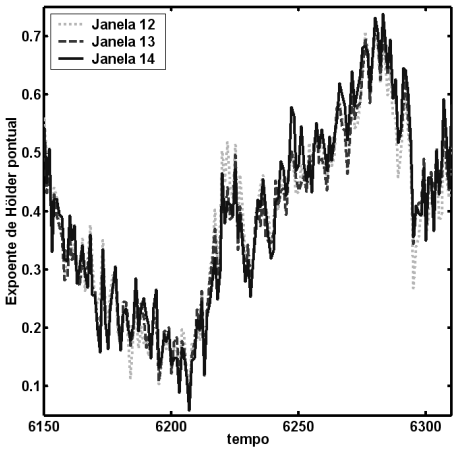


Fig.3. Expoentes de Hölder pontuais de amostras da série lbl-pkt-5 na escala de 100ms, estimados com três tamanhos diferentes de janelas de tempo.

Conforme mostram a Tabela I e a Figura 3, verifica-se que ao se diminuir o tamanho da janela, aumenta-se a imprecisão da estimação do expoente de Hölder pontual em relação aos expoentes de Hölder de referência. Em contrapartida, a estimação se dá de maneira mais rápida, devido à diminuição da quantidade de coeficientes wavelets relativos a cada janela de tempo. Além disso, os erros obtidos apontam que as regularidades locais do tráfego de redes podem ser de fato estimadas adaptativamente usando o algoritmo de estimação do expoente de Hölder pontual em janelas.

IV. PREDIÇÃO DO EXPOENTE DE HÖLDER PONTUAL

Pode-se afirmar que a série temporal formada pelos expoentes de Hölder é mais apropriada para se realizar predição de suas amostras futuras do que a série correspondente de tráfego propriamente dita. Esta afirmação se fundamenta no decaimento da função de autocorrelação das séries formadas pelos expoentes de Hölder, que é mais simples de ser tratada pelos algoritmos de predição. A Figura 4 apresenta a função de autocorrelação das amostras de um processo de tráfego e a série correspondente de expoentes de Hölder pontuais. Pode-se notar um decaimento assintótico mais lento da função de autocorrelação do processo de tráfego, confirmando sua propriedade de dependência de longo prazo [12]. Entretanto, há um decaimento mais acelerado das funções de autocorrelação das séries de expoentes de Hölder pontuais, o que indica que tais séries não apresentam dependência de longo prazo. Isto sugere a possibilidade do uso de técnicas mais simples de predição de séries temporais, tais como o filtro de Mínimos Médios Quadrados (Least-Mean Squares - LMS) e o filtro de Kalman [7] para a predição dos valores dos expoentes de Hölder.

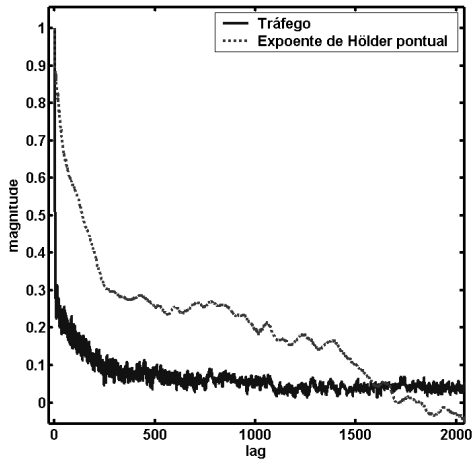


Fig. 4 – Função de autocorrelação das amostras de uma série de tráfego e de seus respectivos expoentes de Hölder pontuais. Esquerda: série lbl-pkt-5 na escala de tempo de 100 ms.

Apresentaremos na próxima seção um algoritmo adaptativo de predição aplicado à predição dos valores dos expoentes de Hölder estimados em janelas de tempo. Pretende-se portanto, estimar com antecedência as intensidades dos surtos de rajadas de cada fluxo de tráfego, disponibilizando essa informação para os mecanismos de controle de tráfego.

A. Algoritmo de Predição com Estimação Adaptativa dos Ruídos do Sistema

Um dos algoritmos adaptativos de predição mais usados é o filtro de Kalman. O filtro de Kalman é um estimador recursivo dos estados de um processo, muito utilizado em várias aplicações do mundo real [7]. Com base nestas equações recursivas de estimação, introduzimos um preditor que apresenta uma nova estratégia de atualização adaptativa dos ruídos do sistema, sendo o mesmo preciso para predições dos expoentes de Hölder.

Seja o modelo de sistema descrito pelas seguintes equações no espaço de estado [7]:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \boldsymbol{\eta}_1(k) \quad (16)$$

$$y(k) = \mathbf{u}^{\mathrm{T}}(k)\mathbf{w}(k) + \eta_2(k) \quad (17)$$

Nestas equações, $w(k)$ é o estado do sistema, $y(k)$ é a saída medida, $\mathbf{u}(k) = [u(k), u(k-1), ..., u(k-N)]^T$ é o vetor de entradas do sistema, N denota a ordem do filtro e k é o instante

de tempo discreto, com $k \in Z$. O vetor $\boldsymbol{\eta}_1(k)$ é conhecido como ruído de processo, o qual se assume como sendo um processo gaussiano com média zero e covariância $Q(k)$ dada por:

$$\mathbf{Q}(k) = E[\boldsymbol{\eta}_1(k)\boldsymbol{\eta}_1^{\mathrm{T}}(k)] \tag{18}$$

De modo semelhante, a variável $\eta_2(k)$ é conhecida como ruído de medida e modelada como um processo gaussiano com média zero e variância $P(k)$ dada por:

$$P(k) = E[\eta_2^{\,2}(k)] \tag{19}$$

Para que seja possível aplicar esse modelo de espaço de estados, é necessário que a covariância $Q(k)$ e a variância $P(k)$ sejam conhecidas. Na prática, é comum encontrar sistemas nos quais $Q(k)$ e $P(k)$ são desconhecidos, ou parcialmente conhecidos. Pode-se verificar que a escolha destes parâmetros influencia no desempenho do preditor. Dessa forma, propomos um método de estimação adaptativa destes parâmetros.

Seja $w(k)$ o vetor de coeficientes de um filtro transversal, o vetor $\boldsymbol{\eta}_1(k)$ pode ser visto como o ajuste dos coeficientes do algoritmo NLMS (Normalized Least-Mean Squares) [7]. Assim,

$$\boldsymbol{\eta}_1(k) = \frac{\tilde{\mu}}{\| \mathbf{u}(k) \|^2}[\mathbf{u}(k)e(k)] \tag{20}$$

em que $\tilde{\mu}$ é o passo de adaptação e $e(k)$ é o erro de estimação. Pode-se dizer que estamos usando um filtro NLMS para se estimar o ruído de processo $\boldsymbol{\eta}_1(k)$.

Segundo a equação (17), o ruído de medida $\eta_2(k)$ é dado por:

$$\eta_2(k) = y(k) - \mathbf{u}^T(k)\mathbf{w}(k) = e(k) \tag{21}$$

Objetivamos o cálculo adaptativo dos ruídos de predição. Então, utilizamos as seguintes equações recursivas para a média $\overline{\eta}_2(k)$ e a variância $r(k)$ de $\eta_2(k)$ [17]:

$$\overline{\eta}_2(k) = \frac{k-1}{k}\overline{\eta}_2(k-1) + \frac{1}{k}\eta_2(k) \tag{22}$$

$$r(k) = \frac{k-1}{k}r(k-1) + \frac{1}{k}(\eta_2(k) - \overline{\eta}_2(k))^2 \tag{23}$$

A matriz de covariância $Q(k)$ pode ser estimada através dos valores de variância do ruído de processo $\boldsymbol{\eta}_1(k)$. Seja o vetor $\boldsymbol{\eta}_1(k) = [\eta_{1,1}(k), \eta_{1,2}(k), ..., \eta_{1,N}(k)]^{\mathrm{T}}$. Cada $\eta_{1,j}(k)$, onde j = 1, 2, ..., N, possui variância $q_j(k)$ que pode ser calculada recursivamente por:

$$q_j(k) = \frac{k-1}{k}q_j(k-1) + \frac{1}{k}(\eta_1(k) - \overline{\eta}_1(k))^2 \tag{24}$$

Dessa forma, a matriz de covariância $Q(k)$ é dada por:

$$\mathbf{Q}(k) = \begin{bmatrix} q_1(k) & \cdots & 0 & 0 \\ \vdots & q_2(k) & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_N(k) \end{bmatrix} \tag{25}$$

Tomando como base as equações recursivas do filtro de Kalman, apresentamos a seguir as equações que regem o preditor proposto cuja saída é $y(k+1)$:

Algoritmo de Predição:

Passo 1) Calcula-se o ganho de Kalman $\mathbf{K}(k)$:

$$\mathbf{K}(k) = \frac{\mathbf{P}(k)\mathbf{u}(k)}{\mathbf{u}^{\mathrm{T}}(k)\mathbf{P}(k)\mathbf{u}(k) + r(k)} \tag{26}$$

Passo 2) Atualiza-se os pesos:

$$\hat{\mathbf{w}}(k+1) = \hat{\mathbf{w}}(k) + \mathbf{K}(k)[y(k) - \mathbf{u}^{\mathrm{T}}(k)\hat{\mathbf{w}}(k)] \tag{27}$$

Passo 3) Calcula-se a variância $P(k)$:

$$\overline{\eta}_2(k) = \frac{k-1}{k}\overline{\eta}_2(k-1) + \frac{1}{k}\eta_2(k) \tag{28}$$

$$P(k) = \frac{k-1}{k}P(k-1) + \frac{1}{k}(\eta_2(k) - \overline{\eta}_2(k))^2 \tag{29}$$

Passo 4) Calcula-se a matriz $Q(k)$:

$$q_j(k) = \frac{k-1}{k}q_j(k-1) + \frac{1}{k}(\eta_1(k) - \overline{\eta}_1(k))^2 \tag{30}$$

$$\mathbf{Q}(k) = \begin{bmatrix} q_1(k) & \cdots & 0 & 0 \\ \vdots & q_2(k) & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_N(k) \end{bmatrix} \tag{31}$$

Passo 5) Atualiza-se $P(k)$:

$$\mathbf{P}(k+1) = \mathbf{P}(k) + \mathbf{Q}(k) - \mathbf{K}(k)\mathbf{u}^{\mathrm{T}}(k)\mathbf{P}(k) \tag{32}$$

Passo 6) A saída do algoritmo de predição é dada por:

$$\hat{y}(k+1) = \mathbf{u}^{\mathrm{T}}(k+1)\hat{\mathbf{w}}(k+1) \tag{33}$$

em que $K(k)$ é chamado de ganho de Kalman e $\hat{\mathbf{w}}(k+1)$ é a estimação do próximo estado do sistema.

B. Avaliação do Preditor Proposto

Avaliamos o desempenho do preditor proposto através de simulações utilizando os expoentes de Hölder pontuais estimados com as janelas 11 (2^{11} amostras), 12 (2^{12} amostras) e 13 (2^{13} amostras) para as séries de tráfego Internet dec-pkt-1, dec-pkt-2 e lbl-pkt-5.

Com relação à configuração do preditor proposto, adotou-se o valor de N=7 para a ordem do preditor; valor acima do qual

o EQM de predição em geral decai infimamente para as séries consideradas. Por meio de simulações, constatamos que o valor de $\tilde{\mu}$ que resultou em um melhor comportamento do preditor proposto foi de $\tilde{\mu}$ = 0,02. Além disso, as condições iniciais consideradas foram:

$$\hat{w}(0) \cong 0 \tag{34}$$

$$K(0) = \hat{w}(0)\hat{w}(0)^T \tag{35}$$

Comparamos o desempenho do algoritmo de predição com o desempenho de outros preditores bastante utilizados na literatura. Consideramos como critério para avaliação do desempenho de predição, duas versões do erro quadrático médio normalizado (EQMN), definidos pelas seguintes equações [16]:

$$EQMN1 = \frac{E[(h_{pred} - h_{jan})^2]}{\sigma^2{}_{h_{jan}}} \tag{36}$$

$$EQMN2 = \frac{E[(h_{pred} - h_{jan})^2]}{E[(h_{ua} - h_{jan})^2]} \tag{37}$$

onde h_{jan} é o expoente de Hölder pontual estimado via janelamento, h_{pred} é o expoente de Hölder predito, $\sigma^2{}_{h_{jan}}$ é a variância dos expoentes de Hölder estimados e h_{ua} é o expoente de Hölder anterior a h_{jan}.

Basicamente, o EQMN nos fornece uma comparação entre o EQM do preditor avaliado com o EQM de um preditor mais simples. No cálculo do EQMN1 considera-se a média amostral do sinal analisado como preditor mais simples. Enquanto no cálculo do EQMN2, compara-se o EQM do preditor avaliado com o EQM de um preditor que considera a última amostra disponível do sinal como valor predito. É desejável que um preditor a ser empregado possua ambos EQMNs menores que a unidade. Caso contrário, pode-se concluir que seu desempenho será, na melhor das hipóteses, similar aos preditores mais simples citados.

Além do preditor proposto, os outros preditores avaliados são o filtro NLMS (preditor 1) e o filtro de Kalman (preditor 2). A ordem do filtro para os preditores 1 e 2 é a mesma para o preditor proposto (N=7). O valor do passo de adaptação $\tilde{\mu}$ do preditor 1 é escolhido de forma a obter o menor erro quadrático médio possível para cada série de tráfego analisada.

TABELA II - EQMNS PARA PREDIÇÃO DE AMOSTRAS DA SÉRIE DE TRÁFEGO *DEC-PKT-1* NA ESCALA DE TEMPO DE 100 MS E DE SEUS EXPOENTES DE HÖLDER PONTUAIS.

	EQMN1	EQMN2
série *dec-pkt-1*	0,7474	0,7129
expoentes de Hölder pontuais (janela 13)	0,3246	0,6317
expoentes de Hölder pontuais (janela 12)	0,3757	0,6347
expoentes de Hölder pontuais (janela 11)	0,3991	0,6350

Observando a Tabela II, pode-se constatar um melhor desempenho de predição para o preditor proposto em relação aos demais analisados. De fato, além de possuir EQMNs menores que a unidade, o preditor proposto também possui EQMNs menores do que aqueles referentes aos preditores 1 e 2.

Outro fator importante na verificação da eficiência de um preditor é o decaimento do erro quadrático com o tempo [2]. O decaimento temporal do erro quadrático de predição para a série lbl-pkt-5 é mostrado na Figura 5. Esta figura nos mostra que o preditor proposto possui o decaimento mais íngreme deste erro, em comparação aos outros preditores analisados. Com poucas amostras iniciais, o erro quadrático do preditor proposto decai e se mantém abaixo dos outros preditores.

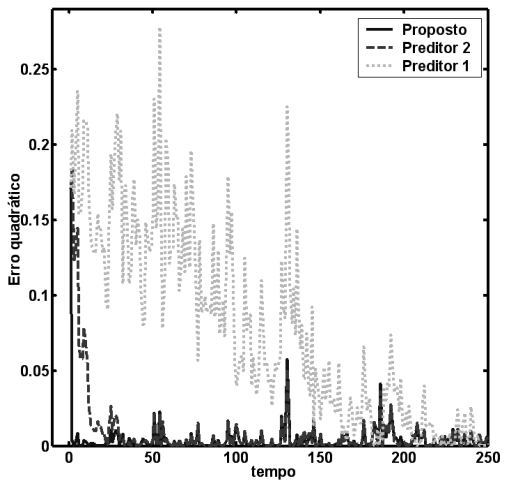


Fig. 5 - Erro quadrático de predição dos expoentes de Hölder de amostras da série *lbl-pkt-5* nas escalas de tempo de 100 ms.

V. DISCIPLINA DE ESCALONAMENTO UTILIZANDO OS EXPOENTES DE HÖLDER PONTUAIS

As disciplinas de escalonamento são úteis em pontos de multiplexação de dados pois permitem que vários fluxos ou sessões compartilhem de uma melhor forma a taxa de transmissão (capacidade) de um enlace. Dentre estas disciplinas, uma das mais conhecidas é a *Generalized Processor Sharing* (GPS) [11]. Uma de suas propriedades mais importantes é a proteção aos fluxos de dados. Esta propriedade refere-se ao isolamento dos fluxos de entrada do escalonador, não permitindo que um fluxo mal-comportado afete o desempenho dos demais fluxos.

Nesta seção propomos um novo esquema de escalonamento de fluxos de tráfego Internet tendo como base o GPS. A inovação deste esquema está na utilização do expoente de Hölder pontual como parâmetro de decisão da prioridade de cada fluxo em cada intervalo de tempo considerado. Nosso objetivo é obter uma melhor distribuição da taxa de transmissão do enlace aos fluxos e como conseqüência uma menor perda de dados.

A. Esquema de Escalonamento de Fluxos de Tráfego Utilizando Predição dos Expoentes de Hölder

O GPS é uma disciplina de escalonamento onde *n* fluxos de entrada compartilham um servidor de taxa fixa *c*. Um conjunto

de parâmetros $\{\varphi_i\}_{1\le i\le n}$ determina a proporção da taxa de serviço que cada fluxo recebe. Cada fluxo *i* de entrada recebe ao menos $c_i = \left(\varphi_i / \sum_{j=1}^{n} \varphi_j \right) c$, que é sua taxa garantida. Caso algum fluxo esteja ocioso em um determinado instante, sua taxa residual é distribuída proporcionalmente aos fluxos ativos.

Considere *n* fluxos de entrada, modelados como fluidos infinitamente divisíveis e cada qual com sua ponderação φ_i, onde $\sum_{i=1}^{n} \varphi_i = 1$. Cada fluxo *i* ($1 \le i \le n$) possui sua própria fila (Figura 6) e parâmetro $\varphi_i(t)$. As duas regras que definem a disciplina de escalonamento GPS são as seguintes:

-o mecanismo de escalonamento é conservativo;

-caso o fluxo *i*, durante o intervalo de tempo [*t*, *t*+*Δt*], esteja com dados no *buffer* (*backlogged*), então

$$\frac{S(i,t,t+\Delta t)}{S(j,t,t+\Delta t)} \ge \frac{\varphi_i(t)}{\varphi_j(t)} \tag{38}$$

para todo *j* = 1, 2, …, *n*. O termo *S*(*i*, *t*, *t*+Δt) é a quantidade de dados do fluxo *i* atendida no intervalo [*t*, *t*+*Δt*). Os parâmetros $\{\varphi_i(t)\}_{1\le i\le n}$, que são reais e positivos, são conhecidos como atribuições GPS (*GPS assignments*) e determinam o grau de prioridade do fluxo *i* em função do tempo *t* e são válidos para o intervalo de tempo [*t*, *t*+*Δt*). Para cada fluxo *i*, sua taxa de serviço garantida g_i é dada por

$$g_i = \frac{tm_i}{\sum_{j=1}^{n} tm_j} c \tag{39}$$

em que tm_i é a taxa média do fluxo *i* e *c* é a taxa total de serviço. Esta taxa garantida é importante para evitarmos o fenômeno da negação de serviço (*starvation*) a fluxos com menores prioridades. Nesse caso, um fluxo mais prioritário monopolizaria a taxa total de serviço. Caso algum fluxo *i* não esteja mais com dados no *buffer* no intervalo de tempo *Δt* após a definição de $\varphi_i(t)$, sua taxa de serviço remanescente é redistribuída para algum fluxo ainda com dados no *buffer* e com maior prioridade em relação a outros na mesma situação. Isto se repete até a utilização total da soma das taxas remanescentes ou até o momento que todos os fluxos não possuam mais dados a serem transmitidos neste intervalo. Generalizando, sempre há a determinação dos parâmetros de prioridade $\varphi_i(t + k\Delta t)$ dos fluxos *i* para o intervalo de tempo $[t + k\Delta t, t + (k+1)\Delta t[$, com $k \in Z$.

Nossa proposta consiste de se obter os parâmetros φ_k de distribuição das taxas residuais dos fluxos de maneira adaptativa e que reflita as necessidades instantâneas de cada fluxo. Para isso utilizamos o expoente de Hölder pontual α para a definição dos parâmetros φ_k em cada instante de tempo, ou seja:

$$\frac{S(i,t,t+\Delta t)}{S(j,t,t+\Delta t)} \ge \frac{\alpha_i(t)}{\alpha_j(t)} \tag{40}$$

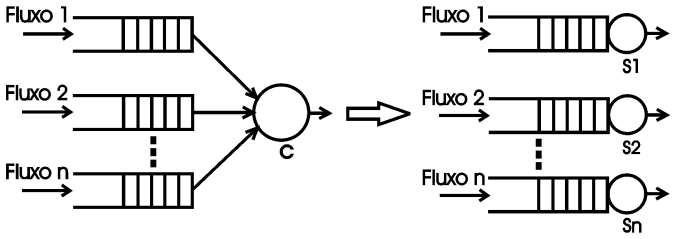


Fig.6. Decomposição e isolamento das filas num sistema GPS com *N* fluxos

O esquema completo do escalonador proposto está ilustrado na Figura 7, que é composto dos módulos de estimação e predição dos expoentes de Hölder e do módulo de escalonamento relativo à alocação de banda para cada fluxo. Novamente realizamos simulações para avaliar o esquema de escalonamento em questão utilizando-se os traços dec-pkt-1, dec-pkt-2 e lbl-pkt-5. Nas simulações, foram considerados diferentes tamanhos de escala de tempo (100 e 200 ms) e de janelas para estimação dos expoentes de Hölder pontuais (janelas 11, 12, e 13), assim como diferentes valores para a capacidade do enlace e tamanhos de *buffers*.

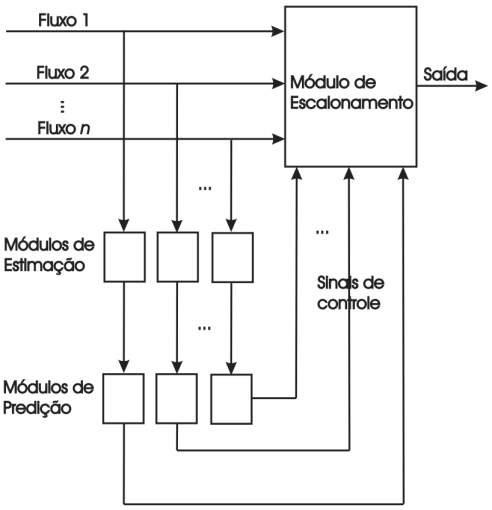


Fig.7 - Esquema completo do escalonador proposto.

Com objetivo de comparação, analisamos outros critérios para a definição dos parâmetros de prioridade $\varphi(t)$ de cada fluxo:

1 - $\varphi_i(t)$ diretamente proporcional à taxa média de chegada de cada fluxo (definido como GPS convencional).

2 - $\varphi_i(t)$ diretamente proporcional à taxa de chegada predita de cada fluxo, para o próximo instante de tempo.

Assim como para os expoentes de Hölder, verificamos através de simulações que os resultados da aplicação do preditor proposto para as séries de tráfegos consideradas são adequados. Por isso, em relação ao critério 2, utilizamos

também o preditor proposto para a predição a um passo da taxa de chegada de cada fluxo.

O intervalo de alocação de uma nova taxa para cada fluxo equivale ao intervalo de tempo entre duas amostras (100 ou 200ms). Neste ínterim, alguns fluxos podem ter uma taxa excessiva às suas necessidades momentâneas, sendo esta taxa redistribuída segundo os níveis de prioridade de cada fluxo. Outro fato que pode ocorrer é o de alguns fluxos terem uma taxa insuficiente às suas necessidades naquele intervalo, mesmo após a realocação das taxas excedentes oriundas de outros fluxos. Neste caso, há a ocorrência de perda de pacotes quando o tamanho do *buffer* não é estipulado adequadamente.

B. Resultados das Simulações com o Esquema de Escalonamento Proposto

As Tabelas III, IV e V resumem os resultados obtidos nas simulações realizadas na escala de tempo de 100 ms, escala na qual os traços de tráfego são multifractais [6]. Nestas simulações, os tamanhos do *buffer* são 16 Kbytes e 25 Kbytes e os valores da taxa de transmissão do enlace são 2,25 Mbps e 2,75 Mbps. Cada tabela retrata a porcentagem de perda de bytes, o ganho de desempenho sobre o GPS convencional e a utilização do enlace, em função dos critérios de escalonamento considerados. A porcentagem de perda de bytes está relacionada a quantidade total de bytes perdida em relação à quantidade total de bytes transmitida pelos fluxos. Definimos o ganho do escalonador proposto sob o GPS como a redução percentual da taxa de perda de Bytes do GPS proposto sobre o GPS comum. O ganho de desempenho sobre o GPS convencional é um indicador da eficiência do uso de um determinado critério de escalonamento em relação ao uso do critério baseado na taxa média (GPS convencional). Por último, a utilização do enlace é a razão da taxa média utilizada pelos fluxos em relação à taxa total de transmissão do enlace.

As seguintes tendências foram observadas nos resultados das simulações:

1. A diminuição do tamanho da janela de tempo na estimação do expoente de Hölder pontual (notação H Pont J_ nas Tabelas III, IV e V) aumenta a porcentagem de perda de bytes e diminui o ganho de desempenho do escalonador proposto em relação ao GPS convencional (notação Taxa Média). Isto se deve ao aumento das imprecisões na estimação dos expoentes de Hölder com a diminuição do tamanho da janela de tempo.

2. Para um mesmo tamanho de *buffer*, um aumento da taxa de transmissão total proporciona uma diminuição da perda de dados e um aumento no ganho de desempenho do escalonador proposto em relação ao GPS convencional. No entanto, uma porcentagem maior da taxa de transmissão total pode ficar inutilizada. Isso pode ocorrer para todos os escalonadores analisados.

3. Para um mesmo valor de taxa de transmissão total, o aumento do tamanho do *buffer* proporciona uma diminuição da perda de dados, um aumento no ganho de desempenho do escalonador proposto em relação ao GPS convencional e uma melhor utilização da taxa de transmissão do enlace.

4. A diminuição da escala de tempo entre amostras proporciona uma diminuição da perda de dados, um aumento no ganho de desempenho do escalonador proposto em relação ao GPS convencional e um melhor aproveitamento da taxa de transmissão do enlace. Isto se deve a uma maior quantidade de ajustes dentro de um mesmo intervalo de tempo, obtendo uma alocação mais coerente de taxa a cada fluxo, apesar do aumento de informações que devem ser processadas na rede.

5. O uso da taxa de tráfego predita pelo preditor proposto (notação Taxa inst. nas tabelas) como critério de escalonamento, demonstrou um desempenho superior ao GPS convencional. No entanto, seu desempenho apresentou-se inferior ao do escalonador proposto na maioria dos casos. Isto ocorre principalmente quando se utiliza janelas de tempo mais largas para a estimação dos expoentes de Hölder pontuais.

TABELA III- DESEMPENHO DO ESCALONADOR PROPOSTO PARA TAXA DO ENLACE DE 2,25MBPS E BUFFER DE 16 KBYTES

Critério de escalonamento	Porcentagem de perda de *bytes*	Ganho sobre o GPS convencional	Utilização do enlace
Taxa média	0,0146	-	0,7664
Taxa inst.	0,0135	7,35 %	0,7673
H Pont Ref	0,0125	14,38 %	0,7681
H Pont J13	0,0127	13,01 %	0,7679
H Pont J12	0,0131	10,27 %	0,7676
H Pont J11	0,0136	6,85 %	0,7672

TABELA IV - DESEMPENHO DO ESCALONADOR PROPOSTO PARA TAXA DE 2,25 MBPS E BUFFER DE 25 KBYTES

Critério de escalonamento	Porcentagem de perda de *bytes*	Ganho sobre o GPS convencional	Utilização do enlace
Taxa média	0,0099	-	0,7701
Taxa inst.	0,0087	12,12 %	0,7710
H Pont Ref	0,0079	20,20%	0,7716
H Pont J13	0,0080	19,19%	0,7716
H Pont J12	0,0083	16,16%	0,7713
H Pont J11	0,0088	11,11%	0,7709

TABELA V - DESEMPENHO DO ESCALONADOR PROPOSTO PARA TAXA DE 2,75 MBPS E BUFFER DE 25 KBYTES

Critério de escalonamento	Porcentagem de perda de *bytes*	Ganho sobre o GPS convencional	Utilização do enlace
Taxa média	0,0017	-	0,6353
Taxa inst.	0,0013	23,53 %	0,6355
H Pont Ref	0,0010	41,18 %	0,6357
H Pont J13	0,0012	29,41 %	0,6356
H Pont J12	0,0012	29,41 %	0,6356
H Pont J11	0,0012	29,41 %	0,6356

VI. CONCLUSÕES

O uso de mecanismos dinâmicos e preventivos de controle de tráfego requer que as características locais do tráfego sejam preditas. A predição (e também o controle e gerenciamento) do tráfego encontra como obstáculo o fato de o mesmo ser

complexo e possuir irregularidades com intensidades elevadas, principalmente em pequenas escalas de tempo.

Neste artigo, apresentamos um estudo sobre a estimação do expoente de Hölder pontual, o qual é capaz de indicar o grau da regularidade do tráfego. Verificamos que a estimação destes expoentes em janelas de tempo é possível e que o algoritmo empregado é adequado para tal, possibilitando um processamento mais rápido das informações.

O preditor proposto se mostrou robusto e preciso em escalas de tempo distintas, mesmo para processos com características complexas, como o tráfego de redes e as séries formadas pelos expoentes de Hölder pontuais.

Este trabalho contribui também com um esquema de escalonamento que incorpora o conhecimento da regularidade local do tráfego por meio dos expoentes de Hölder pontuais. Em nossa abordagem utilizamos o método das janelas de tempo para a estimação dos expoentes das amostras de cada fluxo, tornando a operação do escalonador mais dinâmica. Em situações em que há a necessidade de um menor tempo de processamento das informações, o tamanho da janela utilizada pode ser diminuído, às custas de uma menor precisão das estimativas.

Pode-se concluir que o esquema de escalonamento proposto possui melhor desempenho em relação à perda de bytes e utilização do enlace, comparado a escalonadores que usam como critério de distribuição de taxa, a taxa média (GPS convencional), o expoente de Hölder pontual médio e a predição da intensidade de tráfego. Isso ocorre porque o escalonador proposto utiliza a predição da informação de surtos de tráfego. A predição do comportamento em rajadas é importante para que ações sejam tomadas a tempo de modo a evitar congestionamento. Portanto, o escalonador proposto é uma ferramenta poderosa para engenharia de tráfego nas redes atuais.

Em trabalhos futuros, aplicaremos a disciplina de escalonamento proposta em cenários de rede envolvendo diferentes protocolos.

AGRADECIMENTOS

Os autores agradecem a Fapesp (Proc. 06/60363-6) pelo apoio à pesquisa.

REFERÊNCIAS

[1] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, D. Veitch, "The multiscale nature of network traffic: Discovery, analysis, and modeling", IEEE Signal Processing Magazine, Vol. 19, No. 3, pp.28-46, maio 2002.

[2] V. Alarcon-Aquino, J. A. Barria, "Multiresolution FIR neural-network-based learning algorithm applied to network traffic prediction", IEEE Transactions on Systems, Man, and Cybernetics – part C: Applications and Reviews, Vol. PP, No. 99, pp. 1-13, 2005.

[3] A. Arneodo, E. Bacry, S. Jaffard, J. F. Muzy, "Singularity spectrum of multifractal functions involving oscillating singularities", Journal of Fourier Analysis and Applications, Vol. 4, No. 2, pp. 159-174, 1998.

[4] I. Daubechies, Ten Lectures on Wavelets, Philadelphia: SIAM, 1992.

[5] K. Falconer, Fractal Geometry: Mathematical Foundations and Applications, Nova York: John Wiley & Sons, 1990.

[6] A. Feldmann, A. C. Gilbert, W. Willinger, T. G. Kurtz, "The changing nature of network traffic: scaling phenomena", Computer Communications Review, Vol. 28, No. 2, Abril 1998.

[7] S. Haykin, Adaptive Filter Theory, Nova Jersey: Prentice Hall, 1991.

[8] S. Jaffard, "Exposants de Hölder en des points donnés et coéfficients d'ondelettes", C. R. Acad. Sci, Paris, Vol. 308, pp. 79-81, 1989.

[9] S. Mallat, W. Hwang, "Singularity detection and processing with wavelets", IEEE Transactions on Information Theory, Vol. 38, No. 8, pp. 617-643, mar. 1992.

[10] S. A. N. Ostring, H. Sirisena, "The influence of long-range dependence on traffic prediction", Proceedings of the International Communications Conference, Helsinki, pp. 1092-1101, jun. 2001.

[11] A. K. Parekh, G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case", IEEE/ACM Transactions on Networking, Vol. 1, No. 3, pp. 344-357, jun. 1993.

[12] Y. Qiao, J. Skicewicz, P. Dinda, "An empirical study of the multiscale predictability of network traffic", IEEE Proceedings of the International Symposium on High Performance Distributed Computing, Vol. 0, pp. 66-76, 2004.

[13] R. H. Riedi, J. Lévy-Véhel, "TCP traffic is multifractal: a numerical study", Technical Report, INRIA Rocquencourt, No. 3129, mar. 1997.

[14] A. Sang, S. Q. Li, "A predictability analysis of network traffic", Computer Networks, Vol. 39, No. 4, pp. 329-345, jul. 2002.

[15] S. Seuret, A. C. Gilbert, "Pointwise Hölder exponent estimation in data network traffic", ITC Specialist Seminar, Monterey, set. 2000.

[16] K. Shah, S. Bohacek, E. Jonckheere, "On the predictability of data network traffic", Proceedings of the American Control Conference, Vol. 2, pp. 1619-1624, 2003.

[17] P. Young, Recursive Estimation and Time-series Analysis, Nova York: Springer-Verlag, 1984.

[18] Z. R. Struzik, "Determining Local Singularity Strengths and their Spectra with the Wavelet transform", *Fractals*, 8(1):469-475, Março 2000.

[19] S. Seuret. Detecting and creating oscillations using multifractal methods. Mathematische Nachrichten, 279(11), 1195-1211, 2006.

Flávio Henrique Teles Vieira nasceu em Barra do Garças, MT em 25 de Outubro de 1978. Recebeu o título de bacharel em Engenharia Elétrica pela Universidade Federal de Goiás (UFG) em 2000, o título de Mestre em Engenharia Elétrica e de Computação pela Escola de Engenharia Elétrica da UFG em 2002. Em 2006 recebeu o título de Doutor em Engenharia Elétrica pela Faculdade de Engenharia Elétrica e Computação (FEEC) da UNICAMP. Iniciou em fevereiro de 2007 pós-doutorado no DECOM (Departamento de Comunicações)- FEEC- Unicamp. É membro da Sociedade Brasileira de Telecomunicações e atua nas seguintes áreas de pesquisa: Modelagem, predição e controle de tráfego de redes, redes de comunicações, sistemas inteligentes aplicados a telecomunicações. email: flavio@decom.fee.unicamp.br

Lee Luan Ling obteve o título de engenheiro eletricista pela USP em 1980, mestre em Engenharia Elétrica pela UNICAMP em 1984 e PhD em Engenharia Elétrica pela Universidade de Cornell em 1991. Foi fundador e tem sido o coordenador do Laboratório de Reconhecimento de Padrões e Redes de Comunicações (LRPRC) da FEEC-UNICAMP desde 1994. Foi o chefe do Departamento de Comunicações em 2000. Desde 2002 ele tem sido Professor Titular pela FEEC-UNICAMP. Atualmente ele atua intensamente em duas áreas de pesquisa: Redes de Comunicações e Reconhecimento de Padrões. email: lee@decom.fee.unicamp.br

Mistura de Quatro Ondas: Teoria, Simulações Computacionais e sua importância nas Redes Ópticas Transparentes

Iwanir Araújo da Silva Junior & Maria Regina Campos Caputo

***Abstract* - QoS is an important parameter for transparent optical networks performance and it has dependence with signal quality. When the signal stays on the optical domain (through of a lightpath) it can accumulate problems as nonlinear effects from the optical fibers. Physical impairments are very important when the optimization of the network is considered. In this article is examined the dependence of four wave mixing nonlinear effect in a DWDM system with various parameters, through numerical simulations with a software from VPI - Virtual Photonics Incorporation.**

***Index Terms* - Four Wave Mixing, Transparent Networks, Physical Layer Impairments.**

***Resumo* – Um importante parâmetro de desempenho da rede, a Qualidade de Serviço (QoS – Quality of Service), tem uma dependência com a qualidade do sinal. Na camada física o sinal permanece no domínio óptico (ao longo de uma rota totalmente óptica ou de um "lightpath") e pode acumular degradações, como por exemplo, advindas dos efeitos não-lineares nas fibras ópticas. Assim os impedimentos advindos da camada física constituem-se numa importante métrica para a otimização da rede. Nesse artigo verifica-se a dependência do efeito não-linear MQO – Mistura de Quatro Ondas (FWM - Four Wave Mixing) em um Sistema Densamente Multiplexado por Divisão de Comprimento de Onda (DWDM – Dense Wavelength Division Multiplexing), o qual aqui se designa por sistema multiplexado, com diversos parâmetros através de simulações numéricas utilizando o Módulo de Transmissão do software VPI - Virtual Photonics Incorporation.**

***Palavras Chave* - Mistura de Quatro Ondas, Redes Transparentes, Impedimentos de Camada Física.**

I. INTRODUÇÃO

A otimização de *redes ópticas transparentes* [1], [2], [3], [4], [5], [6] para uma dada rede física de fibras ópticas, pode ser separada em duas etapas**:**

Etapa 1 - o projeto da *topologia virtual* (ou *topologia lógica*);

Etapa 2 - o roteamento e a escolha do comprimento de onda.

Um importante parâmetro de desempenho da rede, a Qualidade de Serviço (QoS – Quality of Service), tem uma dependência com a qualidade do sinal. Na camada física o sinal permanece no domínio óptico ao longo de uma rota totalmente óptica e pode acumular degradações, por exemplo, advindas de efeitos não-lineares em fibras ópticas. Assim os impedimentos advindos da camada física constituem-se uma importante métrica para a otimização da rede.

Neste artigo é feita uma verificação da dependência do efeito não-linear da Mistura de Quatro Ondas – MQO (FWM - Four Wave Mixing) em um sistema multiplexado com diversos parâmetros tais como [7], [8], [9], [10]: comprimento do enlace, separação espectral entre canais, dispersão da fibra na região de transmissão dos canais bem como potência dos mesmos. A verificação foi feita através de simulações numéricas através do Módulo de Transmissão da VPI - Virtual Photonics Incorporation.

A partir dos resultados aqui obtidos pode-se calcular o fator Q (*função erfc* ou *probabilidade gaussiana normalizada)* para cada canal. Muitos algoritmos destinados á definição da Topologia Virtual da rede (escolher as rotas totalmente ópticas que otimizam o seu desempenho) incluem os impedimentos da camada física como uma das métricas, limitando o valor do fator Q na hora de decidir dentre as rotas ópticas canditadas qual a melhor [11], [12], [13]. A escolha do comprimento de onda então é feita pesquisando-se os comprimentos de onda disponíveis e calculando-se as penalidades advindas da potência da MQO gerada pelos canais usados simultaneamente naquela rota. O fator Q de uma rota totalmente óptica pode flutuar quando conexões (canais) são estabelecidas ou desfeitas [13].

Este artigo está assim organizado: a secção 2 mostra o modelo analítico da teoria de MQO, na secção 3 está descrito o modelo numérico adotado nas simulações, na secção 4 são apresentados os resultados das simulações e a secção 5 apresenta-se as conclusões do trabalho.

II. MODELO ANALÍTICO

Através de um processo de MQO, a propagação simultânea de dois campos com freqüências f_i e f_j geram bandas laterais nas freqüências $2f_i - f_j$ e $2f_j - f_i$. No caso de três campos co-propagantes, nas freqüências f_i, f_j e f_k $(i, j \neq k)$, há a geração de um quarto termo com freqüência $f_i + f_j - f_k$, com i, j e k podendo tomar valores entre 1 e 3. Nas redes ópticas transparentes se f_i é a freqüência de uma dada rota totalmente óptica, f_j, f_k são as freqüências da rota que compartilham a mesma rota física e são potencialmente fontes de geração de MQO.

Manuscrito recebido em 27 de março de 2006; revisado em 28 de fevereiro de 2007.

I. A. Silva Jr (iwanir.silva@vale.com) pertence à Companhia Vale do Rio Doce - VALE. Av. Dante Micheline, 5500 - Vitória - ES - Brasil - 29090-900 e M. R. C. Caputo (mreginacaputo@hotmail.com) pertence á Pontifícia Universidade Católica de Minas Gerais – PUC-MG – Belo Horizonte - MG - Brasil – 30000-000.

Para que o processo ocorra, o casamento de fase deve satisfazer a seguinte condição [14]:

$$\Delta\beta = \beta_i(f_i) + \beta_j(f_j) - \beta_k(f_k) - \beta_{ijk}(f_{ijk}) \qquad (1)$$

onde $\Delta\beta$ é a diferença entre as constantes de propagação das ondas envolvidas no processo de MQO. Para um perfeito casamento de fase, $\Delta\beta$ deve ser igual a zero significando uma máxima eficiência do processo. β_i, β_j e β_k são as constantes de propagação respectivamente nas freqüências f_i, f_j e f_k e β_{ijk} é a constante de propagação na freqüência da onda gerada pelo processo de MQO.

No caso de ondas co-polarizadas e considerando-se a não depleção dos sinais de entrada (pela geração das bandas laterais) a potência gerada para uma banda lateral, P_{ijk} , na freqüência f_{ijk}, , é dada por [14]:

$$P_{ijk}(L) = \eta\left(\frac{d_{ijk}\gamma}{3}\right)^2 P_i P_j P_k e^{-\alpha L} L_{eff}{}^2 \qquad (2)$$

onde P_i, P_j e P_k são as potências de entrada dos sinais respectivamente nas freqüências f_i, f_j e f_k ; L é o comprimento do enlace; d_{ijk} é o fator de degenerescência (igual a 3 para $i = j$ e igual a 6 para $i \neq j$); α é o coeficiente de atenuação da fibra; L_{eff} é o comprimento efetivo (no qual assume-se que a potência do sinal permanece constante [15]) e é dado por $L_{eff} = (1 - e^{-\alpha L})/\alpha$. O coeficiente de não-linearidade γ depende da área efetiva da fibra A_{eff} , a qual depende da distribuição do modo propagante fundamental no núcleo da fibra [7], do índice de refração não-linear n_2 e do comprimento de onda λ e é dado por [7]:

$$\gamma = \frac{n_2 \omega_0}{c A_{eff}} \qquad (3)$$

A eficiência MQO é definida pela relação entre a potência do sinal gerado com descasamento de fase e a potência do sinal gerado quando se tem um casamento de fase perfeito, $\Delta\beta = 0$, sendo dada por [14]:

$$\eta = \frac{P_{ijk}(L, \Delta\beta)}{P_{ijk}(L, \Delta\beta = 0)} \qquad (4)$$

Pode-se expressar a eficiência da MQO em termos do comprimento da fibra e do descasamento de fase como:

$$\eta = \frac{\alpha^2}{\alpha^2 + (\Delta\beta)^2} \cdot \left(1 + \frac{4e^{-\alpha L}\,\mathrm{sen}^2\left(\Delta\beta \cdot L \big/ 2\right)}{\left(1 - e^{-\alpha L}\right)^2}\right) \qquad (5)$$

na qual $\Delta\beta$ tem sua expressão dependente da dispersão da fibra e do espaçamento entre canais escrito da seguinte forma [14], [16]:

$$\Delta\beta = \frac{2\pi\lambda_k^2}{c}\Delta f_{ik}\Delta f_{jk}\left[D(\lambda_k) + \frac{\lambda_k^2}{2c}\left(\Delta f_{ik} + \Delta f_{jk}\right)\left(\frac{dD(\lambda_k)}{d\lambda}\right)\right] \qquad (6)$$

onde $D(\lambda_k)$ é a dispersão cromática da fibra no comprimento de onda λ_k, $dD(\lambda_k)/d\lambda$ é a derivada da dispersão também designada por S_0 e $\Delta f_{mn} = |f_m - f_n|$ para $m, n = i, j, k$. Essa equação é chamada de fator de descasamento de fase linear, pois não depende da potência das ondas que interagem no processo de MQO.

A obtenção do descasamento de fase total, o qual inclui o descasamento de fase linear e a dependência da potência, foi realizado por SONG *et al.* (1999) [17] dado por:

$$\Delta\beta_T = \Delta\beta - \gamma\left(P_i + P_j - P_k\right)\left\{\frac{1 - \exp\left(-\alpha L_{eff}\right)}{\alpha L_{eff}}\right\} \qquad (7)$$

A Eq. (7) inclui a dependência do descasamento de fase devido à potência das ondas que interagem na MQO, incluindo também os efeitos não-lineares de Automodulação de Fase – AMF (ou SPM - *Self Phase Modulation*) e Modulação de Fase Cruzada – MFC (ou XPM - *Cross Phase Modulation*).

Define-se um parâmetro designado *comprimento de coerência* L_{coh}, que é a distância que o campo eletromagnético na freqüência f_{ijk} deve propagar na fibra antes que fique fora de fase por 180° com o campo eletromagnético gerado a partir da componente de polarização não-linear. Tal distância fará com que o processo de MQO deixe de ser significante, ou seja, ocorrerá MQO se $L < L_{coh}$. O comprimento de coerência é dado por [17], [18], [19], [20]:

$$L_{coh} = \frac{2\pi}{\Delta\beta_T} \qquad (8)$$

Para uma situação ideal no qual $\Delta\beta_T = 0$, tem-se o resultado de $L_{coh} = \infty$, indicando que um eficiente processo de MQO ocorre para todo L.

Ocorrerá interferência entre canais se os mesmos forem escolhidos com igual espaçamento espectral entre eles. Em caso contrário as freqüências geradas pela MQO não irão se sobrepor com os canais originais. Assim com uma escolha das posições dos canais baseada nesse critério não haverá nenhuma das componentes geradas pela MQO que se situe na mesma posição de qualquer um dos canais originais [1], [20], [21], [22], [23].

A eficiência da MQO se torna significativa quanto melhor for satisfeita a condição de casamento de fase entre os campos ópticos co-propagantes. Essa condição é função da separação espectral entre canais do sistema multiplexado, do valor da dispersão cromática na região de operação do sistema, do valor da potência dos canais e do comprimento do enlace, como visto anteriormente.

III. MODELO NUMÉRICO

Para se comprovar a dependência do processo de MQO com os parâmetros citados anteriormente, considerou-se a solução

numérica da equação de propagação do pulso numa fibra óptica, a qual inclui os efeitos lineares de perda por atenuação e de dispersão cromática, como também os efeitos não-lineares de AMF e MQO. A equação de propagação do pulso é a conhecida equação não-linear de Schrödinger (ENLS) dada por [7], [21]:

$$\frac{\partial A(z,T)}{\partial z} + j\frac{\beta_2}{2}\frac{\partial^2 A(z,T)}{\partial T^2} - \frac{\beta_3}{6}\frac{\partial^3 A(z,T)}{\partial T^3} + \frac{\alpha}{2}A(z,T) = \quad (9)$$
$$j\gamma\left|A(z,T)\right|^2 A(z,T)$$

na qual $A(z, T)$ representa a envoltória do pulso variando lentamente, β_2 representa os efeitos da dispersão cromática de primeira ordem, conhecida como *dispersão de velocidade de grupo*–DVG e β_3 representa a dispersão cromática de segunda ordem, ou derivada da dispersão cromática de primeira ordem. Têm-se incluídos ainda os efeitos da perda, ou atenuação da fibra, através de α e do efeito da não-linearidade da fibra através do parâmetro γ.

A Eq. (9) serve como base para a análise dos efeitos lineares e não-lineares presentes numa fibra óptica e sua solução foi obtida pelo método numérico Split Step Fourier – SSF através de simulações realizadas pela ferramenta computacional Modulo de Transmissão da Virtual Photonics Inc. – VPI.

IV. Resultado e Simulações

A primeira parte deste item trata da influência da potência dos canais originais (rotas totalmente ópticas simultâneas na mesma rota) sobre a potência das componentes geradas pela MQO; depois se estuda a dependência da MQO com o comprimento da rota e também com a separação espectral entre os canais originais do sistema. E, finalmente, mostra-se influência da dispersão no processo de MQO ressaltando uma comparação do valor da potência de uma onda gerada pelo processo de MQO usando-se uma *fibra com dispersão deslocada* FDD [24] e uma *fibra com dispersão deslocada não nula* FDDNN [25], que são duas fibras amplamente utilizadas em sistemas multiplexados. Também foram calculados os valores das potências das componentes geradas pelo MQO no regime de dispersão anômalo e normal de uma fibra FDD.

A - Influência da potência dos canais Multiplexados

Como acontece em outros fenômenos não-lineares, ao se diminuir a potência do canal de transmissão, diminui-se a eficácia do processo não-linear [1], [7], [15]. Assim, a potência da onda gerada pelo processo de MQO cresce com o aumento da potência dos canais originais, como pode ser visto na Eq. (2).

Para confirmar simulou-se a propagação de dois canais sem modulação, um em 1558 nm (f_1 = 192,55 THz) e outro em 1558,8 nm (f_2 = 192,46 THz), numa fibra FDD [24] com 20 km. A Tabela 1 mostra os dados da fibra utilizada na simulação.

TABELA 1
PARÂMETROS DA FIBRA UTILIZADA NA SIMULAÇÃO DA POTÊNCIA GERADA PELA MQO EM RELAÇÃO À POTÊNCIA DE ENTRADA DOS CANAIS

Fibra FDD [24]			
A_{eff}	52 μm²	S_0	0,075 ps/nm²km
α	0,24 dB/km	D	$(\lambda - \lambda_0)S_0$
λ_0	1550 nm	L	20 km
γ	1,3 (w.km)$^{-1}$		

A potência de entrada do canal 1 (P_1) é igual à do canal 2 (P_2), e varia de 1 a 12 mW. A Figura 1 mostra o comportamento das duas componentes geradas pela MQO, sendo P_{112} a potência da primeira componente $f_{112} = 2f_1 - f_2$ (=192,64 THz) e P_{221} a da segunda componente $f_{221} = 2f_2 - f_1$ (=192,37 THz).

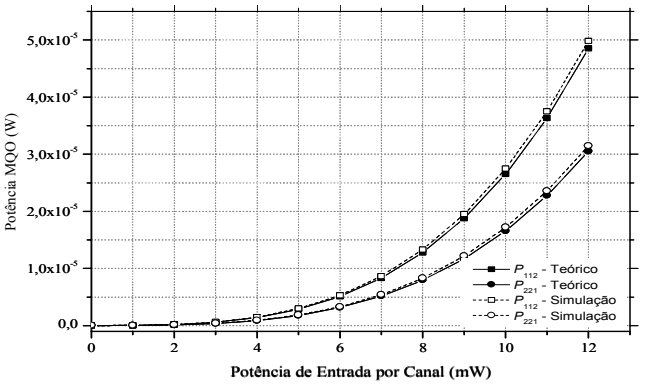


Fig. 1. Potência gerada pelo processo MQO em função da potência de entrada dos canais originais. Figura gerada pela ferramenta Mathcad [30] (teórica) e pelo VPI (simulação).

Assim com o aumento da potência dos canais, aumentou-se a potência das bandas laterais geradas em f_{112} e f_{221}, confirmando-se por simulação a previsão analítica da equação $f_{ijk} = f_i + f_j - f_k$. Pode-se observar que P_{112} é maior que P_{221} devido ao fato de que a componente gerada pela MQO em f_{112} = 1557,3 nm se encontra mais perto do comprimento de onda de dispersão zero, experimentando assim um menor valor de dispersão. Isso proporciona um melhor casamento de fase, visto que quanto menor for o valor da dispersão das ondas que estiverem interagindo no processo, maior será a eficiência do mesmo [1], [14], [15], [17], [20], [26], [27], [27].

As Figuras 2(a) e 2(b) mostram espectros observados na saída da fibra óptica utilizada na simulação para alguns dos valores de potência de entrada dos canais.

Outro fato observado nas figuras 2(a) e 2(b) é que ocorre a geração de novas componentes de freqüência, em 192,28 THz (≈ 1560,23 nm) e 192,73 THz (≈ 1556,58 nm). Isto se deve ao fato de as duas componentes geradas anteriormente f_{112} e f_{221} estarem agora participando do processo de geração de outras componentes de freqüência juntamente com os sinais originais em f_1 e f_2. As novas ondas geradas são as combinações $f' = f_1 + f_{221} - f_{112}$ = 192,28 THz e $f'' = f_1 + f_{112} - f_2$ = 192,73 THz.

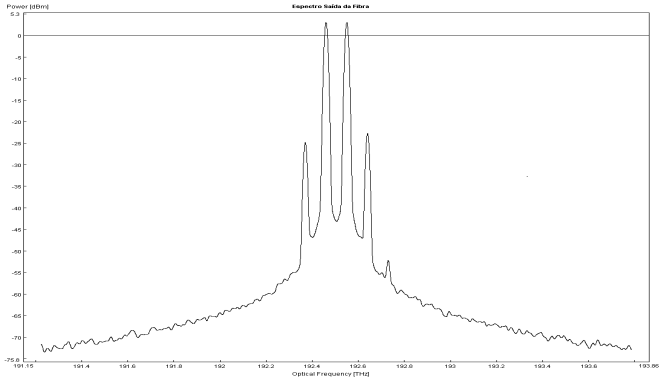


Fig. 2(a). Espectro óptico observado na saída da fibra óptica, mostrando o crescimento da MQO em função do aumento da potência de entrada dos canais para 6 mW .

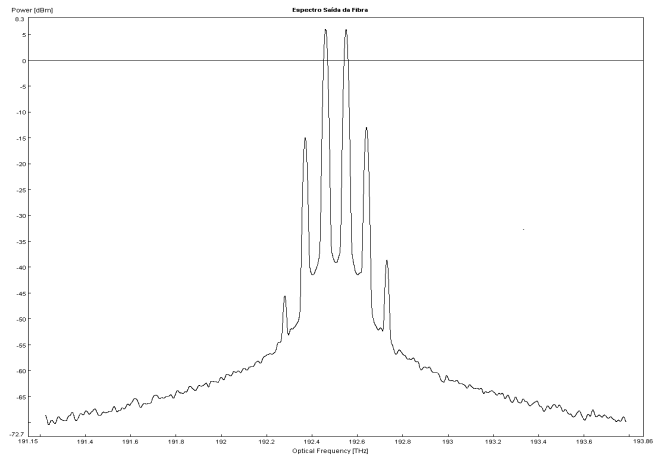


Fig. 2(b). Espectro óptico observado na saída da fibra óptica, mostrando o crescimento da MQO em função do aumento da potência de entrada dos canais para 12 mW (b).

B - Influência do comprimento do enlace

A análise potência da onda gerada pela MQO em relação à variação do comprimento da fibra pode ser feita considerando-se a propagação dos mesmos dois canais (f_1 = 192,55 THz e f_2 = 192,46 THz) utilizados anteriormente. Variando-se o comprimento da fibra, observa-se a variação da potência das ondas geradas. A Figura 3 confirma o comportamento citado.

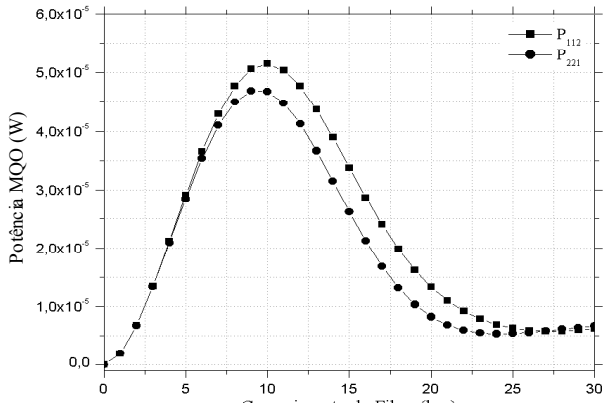


Fig. 3. Potência MQO como função do comprimento da fibra para as duas componentes de freqüência P_{112} e P_{221}.

Usando-se a Eq. (8), pode-se calcular que $L_{coh112} \approx 26$ km para P_{112} e $L_{coh221} \approx 24$ km para P_{221}. Esses resultados também foram confirmados através da simulação de acordo com a Figura 3.

C - Influência da separação espectral entre canais do sistema Multiplexado

Analisa-se aqui a influência do aumento do espaçamento entre os canais originais na eficiência do processo de MQO e também a variação da posição espectral dos canais originais no processo de MQO considerando-se espaçamento igual e diferente entre eles.

Primeiro, analisou-se a influência do valor absoluto do espaçamento entre canais para um sistema com dois canais operando no regime anômalo de dispersão de uma fibra FDD [11]. A potência da onda gerada pela MQO é analisada em relação ao comprimento da fibra para dois valores de separação entre canais Δf_1 = 100 GHz e Δf_2 = 200 GHz, considerando λ_0 = 1550 nm. Para Δf_1, utilizando λ_1 = 1552 nm e λ_2 = 1552,8 nm, mediu-se a potência da onda gerada em f_{112} = 193,4 THz. Já para Δf_2, utilizando λ_1' = 1552 nm e λ_2' = 1553,6 nm, mediu-se a potência da onda gerada em f_{112}` = 193,5 THz. A Tabela 2 apresenta os dados dos canais usados na simulação e a Figura 4, a seguir, mostra os resultados obtidos na simulação.

TABELA 2

PARÂMETROS UTILIZADOS NA SIMULAÇÃO DA POTÊNCIA GERADA PELA MQO EM RELAÇÃO À VARIAÇÃO DO COMPRIMENTO DA FIBRA PARA DIFERENTES VALORES DE SEPARAÇÃO ENTRE CANAIS.

Fibra FDD [24] (λ_0 = 1550 nm)			
Espaçamento	Canais		f_{MQ0}
Δf_1	λ_1 = 1552 nm	f_1 = 193,3 THz	f_{112}
	λ_2 = 1552,8 nm	f_2 = 193,2 THz	193,4 THz
Δf_2	λ_1' = 1552 nm	f_1` = 193,3 THz	f_{112}'
	λ_2' = 1553,6 nm	f_2` = 193,1 THz	193,5 THz

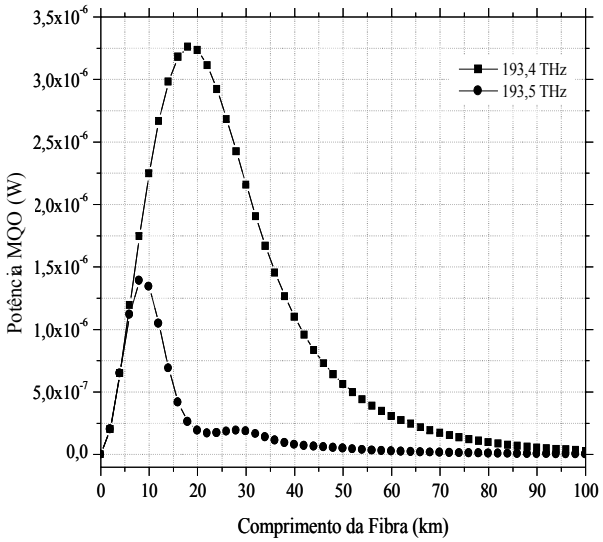


Fig. 4. Comparação de potência da onda gerada pela MQO para f_{112} = 193,4 THz e f_{112}'=193,5 THz, considerando espaçamento entre canais de 100 GHz e 200 GHz respectivamente, para uma fibra FDD operando no regime de propagação Anômalo.

Pode-se constatar que, devido a um maior espaçamento entre canais e conseqüentemente o maior descasamento de fase, a componente de freqüência gerada pela MQO apresenta menor potência do que a onda gerada pela MQO com um espaçamento menor. Observa-se que a componente de freqüência gerada em f_{112}' = 193,5 THz ($\Delta f_2 = 200$ GHz) possui uma eficiência baixa já nos primeiros quilômetros da fibra enquanto que a componente em f_{112} = 193,4 THz ($\Delta f_1 = 100$ GHz) mantém-se com alta eficiência mesmo em comprimentos maiores da fibra. Isso pode ser entendido através da Eq. (8) que define o comprimento de coerência.

Faz-se agora a análise da influência da posição dos canais originais a fim de se constatar a interferência entre componentes espectrais geradas pela MQO e os canais originais. Utilizou-se um sistema com 3 canais com igual espaçamento ($f_1 = 193{,}05$ THz, $f_2 = 193{,}1$ THz $f_3 = 193{,}15$ THz) e com espaçamento diferente ($f_1 = 193{,}02$ THz, $f_2 = 193{,}1$ THz $f_3 = 193{,}15$ THz), uma fibra FDD [24] com 50 km na região de 1550 nm, considerando λ_0 = 1550 nm, e canais com potências iguais a 4 mW.

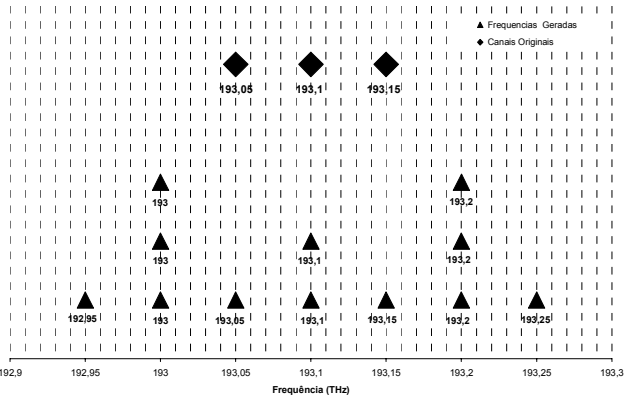


Fig. 5(a). Posição dos produtos gerados pela FWM com espaçamento igual entre canais, resultando em interferência nos canais originais. Cálculos utilizando $f_i + f_j - f_k$, com i, j e k podendo tomar valores entre 1 e 3.

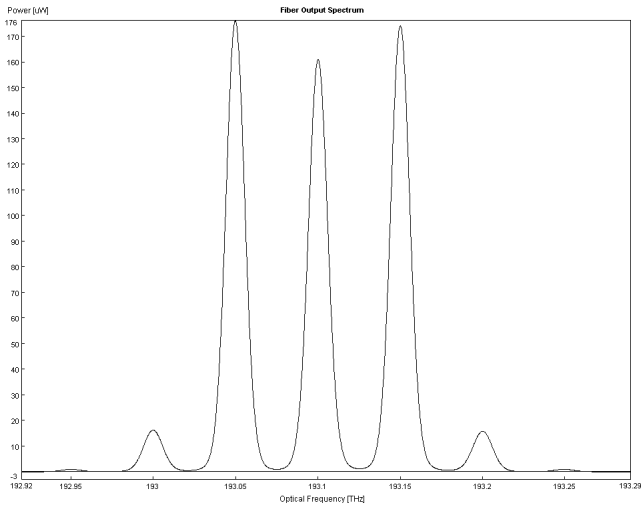


Fig. 5(b). Posição dos produtos gerados pela FWM com espaçamento entre canais igual, resultando em interferência nos canais originais. Gerada pelo VPI (simulação).

Utilizando $f_i + f_j - f_k$, para se calcular a posição das componentes de freqüência geradas pela MQO mostrou-se na Figura 5(a) o caso de igual espaçamento entre canais e na Figura 6(a) o caso de espaçamento diferente. As Figuras 5(b) e 6(b) confirmam o posicionamento indicado pelas Figuras 5(a) e 6(a) respectivamente através de simulação.

Com a utilização de igual espaçamento entre os canais originais, pode-se observar que haverá a geração de novas componentes de freqüência as quais irão se sobrepor com as freqüências dos canais originais. A simulação confirma as posições das componentes espectrais geradas. Com a utilização de espaçamento diferente entre canais, pode-se observar que não mais haverá a geração de novas componentes de freqüência que se situem na mesma freqüência dos canais originais.

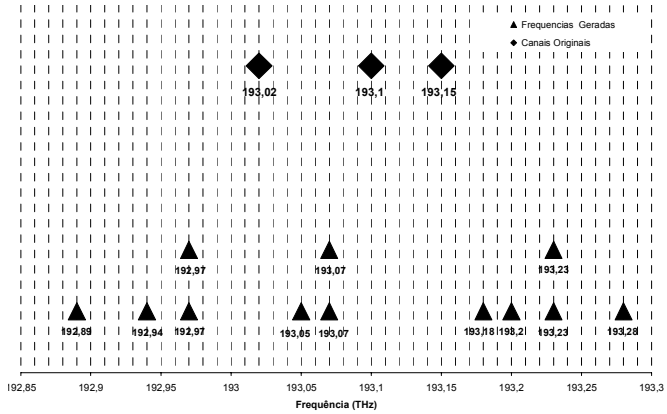


Fig. 6(a). Posição dos produtos gerados pela com espaçamento diferente entre canais. Não há interferência nos canais originais. Cálculos utilizando $f_i + f_j - f_k$, com i, j e k podendo tomar valores entre 1 e 3.

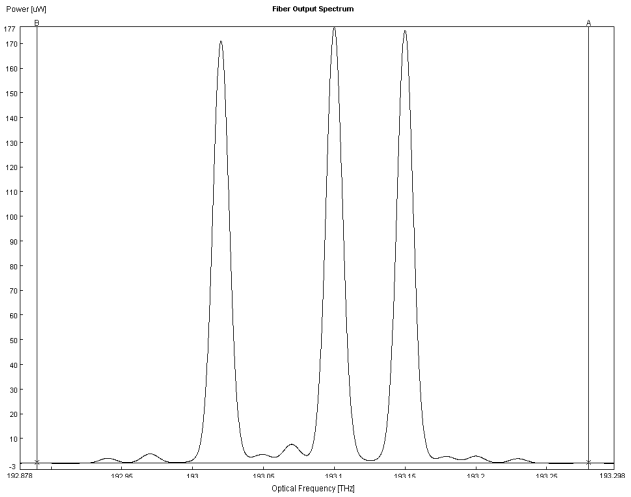


Fig. 6(b). Posição dos produtos gerados pela com espaçamento entre canais diferente. Não há interferência nos canais originais. Gerada pelo VPI (simulação).

É importante salientar que de acordo com os resultados obtidos até aqui ao se aumentar o espaçamento entre canais, aumenta-se o descasamento de fase, provocando menor eficiência na MQO. Em termos de enlaces de telecomunicações, isso tem o inconveniente de aumentar a

largura de banda total do sinal transmitido, exigindo que os amplificadores ópticos possuam ganho plano sobre uma largura de banda mais ampla [1], [20], [27], [29].

D - Influência da dispersão cromática na região de operação do sistema

Mesmo em fibras FDD uma significante quantidade de dispersão estará presente em comprimentos de onda de operação além de 1560 nm, reduzindo assim o efeito da MQO. Quando se opera com comprimentos de onda além de 1560 nm pode-se usar além dos Amplificadores à Fibra Dopada com Érbio (AFDE) também os amplificadores baseados no efeito não-linear Raman [1], [7], [16].

Analisa-se agora a eficiência da MQO quanto à variação da dispersão sob a influência do espaçamento entre canais e o comprimento da fibra. Considera-se a região de operação do sistema afastada do comprimento de onda de dispersão zero da fibra, minimizando assim o efeito de geração de MQO, resultando na derivada da dispersão desprezível.

Na Figura 7 tem-se o comportamento da eficiência da MQO para uma fibra FDD, uma fibra padrão e uma fibra FDDNN todas operando na janela de 1550 nm. Esta figura foi obtida a partir da Eq. (5) fazendo $\Delta\beta_T = \Delta\beta$ e com o auxílio da ferramenta Mathcad [30].

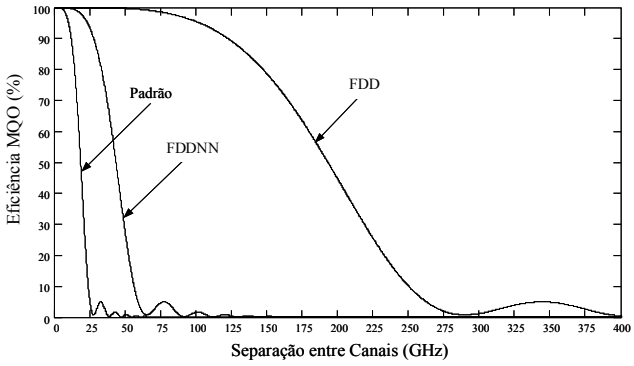


Fig. 7. Comparação da eficiência MQO entre uma fibra Padrão, uma fibra FDD e uma fibra FDDNN, com λ = 1552 nm, α=0,25 dB/km e L =10 km.

Pode ser observado que a influência da dispersão na eficiência da MQO é caracterizada diferentemente para ambas as fibras. Isto se deve ao fato de que quanto maior for a dispersão, mais difícil será de se conseguir um perfeito casamento de fase. Assim com a presença da dispersão, as ondas co-progantes terão menor probabilidade de se interferirem construtivamente num processo significativo de MQO [17], [20], [30]. Como a fibra FDDNN possui um valor de dispersão baixo em relação à fibra Padrão e alto em relação à fibra FDD, a eficiência dessa fibra na geração de produtos de MQO é intermediária entre as fibras Padrão e FDD. Isto é, para um mesmo comprimento da fibra, 10 km, e para uma mesma separação espectral entre os canais originais, a eficiência da MQO para a fibra FDDNN é maior do que para a fibra Padrão, porém menor do que para a fibra FDD na janela de 1550 nm.

V. CONCLUSÕES

Neste trabalho analisou-se o efeito não-linear de Mistura de Quatro Ondas. A ENLS foi solucionada numericamente através de uma ferramenta computacional, o Modulo de Transmissão da Virtual Photonics Inc. – VPI. *Os resultados mostraram* a dependência da eficiência de geração de novas componentes espectrais com a potência dos canais, separação entre eles, comprimento do enlace e nível de dispersão apresentado pela fibra na região de operação do sistema.

Constatou o aumento da potência da onda gerada pela MQO ao aumentar-se a potência dos canais de transmissão (no caso rotas ópticas concorrentes). Observou-se também o comportamento da eficiência MQO com a variação do comprimento da fibra. Pôde-se observar que para grandes espaçamentos entre canais, a onda gerada pelo processo de MQO terá eficiência apenas nos comprimentos iniciais da fibra. Isso evidencia que, ocorrerá o descasamento de fase entre as ondas co-propagantes já no início da fibra. Diminuindo o espaçamento entre canais, o descasamento ocorre para um comprimento da fibra maior.

O nível de dispersão apresentado pela fibra na região de operação do sistema influenciará significantemente nas condições do casamento de fase para que o processo de MQO ocorra, conforme os resultados apresentados.

Os resultados aqui obtidos servem como base para avaliação do desempenho de redes ópticas transparentes, onde o comprimento da fibra, no qual as rotas ópticas seguem concorrendo na mesma rota, definirá a eficiência do processo. Também a escolha de rotas ópticas com diferentes espaçamentos espectrais resultará em diferente eficiência do processo.

REFERÊNCIAS

[1] Ramaswami, Rajiv; Sivarajan, Kumar N. Optical Networks: A Pratical Perspective. 2nd Ed., San Diego – USA: Academic Press, 2002.

[2] Steinkampa, Ansgar; Windmanna, Martin; Pachnicke, Stephan. More accurate simulation of dynamic transparent optical wavelength division multiplex networks. AEU - International Journal of Electronics and Communications vol. 61, Issue 3, p. 147-152, March 2007.

[3] O'Mahony, M. J.; Politi, C.; Klonidis, D.; Nejabati, R.; Simeonidou, D. Future Optical Networks. Journal of Lightwave Technology, vol. 24, Issue12, p. 4684-4696, Dec. 2006.

[4] Nouchi, Pascale; Montmorillon, Louis-Anne de; Sillard, Pierre; Bertaina, Alain; Guenot, P. Optical fiber design for wavelength-multiplexed transmission. Comptes Rendus Physique, vol. 4, Issue 1, Pages 29-39, January-February 2003.

[5] Souza Filho, Agostinho Linhares. Projeto de Redes Translúcidas de Longa Distância. Tese de Mestrado.FEEC/UNICAMP, 2003.

[6] Naves, José Renato de Paula. Estratégias Evolutivas para a Rede Óptica a partir de Anéis WDM. Tese de Mestrado. FEEC/UNICAMP, 2002.

[7] Agrawal, Govind; P. Fiber Optics Communication Systems. 3rd Ed., New York – USA: Wiley-Interscience, 2002.

[8] Faisal, Mohammad; Islam, Mohammed Nazrul; Majumder, Satya Prasad. Performance comparison of wavelength shift keying WDM system and conventional on-off WDM system in presence of four-wave mixing Optik - International Journal for Light and Electron Optics, vol. 117, Issue 12, p. 55-562, December 2006.

[9] Wegener, L. G. L.; Povinelli, M. L.; Green, A. G.; Mitra, P. P.; Stark, J. B.; Littlewood, P. B. The effect of propagation nonlinearities on the information capacity of WDM optical fiber systems: cross-phase modulation and four-wave mixing Physica D: Nonlinear Phenomena, vol. 189, Issues 1-2, p. 81-99, February 2004.

[10] Wong, Kenneth K.Y.; Lu, Guo-Wei; Chen, Lian-Kuan. Experimental studies of the WDM signal crosstalk in two-pump fiber optical parametric amplifiers Optics Communications, vol. 270, Issue 2, p. 429-432, February 2007.
[11] Ramamurthy B et all, Journal of Lightwave Technology, 17(10):1713-1723, outubro 1999.
[12] Fonseca I E, Moisés R N Ribeiro, Hekio Waldman, XXI Simpósio Brasileiro de Telecomunicações, Brasil, Setembro de 2004.
[13] Song, S.; Livas, J. Four-wave mixing induced Q-factor variations in WDM systems. Conference on Lasers and Electro-Optics - CLEO 2000. p. 341-342, 2000.
[14] Shibata, Nori; BRAUN, Ralf P.; WAARTS, Robert G. Phase-Mismatch Dependence of Efficiency of Wave Generation Trough Four-Wave Mixing in a Single-Mode Optical Fiber. IEEE J. Quantum electron., vol.QE-23, Nº 07, p. 1205-1210, July 1987.
[15] Keiser, Gerd. Optical Fibers Communications. 3rd Ed., USA: McGraw-Hill, 2000.
[16] Inoue, Kyo; Toba, Hiromu. Error-Rate Degradation due to Fiber Four-Wave Mixing in Four-Channel FSK Direct-Detection Transmission. IEEE Photonics Technology Letters, vol. 03, Nº 01, p. 77-79, January 1991.
[17] Song, Shuxian et al. Intensity-Dependent Phase-Matching Effects on Four-Wave Mixing in Optical Fibers. J. Lightwave Technology, vol. 17, Nº 11, p. 2285-2290, November 1999.
[18] Hill, K. O. et al. CW three-wave mixing in single-mode optical fibers. J. Applied Physics, vol. 49, Nº 10, p. 5098-5106, October 1978.
[19] Stolen Roger H.; BJORKHOLM, John E. Parametric Amplification and Frequency Conversion in Optical Fibers. IEEE J. Quantum Electron., vol.QE-18, Nº 07, p. 1062-1072, July 1982.
[20] Araújo, Iwanir. Análise do Efeito Mistura de Quatro Ondas em Fibras Ópticas na Janela de 1550 nm. Dissertação de Mestrado. Instituto Nacional de Telecomunicações, 2003.
[21] Abbade, Marcelo Luís Francisco. Contribuição para o Estudo de Não-Linearidades em Fibras Ópticas Monomodo. Tese de Doutorado. FEEC/UNICAMP, 2003.
[22] Grosz, Diego F. Efeitos não lineares em sistemas de comunicação óptica de longas distancias e altas taxas. Tese de Doutorado. IFGW/UNICAMP, 1998.
[23] Chavez Boggio, Jose Manoel. Efeitos não lineares em fibras ópticas de dispersão deslocada. Tese de Doutorado. IFGW/UNICAMP, 2001.
[24] Corning Incorporated. Corning SMF/DS CPC6 Single Mode Dispersion-Shifted Optical Fiber, 1996.
[25] Lucent Technologies. Fibra Óptica de Dispersão Deslocada Não Nula – TrueWave RS. 1998. Disponível em: <http://www.lucent.com>. Acesso em: dezembro 2002.
[26] Freitas, Márcio; CALMON, Luiz de Calazans; Almeida, Renato Tannure R. Mistura de Quatro Ondas em Sistema WDM Utilizando Fibras DS. Anais do 19º Simpósio Brasileiro de Telecomunicações, Fortaleza/CE, setembro de 2001.
[27] Waarts, Robert G. et al. Nonlinear Effects in Coherent Multichannel Transmission through Optical Fibers. Proceedings of the IEEE, vol. 78, Nº 08, p. 1344-1368, August 1990.
[28] Wehmann, C.F.; Fernandes, L.M.; Sobrinho, C.S.; Lima, J.L.S.; Silva, M.G.; Almeida, E.F.; Medeiros Neto; J.A.; Sombra; A.S.B. Analysis of the four wave mixing effect (FWM) in a dispersion decreasing fiber (DDF) for a WDM system. Optical Fiber Technology, vol. 11, Issue 3, p. 306-318, July 2005.
[29] Kojima, S. Numai, T. Theoretical analysis of modified repeated unequally spaced frequency allocations in FDM lightwave transmission systems. Journal of Lightwave Technology, vol. 24, Issue: 7, p. 2786 – 2797, July 2006.
[30] Bogoni, A. Poti, L. Bononi, A. Accurate measurement of in-band FWM power in DWDM systems over nonzero dispersion fibers. IEEE Photonics Technology Letters, vol. 15, Issue 2 p. 260 – 262, February 2003.
[31] Mathcad 2001 Professional for Windows 95/NT, version 6.0: Copyright © 1986-2001 Mathsoft, Inc. All rights reserved.

Maria Regina Campos Caputo: nasceu em São Tiago, MG, em 09 de abril de 1954. Possui os títulos de engenheira (INATEL 75), mestre e doutora (UFMG 93/2000). Foi professor adjunto II do INATEL (2001-2005) ministrando aulas no pós lato-sensu e mestrado, orientando 3 dissertações de mestrado e 4 monografias lato-sensu. Área de pesquisa: Redes Ópticas. É professora (desde 1999) do lato-senso em engenharia de Telecomunicações da UFMG. É revisora de área da Optics Communications (Holanda) e Revista do Inatel. Foi consultora e proprietária da SOLITTONS Engenharia Ltda (1994/2003) prestando consultoria técnica para: CVRD/95, ESCELSA/95 - Espírito Santo Centrais Elétricas e implementando cursos de "Redes Ópticas" para: Teleron/97, Petrobrás- Rio/ Salvador 95-97, Telemig/96, CBTU/97, CEB-Brasília/97, SERCONTEL -Paraná/98, ENERSUL/ MT/98. Trabalhou com planejamento, projeto e implantação de sistemas de grande porte via rádio e fibras ópticas nas empresas de consultoria técnica Engevix Engenharia S.A.- Belo Horizonte (85 à 89) e Main Engenharia S.A.- São Paulo (83 a 85), executando serviços para: Eletronorte (obras de ampliação do sistemas de telecomunicações em subestações no Mato Grosso), Furnas, Itaipu Binacional, Chesf, Eletrosul e Cemig. Trabalhou na Italtel Società Italiana Telecommunicazione - Belo Horizonte (79 a 82) com a fabricação de equipamentos de ondas portadoras e ESCELSA (78 a 81). Integrou comissões da ABNT - Associação Brasileira de Normas Técnicas, 81/82. Possui diversas publicações técnicas em revistas nacionais e internacionais. Obteve os prêmios Telexpo/Equitel (3.º lugar-1995, 2.º lugar-1996, 4.º lugar-1999).

Iwanir Araújo da Silva Júnior nasceu em Formiga, MG, em 27 de setembro de 1977. Possui os títulos: Auxiliar Técnico de Química (Dom Belchior, 1995), Engenheiro Eletricista (CEFET-MG, 2001) e Mestre em Engenharia de Telecomunicações (Inatel, 2003).

Em 2003 participou do corpo docente do Departamento Acadêmico de Engenharia Elétrica – DAEE do CEFET-MG lecionando disciplinas de Eletrônica Digital, Sistemas de Comunicação e Desenho Técnico. De 2004 a 2005 atuou na área de engenharia de manutenção da MRS Logística S.A. desenvolvendo trabalhos no campo de confiabilidade de locomotivas e foi coordenador do Grupo de Controle de Perdas (SGMASST) sendo responsável pelas Oficinas de Manutenção de Conselheiro Lafaiete e Jeceaba. Na mesma empresa, atuou como Coordenador de Qualidade sendo responsável pela implantação do Sistema ISO 9001:2000 das mesmas oficinas e também como coordenador da área manutenção de locomotivas de toda as oficinas da MRS Logística S. A. sendo responsável por integrar toda a parte de implantação do Sistema de Gestão de Qualidade. Foi coordenador de implantação na área de manutenção do projeto piloto do Sistema de Gestão Integrado envolvendo as Normas ISO 9001:2000, OHSAS 18000:2002 e ISO 14001:2004. Desde 2006 trabalha na VALE onde atuou como Supervisor de Programação e Controle da Manutenção e no Grupo de Análise de Falhas ligado à Gerência de Sinalização, Telecomunicações e Energia da Estrada de Ferro Vitória Minas - EFVM. Atualmente atua na implantação do Sistema de Gerenciamento da Manutenção – SGM em toda as ferrovias VALE (EFVM, EFC e FCA), faz parte do Subcomitê de Planejamento e Gestão da Manutenção VALE e está ligado a Gerência de Gestão de Tecnologia. Tem interesse em pesquisas no campo de fibras ópticas e telecomunicações em geral bem como na área de engenharia elétrica, possui também interesse na área de confiabilidade, manutenção elétrica e eletroeletrônica, projetos, qualidade, sistemas de gerenciamento, análise de falha da manutenção, LCC, RCM e Planejamento e Controle buscando conhecimentos relacionados à Excelência da Manutenção e implementação de melhorias em processos de manutenção.

Análise da Intensidade de Campo Elétrico de Estações Rádio-Base

Marco Antonio Brasil Terada

***Abstract*—This work discusses in detail the directional properties of the radiation of radio base antennas in cellular systems with respect to various variables, in addition to the usually employed variable of the distance from the bottom of the radio base tower. It is demonstrated that the electric field intensity peaks at different distances from the bottom of the radio base tower depending on these variables, leading to the conclusion that fixing a minimum distance from populated areas for the installation of radio bases is not the proper way to ensure safety.**

***Index Terms*—Applied Electromagnetics, Antennas and Propagation, Mobile Communications, Biological Effects.**

***Resumo*—Este trabalho[1] investiga em detalhe as propriedades direcionais de irradiação de antenas rádio-base em sistemas de comunicações celulares em função de diversas variáveis, além da usual distância da base da estação rádio-base. É mostrado que o máximo da intensidade do campo elétrico ocorre a distâncias diferentes da base da ERB dependendo destas outras variáveis, levando à conclusão que se fixar uma distância mínima de regiões povoadas para a instalação de ERBs não é a maneira adequada para se assegurar a segurança e o bem-estar da população.**

***Palavras chave*—Eletromagnetismo Aplicado, Antenas e Propagação, Comunicações Móveis, Efeitos Biológicos.**

I. INTRODUÇÃO

O explosivo recrudescimento de serviços baseados em sistemas de comunicações sem-fio representa um mercado que diretamente suporta e viabiliza o processo de globalização através de projetos de valor agregado superior a centenas de bilhões de reais. Correntemente existem cerca de 2 bilhões e 740 milhões de usuários de comunicações móveis celulares e PCS no mundo [1]. Esses dados não incluem usuários que estarão exclusivamente imersos em redes locais e pessoais, apesar de um certo grau de compartilhamento seja esperado. A proliferação de sistemas de comunicações sem-fio com alto grau de sofisticação necessitará do projeto e implementação de novas configurações, a serem usadas no atendimento à demandas técnicas cada vez mais exigentes derivadas de novos serviços e aplicações.

O mercado Brasileiro de comunicações celulares segue um caminho similar, com quase 100 milhões de usuários de serviços de comunicações sem-fio e de voz no final de 2006. No entanto, o desenvolvimento técnico e a implementa-ção de novos serviços têm sofrido atrasos consideráveis nestes últimos anos. Isto se deve ao desconhecimento da população em geral quanto aos fundamentos de funcionamento das comunicações celulares e seus possíveis efeitos em nossas vidas. Estas dúvidas, ainda que justificáveis do ponto de vista da proteção de nosso bem-estar e saúde, levaram à elaboração de diversas leis que tentam restringir a instalação de Estações Rádio-Base (ERBs) perto de áreas povoadas. Estações Rádio-Base são estações de radiocomunicações de base do Serviço Móvel Pessoal (a telefonia móvel), usadas para radiocomunicação com estações móveis, ou seja, com os aparelhos terminais da telefonia móvel (os telefones celulares)[2].

Este trabalho tem por objetivo investigar as propriedades direcionais de radiação da antenas de ERBs. Será também evidenciado que as leis anteriormente mencionadas são arbitrárias e sem consistência científica, podendo inclusive introduzir problemas ao invés de solucioná-los. A maioria destas leis requer que a instalação de ERBs ocorra à uma distância mínima de escolas e unidades imobiliárias, e obrigam o re-posicionamento de ERBs que já estejam instaladas e em operação em distâncias inferiores a estes mínimos. Na realidade, a distância da base da ERB não é a única variável que deve ser levada em consideração, e o re-posicionamento da ERB para uma distância de 50 m [2], por exemplo, pode aumentar a radiação eletromagnética nas unidades imobiliárias que as leis estão tentando proteger. É importante se ressaltar que todas as ERBs conhecidas do autor atendem às especificações da Agência Nacional de Telecomunicações (ANATEL), as quais ao contrário de leis como a [2] não obrigam que as ERBs sejam instaladas à uma distância mínima de unidades imobiliárias, mas requerem que a irradiação eletromagnética em todas as regiões povoadas esteja abaixo de valores mínimos de acordo com a freqüência de operação [3].

II. CONSIDERAÇÕES PRELIMINARES

Inicialmente, esclarece-se que a distância da base da antena não é a única nem a mais importante variável a partir da qual se define qual a intensidade do campo eletromagnético proporcionado por essa mesma antena [4,5]. Há diversas outras variáveis ao lado da distância da base da antena que são relevantes para se apurar qual a intensidade do campo eletromagnético produzido pela antena. Sem o intuito de esgotar o rol dessas variáveis, o que seria desnecessário para o escopo desse trabalho, pode-se destacar dentre elas: (1) a altura de instalação da antena; (2) a altura, em relação ao solo,

[1] Manuscrito recebido em 18 de janeiro de 2007; revisado em 2 abril de 2007.
Departamento de Engenharia Elétrica - Universidade de Brasília
Caixa Postal 4386 – Brasília/DF 70919-970
Este trabalho foi desenvolvido com suporte da Associação Nacional das Operadoras Celulares (ACEL), através de um contrato entre a ACEL e a Universidade de Brasília.

[2] Cf. Art. 3º, XIV, do Regulamento do Serviço Móvel Pessoal, aprovado pela Resolução nº. 316, de 27/09/2002, da Anatel.

do ponto em que se deseja medir a intensidade do campo eletromagnético; (3) o ângulo de inclinação da antena; e (4) a potência efetivamente irradiada pela antena[3].

A representação gráfica que segue (Figura 1) apresenta a geometria dessa questão relativamente às duas primeiras variáveis destacadas:

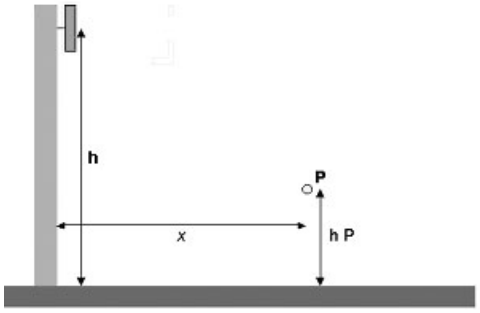


Figura 1 - Distâncias relevantes para apuração do campo eletromagnético.

Considerado um ponto qualquer no espaço (ponto *P*, na Figura 1), a intensidade do campo eletromagnético nesse ponto em decorrência da antena ali representada variará não apenas em função da distância desse ponto até a torre onde está instalada a antena (*x*), mas também em função da distância desse ponto ao chão (*hP*), e da altura de instalação da antena (*h*).

O ângulo de inclinação da antena também é relevante, pois as antenas de telefonia móvel são direcionais[4]. As Figuras 2 e 3 objetivam apontar que ângulo seria este.

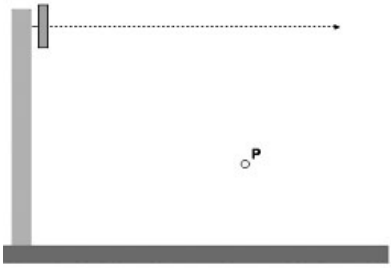


Figura 2 – Antena sem inclinação.

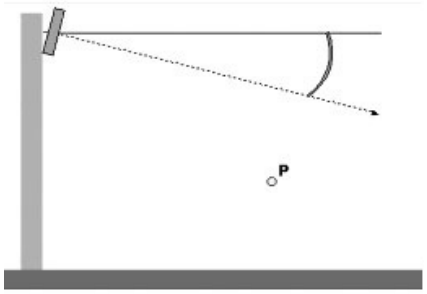


Figura 3 – Antena com inclinação.

Como se constata pela comparação entre as figuras, tem-se na Figura 2 um ângulo de 0º quando a antena é paralela à torre em que está instalada. Na Figura 3 a antena é inclinada em direção ao solo, daí surgindo um ângulo. Trata-se do ângulo de inclinação do eixo da antena, que faz com que esse eixo se aproxime do ponto *P*, no exemplo citado, fazendo com que seja maior a intensidade do campo eletromagnético no mesmo. Para um correto entendimento do funcionamento dos sistemas de comunicações celulares, é poratanto necessário que se esclareça não só como varia a intensidade do campo eletromagnético em função da distância *x* como também como varia em função das demais variáveis apontadas (altura da antena, altura do ponto *P* e ângulo de inclinação da antena).

Esclarece-se, ainda, que seja nos regulamentos tratando da matéria, seja na legislação, ora se fala em *campo elétrico*, ora em *campo eletromagnético*. Os dois conceitos se distinguem, dado que o campo eletromagnético é a co-existência dos campos elétrico e magnético variando no tempo simultaneamente (um gera o outro), mas para o que interessa ao presente trabalho, é importante deixar destacado que os valores da intensidade do campo elétrico e do campo magnético são sempre diretamente proporcionais um ao outro. No escopo deste trabalho, a intensidade do campo magnético é simplesmente a intensidade do campo elétrico dividida por 377 [4,5]. As unidades são diferentes, no entanto, sendo que o campo elétrico é dado em V/m (Volts por metro) e o campo magnético em A/m (Amperes por metro). Não faz sentido se falar em intensidade de campo eletromagnético, a menos que esteja implícita as intensidades dos campos elétrico e magnético separadamente. Na continuidade deste trabalho, serão utilizados sempre os dados do campo elétrico, ficando, porém, claro que a toda e qualquer variação do campo elétrico varia em igual proporção o campo magnético.

Apresentamos inicialmente um gráfico que aponta a variação da intensidade do campo elétrico gerado por uma dada antena à medida que se varia a distância à base da ERB. Para tanto, utilizando técnicas analíticas de análise de antenas, bem como utilizando o programa computacional GRADMAX [6,7], desenvolvido pelo autor, é apresentado o gráfico que segue (Figura 4), calculado a partir de uma antena instalada a dez metros de altura (*h* = 10m), distância do solo ao ponto analisado de um metro e meio (*hP* = 1,5m) e com uma

[3] A potência efetivamente irradiada, *e.r.p.*, é definida como o produto da potência de entrada da antena pelo ganho linear da antena com relação à antena do tipo dipolo. Nesse sentido, também, a bibliografia já citada [4,5].

[4] Antenas direcionais apresentam uma *e.r.p.* que é função da direção, ou ângulo, entre a antena e o ponto de observação. Conforme podemos ver das Figs. 2 e 3, o ângulo entre a seta de linha pontilhada e o ponto P varia de acordo com o ângulo de inclinação da antena, o que resulta em campos eletromagnéticos diferentes no ponto *P* se a antena for direcional. Uma antena omnidirecional produz campos uniformes ao seu redor, independente da direção. Desta forma, os campos produzidos por uma antena omnidirecional no ponto *P* das Figs. 2 e 3 seriam os mesmos independente do ângulo de inclinação da antena, em contraste com as antenas direcionais que dependem do mesmo.

potência efetivamente irradiada de 2512 Watts[5] e um ângulo de inclinação da antena de 8º.

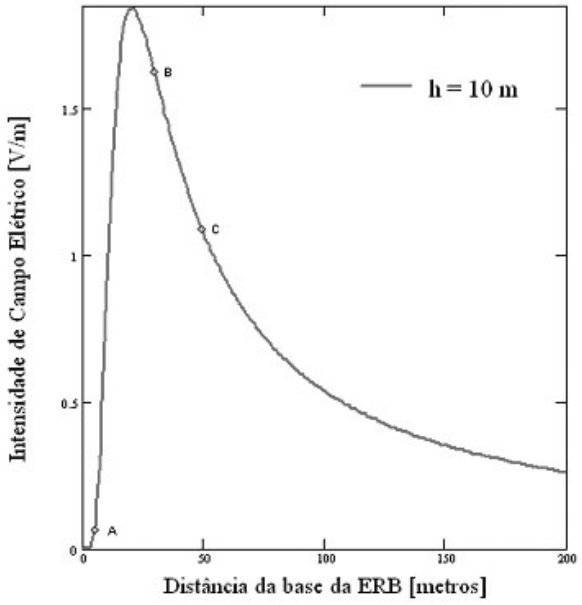


Figura 4 - Intensidade de campo elétrico em função da distância *x* da base da ERB. A análise foi feita para uma antena instalada a 10 metros e com uma inclinação de 8º.

No gráfico em questão (Figura 4), o eixo horizontal mostra a distância *x* em metros da base da ERB e o eixo vertical mostra a intensidade de campo elétrico em V/m (Volts por metro). Da Figura 4, conclui-se que:

- a intensidade do campo elétrico é a mais baixa exatamente quando são menores as distâncias do ponto até a torre da antena (valores de *x* próximos a zero);
- os valores da intensidade do campo elétrico vão aumentando progressivamente, atingindo o valor máximo de 1,850 V/m quando a distância *x* é de 20 m;

A TABELA I mostra os valores da intensidade do campo elétrico para as distâncias *x* = 5 m, 30 m e 50 m. Assim, quando se está a 1,5 m de altura e 5,0 metros de distância de uma torre de antena com as especificações dadas, a intensidade do campo elétrico é de 0,077 Volts por metro. À medida que a distância *x* vai aumentando (à medida que se afasta da base da antena), esse valor aumenta, até atingir 1,850 V/m a 20 metros da base da antena. Começa a diminuir, chegando a 1,615 V/m a uma distância de 30 metros e a 1,076 V/m a cinqüenta metros, prosseguindo em queda.

TABELA I
INTENSIDADE DO CAMPO ELÉTRICO.

	x [metros]	Intensidade do Campo Elétrico [V/m]
A	5	0,077
B	30	1,615
C	50	1,076

III. ALTERAÇÕES DA INTENSIDADE DO CAMPO ELÉTRICO EM FUNÇÃO DAS DEMAIS VARIÁVEIS

A. Variação da intensidade do campo elétrico em função da variação da altura da antena

O gráfico seguinte (Figura 5) mostra o comportamento da intensidade do campo elétrico em função da distância *x* da base da ERB para três diferentes alturas de antena, quais sejam, 10 m, 20 m e 30 m, mantendo-se as demais variáveis constantes[6].

A Tabela II lista os valores da intensidade do campo elétrico para as distâncias *x* = 5 m, 30 m e 50 m, bem como para o pico do campo em cada uma das torres.

A partir da comparação das três curvas da Fig. 5 e dos valores listados na Tabela II, constata-se que a intensidade do campo elétrico varia sensivelmente em função da variação da altura de instalação da antena, levando às seguintes conclusões:

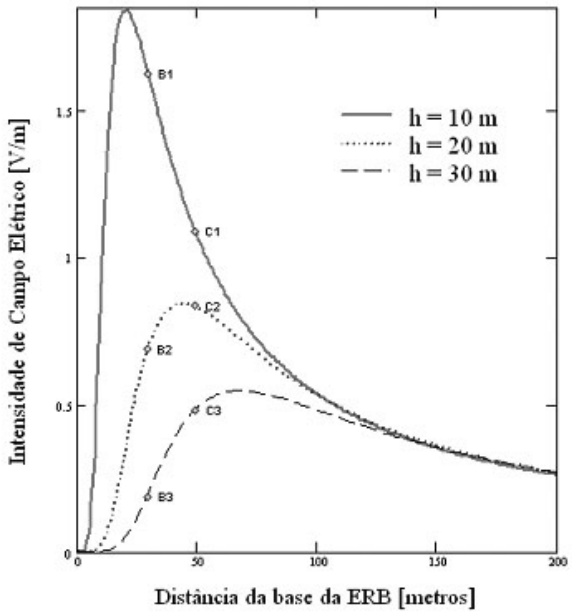


Figura 5 - Intensidade de campo elétrico em função da distância *x* da base da ERB considerando três diferentes alturas de instalação da antena**.**

[5] Essa especificação corresponde ao valor máximo admitido na regulamentação editada pela Anatel para o Serviço Móvel Pessoal [8], ou seja, corresponde à especificação de *e.r.p. (potência efetivamente irradiada)* de 64 dBm para antenas que operam na freqüência de 869 a 894 MHz (Banda A da telefonia móvel). com uma potência de entrada de 100 W. Por definição, *e.r.p.* [dBm] = 10 log (*e.r.p.* em mW), daí resultando que 64 dBm correspondem a 2512W (Nesse sentido, também, a bibliografia já citada nas notas [4,5]).

[6] ponto de observação (ponto "*P*") distante 1,5 m do nível do solo, inclinação da antena de 8º, potência efetivamente irradiada de 2512 W.

TABELA II
INTENSIDADE DE CAMPO ELÉTRICO EM FUNÇÃO DA DISTÂNCIA X DA BASE DA ERB PARA DIVERSAS ALTURAS DA ANTENA E UMA INCLINAÇÃO DA ANTENA DE 8º.

		Diferentes alturas de instalação da Antena		
		h = 10 m	*h* = 20 m	*h* = 30 m
	Distância da torre	Intensidade do Campo Elétrico em Volts por metro		
A	5m	0,077	0,000635	0,00004749
B	30m	1,615	0,705	0,207
C	50m	1,076	0,833	0,492
Pico	variável	1,850	0,847	0,550

- Um ponto situado a 50 metros de uma antena instalada a 10 metros de altura está sujeito a uma intensidade do campo elétrico superior ao pico do campo que pode ser produzido por antenas semelhantes instaladas a 20 ou 30 metros de altura.

- O valor máximo da intensidade do campo elétrico (pico de cada curva) é menor à medida que se aumenta a altura de instalação da antena:
  - para uma antena instalada a 10 metros de altura, com as demais especificações já apontadas[7], o pico é de 1,850 Volts por metro;
  - Caso essa mesma antena fosse instalada a 20 metros de altura, o pico seria reduzido a 0,847 Volts por metro;
  - Caso a antena fosse instalada a 30 metros de altura, o pico passaria a ser de 0,550 Volts por metro.

- Os valores máximos (o pico de cada curva), a medida em que a altura de instalação da antena é maior, são atingidos em pontos mais distantes da torre:
  - para a antena instalada a 10 metros de altura, o ponto a 1,5m de altura do solo (ponto *P*, de análise) em que se observa a maior intensidade do campo elétrico (pico) está a 20 metros de distância da torre;
  - para a mesma antena instalada a 20 metros de altura, o pico passa a ser registrado a 44 metros de distância da torre;
  - para a antena instalada a 30 metros de altura, o pico está a 68 metros de distância da torre.

- Tomando-se em consideração uma distância específica da torre de instalação da antena (de 50m, por exemplo), constata-se que a intensidade do campo elétrico não só tem valores diferentes como está em momentos distintos da curva:
  - para a antena instalada a 10 metros de altura, a intensidade do campo elétrico a 50 metros da antena é de 1,076 V/m e está em momento de curva descendente acentuada;
  - para a antena instalada a 20 metros de altura, a intensidade do campo elétrico a 50 metros da antena é de 0,833 V/m e está praticamente no pico da curva, que é apenas 1,7% superior a ele (0,847 V/m) e foi atingido seis metros antes (44 m);
  - para a antena instalada a 30 metros de altura, a intensidade do campo elétrico a 50 metros da antena é de 0,492 V/m e ainda não atingiu o pico, que se verificará a 18 metros adiante (68 m).

[7] ponto de observação (ponto "P") distante 1,5 m do nível do solo, inclinação da antena de 8º, potência efetivamente irradiada de 2512 W.

B. Variação da intensidade do campo elétrico em função da variação do ângulo de inclinação da antena

O próximo gráfico (Figura 6) mostra o comportamento da intensidade do campo elétrico em função da distância *x* da base da ERB para três diferentes ângulos de inclinação da antena, quais sejam, 8º, 4º e 2º, mantendo-se as demais variáveis constantes[8]. Em seguida, a Tabela III aponta os valores da intensidade do campo elétrico para as distâncias *x* = 5 m, 30 m e 50 m da torre da antena, bem como para o pico da intensidade do campo elétrico correspondente a cada uma das alturas de instalação das antenas.

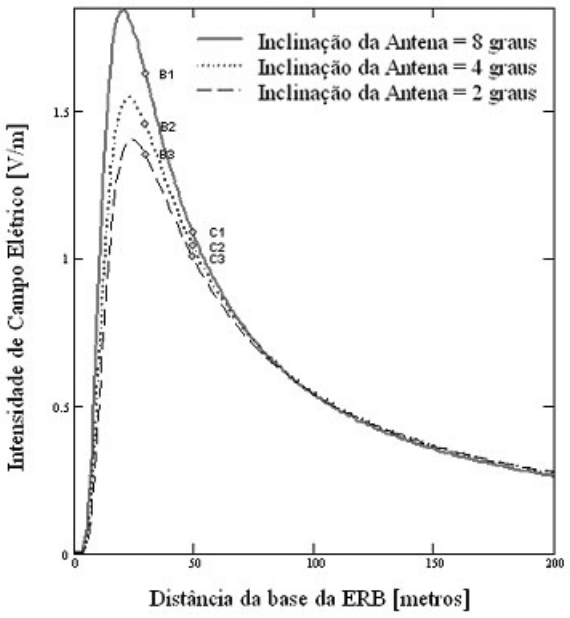


Figura 6 - Intensidade de campo elétrico em função da distância *x* da base da ERB considerando três diferentes ângulos de inclinação da antena**.**

Aos dados mostrados, acrescenta-se que ainda que na representação gráfica isso não transpareça de modo tão claro, também há variação do pico de cada uma das curvas (ponto em que há maior intensidade do campo elétrico proporcionado pela antena) quanto à distância em que o mesmo se verifica. Assim, os valores máximos para as inclinações de 8º, 4º e 2º são verificados respectivamente à distância de *x* = 20 m, 23 m e 24 m.

[8] ponto de observação (ponto "P") distante 1,5 m do nível do solo, altura da antena de 10m, potência efetivamente irradiada de 2512 W.

Novamente se pode constatar que à medida que a inclinação é menor, menores são os picos das curvas, o que significa que são menores as intensidades dos campos elétricos ao nível do solo ou, como nos dados apurados, a 1,5 metros de altura do solo e tais picos são atingidos a distâncias mais distantes da torre.

TABELA III
INTENSIDADE DE CAMPO ELÉTRICO EM FUNÇÃO DA DISTÂNCIA *X* DA BASE DA ERB PARA UMA ALTURA DA ANTENA DE 10 M E DIVERSAS INCLINAÇÕES DA ANTENA.

		Diferentes ângulos de inclinação da Antena		
		8º	4º	2º
	Distância da torre	Intensidade do Campo Elétrico em Volts por metro		
A	5m	0,077	0,033	0,021
B	30m	1,615	1,448	1,349
C	50m	1,076	1,034	0,997
Pico	variável	1,850	1,545	1,404

C. Variação da intensidade do campo elétrico em função da variação da potência efetivamente irradiada da antena

O gráfico apresentado a seguir (Figura 7) mostra o comportamento da intensidade do campo elétrico em função da distância *x* da base da ERB para três diferentes potências efetivamente irradiadas da antena, 2512 W, 1256 W e 754 W, mantendo-se as demais variáveis constantes[9]. Em seguida, a Tabela IV aponta as intensidades dos valores do campo elétrico para as distâncias *x* = 5 m, 30 m e 50 m da torre da antena, bem como para o pico do campo em cada uma das torres.

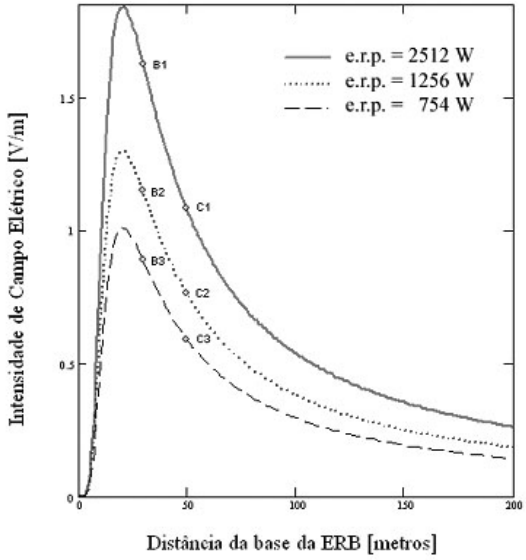


Figura 7 - Intensidade de campo elétrico em função da distância *x* da base da ERB considerando três diferentes potências efetivamente irradiadas da antena.

[9] ponto de observação (ponto "P") distante 1,5 m do nível do solo, altura da antena de 10m, ângulo de inclinação de 8º

TABELA IV
INTENSIDADE DE CAMPO ELÉTRICO EM FUNÇÃO DA DISTÂNCIA *X* DA BASE DA ERB PARA UMA ALTURA DA ANTENA DE 10 M E DIVERSAS POTÊNCIAS DE ENTRADA.

		Diferentes potências efetivamente irradiadas		
		2512 W	1256 W	754 W
	Distância da torre	Intensidade do Campo Elétrico em Volts por metro		
A	5m	0,077	0,018	0,014
B	30m	1,615	1,142	0,885
C	50m	1,076	0,761	0,589
Pico	20m	1,850	1,303	1,009

Nota-se, a partir da Figura 7 que o pico das curvas, representando o máximo de intensidade do campo elétrico, ocorre sempre a uma mesma distância da base da ERB independente da potência da antena.

- As três curvas têm esse pico a 20 m de distância da torre.

Conclui-se, também, que quanto menor é a potência da antena, menor é o valor da intensidade do campo elétrico a uma mesma distância da base da ERB:

- para uma antena com potência efetivamente irradiada de 2512 W o pico é de 1,850 Volts por metro;
- para uma antena com potência efetivamente irradiada de 1256 W o pico é de 1,303 Volts por metro;
- para uma antena com potência efetivamente irradiada de 754 W o pico é de 1,009 Volts por metro.

IV. CONCLUSÕES

As propriedades direcionais de radiação de antenas rádio-base foram apresentadas e discutidas em detalhe neste trabalho. Foram consideradas diversas variáveis além da usual distância à base da torre da estação rádio-base (ERB). Foi demonstrado que o máximo da intensidade do campo elétrico ocorre a distâncias diferentes da base da ERB dependendo destas outras variáveis, levando à conclusão que se fixar uma distância mínima de regiões povoadas para a instalação de ERBs não é a maneira adequada para se assegurar a segurança e o bem-estar da população. O re-posicionamento de uma ERB para uma distância de 50 m, por exemplo, pode na realidade aumentar a intensidade do campo elétrico em mais de mil vezes na área que se deseja proteger. Isto demonstra a arbitrariedade e a falta de consistência técnica na imposição de distâncias mínimas de proteção. O critério científico correto e seguro é o de se limitar o valor da intensidade do campo elétrico máximo resultante de todas as antenas em operação na faixa de freqüências considerada [3].

REFERÊNCIAS

[1] GSM World e Wireless Inteligence (http://www.gsmworld.com e https://www.wirelessintelligence.com) obtido em 18 de Janeiro de 2007.

[2] Lei nº. 3.446, de 23 de setembro de 2004, do Distrito Federal, publicada no Diário Oficial do Distrito Federal de 7 de outubro de 2004.

[3] "*Regulamento sobre limitação da exposição a campos elétricos, magnéticos e eletromagnéticos na faixa de radiofreqüências entre 9 kHz e 300 GHz*" aprovado pela Resolução Anatel nº. 303, de 2 de julho de 2002.

[4] L.C. Godara (ed.), *Handbook of Antennas in Wireless Communications*, CRC Press, 2002.

[5] W.L. Stutzman e G.A. Thiele, *Antenna Theory and Design*, 2ª Edição, John Wiley & Sons, 1998.

[6] M.A.B. Terada, A.J.M. Soares, F.C. Silva, e S.B.A. Fonseca, "GRADMAX: Um Programa Para Análise e Otimização de Antenas de Fios", *TELEMO92 (Software Educacional): X Simpósio Brasileiro de Telecomunicações e V Simpósio Brasileiro de Microondas,* Brasília - DF, pp. 15-18, Julho de 1992.

[7] M.A.B. Terada, A.J.M. Soares, F.C. Silva, e S.B.A. Fonseca, "Otimização de Antenas de Fios Através do Método do Gradiente Modificado", *TELEMO92: X Simpósio Brasileiro de Telecomunicações e V Simpósio Brasileiro de Microondas,* Brasília - DF, pp. 7-12, Julho de 1992.

[8] Resolução nº. 315 da ANATEL, de 19 de Setembro de 2002.

Marco Antonio Brasil Terada nasceu em São Paulo, SP, em Novembro de 1966. Recebeu os títulos de Engenheiro Eletricista e Mestre em Engenharia Elétrica pelo Departamento de Engenharia Elétrica da Universidade de Brasília (UnB) em 1989 e 1991, respectivamente. Em 1995 recebeu o título de Ph.D em Engenharia Elétrica pela Universidade Virginia Tech - EUA.

Desde Agosto de 1995 é professor da UnB, onde atua nas áreas de Antenas e Sistemas de Comunicações Sem-Fio. Foi Engenheiro Sênior e Gerente de Antenas do programa INTELSAT IX (sete satélites operando nas bandas C e Ku) na empresa INTELSAT (1998 a 2001), Washington – EUA, e professor da Universidade Estadual do Novo México – EUA. (2001 a 2004). Seus interesses correntes incluem a pesquisa e desenvolvimento de antenas e sistemas de comunicações com alta taxa de rejeição sistêmica (isolamento). É bolsista do CNPq em desenvolvimento tecnológico e extensão inovadora.

Convergência Tecnológica Aplicada à Integração de Sistemas em Telecomunicações

Emílio J. M. Arruda Filho[*/**], Elionai G. de Almeida Sobrinho[**], Silvia Biffignandi[*], Alberto Marino[*]

Abstract **– Drawing on LCR technology (Least Cost Routing) services that define the best route on communication using costs, this article presents the economic analyses made about the internal and external communication behavior of an evaluated case. The mathematic modeling validation based on LCR applied to the system integration on telecommunications is presented and proved by the integrations diversity on the technologies convergence. The convergent algorithm was developed using alternative technologies introduced in the previous article, which was presented as a starting point to develop the general analyses of the costs in the telecommunications service. The services convergence and the control failure in the telecommunications sector - based on the technologies complexity and the lack of technical knowledge about these - are focused viewing the economic sector, public utilities regulation and marketing of these products and services.**

Index Terms **– Technology Marketing, Convergence, VoIP, Cellular, LCR, Telecommunications, Cost benefits and logistic.**

Resumo **— Baseado na tecnologia de LCR (rota de menor custo), serviço que define a melhor rota de comunicação dado os custos, este artigo apresenta a análise econômica realizada sobre o comportamento de comunicação, interno e externo, em um caso avaliado. A validação do modelamento matemático, utilizando o LCR para integração de sistemas em telecomunicações, é apresentada e comprovada pela variação de integrações nas convergências tecnológicas. O algoritmo de convergência foi desenvolvido, focando-se os tipos de tecnologia existente em um trabalho precedente, onde foi apresentado como ponto de partida um *case* para desenvolver a análise geral de custos nos serviços de telecomunicações. A convergência dos serviços e a falha de controle no setor de telecomunicações, dada a complexidade e o desconhecimento técnico das tecnologias, são focadas sobre o ponto de vista do setor econômico, regulação do serviço público e marketing de produtos e serviços.**

Palavras-chaves **– Marketing Tecnológico, Convergência, Telefonia, VoIP, Celular, LCR, Custo-benefício e logística.**

I. INTRODUÇÃO

O uso dos serviços de telecomunicações vem, a cada dia, se tornando inevitável e necessário, porém o conhecimento sobre o aproveitamento do melhor desempenho, baseado em qualidade e custos, não foi até o momento muito bem definido [1, 2]. Para consumidores, sejam eles empresas ou clientes individuais, essa escolha vêm sendo definida por especialistas em tecnologia ou especialistas em economia. Como este assunto ainda não foi tratado de forma convergente [3], sem estar vinculado a uma única aplicação, não existe em funcionamento um modelo que possa validar a melhor solução em tempo real, onde os custos dos serviços utilizados estejam ligados diretamente a uma qualidade técnica satisfatória, definindo as tecnologias necessárias para desenvolvimento de um serviço com qualidade.

O objetivo deste artigo é validar estatisticamente o modelamento proposto em [4], para uma empresa com 6 filiais e com diversos serviços de telecomunicações existentes (Universidade de Bergamo), e garantir que este sirva para qualquer tipo de empresa analisada.

O modelo utilizado propõe uma solução genérica baseada em um caso amplo, o qual inclui todas as possibilidades tecnologicamente conhecidas, onde os resultados gerados pelo software de tarifação foram organizados e normalizados tal que pudessem ser inseridos no software de estatística SAS [5]. Estes mesmos dados, em [4], foram analisados em valores percentuais, já neste trabalho, os valores medidos individualmente para cada tipo de chamada, horário, custo, tecnologia e outros, são então confrontados com as avaliações de correlação e histograma de comportamento percebido pelo método.

Com esta avaliação bem mais detalhada, os dados de seis meses de contas dos sistemas de telefonia fixa, celular, internet e intranet (ISDN – Integrated Services Digital Network) da empresa, transformam-se em uma média de 60000 valores por mês a serem trabalhados. Outra diferença importante com relação ao caso, é que agora não se trabalha apenas pelos resultados de rota de comunicação, mas pela redução de canais destas rotas a partir dos horários de utilização, já que é possível definir os valores máximos de canais necessários na rede, de acordo com os padrões dos tipos de serviços estabelecidos [6].

Com o modelamento de um caso amplo e genérico, avalia-se que os tipos de serviços existentes nesta universidade, são muito semelhantes ao de qualquer outra empresa, pois não apresentam comportamentos irregulares ou que sejam definidos fora dos padrões de mercado para qualquer empresa. Fazendo-se a correlação dos dados, pode-se avaliar que pontos são comuns entre os valores medidos, e que pontos têm maior influência sobre o comportamento dos custos. Já com os histogramas desenvolvidos, é possível avaliar o comportamento dos serviços de acordo com a hora, dia da semana e os meses (período), podendo ainda se obter os pontos críticos de utilização dos serviços na rede de comunicação.

A principal contribuição deste artigo é a de integrar rotas [7, 8] entre tecnologias homogêneas e heterogêneas, porém com

Manuscrito recebido em 13 de julho de 2007; revisado em 23 de novembro de 2007.

[*] UNIBG – Università degli Studi di Bergamo – Via dei Caniana 2, 24127 Bergamo – Itália

[**] IESAM – Instituto de Estudos Superiores da Amazônia – Av. Governador José Malcher, 1148 Belém – Pará – Brasil

as mesmas possibilidades de serviços, onde em [4], a utilização de diversas rotas era apenas entre tecnologias heterogêneas. Como exemplo, no projeto anterior o sistema poderia escolher entre efetuar uma ligação para um celular externo, saindo pela rede fixa ou celular, ao interno da empresa. Já no sistema proposto neste trabalho, além de se poder escolher entre tecnologias diferentes, também se poderá escolher na mesma tecnologia, entre os concorrentes com o mesmo serviço, aumentando assim a quantidade de rotas possíveis para o sistema. Em outras palavras, além das rotas de uma tecnologia à outra, como efetuar a ligação do sistema celular ou fixo internos, existirá a partir deste momento, dentro da rede fixa e celular coorporativa da empresa, duas ou três rotas concorrentes.

Outro fator interessante, incluído na validação, é o ponto de fronteira [9], o qual faz com que o modelamento interno seja dinâmico, podendo sempre atualizar-se para que no instante em que a relação tempo x preço ultrapassasse certo limite, a gerência do sistema seja informada, fazendo que o sistema receba novas imposições.

A diferença deste artigo, não está em realizar o que um LCR comum faz, mas sim em determinar, entre diferentes serviços como: telefonia celular, telefonia fixa, VoIP (Voice over IP), acesso à internet, "wireless" e outros, qual o investimento lógico para realizá-los [10, 11] com um valor ótimo baseado na dupla necessidade: qualidade e custo. Deve-se estar atento para que os custos; como por exemplo, de operadoras de telefonia celular, as quais mudam constantemente seus planos de serviços, dada a acirrada competição existente entre as mesmas*;* possam ser atualizados no sistema sem a necessidade de se mudar o "hardware", pois não se pode mudar constantemente os equipamentos, fornecedores ou instalações, a cada momento em que se apresentam mudanças no mercado.

Desta forma, com o estudo da média e predição em função da avaliação do período de análise da universidade (seis meses), busca-se a otimização no uso das rotas e estruturas existentes [12]. Tendo em vista todas as possibilidades, em função da estrutura disponível, projeta-se o modelo em sua forma intermediária com o uso conjunto de diversas operadoras, mesmo sendo da mesma tecnologia, o que permitirá a redução de custos futuros com mudanças de operadoras, por exemplo; porém obtém-se certa margem de segurança na largura de banda, onde deve-se utilizar simultaneamente as operadoras existentes tendo como base os horários críticos de utilização; isto obviamente sobrecarregará a operadora com o menor preço, porém haverá uma rota alternativa (segunda opção) para os horários de maior fluxo, onde se pode escolher qual operadora utilizar, uma vez que pelo menos duas destas estarão instaladas na empresa.

O nível de otimização [13] a partir da arquitetura apresentada, deve ser a melhor possível, tal que, caso haja necessidade de alocação de um novo equipamento, esta aquisição seja de baixo custo. Na maioria das vezes, este investimento representa a aquisição de "chips" de telefonia celular, cabo de entrada da operadora fixa ou internet e configuração no sistema existente, dado o número de canais para cada caso.

As mudanças constantes de valores de planos e custos de promoções, além das tecnologias nos sistemas de telecomunicações, fazem com que os preços variem muito sensivelmente [14] e, como estratégia de marketing de cada operadora, a cada momento existe alguma promoção baseada no tipo de cliente ou em novos modelos e/ou serviços praticados pelas operadoras.

Utilizando produtos básicos de telecomunicações e convergências nos serviços prestados [15], pode-se agora comutar automaticamente a utilização dos serviços com o algoritmo de LCR, porém deve-se ter sempre uma quantidade de rotas maior do que o necessário [16]. Esta quantidade será uma rota alternativa a qual pode ter sua utilização medida dado o porte de cada empresa e ainda poderá ser modificado de acordo com a dinâmica do uso dos serviços de telecomunicações.

A escolha, por exemplo, de duas ou mais operadoras e serviços, traz como vantagem a possibilidade da rota alternativa onde, em caso de falha de uma operadora, tem-se uma ou mais saídas extras na empresa, o que garante sempre a disponibilidade desta, sem deixá-la isolada do ponto de vista da comunicação. O modelamento matemático é agora garantido [12] sendo usual para qualquer tipo de empresa, pois este funcionará sob uma forma lógica de conectividade, sem a interferência de dados externos.

II. Modelamento da Integração de Sistemas

A necessidade da integração de sistemas [17] é devida basicamente ao custo destes, pois a qualidade de funcionamento de cada um, separadamente, pode tranquilamente atender as expectativas de comunicação entre os demais. A partir de um telefone fixo, pode-se ligar para outro número fixo ou celular, realizar um chamada nacional ou internacional e, até mesmo, para número virtual na internet ou um computador com VoIP, por exemplo; sendo que dificilmente alguém notaria a diferença destas, em caso de mudança na origem da chamada.

A vantagem que se tem em uma integração de sistemas, é determinada pela possibilidade de garantir sempre a disponibilidade de algum canal de comunicação ("backup") para um grupo de pessoas ou clientes. Para alguns, o fato de sempre estar disponível para comunicação, pode parecer fantástico; já para outros, isto pode ser um verdadeiro problema.

O interessante da integração, é que ela se apresenta como o número único, onde o telefone celular de um funcionário passa a ser o seu número interno (ramal) dentro de sua empresa, ou seja, quando este entra na empresa, automaticamente seu celular funciona como DECT – "Digital Enhanced Coordless Telecommunications" [18] e recebe o número interno desviado para o telefone móvel ou vice-versa. Em empresas onde os funcionários se movimentam bastante, este sempre estará comunicável quando estiver fora de sua sala, e ainda no momento em que este estiver em outra filial da empresa, o seu ramal é deslocado automaticamente com ele, dado a conexão automática das ERBs (Estações Rádio Bases) DECT da empresa.

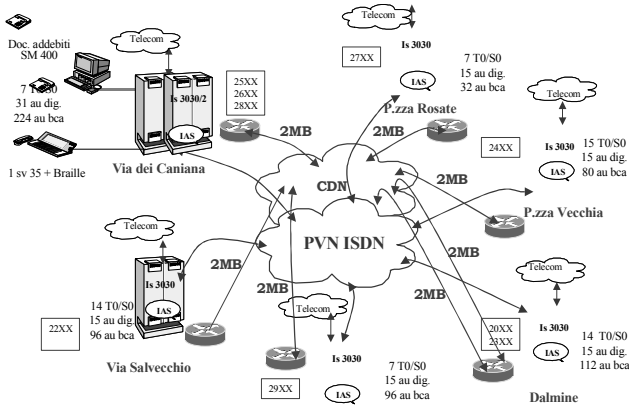
Além disto, o sistema integrado é capaz de gerenciar situações em que o telefone pode estar fora da área da empresa. Neste caso, o usuário pode estar em deslocamento de uma filial para outra e por isso o ramal não estará funcionando, porém a rede fixa interna da empresa detectará esta situação e, como

procedimento em tempo real, desviará automaticamente a ligação ou solicitará ao usuário se este pretende deixar um recado, enviar um fax, deixar uma mensagem via e-mail ou ainda conectar-se ao celular, podendo assim ser chamado através da operadora celular enquanto se desloca entre as empresas.

De forma geral, a combinação da quantidade de entradas e saídas de comunicações, se expande gerando diversas oportunidades de conexões, ampliando a possibilidade de redução dos custos em cada uma destas combinações. No caso dos custos serem altos para uma chamada entre telefonia fixa e telefonia celular, o sistema, inteligentemente, utiliza a melhor rota para realizar a chamada, efetuando o desvio em busca do número interno ou externo, ou o destinatário pode até mesmo ser um número IP, no caso do usuário estar com o celular desligado, porém conectado a computador com sua senha pessoal.

Os recursos e flexibilidade desta tecnologia já existem e estão em normal funcionamento no mercado [19], porém a integração atual é feita a partir da marca do produto, onde é necessário que todos os produtos sejam de uma mesma marca (homogêneos).

A figura 1 apresenta a rede atual do caso avaliado [4], onde a



plataforma deverá gerenciar todo o acesso conjunto das tecnologias citadas, otimizando as rotas e os custos do sistema implantado.

Fig. 1. Rede de dados e de telefonia atualmente existentes entre as filiais da Universidade de Bergamo.

A rede é composta por seis nós (sedes) com a topologia do tipo estrela. No centro é a sede de Caniana, onde as comunicações dedicadas de CDN (Conexão Numérica Digital), são todas efetuadas por um roteador 7200 Cisco, com 6 portas WAN ativas. A sede de Dalmine possui dois *links* de comunicações de 2MB enquanto que as outras sedes possuem um *link* de comunicação de 2MB cada uma.

Todas as sedes possuem centrais telefônicas, onde a entrada de todas as ligações, são realizadas por dois Links ISDN PRI na sede Caniana, que depois são direcionadas por link's ISDN BRI para as outras sedes.

Em [4], criou-se um modelamento para redução dos custos da rede apresentada. Neste trabalho foi realizada uma atualização deste modelamento. O modelo básico de custos nos serviços de telecomunicações é derivado de quatro pontos principais, que são: o tráfego, a assinatura das linhas telefônicas, o link digital de dados e voz, e o aluguel ou manutenção da central de telecomunicações utilizada. A tabela 1 apresenta os serviços de telecomunicações com os custos fixos e o tráfego (custo variável), antes e depois da integração de sistemas que levaram a uma projeção de mais de 30% na redução dos custos iniciais. O estudo inicial partiu da consideração do tráfego como sendo de custo variável no sistema, para depois integrar os custos fixos com suas variações em função do tráfego.

TABELA I
DISTRIBUIÇÃO PERCENTUAL DOS SERVIÇOS DE TELECOMUNICAÇÕES DA UNIV. DE BERGAMO ANTES E DEPOIS DA INTEGRAÇÃO DAS TECNOLOGIAS.

Ano	Assinat. telefonia fixa %	Bem material PABX %	Link Dados %	Custos Fixos%	Tráfego %	TOTAL %
Antes	12,69	32,91	20,02	65,62	34,38	100,00
Depois	8,30	20,00	24,02	32,32	15,15	67,47

Como ponto de partida, a equação (1) apresenta o estudo do tráfego baseado no índice de custo, resultante da multiplicação do preço por tipo de serviço, pela quantidade de utilização deste.

$$K_t = \sum_{m=1}^{S} \sum_{n=1}^{T} P_m . q_n \tag{1}$$

P é o preço do serviço de telecomunicações, q a quantidade do produto [20], que neste caso é o tempo de realização da chamada. Define-se m como o tipo de serviço que varia seu valor para o tipo de chamada realizada, seja uma ligação local até uma ligação internacional, e n a variação de tempo, que é desenvolvida a ligação, podendo variar de 1 segundo até um valor T, de tempo.

Da mesma forma que o sistema de telecomunicações é dividido em diversas componentes, o tráfego também se apresenta dividido em cinco principais componentes de comunicação de voz, sendo do tipo local, comunicação nacional, internacional e comunicação móvel nacional ou internacional, além de ligações especiais como serviços das operadoras e parceiros. Seguindo a estrutura da equação (1), encontra-se a equação (2) que é uma expansão detalhada dos serviços.

$$\begin{aligned} K_t = & \sum_{t=1}^{t=t'} P_{Local} . \left(q_{L1}(t) + q_{L2}(t) \right) + \\ & \sum_{i=1}^{R} P_{Nacional} \left[i \,.\, s(Q) \right] . \left\{ q_{N1} \left[i \,.\, s(Q) \right] + q_{N2} \left[i \,.\, s(Q) \right] \right\} + \\ & \sum_{n=1}^{T} P_{Mobile} . \left(q_{M1}(n) + q_{M2}(n) \right) + \\ & \sum_{j=1}^{T'} P_{Intern} \left[j \,.\, s(Q) \right] . \left\{ q_{I1} \left[j \,.\, s(Q) \right] + q_{I2} \left[j \,.\, s(Q) \right] \right\} + \\ & \sum_{v=1}^{X} P_{Others}(v) . q_{O1}(v) \end{aligned} \tag{2}$$

Na equação (2), os índices L, N, M, I e O conectados ao valor de q (tempo de conectividade), significam ligações locais, nacionais, móveis, internacionais e outros respectivamente, definindo o tipo e tempo de comunicação para cada serviço. Os índices 1 e 2, aliados aos índices de tipo de ligação,

identificam o tipo de destinatário para as ligações efetuadas, sendo 1 para empresa ou cliente genérico e 2 para empresa filial, a qual pode ser eliminada da equação, caso esta situação não exista.

As variáveis s(Q), i, j e v são, respectivamente, os operadores fixos (dada a região de comunicação), região nacional que efetuará a comunicação, região internacional (dadas as regras de divisão de telecomunicações internacionais) e tipos de serviços especiais em telefonia. Todas estas variáveis são necessárias para se definir mudanças de rota e classificação de preços pelo sistema.

Seguindo o mesmo raciocínio realizado para a expansão de (1) em (2), distinguem-se os preços e expande-se a equação em (3) para que se possam analisar os custos dos serviços de comunicação com outras empresas e os custos de comunicação entre as filiais da empresa analisada. Este fator é bastante fundamental para se definir a implantação ou não de algum investimento em novas tecnologias e criando novas rotas de comunicação entre as filiais como rotas do tipo VoIP, por exemplo.

$$\begin{aligned} K_t = & \sum_{t=1}^{t=t'} P_{L1}\, x\, q_{L1}(t) \; + \; \sum_{t=1}^{t=t'} P_{L2}\, x\, q_{L2}(t) \; + \\ & \sum_{i=1}^{R} P_{N1}\left[i \,.\, s(Q)\right] x\, q_{N1}\left[i \,.\, s(Q)\right] \; + \\ & \sum_{i=1}^{R} P_{N2}\left[i \,.\, s(Q)\right] x\, q_{N2}\left[i \,.\, s(Q)\right] + \\ & \sum_{n=1}^{T} P_{M1}(n)\, x\, q_{M1}(n) + \sum_{n=1}^{T} P_{M2}(n)\, x\, q_{M2}(n) + \\ & \sum_{j=1}^{T'} P_{I1}\left[j \,.\, s(Q)\right] x\, q_{I1}\left[j \,.\, s(Q)\right] + \\ & \sum_{j=1}^{T'} P_{I2}\left[j \,.\, s(Q)\right] x\, q_{I2}\left[j \,.\, s(Q)\right] + \sum_{v=1}^{X} P_{O1}(v)\, x\, q_{O2}(v) \end{aligned} \tag{3}$$

Em [4], cada uma destas equações são avaliadas separadamente criando o algoritmo de decisão, que define continuar com a operadora ou modificá-la, dado os parâmetros de custo com relação ao horário e o tipo de chamada. Neste trabalho, não se apresenta análise individual de cada comunicação, mas a inserção de uma nova componente que é a possibilidade de concorrentes simultâneos no sistema (equação 4).

A equação (4) é basicamente a mesma equação (3), com as inclusões de dados específicos para cada tipo de tecnologia, logo, apresentando certas particularidades internas em função do preço e do tempo.

Os valores de L, N, M, I e O são os mesmos assim como os índices 1 e 2, sendo que, neste momento, insere-se um índice *y*, o qual incluí uma segunda somatória garantindo a funcionalidade de duas ou mais operadoras do mesmo serviço simultaneamente, ou seja, operadoras concorrentes.

Nas equações, para todo tipo de comunicação com termos de índice 2, estão incluídas apenas pela possibilidade de existir um parceiro, funcionário ou filial da empresa, permitindo separar ou não estes da conta convencional, logo, pode-se definir uma filial distante como uma ligação normal ou como um ramal interno, sem custo.

Como a somatória incluída na fórmula significa a inclusão de operadoras para cada solução de telecomunicações, o custo individual de cada serviço poderá aumentar, porém não pelo fato de existir uma operadora a mais, e sim por existir alguma diferença de preço entre estas concorrentes que, por uma simples lógica de ampliação de rotas, redundará em uma máxima utilização dos canais da operadora com o menor preço, e complementará as ligações que estiverem em linha (fila), podendo utilizar os canais de outra operadora. O importante é que desta forma possa-se garantir um backup extra e, em caso de mudanças de custos ou promoções entre operadoras, o sistema automaticamente providencie uma saída de segurança, mantendo o preço baixo ou possuindo rotas alternativas a preços igualmente competitivos.

$$\begin{aligned} K_t = & \sum_{t=1}^{t=t'} \sum_{y=1}^{n} P_{L1y} \,.\, q_{L1y}(t) \; + \; \sum_{t=1}^{t=t'} \sum_{y=1}^{n} P_{L2y} \,.\, q_{L2y}(t) \; + \\ & \sum_{i=1}^{R} \sum_{y=1}^{n} P_{N1y}\left[i \,.\, s(Q)\right].q_{N1y}\left[i \,.\, s(Q)\right] \; + \\ & \sum_{i=1}^{R} \sum_{y=1}^{n} P_{N2y}\left[i \,.\, s(Q)\right].q_{N2y}\left[i \,.\, s(Q)\right] + \\ & \sum_{n=1}^{T} \sum_{y=1}^{n} P_{M1y}(n).q_{M1y}(n) + \sum_{n=1}^{T} \sum_{y=1}^{n} P_{M2y}(n).q_{M2y}(n) + \\ & \sum_{j=1}^{T'} \sum_{y=1}^{n} P_{i1y}\left[j \,.\, s(Q)\right].q_{i1y}\left[j \,.\, s(Q)\right] + \\ & \sum_{j=1}^{T'} \sum_{y=1}^{n} P_{i2y}\left[j \,.\, s(Q)\right].q_{i2y}\left[j \,.\, s(Q)\right] + \\ & \sum_{v=1}^{X} \sum_{y=1}^{n} P_{o1y}(v).q_{9y}(v) \end{aligned} \tag{4}$$

Como visto, o índice *y* significa o tipo de operadora, permitindo a co-existência de duas ou mais destas no mesmo sistema, e n é o número máximo destas operadoras. Assim P_{L1y}, por exemplo, significa dizer que o preço da ligação local para empresas que não sejam filiais da empresa, tenha um valor para P_{11}, ligado a operadora 1 e outro para P_{12}, ligado a operadora 2, onde P_{11} sempre será menor ou igual ao P_{12}, pela lógica de determinação e implementação do LCR.

Outra situação é na somatória de P_{N1y}, P_{N2y}, P_{I1y} e P_{I2y}, onde na estrutura existe o componente S(Q), o qual identifica as operadoras fixas, porém estas significam a rota a ser seguida externamente, e não a conexão interna do usuário. Um exemplo claro para esta situação, seria quando de um telefone celular ou fixo é realizada uma ligação para outro estado ou país, pois neste momento, por mais que se tenha a componente operadora diferenciada (celular ou fixa), se pode e deve ainda, escolher que co-operadora facilitará a ligação para longa distância. Por exemplo, ligando para o Rio de Janeiro a partir de São Paulo, pode-se ter, com dois telefones celulares ou fixos diferentes, muitas possibilidades como demonstrado na tabela 2.

Desta forma o valor S(Q) indicando uma operadora diferente na equação (4), é apenas pelo fato de relacionar a rota para o LCR, enquanto que a inclusão de uma conexão (y), liga a

operadora local com o sistema interno do usuário, sendo uma conexão física intrínseca ao sistema de telecomunicações, que pode ser uma central telefônica, por exemplo.

A figura 2 apresenta o esquema do sistema proposto, sem inclusão de operadoras simultâneas (linha cheia) e com a inclusão de múltiplos concorrentes (pontilhado), definindo o aumento das possibilidades de LCR sem a necessidade de mudança de hardware em um breve período.

TABELA II
OPERADORES DE TELEFONIA COM CONEXÕES A LONGA DISTÂNCIA

Operadora interna ao cliente (Físico)	Operadora Externa para longas dist. (Rota) S(Q)	Convergência de chamadas	Número a chamar
TIM (y=1)	031 S(1) 021 S(2) 015 S(3) 041 S(4) ...	TIM + 031	4260-1000
		TIM + 021	
		TIM + 015...	
OI (y=2)		OI + 031	
		OI + 021	
		OI + 015...	
Telemar (y=3)		Telemar+031	
		Telemar+021	
		Telemar+015...	
Embratel (y=4)		Embratel+031	
		Embratel+021	
		Embratel+015...	

Na realidade, a inclusão de um ou mais provedores internos nos hardwares já pré-existentes, sinaliza a redução da quantidade de canais já existentes. Para empresas que já possuem o sistema funcionando, seria como se dentro de uma central celular existente na empresa, se esta possuísse 10 canais de voz (10 chips de uma operadora), por exemplo, no momento da ampliação citada, esta central agora continuaria com os 10 chips de conexão celular, porém quatro destes passariam a ser de uma operadora diferente. Desta forma a quantidade de canais necessária para atender o sistema e o grupo de usuários continuaria a mesma, porém seria dividida em valores desiguais de forma a suprir, nos momentos de pico, as conexões necessárias e nos momentos de baixa utilização, seriam usados os canais da melhor operadora (menos oneroso).

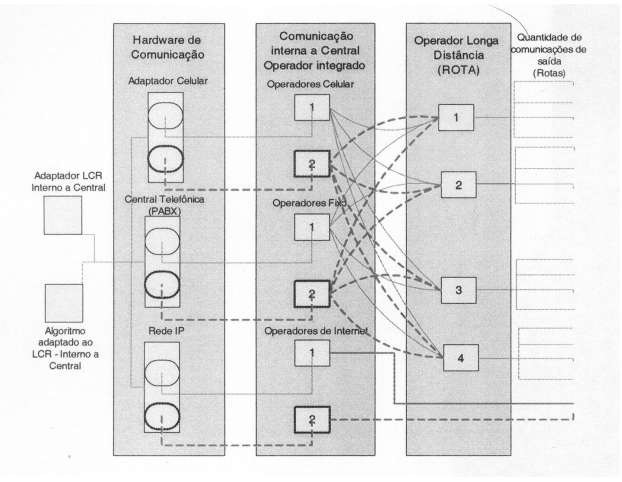


Fig. 2. Design de ampliação das rotas de LCR, dado a inclusão de um operador a mais para cada sistema de telecomunicações.

É importante ressaltar que no sistema apresentado, a atualização de preços das operadoras, é feita de forma automática no software de tarifação, onde apenas a interface entre o hardware que executa a chamada, e o software que gerencia o processo, utilizaria o modelo desenvolvido.

Verifica-se que com a inclusão no modelo proposto de uma operadora a mais para cada tecnologia de telecomunicações, a quantidade de canais de saídas aumenta de nove para dezoito, obtendo-se agora o dobro das rotas para a escolha do menor custo a ser alcançado, baseado no horário da chamada, tipo de operador, local de destino da chamada e outros parâmetros referentes ao LCR.

III. ANÁLISE ESTATÍSTICA DO SISTEMA DE TELECOMUNICAÇÕES

Baseado nos serviços empregados, no caso da Universidade de Bergamo, trabalhou-se diretamente sobre pontos centrais do tipo, hora da realização da chamada, custo individual das chamadas, duração de cada conexão, número chamado, linhas (canais) utilizadas e data da realização destas. Trabalhando em cada filial separadamente, foram realizados os modelos de correlação e histograma em relação aos pontos de convergência, para se levantar o comportamento do sistema, podendo-se então propor a correção da integração dos sistemas com o investimento mínimo requerido.

A tabela 3 apresenta o modelo de correlação entre a data das ligações, a hora das ligações efetuadas, custo efetivo, duração da chamada e dia das ligações efetuadas, além de um *Dummy* (simulação) dos dias da semana, para poder avaliar possíveis correlações ou comportamentos individuais para períodos específicos. Na correlação pode-se verificar que não existe nenhum tipo de convergência ou plano de tarifação especial para esta empresa, de forma a reduzir seus custos. A operadora fixa analisada provê ligações para os números locais, nacionais, internacionais, celulares e outros, sem diferenciá-los, pois não existe correlação entre os valores apresentados que não seja apenas entre o custo da ligação e a duração da chamada, o que é óbvio pois, quanto mais se fala, mais caro é o valor final da ligação na forma linear.

A ausência de correlação, por exemplo, entre a data e o custo, mostra a ausência de planos especiais.

Em alguns países, certas operadoras analisam em detalhes o funcionamento de seus clientes, de forma a oferecer-lhes certos planos tarifários de acordo com o tipo de funcionamento da empresa, onde poderia ser convencional analisar o horário das ligações, verificando que certas empresas possuem muitas chamadas para o exterior, horários fixos de maior utilização, números repetidos (muitas ligações para mesmos números) e outros do tipo, oferecendo descontos pela continuação deste modelo, chamando isto de fidelização pela manutenção do cliente.

No gráfico da figura 3, pode-se visualizar pelo histograma dos horários das ligações, que existem dois momentos de concentração de chamadas gerando dois máximos, um pelo horário da manhã e outro pelo horário da tarde. Ao meio-dia encontra-se o menor valor entre os dois máximos.

As ligações no horário da tarde possuem seu valor máximo em média com 40% a menos que no horário da manhã. Isto significa que a quantidade de canais prevista na instalação da central telefônica é mal utilizada, pois em cerca de 50% do tempo (horário da tarde), esta utiliza entre 60% e 70% de sua capacidade, o que poderia ser aproveitado para outros clientes.

TABELA 3 – MATRIZ DE CORRELAÇÃO DE PEARSON DE UMA DAS FILIAIS DA EMPRESA ANALISADA

TIPO	DATA	HORA	CUSTO	DUR HMS	LINHA NORM	SEG	TER	QUA	QUI	SEX	SAB
MEAN	10,69	11,59	0,025	2,006	4,294	0,131	0,208	0,186	0,25	0,208	0,016
STD	8,875	2,638	0,041	3,170	2,282	0,338	0,406	0,389	0,433	0,406	0,126
N	3667	3667	3667	3667	3667	3667	3667	3667	3667	3667	3667
Data	1	-0,055	0,031	0,031	-0,029	-0,014	-0,041	-0,264	-0,127	0,433	0,023
Hora	-0,055	1	0,025	0,025	-0,004	0,104	0,039	0,117	-0,021	-0,201	-0,047
Custo	0,031	0,025	1	0,999	-0,025	0,049	-0,002	0,008	-0,029	-0,017	0,010
dur	0,031	0,025	0,999	1	-0,025	0,049	-0,002	0,008	-0,029	-0,017	0,010
linha	-0,029	-0,004	-0,025	-0,025	1	0,034	-0,022	0,006	0,024	-0,030	-0,027
Seg	-0,014	0,104	0,049	0,049	0,034	1	-0,200	-0,186	-0,224	-0,200	-0,050
Ter	-0,041	0,039	-0,002	-0,002	-0,022	-0,200	1	-0,245	-0,296	-0,263	-0,066
Qua	-0,264	0,117	0,008	0,008	0,006	-0,186	-0,245	1	-0,276	-0,245	-0,061
Qui	-0,127	-0,021	-0,029	-0,029	0,024	-0,224	-0,296	-0,276	1	-0,296	-0,074
Sex	0,433	-0,201	-0,017	-0,017	-0,030	-0,200	-0,263	-0,245	-0,296	1	-0,066
Sab	0,023	-0,047	0,010	0,010	-0,027	-0,050	-0,066	-0,061	-0,074	-0,066	1

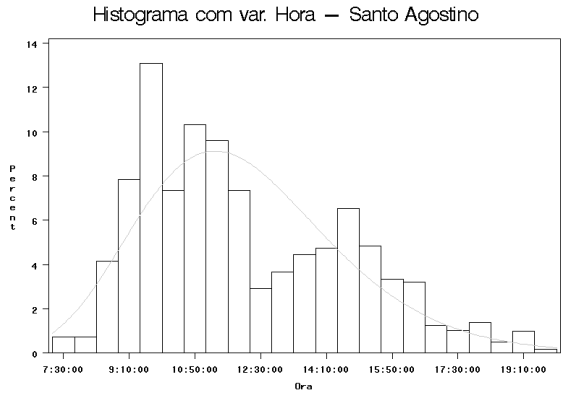


Fig. 3. Análise percentual das chamadas realizadas pelo horário de funcionamento da empresa.

Como na tecnologia digital, a quantidade de canais é indiferente da quantidade física de cabos ou ligações, poderia ser proposto pela operadora, que no horário da tarde os canais de comunicação fossem reduzidos para 70%, onde o custo de assinatura seria também reduzido para o cliente automaticamente.

Para a operadora, esta redução poderia servir para desviar os canais em excesso, para utilização ou ampliação de outra empresa, que possua necessidade de mais canais pelo horário da tarde. Em geral, isto reduz custos de equipamentos, pois pode-se reduzir a quantidade de equipamentos, que devem fornecer acesso a rede de voz ou dados, podendo utilizar estes equipamentos internamente para backup da empresa operadora.

A figura 4 apresenta a duração das chamadas realizadas. Pode-se então avaliar que, em geral, as ligações são de curta duração e desta forma o número de canais é maior pelo uso de diversas chamadas efetuadas em curtos períodos de tempo. Neste caso poderia ser apresentado um modelo da operadora fixo, que não cobre taxa pela resposta a ligação, pois na conta telefônica existe a divisão do preço em tempo de chamada e um valor fixo para o primeiro minuto quando a chamada é completada. Uma ligação de dois minutos não custa o dobro de uma ligação de um minuto. Simples de analisar, este valor fixo que existe ao primeiro minuto, no seu ínfimo valor de alguns centavos pode significar, para uma empresa que efetua alguns milhares de ligações mensais de curta duração, a economia de valores muito significativos, além das demais reduções já comentadas.

A inclusão de duas operadoras fixas no projeto atual, cria a possibilidade de efetuarem-se as ligações, de curta duração, pela operadora com valor determinado ao minuto, sem o custo adicional de atendimento, e para as ligações de durações um pouco mais longas, utilizarem a operadora que cobra também pelo atendimento a chamada, pois seus minutos consecutivos são mais baratos do que das operadoras que não cobram uma taxa para cada chamada atendida (completada). Todas estas possibilidades podem ser automaticamente decididas pelo algoritmo, informando e executando desta maneira, mudanças nas rotas dado uma infinidade de soluções, que são atualizadas pelas promoções e custos das operadoras.

O gráfico da figura 5 mostra que os canais de saída para as linhas fixas estão mal distribuídos, pois percentualmente existem linhas muito mais utilizadas que outras, logo pela média entre os valores máximos e mínimos de chamadas em função da quantidade de canais, pode-se verificar que poderia ser reduzido o número destas linhas de comunicação, reduzindo assim a assinatura (em excesso) das mesmas.

Dada a projeção de duas operadoras simultâneas, deve-se ter muito cuidado ao projetar o número de canais tal que correspondam ao perfil normal de funcionamento do sistema, e não em um momento atípico pois, uma vez realizada a redução de canais, estes não são automaticamente ampliados quando necessário, o que poderia provocar uma pane no sistema.

Na realidade deve-se cruzar as informações das figuras 3 e 5, ainda verificando a freqüência de chamadas, dado os horários

de pico da figura 3, para se alcançar o número ideal de canais necessários para aquele sistema.

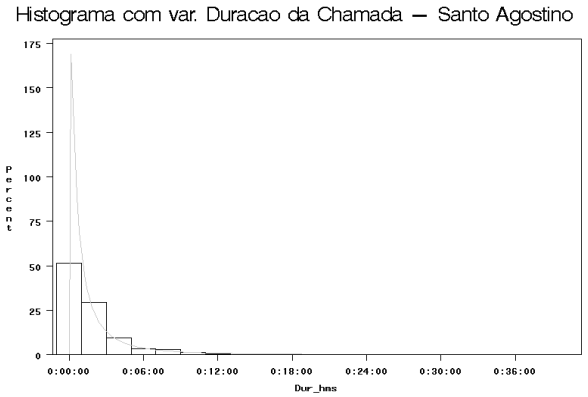


Fig. 4. Análise percentual da duração das chamadas realizadas pelo número total de chamadas mensal.

A definição de canais necessários, dado o perfil de comunicação realizado pela empresa, é apresentada na literatura sobre o setor de telecomunicações, porém este trabalho não visa este tipo de solução, mas propõem-se analisar se estes valores são adequados, dado uma convergência de diversos serviços simultâneos com rotas diferentes.

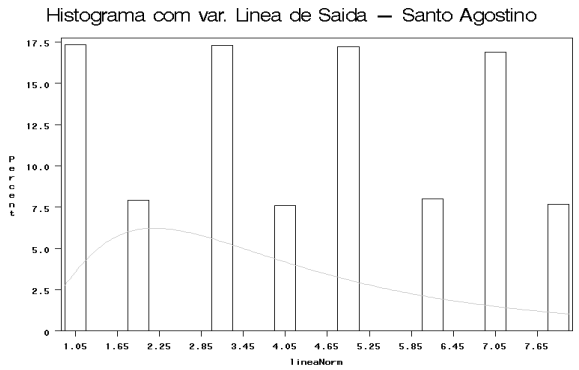


Fig. 5. Análise percentual das linhas de comunicação da empresa pela quantidade total de chamadas realizadas.

Pode-se agora definir quantos canais simultâneos serão reduzidos ou ampliados dado o perfil de cada empresa, disposto pela equação (4) e controlado pela tarifação e o histograma de utilização da rede.

IV. MODELAMENTO DO INVESTIMENTO

A redução de assinatura da linha telefônica ou do sistema de rede de dados não é considerado primordial, até mesmo porque este não é percentualmente significativo para empresas de porte médio ou grande. No caso destas linhas ou canais estarem sendo mal utilizadas, este custo desnecessário deve ser analisado, porém o intuito principal deste artigo é o de garantir que o investimento seja o suficiente de forma a permitir a funcionalidade da convergência a um valor médio aceitável.

A análise entre o desempenho e o custo é o ponto base do projeto, pois do ponto de vista da engenharia, podem-se criar diversas convergências e somar diferentes tipos de seguranças, rotas de backup e outras, para depois se calcular em quanto estará se reduzindo os custos e ganhando em desempenho. Do ponto de vista do setor administrativo, economicamente pode-se cortar custos em todos os limites, de forma a reduzir a qualidade do atendimento e serviços prestados pela empresa, porém mantendo uma comunicação mínima necessária.

Propõe-se que o resultado seja uma média baseada na ocorrência, ou seja, matematicamente não se consegue definir qual a melhor convergência para uma empresa que ainda não exista, pois seria improvável a predição de valores e comportamentos da mesma.

Baseado no modelamento da convergência pode-se definir serviços, equipamentos e um número médio de canais, para funcionamento da empresa, porém a otimização destes só será possível após certo tempo de funcionamento da mesma. O comportamento destas tecnologias, ainda assim estarão sempre sob o efeito do crescimento e evolução da empresa, pois em uma empresa em expansão, não se pode prever sua estratégia logística. Logo, o uso da tecnologia que auxilia este funcionamento, não pode ser medido.

O projeto de LCR adaptado com um algoritmo de convergência entre diversas tecnologias e serviços, deve ser aplicado em empresas com certo nível de consolidação, para não se definir produtos inadequados ou sub-dimensionados para a mesma.

De forma clara, não se pode definir que a inserção de VoIP entre todas as filiais de uma empresa seja o caminho certo para esta, porém pode-se predispor a tecnologia nestas empresas de forma que, a partir do momento em que seja economicamente viável esta inserção, a adaptação seja simples e sem mudança ou cancelamento do sistema de comunicação atualmente instalado.

Entre o tipo de tecnologia e a sua utilização, existe uma fronteira que define a partir de que momento a mudança ou inserção de novos produtos são eficientes à empresa. Esta fronteira possui valores ainda não totalmente mensuráveis no setor econômico [11]. Os pontos de indecisão se apresentam principalmente como desempenho da empresa, dado a utilização de novos recursos tecnológicos, onde isto deve ser medido pela mudança no lucro da empresa após certo tempo de aplicação deste modelo, ou ainda pelo fato da aplicação do novo modelo garantir certo nível de fidelização ou estabilidade no mercado da empresa por um período requisitado.

Segundo Chircu and Kauffman [21], é possível avaliar os fatores de impacto sobre os investimentos em tecnologia da informação (TI), onde Boltin [22] coloca isto como uma preocupação e possível problema. Desta forma baseia-se esta proposta para o investimento, seguido pela inserção de um contrato de manutenção nas ampliações de novas tecnologias, mais a redução dos custos apresentados pela solução tecnológica proposta.

Poucos são os casos que não possuem uma redução de custos na convergência e integração de sistemas tecnológicos, pois o projeto modifica o sistema baseado nas melhores rotas de custos, além de ampliar as formas de comunicações. Por

exemplo, no caso da implantação de um sistema ter sido efetuada na proximidade da fronteira de migração tecnológica, isto poderia acontecer, como exemplo, na implantação de um sistema VoIP que não reduziu os custos de comunicação, porém aumentou consideravelmente a utilização da comunicação que, caso fosse cobrada, no sistema anterior seria muito mais onerosa.

Quando é projetado um sistema VoIP, o tempo de comunicação (duração entre as chamadas) utilizando certa quantidade de canais, não interfere no custo do sistema, ou seja, se eles mantiverem o tempo de comunicação anterior ou falarem diversas horas a mais, nada implicará no novo sistema, enquanto que no sistema de pagamento por pulso, a cada novo minuto, novos centavos são incluídos ao custo existente.

Dada algumas exigências, definiu-se um período menor ou igual a 5 anos, para a atualização ou ampliação total dos equipamentos, baseado nos modelos de Plano de Negócios, com relação a vida útil dos produtos tecnológicos em uma empresa.

Avaliando-se de um ponto bem amplo para este desenvolvimento, foi definido o pior caso e restrito a possibilidade de mudanças bruscas no mercado.

Desta forma, definido K como o custo do sistema de tecnologia em funcionamento, que seria a soma dos valores de assinatura para a rede atualizada, os custos de manutenção dos bens materiais, os custos da rede de dados e o custo do tráfego (K_t) visto por (4), encontra-se um novo valor de custo $(1 - \xi)$ K, onde ξ é a redução percentual do custo anterior visto na diferença dos totais da tabela 1.

Encontra-se desta forma σ "Savings", como sendo o benefício durante certa quantidade de tempo n, que define o momento máximo que se pode suportar a mesma tecnologia sem necessidade de troca desta [23, 24].

$$\sigma = \sum_{t=0}^{n} \xi(t).\mathrm{K}(t) \tag{5}$$

Onde, para
K(0) = K
$\xi(0) = 0$

O custo do investimento também deve ser baseado no tempo máximo para que este seja pago [25], o qual não pode ultrapassar o tempo de mudança total da tecnologia, de forma a não existir dois investimentos simultâneos e automaticamente dois financiamentos para uma mesma tecnologia, que em um tempo t(n+1) estará desvalorizada. O custo deve contar com os fatores da tecnologia empregada, o valor do contrato de manutenção necessário para garantir a sua atualização até o tempo final real do produto, e mais as taxas de financiamento empregadas pelo sistema, criando desta forma (6).

$$\mathrm{K}_I(t) = \sum_{t=0}^{n} I\left(1 + \delta(t)\right) + \theta(t) \tag{6}$$

onde
K_I = Custo total do Investimento
I = Investimento base da tecnologia
$\delta(t)$ = taxa de interesse do órgão financiador
$\theta(t) = \tau$ (t) x I = Valor percentual do investimento total para garantir o contrato de manutenção

Assim, chega-se a:

$$\mathrm{K}_I(t) = \sum_{t=0}^{n} I\left(1 + \delta(t) + \tau(t)\right) \tag{7}$$

Onde, para $\theta(0) = 0$ e $\delta(0) = 0$,

$K_I(0) = I$

Desta forma, para que o investimento seja viável ao sistema de telecomunicações a ser implantado, precisa-se que (8) seja satisfeita.

$$\sigma \geq K_I(t) \tag{8}$$

$$\sum_{t=0}^{n} \xi(t).\mathrm{K}(t) \geq \sum_{t=0}^{n} I\left(1 + \delta(t) + \tau(t)\right) \tag{9}$$

Isto significa dizer que, os valores reduzidos nos custos da rede de telecomunicações pela implantação de novas tecnologias, devem em um tempo n suficientemente atual ao mercado e as necessidades da empresa, ser maior ou igual à somatória dos custos dos investimentos necessários mais a manutenção do sistema e as taxas de financiamento do investimento.

Pode-se re-escrever (9) na forma simplificada (10).

$$\sum_{t=0}^{n} \frac{\xi(t).\mathrm{K}(t)}{I\left(1 + \delta(t) + \tau(t)\right)} \geq 1 \tag{10}$$

No caso da condição (equação 10) não ser satisfeita, não se deve realizar este investimento, mas deve-se analisar a solução proposta e buscar uma outra, com custos ajustados até que satisfaça o apresentado.

V. Integração de Sistemas

Sistemas tecnológicos de alto risco introduzem um potencial para eventos catastróficos [26], sendo assim deve-se aplicá-los de forma bem clara aos seus usuários, para se garantir a confiança e aceitação da mudança no setor conforme proposto no modelamento [27].

Diferente do projeto anteriormente proposto, a integração de sistemas levará em conta, como ponto fundamental, o investimento necessário para funcionamento da arquitetura planejada, porém com a inclusão de mais rotas de escape. Esta implantação de novas rotas amplificará a qualidade do investimento realizado, além de possibilitar no setor de marketing, valores agregados de mercado com parceiros diversificados.

É importantíssimo verificar que em nenhum momento apresenta-se uma solução como única ou máxima, pois o principal aqui descrito é sobre qual o fator de integração e flexibilidade da tecnologia [28], onde os serviços podem

trabalhar conjuntamente, de forma que um algoritmo criado a partir das condições estabelecidas, decida baseado em dados atuais ou passados, por análise estatística, predição ou regra de otimização, qual o sistema a ser utilizado naquele momento para o serviço requerido.

A figura 6 apresenta o modelo de integração de sistemas atualizado, com sobre-rotas nos serviços de mesma tecnologia, possibilitando agora rotas nos operadores de mesmo serviço e entre operadores de tecnologias e serviços diferentes, vistos na figura 2. Na realidade, pode-se agora definir que em certo dia, horário e para aquela comunicação devida e definida seja, por exemplo, mais adequado ligar para um celular externo da empresa pelo sistema de telefonia fixa ao invés do sistema de telefonia celular.

Por mais que pareça ilógico, pois se sabe que para sistemas de mesma tecnologia é melhor realizar comunicações idênticas, no mercado atual existem possibilidades irregulares pela utilização de convergências entre os operadores, e pela falta de normas adequadas dos órgãos de autorização do governo, deixando que diversos modelos de preços possam ser aplicados.

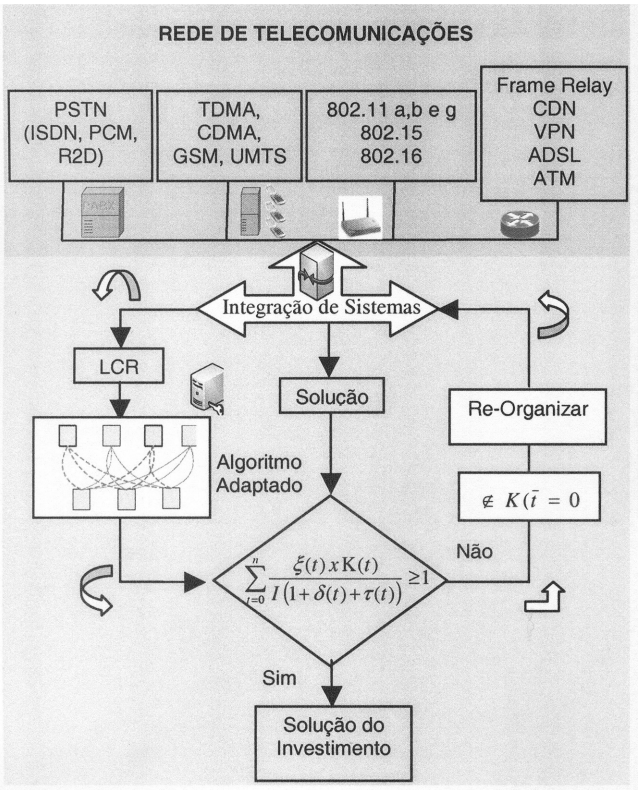


Fig. 6. Integração de Sistemas baseado em redes Multi-Rotas.

A integração de sistemas baseado em redes multi-rotas definida pela figura 6, apresenta as diversidades de tecnologias na parte superior da figura, seguida de um "*loop*" sobre o algoritmo apresentado, para melhor solução de integração com a dependência do investimento. Neste momento não basta definir o custo sobre as diversas possibilidades de rotas concorrentes ou multi-rotas, como definido pelos autores, mas todo este conjunto deve ser garantido pela infra-estrutura existente ou pela nova infra-estrutura definida pela atualização do projeto.

Em caso deste sistema não ser satisfeito, o algoritmo fica em *looping* baseando-se em um custo diferente do custo inicial do projeto, K, buscando valores sempre menores que este que satisfaça simultaneamente as duas condições, as quais são: as rotas de menores custos e o investimento baseado nestas rotas. O que garante um investimento para mudanças de tecnologias, pagos pela diferença do custo anterior, K, e do custo atual, K', após a integração de sistemas.

Uma regulação mais eficaz por parte do governo [17] seria ideal para garantir certa harmonia nos serviços existentes no mercado, porém ainda falta à figura do próprio integrador de sistemas, que seria um especialista para analisar a necessidade do cliente, criar e determinar as rotas de menores custos para os serviços deste (LCR) e, por fim, determinar o investimento para desenvolver o estabelecido, com garantia de custo/benefício ao cliente, conforme se mostra na figura 6.

Em relação ao modelo anterior, desenvolvido para integração de sistemas, o modelo atual baseia-se em um processo adjunto de rota, simplificando-se tudo por IP, logo a integração é feita por meio de um "gateway". Outros dois pontos desenvolvidos são: o algoritmo adaptado para novas rotas suplementares, dada as mesmas tecnologias implantadas e a valorização do investimento pelo incremento ou cancelamento de um custo K médio, definido em um valor de tempo X, ao qual existem dois sistemas simultâneos e independentes que podem ser somados, integrados ou desenvolvidos individualmente.

VI. Aplicação de *Price Cap* no Setor de Telecomunicações

A necessidade da utilização de um modelamento tão detalhado e com diversas rotas para redução de custos, existe devido à insuficiência de controle e regularização desenvolvida pelo órgão regulador no setor de telecomunicações.

Para definir os padrões e preços dos setores de utilidade pública, é estudado a taxa de retorno RoR (Rate of Return) [29], onde um "*proxy*" do aumento dos custos no setor, RPI, e a eficiência da empresa, X, são os pontos responsáveis pelo mecanismo de controle do preço do produto apresentado por:

$$P_t = \left(1 + RPI_t - X\right) P_{t-1} \tag{11}$$

O controle é baseado no desenvolvimento passado, ocorrido ao ano anterior daquele produto no mercado [30]. O plano tarifário ou comumente chamado pelo termo em Inglês de "*tariff basket*", por mais que desenvolvido pela análise de diversas tarifas do produto neste mercado [31] não analisa, neste modelo, a influência de serviços diferentes, que atendem as mesmas funções de tecnologia, com preços ou aplicações diferentes. Por exemplo, não existe a diferenciação entre a ligação telefônica normal e a ligação telefônica de voz sobre protocolos de internet (VoIP), ou ainda, ligações internacionais via um "*Call Center*" Internacional de voz, onde estes utilizam sistemas tecnológicos híbridos, capazes de com certa integração realizar a mesma ligação desenvolvida por um sistema convencional de telefonia, porém com preços muito mais baixos.

Os custos do sistema VoIP é diferente dado o sistema telefônico possuir um custo variável, baseado no tempo de acesso com pagamentos de assinatura e resposta a ligação (atendimento), onde o sistema internet possui apenas o pagamento da assinatura, que é um valor fixo mais barato que o valor do sistema telefônico [32]. Possuindo uma maior capacidade e utilização de integradores de serviços, estes sistemas são compartilhados de forma a utilizar o máximo da eficiência no tempo pela ocupação da banda com diversos usuários.

A análise de *tariff basket* segundo *Cambini et al.* [33] é dado por:

$$\sum_{i=1}^{t} P_i^t \; q_i^{t-1} \leq \left(1 + RPI_t - X\right) \sum_{i=1}^{t} P_i^{t-1} \; q_i^{t-1} \tag{12}$$

Esta equação possui uma deficiência no controle de serviços ou produtos tecnologicamente diferentes, os quais respondem a uma mesma funcionalidade, porém utilizando uma via de acesso alternado. Dentro da somatória do produto pela quantidade de tempo, falta a inclusão de diversos produtos baseados na mesma solução ou uma sub-fórmula que integre separadamente cada caso.

A distribuição dos produtos de tecnologia, de forma a alcançar um percentual muito grande de clientes com comunicação telefônica, é o ponto chave de marketing das operadoras de telecomunicações, pois estas necessitam aumentar a distribuição, o que significa dizer, aumentar o serviço/produto de forma a melhorar a economia de escala de cada região. Segundo Schmalensee and Willig [34], o grau de Economia de Escala Tecnológica é definido por,

$$\overline{S}(X,Y) = -\left\{ \frac{\sum x_i \left(\partial\varphi \big/ \partial x_i \right)}{\sum y_i \left(\partial\varphi \big/ \partial y_i \right)} \right\} \tag{13}$$

Esta fórmula muda para a elasticidade de escala, porém sempre para a mesma função de x e y, ou seja, para o mesmo produto que, no caso do setor de telecomunicações, é um conjunto de serviços. Seja este serviço uma chamada internacional, nacional, local ou serviços específicos, eles são analisados separadamente, sendo compostos somente no momento da média total do "Índice dos preços relativos dos serviços telefônicos". A influência da superposição dos serviços, sobre o mesmo aspecto de valores deve, novamente, ser incluída sobre a valorização de qualquer serviço de tecnologia avaliado.

Seguindo esta lógica, o setor de telecomunicações controla o acesso e os preços dos produtos no mercado pelo desconhecimento técnico do produto e da tecnologia pelos órgãos responsáveis. O exemplo básico para amostragem é visto no caso de uma ligação entre dois estados ou dois países, onde por mais que o governo tenha um medidor que forneça as comunicações realizadas, este não tem como controlar a tecnologia utilizada.

Uma ligação sendo realizada de Belém para o Rio de Janeiro, é taxada pelo governo como uma ligação interurbana, dado a área definida e sendo cobrada pela operadora pelo tipo de comunicação de voz, a qual segue o preço definido pela análise de "*Price Cap"*. "*Price Cap Regulation*" foi projetado para proteger o consumidor das excessivas taxas aumentadas pela limitação ou sobre-carga das taxas cobradas para serviços de telefonia local. Ao mesmo tempo, foi projetado para companhias provedoras de telefonia, como incentivo para reduzir custos e desenvolver serviços, enquanto permitem grande competição nos serviços de telefonia local. Os valores de *Price Cap* são definidos pela tecnologia, infra-estrutura, região, utilização do serviço e outros parâmetros, chegando-se ao valor final que é pago pelo usuário.

Na ligação citada para o Rio de janeiro, quem poderá confirmar e garantir que esta ligação quando realizada, ao invés de trafegar nos *links* de operadoras de longa distância, não trafega pela rede de dados WAN da própria operadora fixa local fornecedora do serviço, mudando o protocolo e realizando a chamada via VoIP de um estado para o outro. Neste caso, após ser desenvolvida a comunicação por VoIP, a chamada poderá ser retornada ao sistema comutado na central telefônica local de destino e então endereçada entre as casas dos dois estados.

Por mais que seja cobrado e controlado que foi realizada uma ligação, esta foi cobrada por uma tecnologia e utilizada por outra, onde a tecnologia VoIP pode comprimir 2 vezes mais o canal de voz que o protocolo de voz comum, mesmo este sendo digitalizado. Isto significa dizer que a operadora de telefonia reduz seus custos pela metade, dada a utilização de uma tecnologia mais moderna e custo de uso inferior, sendo que esta redução não é repassada para o consumidor final. Logo, o lucro está na redução da infra-estrutura requerida, que agora pode realizar o dobro de ligações que a infra-estrutura anterior ou o mesmo número de ligações, utilizando apenas metade de seus canais de comunicação.

A convergência de voz e dados proposta pelo sistema da figura 6, visa neste trabalho o consumidor, e é há muito tempo utilizado pelas operadoras que ganham sem repassar ou dividir conceitos. Sendo assim, as mudanças corriqueiras nos preços das operadoras de telefonia móvel e fixa são comuns, pois a cada momento, um delta entre o valor real e o praticado é desenvolvido por estas, dado o alto índice de inovações tecnológicas. A redução não é repassada ao consumidor diretamente, mais indiretamente na tentativa de inserir novos clientes por uma promoção, onde no contrato base, este valor promocional some após certo período, retornando ao alto e irregular valor praticado no mercado.

A figura 7 apresenta o ciclo de marketing vicioso [35], no qual o consumidor apoiado em um desconhecimento da tecnologia e encantado (e enganado) por sua beleza, tende a ser motivado para buscar o que existe de melhor ou mais aceitável, porém sem desenvolver uma clarificação da sua necessidade. O modelo apresentado de integração de sistemas deve manter uma regularidade, para não focar-se inteiramente em desenvolvimento tecnológico e pouco em desenvolvimento econômico para a empresa, pois esta tecnologia deve ser favorecida por um resultado.

É definido como marketing auto-realizativo [36], para este modelo de marketing de produtos para alta tecnologia, o qual tende a criar uma falsa necessidade ao consumidor,

aproveitando-se de sua posição social em vez da utilidade de fato do produto ou serviço que está sendo contratado por este.

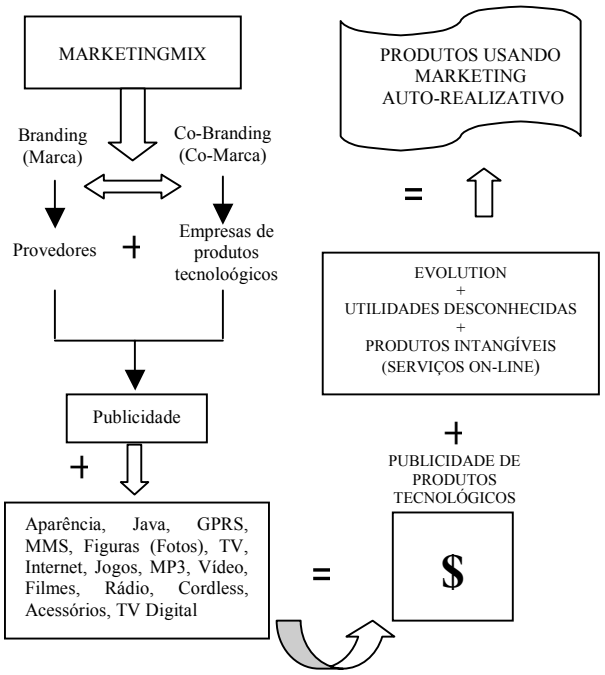


Fig. 7. Marketing Auto-Realizativo evoluído da integração de valores, serviços e produtos tecnológicos.

Este marketing auto-realizativo, pode ser explicado pela utilização de algo que não é importante ao cliente ou não possui prioridade para tal, como o fato de uma criança possuir um telefone celular que possui joguinhos. Não seria melhor ele ter um game portátil que possui qualidade superior e foi desenvolvido para ser utilizado como um *game* e não celular? Logo, a robustez deste produto é bem maior para tal utilização [37].

Outro ponto simples de se verificar seria, por exemplo, os celulares com câmera de vídeo, foto, GPRS, MMS, Java e outras tecnologias, que não são usadas por cerca de mais de 90% dos consumidores, pois mesmo a câmera fotográfica, está provada que é utilizada no início da compra do celular, porém quando existe uma festa, viagem e outro evento importante, a necessidade dos consumidores, é voltada para sua máquina de fotografar digital de alta tecnologia.

O caso é que não adianta melhorar o nível dos celulares com estas tecnologias, pois de que serve pagar um valor a mais por outras tecnologias que não se utilizam?

Aquele a mais que o profissional do marketing utiliza para dar o seu diferencial ao produto ou serviço na hora da venda, em muitos casos é considerado desnecessário, logo, quando o consumidor utiliza, de certo ponto de referência para algo que ele não necessita, porém para ele, foi este ponto que o fez decidir pelo produto, a isto chamamos de auto-realizativo. Sendo assim este marketing é algo que desenvolve a necessidade de realizar, sem existir uma verdade ou conhecimento por trás deste acontecimento.

No setor tecnológico o marketing auto-realizativo é bastante comum, pelo simples fato de ser difícil compreender detalhes muito tecnológicos para grande maioria dos clientes que não são do setor. Outro fator é a beleza e a sensação de poder fazer tudo, que excitam estes consumidores. Nada melhor do que uma jóia no pescoço de uma bela mulher, logo, nada melhor do que um belo telefone celular, carro e outros do tipo na mão de um belo consumidor.

VII. Conclusão

Este trabalho utilizou o método de aproximação orientada aos fenômenos, onde baseado no artigo do case da universidade avaliada em [4], propôs-se uma solução de ampliação tecnológica, para manutenção dos custos existentes, avaliando diversos modelos de convergência da tecnologia e, ainda como suplemento, a inclusão de sua imagem concorrente, para dispor de diversas possibilidades em função do mercado.

Como contribuição, este trabalho inseriu na convergência discutida e definida [4], uma imagem concorrente para cada tipo de tecnologia de telecomunicações avaliada onde, em vez de definir previamente quais operadoras de cada setor como telefonia móvel, fixa, internet e outros que seriam usadas; foram ampliados para cada um destes uma segunda opção, que no momento do uso, a melhor escolha (preço e desempenho) esteja disponível.

Isto geraria uma possibilidade para o software, onde além das escolhas de rota entre as tecnologias, este teria prioritariamente a definição de qual concorrente é o melhor naquele momento, onde as rotas foram duplicadas diretamente e os problemas de modificação de hardware foram reduzidos.

O algoritmo não é mais baseado apenas no planejamento e na estrutura em tempo real para diversas tecnologias, mas também com o novo modelamento este amplia a vida útil do projeto.

Algo que parece simples de visualizar é mais complexo na determinação, pois não se pode dividir igualmente em dois, os canais de uma tecnologia, senão o projeto será um problema em vez de solução. É necessário que seja definido, dada a predição do uso, o percentual de divisão de canais entre os operadores, em função da quantidade total de canais, utilização diária, valores máximos e mínimos, desempenho e relação da mediana, dentre outros pontos.

O fator determinante na limitação desta divisão de canais entre os concorrentes instalados é para que, no horário de pico das chamadas, não se tenha uma perda dada a utilização dos dois concorrentes com preços distantes entre si, e para que nos horários de mínimo, se possa utilizar apenas o operador com menor preço. Podem existir ainda, casos de que uma outra tecnologia possa ser auto-suficiente para necessitar de três operadoras e, desta forma, reduzir uma operadora de uma tecnologia diferente, ficando esta apenas como *backup* do sistema.

Outra contribuição do artigo em questão, foi à análise estatística dos dados de tarifação disponibilizados, onde estes foram essenciais para definir e confirmar o modelamento realizado, principalmente em função do investimento, que em [4] era linear, porém no artigo atual apresentou comportamentos gaussianos e dependentes de predição. Poderia ser interessante inserir em um modelo futuro, um equacionamento estatístico dos valores máximos e mínimos de comunicação, para garantir uma padronização da divisão dos canais entre as operadoras de mesma tecnologia.

O terceiro ponto de contribuição do artigo foi o modelamento da integração de sistemas na forma gráfica, definindo o investimento em função dos fatores relacionados e citados, incluindo uma proposta de centralização do projeto por IP, ou seja, um gateway IP que acelerasse o encaminhamento e fosse processado para comunicação digital após ser escolhida a tecnologia. Baseado nas redes multi-rotas, esta integração foi desenvolvida como a junção do modelamento da integração de sistemas e o modelamento do financiamento disponível.

No lugar do tópico "Discussão", o artigo apresentou um tópico definido como "Aplicação de Price Cap", onde baseado no setor de estudo, tentou-se de forma simples e clara, informar uma deficiência do mercado e do governo no setor de telecomunicações. Apresenta-se que os preços definidos, assim como a transparência no processo, não são bastante evidentes, desta forma, o obscuro para profissionais da área se torna ainda pior para os consumidores que não possuem conhecimentos técnicos, tornando assim mais simples a incitação para desenvolver no mercado, produtos baseados na aplicação de lucros para a empresa, e não na real necessidade do consumidor.

Acredita-se que estes valores e validações do mercado de telecomunicações, devam ser mais bem explorados, tendendo a colocar os consumidores a par da situação e de seus verdadeiros direitos, para que o mercado se ajuste e os preços baixem de forma organizada, sem prejudicar a qualidade da informação. Algumas tecnologias ainda podem satisfazer certas necessidades de alguns clientes que, entretanto, a operadora diz não poder desenvolver [38], pelo fato de não trazer lucro e ainda necessitar de certo trabalho, sendo assim, onde estará nosso órgão responsável pela autorização dos serviços, para controlar e forçar o bom uso das tecnologias desenvolvidas até o momento?

REFERÊNCIAS

[1] Bugamelli, M. & Pagano, P., "Barriers to Investment in ICT", Applied Economics, 36, 2275-2286, 2004.

[2]Desanctis, G. & Jackson, B. M., "Coordination of Information Technology Management: Team-Based Structures and Computer-Based Communication Systems", Journal of Management Information Systems, Vol. 10, No.4, 85-110, 1994.

[3]Douglas, A. & Glen, D., "Integrated Management Systems in small and medium Enterprises", Total Quality Management, Vol. 11, No. 4/5&6, S686-S690, 2000.

[4] Arruda Filho, E. J. M.; Biffignandi, S.; Moriggia, V. & Marino, A., "Least Cost Routing Applied to Telecommunications Systems Integration", Over review process at the Telecommunication Systems Journal (Springer), December, 2007.

[5] Documentation for SAS®9 Products, Access on line at http://support.sas.com/documentation/onlinedoc/sas9doc.html, 2007.

[6] Melody, W. H., "Economic Analysis for Changing Times", Telecommunications Policy, 23, 601-602, 1999.

[7] Stavridou, V., "Integration in Software Intensive Systems", The Journal of Systems and Software, 48, 91-104, 1999.

[8] Brown, S. W., "The Move to Solutions Providers", Marketing Management, 9:1, 10-11, 2000.

[9] Cantwell, J. & Santangelo, G. D., "Capitalism, profits and innovation in the new techno-economic paradigm", Journal of Evolutionary Economics, 10:1/2, p131, 2000.

[10] Pollalis, Y. A., "Patterns of co-alignment in information-intensive organizations: business performance through integration strategies", Int. Journal of Information Management, 23, 469-492, 2003.

[11] Kumar, R. L., "A Framework for Assessing the Business Value of Information Technology Infrastructures", Journal of Management Information Systems, 21:2, 11-32, 2004.

[12] Barnhart, C. & Ratliff, H. D., "Modeling Intermodal Routing", Journal of Business Logistics, Vol. 14, No. 1, 1993.

[13] Gonçalves, R. J. & Garção, A. S., "Implicit Multilevel Modeling in Flexible Business Environment", Communications of the ACM, 45:10, 2002.

[14] Malhotra, N. K., Citrin, A.V. & Shainesh, G., "Editorial: The Marketing of Technology Oriented Products and Services: An Integration of Marketing and technology", Int. Journal of Technology Management, 28:1, 1-7, 2004.

[15] Jovanovic, B. & MacDonald, G. M., "Competitive Diffusion", Journal of Political Economy, vol. 102, No. 1, 1994.

[16] Alkahtani, A. M. S.; Woodward, M. E. & Al-Begain, K., "Prioritised best effort routing with four quality of service metrics applying the concept of the analytic hierarchy process", Computers & Operations Research, 33, 559-580, 2006.

[17] Hobday, M.; Davies A. & Prencipe, A., "Systems Integration: a core capability of the modern corporation", Industrial and Corporate Change, Volume 14, Number 6, pp.1109-1143, Advance Access Published, November 7, 2005.

[18] Bejusic, D.; Rozic, N. & Dujmic, H., "Development of the communication/information infrastructure at the academic institution", Computer Communications, 26, 472-476, 2003.

[19] Pucker, L., "Does the Wireless Industry Really Need all these Digital if Standards?", IEEE Communications Magazine, Vol. 43, Issue 3, Mar 2005.

[20] International Labour Organization, "Sampling and Quality Adjustment", Joint UNECE/ILO Meeting on Consumer Price Indices, 4-5 December 2003, Geneva, Switzland.

[21] Chircu, A. M. & Kauffman, R., "Limits to Value in Electronic Commerce-Related IT Investments", Journal of Management Information Systems, 17:2, 59-80, 2000.

[22] Boltin, J., "Ask ferf about...the seventh annual Technology Issues Survey", Financial Executive, 21:6, 2005.

[23] Fusa, E. & Pisoni, P., "La valutazione degli Investiment", EGEA S.p.A., 2001

[24] Testa, F., "Gli Studi di Fattibilità di Investimenti Industriali", Pàdova, 1984.

[25] Tsai, Y. T. & Hsieh, L. F., "An Innovation Knowledge Game Piloted By Merger and Acquisition of Technological Assets: A Case Study", Journal of Engineering and Technology Management, 23, 248-261, 2006.

[26] Greening, D. W. & Johnson, R. A., "Do Managers and Strategies Matter? A Study in Crisis", Journal of Management Studies, 33:1, 25-51, 1996.

[27] Phillips, A. L., "Migration of Corporate Payments from Check to Electronic Format: A Report on the Current Status of Payments", Financial Managements, 27:4, 92-105, 1998.

[28] Knot, J. M. C.; van den Ende, J. C. M. & Vergragt, P.J., "Flexibility Strategies for Sustainable Technology Development", Technovation, 21, 335-343, 2001.

[29] Taylor, W. E. & Taylor L., "Postdivestiture long-distance competition in the United States", American Economic Review, vol. 83, issue 2, p185, 6p, May 1993.

[30] Bernstein, J. I., "Price Cap Regulation and Productivity Growth. Journal of Regulatory Economics", International Productive Monitor, Carleton University, 2001.

[31] Law, P. J., "Welfare Effects of Pricing in Anticipation of Laspeyres Price-Cap Regulation: An Example", Bulletin of Economic Research, 49:1, p17, 11p, January 1997.

[32] Yannelis, D., "On Access Pricing with Network Externalities", Atlantic Economic Journal, vol. 30, issue 2, p1, 5p, June 2002.

[33] Cambini et al., C.; Ravazzi, P. & Valletti, T., "Il Mercato delle Telecomunicazioni", Mulino, Bolgna, 2003.

[34] Schmalensee, R. & Willig R.D., "Handbook of Industrial Organization", Volume I, Elsevier Publisher, 1989.

[35] Wind, Y. & Mahajan, V., "Marketplace: Convergence Marketing", Journal of Interactive Marketing, 16:2, 2002.

[36] Arruda Filho, E.J.M.; Cássia, F. & Marino, A., "Beyond The Interoperability of Telephony, VoIP and Networking: Self-Realization Marketing Contribution to Value Creation in Telecommunications Sector", International Journal of Technology Marketing – IJTMkt, V. 3, I. 1, 2008.

[36] Harm-Jan Steenhuis, Erik J. De Bruijn, "Exploring the influence of technology size on the duration of production technology transfer implementation", International Journal of Technology Transfer and Commercialisation , Vol. 4, No.2 pp. 172 - 193, 2005.

[37] Brusoni, S.; Prencipe, A. & Pavitt, K., "Knowledge Specialization, Organizational Coupling, and the Boundaries of the Firm: Why Do Firms Know More Than They Make?", Administrative Science Quarterly, 46, 597-621, 2001.

Emílio Arruda Filho was born in Belém, PA, Brazil, on November 07, 1972. His Bachelors and Masters Degrees were obtained at the UFPA, Belém, PA, Brazil, in 1995 and 1998 respectively in the Electronic Engineer course with focus on Telecommunications. He is a PhD Student in Marketing and E-commerce at University of Bergamo (UNIBG – 2005/2008), Italy. At the moment he is a Visiting Scholar at University of Rhode Island (USA – 2007/2008). He is an Associate Professor of Computer Engineering and Telecommunications Engineering in the Department of Engineering at the Amazon Studies Institute (IESAM), Brazil. His research involves two areas: (a) telecommunications solution (publishing in IEEE Society), and (b) studies of economy and marketing tools, (publishing in the business and marketing journals) about technology marketing. He has been reviewing for the IEEE Transactions on Engineering Management (R&D) and for the International Journal of Technology Marketing (IJTMkt). Address: College Business Administration, Office 211, URI, Zip code: 02881 – Kingston – RI, USA. Phone: +1 401 2120498, fax number: +39 035 2052549, e-mail: earruda@prof.iesam-pa.edu.br / earruda@etal.uri.edu.

Elionai Sobrinho was born in Belém, PA, Brazil, on March 02, 1970. His Bachelors and Masters Degrees were obtained at the UFPA, Belém, PA, Brazil, in 1994 and 2002 respectively in the Electronic Engineer course with focus on Telecommunications. He is a PhD Student in Application Computing at Federal University of Pará (UFPA – 2003/2008), Brazil. He is an Associate Professor of Telecommunications Engineering, Computer Engineering and Control and Automation Engineering in the Department of Engineering at the Amazon Studies Institute (IESAM), Brazil. His PhD is a research in Dynamic Bayesian Network with application in Telecommunication Performance. Address: Instituto de Estudos Superiores da Amazônia, Av. Gov. José Malcher, 1148, Zip code: 66055-260 – Belém – PA, Brazil. Phone: +55 91 4005 5400, fax number: +55 91 4005 5407, e-mail: elionai@prof.iesam-pa.edu.br.

Silvia Biffignandi is full professor on Economic and Business Statistics at the University of Bergamo since 1990 and head of the Department of Mathematics, Statistics, Informatics and Applications. Since 2007 she is also Director of the new Interdepartmental Center for Statistical Analyses and Interviewing-surveys (CASI) of the University of Bergamo. She has been visiting by statistical bodies in USA, Canada and France. She has been expert for projects of many national and international bodies (National Statistical Institute, Eurostat, Italian Statistical Advisory Board, NWO-Large Investment Program Netherlands). She has been coordinating various research projects at national and international level. Main research area is on surveying and statistical modelling and on the use of IT tools for information supporting economic, business analyses and decision. Some special research focuses has been on the analysis of small business and economic domains (small areas, specific sectors or business function analyses). Address: Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Facoltà di Economia, Via Caniana n. 2, 24127 Bergamo, Italia. Phone: 0039 035 2052516, fax number: +39 035 2052549, e-mail: silvia.biffignandi@unibg.it

Alberto Marino is a full Professor of Marketing and E-Commerce and Coordinator of the Ph.D. Program in Marketing at University of Bergamo. Main research area is on marketing and business modelling and on the use of IT tools for information supporting economic, business analyses and decision. Some special research focuses has been on the analysis of small business and economic domains (small areas, specific sectors or business function analyses). During thirty years on the faculty he has taught courses on marketing, Marketing and E-commerce and retailing at both MBA and executive levels. In dual finance and value marketing, Marino is particularly interested in the demand side of the industrial sector. Marino has written many books on competitors, business mix, trade marketig plans and marketing as a tool based on differents sub sectors. His research in this area has centered mostly around the analysis of risk. Address: Dipartimento di Economia Aziendali, Facoltà di Economia, Via Caniana n. 2, 24127 Bergamo, Italia. Phone: 0039 035 2052509, fax number: +39 035 2052549, e-mail: alberto.marino@unibg.it.