Channel Estimation by Symmetrical Clustering

Jugurta R. Montalvão Filho, Bernadette Dorizzi, and João Cesar M. Mota, *Member, IEEE*

***Abstract*—A new blind channel estimation algorithm is presented in this paper. This algorithm comes from the well-known maximum likelihood estimation approach. However, we intentionally "smooth" the joint probability density function (pdf) of a finite set of observations in order to reduce the computational burden. As a result, we obtain an online clustering algorithm whose main characteristic is the constraint of symmetry among cluster centers. Computational simulations are used to evaluate this algorithm.**

***Index Terms*—Channel estimation, deterministic annealing, symmetrical clustering.**

I. Introduction

THIS work concerns the blind estimation of the channel impulse response of a sampled system. We assume that the system is linear and time-invariant, the emitted signal is digital, and that the receiver is synchronously sampled. In such a case, the joint probability density function (pdf) of all the received data $x(0), \ldots, x(n)$ is a parametric function of the channel impulse response. Hence, the inference of channel parameters based on the received data is a standard parametric problem for which the maximum likelihood estimation (MLE) may be asymptotically efficient [13], [18].

However, in some cases, the joint pdf may have an intricate dependency on the channel parameters, and the MLE cannot be used because of computational limitations.

A possible suboptimal solution is based on the hypothesis that, given a *window* of M synchronously sampled channel outputs, if M is sufficiently large, we may assume that statistical dependency of the samples $x(n), \ldots, x(n-M+1)$ on a specific emitted symbol is much stronger than its dependency on any sample outside this block [16].

Roughly speaking, the idea of assuming statistical independence between blocks of observations has been successfully applied in problems involving hidden Markov models (HMMs) [25] (conditional *split data likelihood*), deconvolution problems [18], and channel identification [1] (*partial likelihood* estimation).

In our work, we have taken the same path but with some additional constraints. For instance, the pdf estimator is composed of a very small number of Gaussian kernels, which corresponds to using a smoothed model of likelihood function.

Indeed, using a smoothed pdf estimator is the main feature of our proposal because it dramatically reduces the computational load of the resulting algorithm.

Another important aspect of the proposed blind estimation algorithm is the explicit use of *a priori* knowledge about the finite modulation alphabet. Indeed, it is well known that since a discrete-alphabet signal is obviously non-Gaussian, it has led to the development of numerous identification/equalization methods based on implicit high-order statistics (HOS) [4], [8], [20], [24], [26], [27], explicit HOS [5], [9], [21], [29], [31], maximum likelihood [11], [12], [15], [28].

For instance, among the recently proposed approaches based on the exploitation of the finite alphabet of the transmitted symbols, we can cite [32], where the authors propose an algebraic strategy, and [3], where a blind subspace criterion is combined to a decision-directed one.

Furthermore, concerning single input multiple output (SIMO) systems,[1] in [2], a blind recursive algorithm based on deterministic maximum likelihood methods is proposed, where the knowledge about the finite alphabet of symbols is also exploited.

All those recent articles confirm that algorithms based on the *a priori* knowledge about the finite-modulation alphabet are able to improve the blind equalizer performance while keeping low computational burden. More specifically, special attention has been devoted to hybrid algorithms that combine decision-directed (DD) equalization with another blind equalizer technique, which is able to open an initially closed channel eye (see, for instance, [19] and references therein).

Fortunately, as it is shown in Section III-A, our proposed algorithm has such a characteristic of a hybrid cost function, which enables it to open an initially closed eye as well as to converge asymptotically toward a DD equalization.

This paper is organized as follows. The system model is defined in Section II. In Section III, we show why channel identification can be seen as a symmetric clustering task, and then, we present our probabilistic approach, as well as the resulting cost function. In Section III-B, the symmetrical clustering algorithm is presented, whereas in Section IV, we present some illustrative simulation results with the 4-QAM modulation scheme and compare the performance of this algorithm to that of the higher order statistics (HOS) algorithm proposed by Porat and Friedlander [21]. Furthermore, this algorithm is also compared with the algebraic approach proposed by Yellin and Porat [32]. Finally, some more illustrative simulations are presented.

II. System Model

We consider a communication scheme where digital data is represented by a stochastic process $\{a(n)\}$, $n \in \{\ldots,$

Manuscript received December 7, 2000; revised February 22, 2002. The associate editor coordinating the review of this paper and approving it for publication was Prof. Nicholas D. Sidiropoulos.

J. R. Montalvão Filho is with the EPH Department, Institut National des Télécommunications, Evry, France, and also with Tiradentes University, Aracaju, Brazil (e-mail: jugurta_montalvao@unit.br).

B. Dorizzi is with the EPH Department, Institut National des Télécommunications, Evry, France (e-mail: Bernadette.Dorizzi@int-evry.fr).

J. C. M. Mota is with the Federal University of Ceará (UFC), Fortaleza, Brazil (e-mail: mota@dee.ufc.br).

Publisher Item Identifier S 1053-587X(02)04387-8.

[1] A SIMO system is also addressed in [3].

$-1, 0, 1, \ldots\}$, which is drawn with equal probability from a finite and symmetric alphabet of S complex symbols $a(n) \in \{a_1, a_2, \ldots, a_S\}$, forming an independent and identically distributed (i.i.d.) sequence of variance σ_a^2. Furthermore, the noise, which is represented by $\{b(n)\}$, is additive white Gaussian with zero mean and variance σ_b^2. Finally, the channel model is a time-invariant finite impulse response (FIR) filter with N taps.

Let $\mathbf{x}(n) = [x(n)x(n-1)\cdots x(n-M+1)]^T$ denote a *window* of M consecutive channel outputs (the superscript T stands for matrix-transpose). This output is modeled as

$$\mathbf{x}(n) = \mathbf{F}^T\mathbf{a}(n) + \mathbf{b}(n)$$

where

$$\mathbf{F} = \begin{bmatrix} \mathbf{f} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{f} \end{bmatrix}_{(N+M-1)\times M}$$

is the channel convolution matrix, $\mathbf{f} = [f_0 \ f_1 \ \cdots \ f_{N-1}]^T$ is the channel impulse response; $\mathbf{a}(n) = [a(n) \ \cdots \ a(n-d) \ \cdots \ a(n-N-M+2)]^T$ is a *window* of $M+N-1$ symbols, and $\mathbf{b}(n) = [b(n)\, b(n-1) \ \cdots \ b(n-M+1)]^T$ is a *window* of M noise samples.

From a geometric point of view, each realization of the random vector $\mathbf{x}(n)$ defines a point in an M-dimensional complex space C^M, and such points can be labeled according to the corresponding realization of the random variable $a(n-d)$, where d is an arbitrary decision delay. Moreover, it is straightforward to show that there are S clusters of points of the same label whose means—or "barycenters"—are given by the conditional means:

$$\tilde{\mathbf{c}}_{a_s} = E_{\mathbf{x}}\{\mathbf{x}(n)|a(n-d) = a_s\} = a_s \begin{bmatrix} f_d \\ f_{d-1} \\ \vdots \\ f_{d-M+1} \end{bmatrix} \quad (1)$$

where $f_i = 0$ if $i > N-1$ or $i < 0$. For instance, in a 4-QAM scheme, there are four clusters where each such cluster is associated with one complex symbol in the modulation alphabet.

Therefore, according to (1), finding these barycenters corresponds to identifying the channel model coefficients. Furthermore, if $M \geq N$ and $N-1 \leq d < M$, then $\tilde{\mathbf{c}}_{a_s}$ is a function of all the channel coefficients (inversely ordered).

If the symbol alphabet is symmetrically valued, as in the S-PSK or S-QAM schemes, then it is clear from (1) that the barycenters are also symmetrically placed in the M-dimensional complex space. As a consequence, a straightforward strategy to perform a blind channel identification is to apply a clustering algorithm over the M-dimensional observations $\mathbf{x}$. Moreover, in order to take advantage of the *a priori* knowledge about the symmetry between centers, we can also impose the constraint of symmetry between estimates, i.e.,

$$\hat{\mathbf{c}}_{a_s} = a_s \begin{bmatrix} \hat{f}_d \\ \hat{f}_{d-1} \\ \vdots \\ \hat{f}_{d-M+1} \end{bmatrix}, \qquad a_s \in \{a_1, a_2, \ldots, a_S\}.$$

In [17], we studied two approaches to perform channel identification using symmetrical clustering. The first one corresponds to a geometrical strategy where the estimated centers are stochastically adapted according to the "winner-takes-all" rule, like an on-line K-means algorithm. We have shown that the simplest geometrical clustering algorithm is strongly related to the Bussgang [10] algorithms for blind equalization. This is due to the fact that finding the label of the closest center—the winner—is equivalent to estimating $a(n-d)$.

The second approach is better because it is based on a probabilistic formalization of the problem, where, in fact, the clustering algorithm comes from a pdf fitting strategy. This paper is concerned with a special case of this second approach, where conditional pdf smoothing is maximal.

III. Probabilistic Clustering Approach

It is straightforward to show that the pdf of the random vector $\mathbf{x}(n)$ is given by a Gaussian mixture [22], which is parameterized by the channel coefficients and the noise variance, i.e.,

$$p(\mathbf{x}(n); \mathbf{f}, \sigma_b^2) = \frac{1}{S^{N+M-1}} \sum_{i=1}^{S^{N+M-1}} \phi_i(\mathbf{x}(n))$$

where

$$\phi_i(\mathbf{x}(n)) = \frac{1}{\sqrt{(2\pi\sigma_b^2)^M}} \exp \frac{-\|\mathbf{x}(n) - \mathbf{F}^T\mathbf{a}_i\|_2^2}{2\sigma_b^2}$$

$\|\cdot\|_2$ stands for l_2-norm, and $\mathbf{a}_i$ is the ith combination of $N+M-1$ symbols.

One possible strategy to estimate $\mathbf{f}$ and σ_b^2 can be the adaptation of a parametric likelihood function $l(\hat{\mathbf{f}}, \hat{\sigma_b^2})$ with S^{N+M-1} Gaussian kernels toward the maximization of $E_{\mathbf{x}}\{\ln(l(\mathbf{f}, \sigma_b^2))\}$. Nevertheless, it is well known [22] that the practical application of this maximum-likelihood approach to blind identification is discouraged by the possibly prohibitive number of Gaussian kernels of $l(\cdot)$.

In this paper, we propose a specific smoothed likelihood function that highlights the link between likelihood maximization and symmetrical clustering.

This smoothed likelihood function is formed by only S Gaussian kernels

$$q\left(\hat{\mathbf{f}}, \hat{\sigma}_b^2\right) = \frac{1}{S} \sum_{i=1}^{S} \tilde{\phi}_{a_i}(\mathbf{x}(n))$$

where

$$\tilde{\phi}_{a_i}(\mathbf{x}(n)) = \frac{\exp\left(\frac{-\left(\mathbf{x}(n)-\hat{\mathbf{c}}_{a_i}\right)^T \hat{\mathbf{R}}_s^{-1} \left(\mathbf{x}(n)-\hat{\mathbf{c}}_{a_i}\right)^*}{2}\right)}{\sqrt{(2\pi)^M \left|\hat{\mathbf{R}}_s\right|}}$$

$$\hat{\mathbf{R}}_s = \sigma_a^2\left(\hat{\mathbf{F}}^H\hat{\mathbf{F}} - \hat{\mathbf{c}}_1^*\hat{\mathbf{c}}_1^T\right) + \hat{\sigma}_b^2\mathbf{I}_{M\times M} \quad (2)$$

$$\hat{\mathbf{F}} = \begin{bmatrix} \hat{f}_0 & & \mathbf{0} \\ \vdots & \ddots & \\ \hat{f}_{\hat{N}-1} & \ddots & \hat{f}_0 \\ & \ddots & \vdots \\ \mathbf{0} & & \hat{f}_{\hat{N}-1} \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{c}}_1 = \begin{bmatrix} \hat{f}_d \\ \hat{f}_{d-1} \\ \vdots \\ \hat{f}_{d-M+1} \end{bmatrix}.$$

Superscript H stands for Hermitian transposition, and $*$ stands for conjugation.

Note that $\hat{\mathbf{R}}_s$ is the estimate of the cluster covariance matrix [see Appendix A for deduction of (2)]

$$\mathbf{R}_s = E_x\left\{(\mathbf{x}(n)-\tilde{\mathbf{c}}_{a_s})^*\,(\mathbf{x}(n)-\tilde{\mathbf{c}}_{a_s})^T \,\middle|\, a(n-d)=a_s\right\}$$

which is identical for all the S clusters.

Now, a natural cost function for adjusting $q(\cdot)$ should be $J(\hat{\mathbf{f}}, \hat{\sigma}_b^2) = -E_{\mathbf{x}}\{\ln(q(\hat{\mathbf{f}}, \hat{\sigma}_b^2))\}$. In fact, we first derived a stochastic gradient algorithm to minimize this cost function, and by computational simulation, we observed the following.

a) The cost function J is not necessarily convex. Moreover, the number and the "deepness" of local minima of J are related to the estimated variance $\hat{\sigma}_b^2$. This means that whenever $\hat{\sigma}_b^2$ is reduced, local minima of J became deeper, and we can also observe the "birth" of new minima.
b) Consequently, we observed that simulation trials starting from a relatively high parameter $\hat{\sigma}_b^2$ (high w.r.t. the actual noise variance) have a stronger probability of finding a good channel estimates.

Fortunately, these empirical results can be explained by the theory behind deterministic annealing [23], [30], which is a technique inspired by the existing analogy between optimization problems and statistical physics.

In order to take advantage of this theoretical background, we adapted deterministic annealing to our specific problem. This adaptation is straightforward [17]: We simply replace the noise variance estimate by a "temperature parameter," and then, we control this parameter while the stochastic minimization of the cost function is going on. Furthermore, the resulting cost function is a slightly modified version of J, i.e.,

$$J_{\mathcal{E}}\left(\hat{\mathbf{f}}, T_p\right) = -T_p \cdot E_{\mathbf{x}}\left\{\ln\left(q\left(\hat{\mathbf{f}}, T_p\right)\right)\right\}. \tag{3}$$

Note that in this case, we do not estimate the noise variance.

According to the physical analogy, we have the *a priori* knowledge that the number of local minima of $J_{\mathcal{E}}$ increases whenever the temperature goes down [23]. Actually, the number of local minima changes only at specific values of the temperature. For example, at very high temperatures, $J_{\mathcal{E}}$ has only one minimum at the origin, and the first transition occurs when $T_p = 2\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{R}_x = \sigma_a^2\mathbf{F}^H\mathbf{F} + \sigma_b^2\mathbf{I}_{M\times M}$. On the other hand, very low temperatures may cause an excessive number of local minima.

Operating between these two extremes, the annealing process tries to reach the global minimum by starting the optimization process at high temperatures and keeping the process going while the temperature is slowly lowered.

A. Cost Function Analysis

According to [23], the cost function in (3) is convex when $T_p > 2\lambda_{\max}$, where $\lambda_{\max}$ is the maximum (in absolute value) eigenvalue of $\mathbf{R}_x$. Moreover, in our specific case, due to the cluster symmetry, the minimum of this cost function is at the origin (a null vector).

Then, it is evident that we are looking for minima of the cost function for low temperatures. Unfortunately, in this case, the cost function is not convex. That is why we use an annealing strategy in order to track the deepest minima while the cost function changes.

However, what can we say about minima in low temperatures? Though a rigorous study of all these minima seems to be too difficult (due to the imbricated dependence of $\hat{\mathbf{R}}_s$ on $\hat{\mathbf{f}}$), we are able to approximately analyze one class of such a minima. This is made possible thanks to the following property of matrix $\hat{\mathbf{R}}_s$.

Property: Given $0 \le d \le \hat{N} + M - 2$, the minimum eigenvalue of $\hat{\mathbf{R}}_s$, $\hat{\lambda}_0$ and its corresponding eigenvector $\hat{\mathbf{h}}_0$ can be approximated by

$$\text{A1)}\ \hat{\lambda}_0 \approx \lambda_z = \sigma_a^2\frac{\|\hat{\mathbf{g}}_z - \hat{g}_{z_d}\boldsymbol{\delta}_d\|_2^2}{\|\mathbf{h}_z\|_2^2} + \frac{T_p}{2}$$

$$\text{A2)}\ \hat{\mathbf{h}}_0 \approx \mathbf{h}_z = \left(\hat{\mathbf{F}}^H\hat{\mathbf{F}}\right)^{-1}\hat{\mathbf{F}}^H\boldsymbol{\delta}_d$$

where $\hat{\mathbf{g}}_z = \hat{\mathbf{F}}\cdot\mathbf{h}_z$, and $\boldsymbol{\delta}_d = [0\ \cdots\ 0\ 1\ 0\ \cdots\ 0]^T$ (the 1 is at the dth position).

Proof: See Appendix B.

Thanks to these approximations, it is clear that for a channel without spectral nulls (for which a FIR zero forcing equalizer may provide good equalization for finite values of M) for values of M not too small (which corresponds to the length of a "virtual" zero forcing equalizer $\mathbf{h}_z$) the second term of A1 predominates. Moreover, since T_p is a control parameter, when $T_p/2$ is made much smaller than any other eigenvalue of $\hat{\mathbf{R}}_s$, thanks to the singular value decomposition theorem, for most of the channels, we have (see discussion in Appendix D)

$$\text{A3)}\ \hat{\mathbf{R}}_s^{-1} \approx \mathbf{h}_z\mathbf{h}_z^H/\lambda_z.$$

Now, we are able to study the cost function $J_{\mathcal{E}}$ throughout approximations A1–A3. First, we expand the cost function as

$$J_{\mathcal{E}}\left(\hat{\mathbf{f}}, T_p\right) = -T_p\left(E\left\{\ln\left(\hat{\phi}_1(\mathbf{x}(n)) + \cdots + \hat{\phi}_S(\mathbf{x}(n))\right)\right\} - \ln\left(S\sqrt{\left|\hat{\mathbf{R}}_s\right|(2\pi)^M}\right)\right). \tag{4}$$

In the stochastic adaptation algorithm, we intentionally neglect the dependence of $\hat{\mathbf{R}}_s$ on $\hat{\mathbf{f}}$; see Section III-B. Then, we should concentrate our attention only on the first term of (4), where the application of A3 to the Gaussian kernels $\hat{\phi}_i(\cdot)$ leads to

$$\hat{\phi}_s(\mathbf{x}) \approx \exp\left(-(\mathbf{x}-\hat{\mathbf{c}}_s)^T\mathbf{h}_z\mathbf{h}_z^H(\mathbf{x}-\hat{\mathbf{c}}_s)^*/(2\lambda_z)\right).$$

Given that $\hat{\mathbf{c}}_s^T\mathbf{h}_z = a_s\hat{g}_{z_d}$, then

$$\hat{\phi}_s(\mathbf{x}) \approx \exp\left(-|y(n) - a_s\hat{g}_{z_d}|^2/(2\lambda_z)\right), \qquad s = 1, \ldots, S$$

where $y(n) = \mathbf{x}(n)^T\mathbf{h}_z$ is the output of a "virtual" zero-forcing equalizer.

Now, since we get a winner center[2] for each realization of the random vector $\mathbf{x}(n)$, it is useful to associate this center to the random variable $\hat{\phi}_{\max}(\mathbf{x}(n)) = \max_{s=1,\ldots,S}(\hat{\phi}_s(\mathbf{x}(n)))$, and to be coherent with usual notation, we should say that the symbol a_s associated with the winner center is an emitted symbol estimate $\hat{a}(n-d)$. That is to say that

$$\hat{a}(n-d) = \arg\max_{a_s\in\mathrm{A}}\left(\hat{\phi}_s(\mathbf{x}(n))\right).$$

[2]The closest one in Mahalanobis distance.

Then, we obtain

$$J_{\mathcal{E}} = -T_p \left(E\left\{ \ln\left(\hat{\phi}_{\max} + \sum_{s \neq \max} \hat{\phi}_s \right) \right\} - \ln\left(\left| \hat{\mathbf{R}}_s \right| S\sqrt{(2\pi)^M} \right) \right)$$

where $\hat{\phi}_{\max} \approx \exp(-|y(n) - \hat{a}(n-d)\hat{g}_{z_d}|^2/(2\lambda_z))$.

From these expressions, it is straightforward to observe the following.

- When the temperature T_p is high, relative to the variance of $\mathbf{x}(n)$, distances computed in

$$-T_p \cdot E\left\{ \ln\left(\hat{\phi}_{\max} + \sum_{s \neq \max} \hat{\phi}_s \right) \right\}$$

  tend to be numerically closer to each other, and then, the "winner center" is not clearly observed.
- On the other hand, for low temperatures, and when M, which is length of the virtual zero-forcing equalizer $\mathbf{h_z}$, is big enough to allow a good zero-forcing inversion of the channel, if the estimate $\hat{\mathbf{f}}$ is close to the channel $\mathbf{f}$, up to a phase rotation and/or a delay shift (for $\hat{N} > N$), then the channel eye is open, and the summation in the cost function is dominated by the term associated to the winner center

$$-T_p \cdot E\left\{ \ln\left(\hat{\phi}_{\max} + \sum_{s \neq \max} \hat{\phi}_s \right) \right\} \approx -T_p \cdot E\left\{ \ln\left(\hat{\phi}_{\max} \right) \right\} \approx E\left\{ (y(n) - \hat{a}(n-d)\hat{g}_{z_d})^2 \right\}.$$

  It is worth noting that the expression on the right side of the equation, for $\hat{g}_{z_d} \approx 1$, coincides with a mean squared error decision directed (MSE-DD) cost function since it measures the mean squared deviation of the virtual ZF equalizer output from the implicit symbol decision $\hat{a}(n-d)$.

It is an interesting result because the MSE-DD-based algorithm leads to the lowest steady-state error among all blind algorithms [19], although it is not able to open an initially closed channel eye.

In contrast, we have observed by simulations that the cost function $J_{\mathcal{E}}$ is able to open a channel eye at high and medium temperatures, and according to the approximations shown previously, it converges to a MSE-DD cost function when the eye is open.

That is a very worthwhile feature for a blind cost function. Indeed, as it was commented in Section I, some algorithms proposed in the references force such a mixed behavior in their solutions. Here, we have obtained it naturally, as a consequence of the chosen cost function and thanks to a property of the estimated covariance matrix $\hat{\mathbf{R}}_s$.

B. Algorithm

We have used a stochastic gradient method to find a minimum of $J_{\mathcal{E}}$ by adapting the estimated channel coefficients. Therefore, we need to calculate the gradient

$$\nabla_{\hat{\mathbf{f}}} J_{\mathcal{E}} = -2T_p E_{\mathbf{x}} \left\{ \frac{\partial \ln\left(q\left(\hat{\mathbf{f}}, T_p \right) \right)}{\partial \hat{\mathbf{f}}^*} \right\}. \tag{5}$$

Nevertheless, it is worth noting that $\hat{\mathbf{R}}_s$ depends on $\hat{\mathbf{f}}$ in such a way that determining this gradient is quite a difficult task. Then, we preferred to adapt the algorithm as follows.

- Fix $\hat{\mathbf{R}}_s$ while adapting $\hat{\mathbf{f}}$ (stochastically, one step).
- Recalculate $\hat{\mathbf{R}}_s$ (analytically).

The stochastic adaptation of $\hat{\mathbf{f}}$ is done by

$$\hat{\mathbf{f}}^{(k+1)} = \hat{\mathbf{f}}^{(k)} + 2\gamma T_p \frac{\partial \ln\left(q\left(\hat{\mathbf{f}}, T_p \right) \right)}{\partial \hat{\mathbf{f}}^*} \tag{6}$$

where

$$\left. \frac{\partial \ln(q)}{\partial \hat{\mathbf{f}}^*} \right|_{\hat{\mathbf{R}}_s^{-1} = cte} = \mathbf{P}(\hat{\mathbf{R}}_s^{-1})^T \frac{\sum_{i=1}^{S} a_i^* \mathbf{d}_i \exp\left(\frac{-\mathbf{d}_i^T \hat{\mathbf{R}}_s^{-1} \mathbf{d}_i^*}{2} \right)}{2 \sum_{i=1}^{S} \exp\left(\frac{-\mathbf{d}_i^T \hat{\mathbf{R}}_s^{-1} \mathbf{d}_i^*}{2} \right)}$$

where

$$\mathbf{d}_i = \mathbf{x} - \hat{\mathbf{c}}_i, \qquad \text{and} \qquad \mathbf{P} = \begin{bmatrix} \mathbf{0} & & 1 \\ & \iddots & \\ 1 & & \mathbf{0} \end{bmatrix}_{M \times M}.$$

In order to underline the clustering nature of this algorithm, we can alternatively rewrite (6) as

$$\hat{\mathbf{c}}_1^{(k+1)} = \hat{\mathbf{c}}_1^{(k)} + \gamma T_p \left(\hat{\mathbf{R}}_s^{-1} \right)^T \frac{\sum_{i=1}^{S} a_i^* \mathbf{d}_i \exp\left(\frac{-\mathbf{d}_i^T \hat{\mathbf{R}}_s^{-1} \mathbf{d}_i^*}{2} \right)}{\sum_{i=1}^{S} \exp\left(\frac{-\mathbf{d}_i^T \hat{\mathbf{R}}_s^{-1} \mathbf{d}_i^*}{2} \right)}. \tag{7}$$

Then, we can see that in this clustering algorithm, the "winner-takes-all" rule is replaced by something more powerful, i.e., all distances $\mathbf{d}_i$ are linearly transformed by the inverse of the cluster covariance matrix $\hat{\mathbf{R}}_s^{-1}$. Then, these "corrected" metrics are weighted by the negative exponential of the Mahalanobis [6] distance between the observation $\mathbf{x}$ and each center $\hat{\mathbf{c}}_i$. Finally, a "generating center" $\hat{\mathbf{c}}_1$ is updated according to this weighted summation of corrected distances. Note that symmetry between centers is assured because the adjustment of all centers is made in one go, through the adjustment of the generating center.

On the other hand, applying deterministic annealing approach, we exponentially lower the "system temperature" (in place of the estimated noise variance) while using the stochastic gradient to adapt parameter $\hat{\mathbf{c}}_1$ (or, equivalently, $\hat{\mathbf{f}}$). Indeed, according to some heuristics that we are not going to present here, we usually apply an exponential annealing rate where $T_p^{(0)} < 2M\hat{\sigma}_x^2$ and $T_p^{(NS)} > 2\hat{\sigma}_b^2$, and NS is the number of observations to be considered in each on-line annealing trial.

The exponential annealing rate leads to the following temperature updating expression:

$$T_p^{(k)} = T_p^{(0)}(1-\eta)^k \tag{8}$$

where $\eta = 1 - 10^{(\log_{10} T_p^{(NS)} - \log_{10} T_p^{(0)})/NS}$. It is directly computed from (8) for $k = NS$.

Clearly, in such an approach, we need first to estimate the received signal variance σ_x^2, as well as the noise variance σ_b^2, which are both supposedly stationary in a frame of NS symbols. Fortunately, we have observed that even if we have a reasonable "guess" about σ_x^2 and σ_b^2, the effect of bad guesses seems to be negligible for practical purposes.

Concerning the channel length estimate, since we suppose that all coefficients to be estimated are nonzero, we fix the estimated channel order equal to the length of $\mathbf{x}(n)$: $\hat{N} = M$. Furthermore, we also fix $d = \hat{N} - 1$.

The resulting algorithm can be summarized as follows.

Initialization:

Set $\hat{\mathbf{f}}^{(0)} = [0 \ \cdots \ 0 \ 1 \ 0 \cdots \ 0]^T$ (the 1 is at the dth position), γ, $T_p^{(0)}$ and $T_p^{(NS)}$.

Compute $\eta = 1 - 10^{(\log_{10} T_p^{(NS)} - \log_{10} T_p^{(0)})/NS}$.

For $k = 1$ to NS, do the following.

- Analytically compute $\hat{\mathbf{R}}_s^{(k)}$ from (2).
- Stochastically update $\hat{\mathbf{f}}^{(k)}$ from (6) [or $\hat{\mathbf{c}}_1^{(k)}$ from (7)] for every realization of $\mathbf{x}(k_0 + k)$, where k_0 is an arbitrary offset index.
- Lower temperature according to $T_p^{(k)} = T_p^{(k-1)}(1 - \eta)$ [equivalent to (8)].

End.

Thanks to the *overall* learning-rate parameter adaptation [the resulting value of γT_p in (7)], the proposed algorithm is somewhat similar to Kohonen's self-organizing map (SOM) [14], which is a classical neural network algorithm. This equivalence is helpful because it can provide us with some insights into the performance of this new identification algorithm. Nevertheless, the proposed algorithm has a significant particularity: the symmetry constraint, which is obtained by the adaptation of all the centers at once, through the adaptation of the "generating center."

C. Class of Channels That are Identifiable

Given the clustering nature of the proposed algorithm, if clusters of observations $\mathbf{x}(n)$ are not overlapped (or just slightly overlapped), the algorithm is able to find the cluster centers or, equivalently, a channel estimate. Moreover, if the channel does not have spectral nulls, as shown in Appendix E (see also [17] for more details), we can improve the separation between cluster by changing the dimension of the observation vector $\mathbf{x}(n)$, regardless of whether the channel is minimum, maximum, or nonminimum phase.

Then, the proposed algorithm is theoretically able to deal with any kind of channel without spectral nulls, but special attention must be paid to the dimensions of the observation vector. For instance, channels with near-nulls demand higher values of $M(=\hat{N})$ for the algorithm to work properly.

Since we assume that no *a priori* information about the channel is available, we arbitrarily choose a "not too small" length for $\hat{\mathbf{f}}$, keeping in mind that it also corresponds to the dimension of the observation vector. On the other hand, even if we know the length of $\mathbf{f}$, we typically use $\hat{N} = 2N$ in order to leave enough places for eventual delay shifts on the estimate (we initialize the estimator with $[0 \ \cdots \ 1 \cdots \ 0]^T$; see Section III-B.

IV. Simulations

In this section, the proposed algorithm is compared by simulation to two different algorithms for blind channel identification. First, we compare our algorithm with the well-known algorithm by Porat and Friedlander [21] based on higher order statistics (HOS). In fact, in [21], two off-line (block) algorithms are proposed. We are going to compare our algorithm with the most powerful of them: the *nonlinear least square estimation algorithm*. After that, we also compare the proposed algorithm with the algorithm provided by Yellin and Porat [32], which uses the finite alphabet constraint (as we do). Finally, we address identifiability and convergence issues by means of illustrations.

In order to make the first comparison, we use the same performance measure: the residual intersymbol interference (RISI), which is defined by

$$\text{RISI} = 10\log_{10}\left(\left(\mathbf{g}^H\mathbf{g} - \max_i |g_i|^2\right)\Big/ \max_i |g_i|^2\right)$$

where $\mathbf{g} = \mathbf{F}\mathbf{h}_W$ corresponds to the combined impulse response of the channel and a Wiener transversal equalizer given by

$$\mathbf{h}_W = \left(\sigma_a^2\hat{\mathbf{F}}_W^H\hat{\mathbf{F}}_W + \hat{\sigma}_b^2\mathbf{I}_{M\times M}\right)^{-1}\sigma_a^2\hat{\mathbf{F}}_W^H\boldsymbol{\delta}_d$$

where d is an arbitrary decision delay set in this first comparison with $d = 36$. Note that like $\hat{\mathbf{F}}$, $\hat{\mathbf{F}}_W$ is also a convolution matrix, but in the following simulations, $\hat{\mathbf{F}}$ has only five columns (because $\hat{N} = 5$), whereas $\hat{\mathbf{F}}_W$ has 65 columns (because the Wiener equalizer has 65 inputs). We emphasize that the Wiener transversal equalizer is a separate block since the proposed algorithm only performs channel estimation and that the choices of 65 taps and $d = 36$ are made in order to fairly compare our channel estimates with those in [21] since it was what they also used.

The modulation scheme is the 4-QAM, and the channel model is represented by $\mathbf{f} = [2 - 0.4j \ \ 1.5 + 1.8j \ \ 1 \ \ 1.2 - 1.3j \ \ 0.8 + 1.6j]^T$. Furthermore, the signal-to-noise ratio $(\text{SNR} = 10\log(\sigma_x^2/\sigma_b^2))$ is 40 dB.

Note that since this channel presents two in-band near nulls (see Fig. 4), blind equalization of this channel is quite a difficult task.

Figs. 1 and 2 illustrate, respectively, the two identification algorithms coupled with two linear transversal Wiener equalizers with the same number of taps.

We would like to underline that unlike Porat and Friedlander's algorithm, the symmetrical clustering algorithm does not estimate the noise variance $\hat{\sigma}_b^2$. Instead, we set it to a small value just to get $\hat{\mathbf{R}}_s$ nonsingular, even when $\hat{\mathbf{f}} \approx \mathbf{0}$. In the presented simulations, we set $\hat{\sigma}_b^2 = 0.01$.

Simulation results are shown in Fig. 3, where each point was averaged over 100 Monte Carlo independent runs.

In these trials, the initial temperature was set to $T_p^{(0)} = 6$, the final temperature was set to $T_p^{(NS)} = 0.05$, and the gradient step was set in the range $0.01 \le \gamma \le 0.03$. These choices were made empirically.

We can see that for more than 4000 samples, the on-line clustering algorithm provides a better channel equalization than the

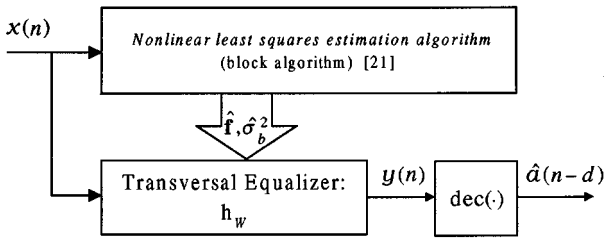

Fig. 1. Nonlinear least square estimation algorithm coupled to a 65-tap Wiener linear equalizer.

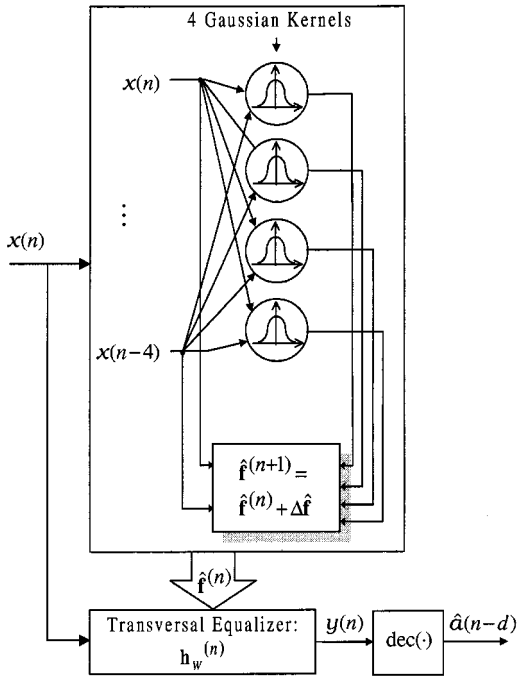

Fig. 2. Symmetrical clustering algorithm coupled to a 65–tap Wiener linear equalizer.

off-line (block) nonlinear HOS algorithm of Porat and Friedlander. However, although we observed a high convergence ratio among the independent simulation trials (100 trials per point), we note that the clustering algorithm may eventually have a poor convergence (toward spurious minima of the cost function) in terms of RISI.

Moreover, since our algorithm is on-line, the period of time of one annealing trial is $NS \cdot \Delta t$, where Δt is the symbol period. For instance, according to the simulation presented in Fig. 3, we can get an RISI of about -17 dB in $8000\Delta t$ s. Therefore, it is clear that in this algorithm, deterministic annealing does not result in long bath simulations (which is frequently the case with probabilistic annealing).

Fig. 4 shows the frequency responses of the channel before and after equalization for 8000 samples.

Results presented in Fig. 3 were obtained for a SNR of 40 dB, which is convenient for comparison with the results presented in [21]. Nevertheless, Fig. 5 illustrates the symmetrical clustering algorithm for the same channel but with a SNR $= 15$ dB, compared to that for 40 dB. This comparison provides a rough idea about the effect of the SNR on the algorithm performance.

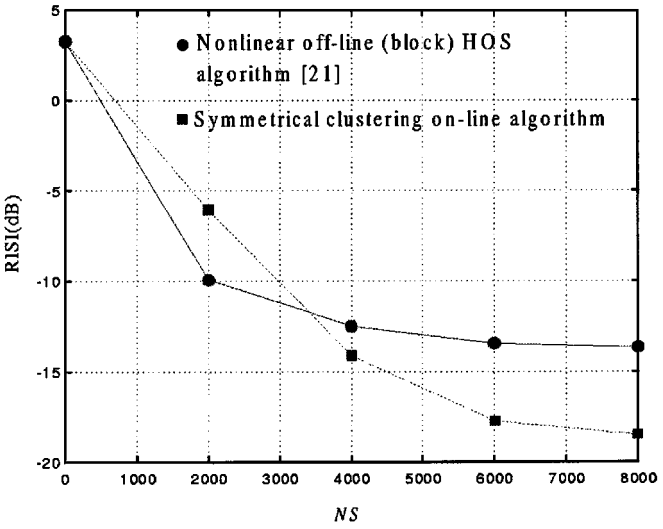

Fig. 3. Simulation results. Performance versus number of symbols. Two algorithms.

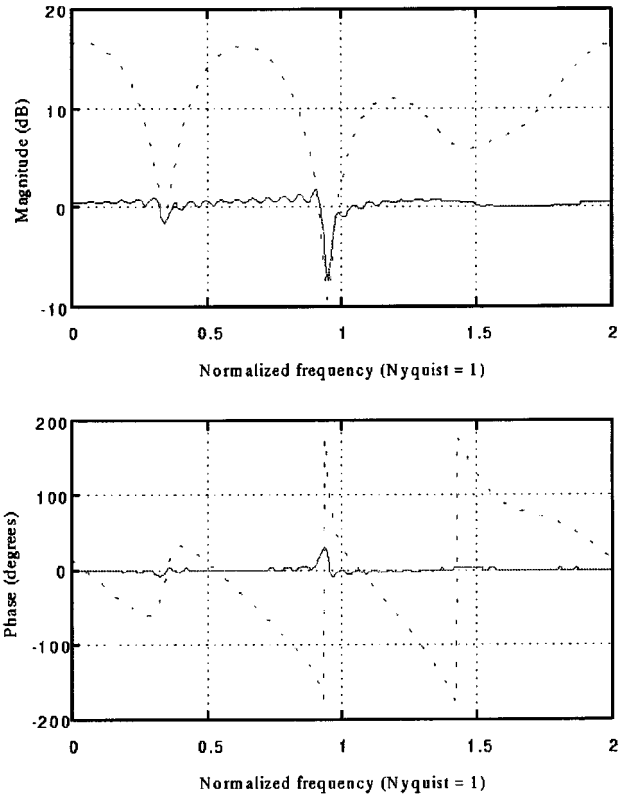

Fig. 4. Frequency responses of the channel before (dashed line) and after equalization.

For the second comparison, we got the longest channel used in [32], which illustrates a telephone channel: $\mathbf{f} = [0.06\ 0.02\ -0.60\ -0.05\ 1.30\ 0.01\ 0.36\ 0.02\ 0.10\ 0.01\ 0.02]^T$ and BPSK modulation.

By using their algebraic approach and then applying their channel estimate on a least squares FIR approximation of the inverse system (a transversal equalizer), Yellin and Porat obtained a mean square error (MSE $= E\{y(n) - a(n-d)\}$) of -17.2 dB for an SNR of 45 dB.

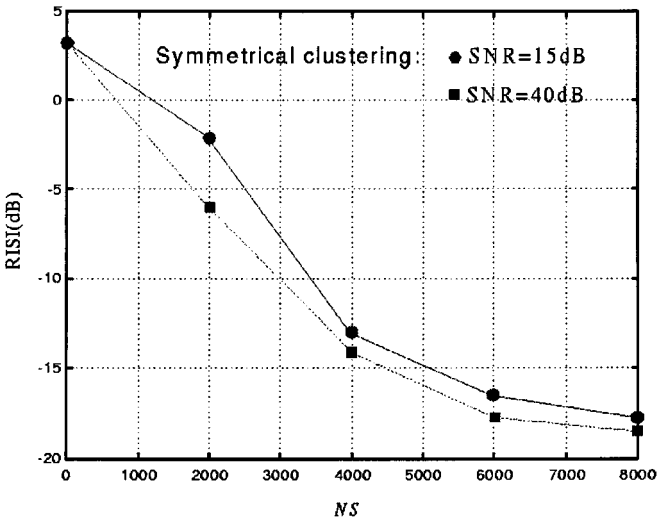

Fig. 5. Simulation results. Performance versus number of symbols. Two SNR values.

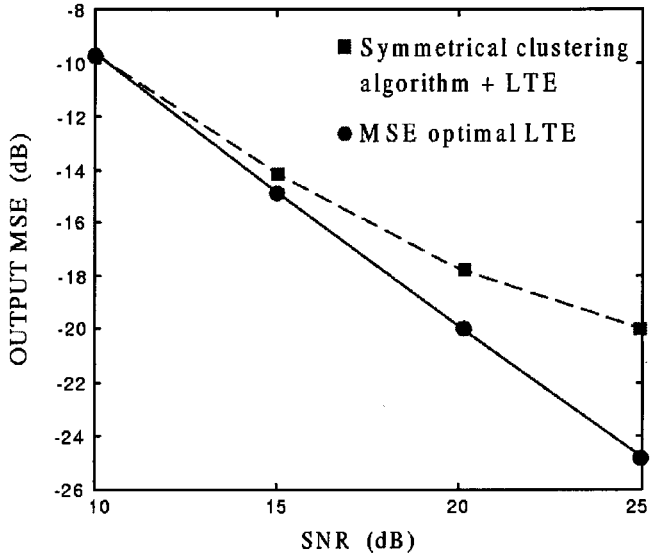

Fig. 6. BPSK, telephone channel, $\hat{N} = M = 11$, $NS = 2000$. Wiener equalizer length: 11 taps, $T_p^{(0)} = 1$, $T_p^{(NS)} = 0.01$, $\gamma = 0.02$. Fifty independent trials per point.

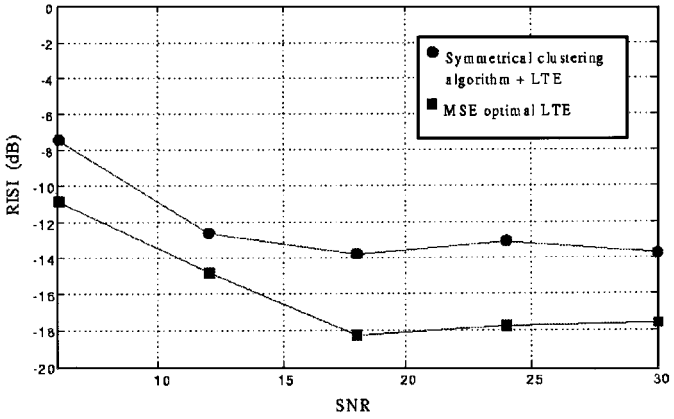

Fig. 7. BPSK, 150 randomly drawn channels (with three coefficients) per point, ten independent trials per channel, $\hat{N} = M = 7$, $NS = 4000$, linear equalizer length: 11 taps, $T_p^{(0)} = 2$, $T_p^{(NS)} = 0.01$, $\gamma = 0.01$.

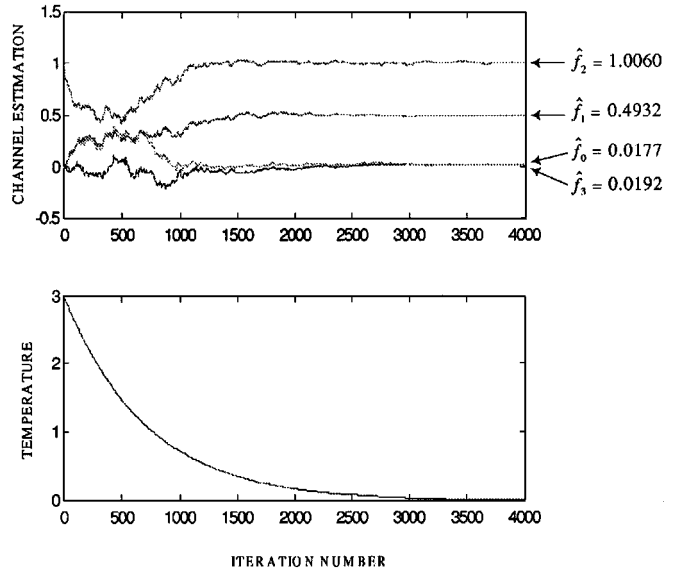

Fig. 8. BPSK, $\mathbf{f} = [0.5\ 1]^T$, $\hat{N} = M = 4$, SNR = 40 dB, $NS = 4000$, $T_p^{(0)} = 3$, $T_p^{(NS)} = 0.01$, $\gamma = 0.02$. Oone single trial.

In contrast, we tested our algorithm with the same channel for a range of noise levels. Fig. 6 shows the performances we obtained, as well as the optimum MSE for a linear transversal equalizer with 11 taps. Note that for a SNR beyond 25 dB, our algorithm provides a MSE lower than -20 dB.

On the other hand, the algorithm by Yellin and Porat converges after about 580 symbols were emitted, whereas we have used 2000 channel output samples to get the results shown in Fig. 6.

Evidently, the convergence of our algorithm depends on channel characteristics, cooling rate, and the stochastic gradient step. In fact, a theoretical analysis of the dynamic behavior of this algorithm has yet to be made.

In order to get a feeling about identifiability and convergence issues, the average RISI performance was computed over 150 randomly drawn channels (ten independent estimation trials per channel) versus SNR. Channel estimates were then used to implement MSE linear transversal equalizers. Finally, the MSE equalizer performances for the actual channel and its estimate were put together in Fig. 7.

Note that channel coefficients were independently drawn from a Gaussian random source with zero mean and unit variance. It means that no restrictions were imposed, and all classes of channels were potentially used in this simulation, including those with spectral nulls and near nulls.

Finally, to illustrate the effect of temperature parameter on the channel estimation process, as well as to illustrate the on-line evolution of estimated coefficients, one single estimation trial for $\mathbf{f} = [0.5\ 1]^T$ and $\hat{\mathbf{f}}^{(0)} = [0\ 1\ 0\ 0]^T$ is presented in Fig. 8.

In this trial, we set a high initial temperature. Consequently, we can easily observe that for the initially high temperatures, all coefficient estimates roughly tend to zero (500 initial iterations). We also can observe that the two additional coefficients of $\hat{\mathbf{f}}$ converge to zero at the end of the annealing process.

V. Conclusions and Perspectives

A new blind channel estimation algorithm has been presented in this paper. This algorithm uses a smoothed estimator of the sampled signal pdf, which provides control over the computational burden commonly associated with the maximum likelihood estimator in these cases.

The chosen cost function came from the analogy between the formulation of our parametric estimation problem and physical systems free energy minimization. Such an analogy provides a straightforward implementation of the deterministic annealing scheme in order to cope with local minima problems.

We have shown that the final algorithm is, from a certain point of view, a clustering algorithm, where a constraint of symmetry between clusters centers is imposed. Furthermore, the "winner-takes-all" rule commonly associated with most clustering algorithms is replaced by an adaptation rule that takes into account all centers. As a result, this clustering algorithm is similar to Kohonen's self-organizing map (SOM), which is a classical neural network algorithm.

Nevertheless, despite this helpful similarity and the good simulation results obtained up to now, some new aspects of this approach give rise to questions concerning the optimization of the algorithm parameters. Clearly, in order to optimize such parameters, a theoretical analysis of the proposed cost function seems to be necessary. Indeed, this is an open subject toward which our future work will be oriented.

Appendix

A. Calculation of $\hat{\mathbf{R}}_s$

For the calculation of $\hat{\mathbf{R}}_s$, an useful auxiliary matrix can be obtained from $\hat{\mathbf{F}}$ by keeping only the dth row of $\hat{\mathbf{F}}$ and setting to zero all the rest:

$$\hat{\mathbf{F}}_c = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \hat{f}_d & \cdots & \hat{f}_{d-M+1} \\ \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}_{(N+M-1)\times M}.$$

Thanks to this auxiliary matrix, we can write

$$(\mathbf{x}(n) - \tilde{\mathbf{c}}_{a_s})|_{a(n-d)=a_s} = \hat{\mathbf{F}}_a^T \mathbf{a}(n)\Big|_{a(n-d)=a_s} + \mathbf{b}(n)$$

where

$$\mathbf{a}(n)|_{a(n-d)=a_s} = [a(n) \ \cdots \ a_s \ \cdots \ a(n-N-M+2)]^T$$

and $\hat{\mathbf{F}}_a = (\hat{\mathbf{F}} - \hat{\mathbf{F}}_c)$. Substituting it in

$$\hat{\mathbf{R}}_s = E_x\left\{ (\mathbf{x}(n) - \tilde{\mathbf{c}}_{a_s})^* (\mathbf{x}(n) - \tilde{\mathbf{c}}_{a_s})^T \Big| a(n-d) = a_s \right\}$$

gives

$$\hat{\mathbf{R}}_s = \hat{\mathbf{F}}_a^H E_x\left\{\mathbf{a}(n)^* \mathbf{a}(n)^T | a(n-d) = a_s\right\} \hat{\mathbf{F}}_a + E_x\left\{\mathbf{b}(n)^* \mathbf{b}(n)^T\right\}$$

$$\hat{\mathbf{R}}_s = \sigma_a^2 \hat{\mathbf{F}}_a^H \begin{bmatrix} 1 & & & \mathbf{0} \\ & \ddots & & \\ & & |a_s|^2/\sigma_a^2 & \\ & & & \ddots \\ \mathbf{0} & & & 1 \end{bmatrix} \hat{\mathbf{F}}_a + \sigma_b^2 \mathbf{I}_{M\times M}.$$

Since matrix $\hat{\mathbf{F}}_a$ has a null row that multiplies the only term depending on a_s in the equation, i.e., $|a_s|^2/\sigma_a^2$, the resulting matrix does not depend on a_s, and it follows that

$$\hat{\mathbf{R}}_s = \sigma_a^2 \left(\hat{\mathbf{F}}^H - \hat{\mathbf{F}}_c^H\right)\left(\hat{\mathbf{F}} - \hat{\mathbf{F}}_c\right) + \sigma_b^2 \mathbf{I}_{M\times M}.$$

Finally, thanks to the following identities:[3] $\hat{\mathbf{F}}_c^H \hat{\mathbf{F}} = \hat{\mathbf{F}}^H \hat{\mathbf{F}}_c = \hat{\mathbf{F}}_c^H \hat{\mathbf{F}}_c = \hat{\mathbf{c}}_1^* \hat{\mathbf{c}}_1^T$, it yields

$$\hat{\mathbf{R}}_s = \sigma_a^2 \left(\hat{\mathbf{F}}^H \hat{\mathbf{F}} - \hat{\mathbf{c}}_1^* \hat{\mathbf{c}}_1^T\right) + \sigma_b^2 \mathbf{I}_{M\times M}.$$

B. Proof of the $\hat{\mathbf{R}}_s$ Matrix Property

Proof: By definition, the minimum eigenvalue of $\hat{\mathbf{R}}_s$ must satisfy

$$\hat{\lambda}_0 = \min_{\mathbf{h}} \left(\frac{\mathbf{h}^H \hat{\mathbf{R}}_s \mathbf{h}}{\mathbf{h}^H \mathbf{h}} \right). \tag{9}$$

Given the definition of $\hat{\mathbf{R}}_s$, after some algebraic manipulations (detailed in Appendix C), we obtain

$$\hat{\lambda}_0 = \min_{\mathbf{h}} \left(\frac{\sigma_a^2 \sum_{i=0}^{\hat{N}+M-2} |\hat{g}_i|^2 - |\hat{g}_d|^2}{\|\mathbf{h}\|_2^2} + \frac{T_p}{2} \right)$$

where $T_p/2$ replaces $\hat{\sigma}_b^2$ in the definition of $\hat{\mathbf{R}}_s$ [see (2)].

An approximate solution $\mathbf{h}_z$ can be obtained by forcing the dth element of $\hat{\mathbf{g}}$ to be as close as possible to 1 while assuming that $\|\mathbf{h}_z\|_2^2 = 1$. That is, the minimization of the right side of (9) is approximately equivalent to the minimization of $\|\hat{\mathbf{g}} - \boldsymbol{\delta}_d\|_2^2$. Consequently, the vector $\hat{\mathbf{h}}_z$ that satisfies it is given by $\hat{\mathbf{h}}_z = \left(\hat{\mathbf{F}}^H \hat{\mathbf{F}}\right)^{-1} \hat{\mathbf{F}}^H \boldsymbol{\delta}_d$, and $\hat{\mathbf{g}}_z = \hat{\mathbf{F}} \hat{\mathbf{h}}_z$. □

C. Rayleigh Quotient Manipulation

Given the Rayleigh quotient

$$r(\mathbf{h}) = \frac{\mathbf{h}^H \hat{\mathbf{R}}_s \mathbf{h}}{\mathbf{h}^H \mathbf{h}} \tag{10}$$

where $\hat{\mathbf{R}}_s = \sigma_a^2 (\hat{\mathbf{F}}^H \hat{\mathbf{F}} - \hat{\mathbf{c}}_1^* \hat{\mathbf{c}}_1^T) + \hat{\sigma}_b^2 \mathbf{I}_{M\times M}$ and $\hat{\mathbf{c}}_1^T$ is the dth row of $\hat{\mathbf{F}}$, i.e., $\hat{\mathbf{c}}_1^T = [\hat{f}_d \ \hat{f}_{d-1} \ \cdots \ \hat{f}_{d-M+1}]$, we can handle the numerator as follows:

$$\begin{aligned} \mathbf{h}^H \hat{\mathbf{R}}_s \mathbf{h} &= \sigma_a^2 (\mathbf{h}^H \hat{\mathbf{F}}^H \hat{\mathbf{F}} \mathbf{h} - \mathbf{h}^H \hat{\mathbf{c}}_1^* \hat{\mathbf{c}}_1^T \mathbf{h}) + \sigma_b^2 \mathbf{h}^H \mathbf{h} \\ &= \sigma_a^2 \left(\|\hat{\mathbf{g}}\|_2^2 - |\hat{g}_d|^2 \right) + \sigma_b^2 \|\mathbf{h}\|_2^2. \end{aligned} \tag{11}$$

Applying (11) in (10) and replacing σ_b^2 by $Tp/2$ (temperature parameter), we obtain

$$r(\mathbf{h}) = \frac{\sigma_a^2 \left(\|\hat{\mathbf{g}}\|_2^2 - |\hat{g}_d|^2 \right) + (Tp/2) \|\mathbf{h}\|_2^2}{\|\mathbf{h}\|_2^2}$$

$$r(\mathbf{h}) = \frac{\sigma_a^2 \left(\|\hat{\mathbf{g}}\|_2^2 - |\hat{g}_d|^2 \right)}{\|\mathbf{h}\|_2^2} + \frac{Tp}{2}.$$

[3] These are easily verifiable since, by definition, $\hat{\mathbf{F}}_c$ has a single non-null row.

D. Some Considerations About the Eigenvalues of $\hat{\mathbf{R}}_s$

It is easy to show with examples that given $\hat{\mathbf{f}}$, M, and d, if a zero forcing (ZF) equalizer provides a cascade (estimated channel + ZF equalizer) effect close to the ideal (i.e., $\mathbf{g}_z \approx \boldsymbol{\delta}_d$) then, for $T_p = 0$, the smallest eigenvector of $\hat{\mathbf{R}}_s$, $\hat{\lambda}_0$, is much smaller than any other.

For instance, let $\hat{\mathbf{f}} = [-0.4+0.3j\ 1+0.5j\ 0.6-0.1j]^T$, $M = 10$, and $d = 7$ (arbitrarily chosen), we obtain $\hat{\lambda}_0 = 0.0011\sigma_a^2$, whereas the minimum eigenvalue among the remaining is $\hat{\lambda}_1 = 0.5144\sigma_a^2$. It is also worth observing that the ratio between $\hat{\lambda}_1$ and $\hat{\lambda}_0$ increases with M. For instance, by increasing M from 10 to 12 in the previous example, we have $\hat{\lambda}_0 = 0.0001\sigma_a^2$ and $\hat{\lambda}_1 = 0.5091\sigma_a^2$.

Unfortunately, producing a rigorous proof of this property turns out to be quite difficult since it depends on the characteristics of each channel model.

However, some analytical clues will provide us with a little more than just numerical evidences.

From (2), matrix $\hat{\mathbf{R}}_s$ is clearly Hermitian, and thus, its eigenvectors form a vector basis (i.e., they are mutually orthogonal).

Now, let $\hat{\mathbf{h}}_0$ be the eigenvector of $\hat{\mathbf{R}}_s$ corresponding to $\hat{\lambda}_0$, which is its minimum eigenvalue. According to approximation A2, $\hat{\mathbf{h}}_0$ approximates the ZF equalizer $\mathbf{h}_z$.

On the other hand, for the eigenvector $\hat{\mathbf{h}}_1$ of $\hat{\mathbf{R}}_s$ corresponding to $\hat{\lambda}_1$ (the eigenvalue of $\hat{\mathbf{R}}_s$ closest to $\hat{\lambda}_0$), it can be shown that

$$\hat{\lambda}_1 = \frac{\hat{\mathbf{h}}_1^H \hat{\mathbf{R}}_s \hat{\mathbf{h}}_1}{\hat{\mathbf{h}}_1^H \hat{\mathbf{h}}_1}.$$

According to Appendix C, it is equivalent to

$$\hat{\lambda}_1 = \frac{\sigma_a^2 \left(\|\hat{\mathbf{g}}_1\|_2^2 - |\hat{g}_{1_d}|^2 \right)}{\|\mathbf{h}_1\|_2^2} + \frac{Tp}{2}$$

where $\hat{\mathbf{g}}_1 = \hat{\mathbf{f}} * \mathbf{h}_1$.

Now, let $T_p = 0$ and assume by hypothesis that $\hat{\lambda}_1 \approx \hat{\lambda}_0$; then

$$\frac{\|\hat{\mathbf{g}}_0\|_2^2 - |\hat{g}_{0_d}|^2}{\|\mathbf{h}_0\|_2^2} \approx \frac{\|\hat{\mathbf{g}}_1\|_2^2 - |\hat{g}_{1_d}|^2}{\|\mathbf{h}_1\|_2^2}.$$

Since $\|\mathbf{h}_0\|_2^2$ and $\|\mathbf{h}_1\|_2^2$ are irrelevant here, then, without loss of generality, let us assume that $\|\mathbf{h}_0\|_2^2 = \|\mathbf{h}_1\|_2^2 \neq 0$, yielding

$$\|\hat{\mathbf{g}}_0\|_2^2 - |\hat{g}_{0_d}|^2 \approx \|\hat{\mathbf{g}}_1\|_2^2 - |\hat{g}_{1_d}|^2. \tag{12}$$

A special interpretation of (12) corresponds to those cases where good ZF equalizers are obtained, i.e., when $\mathbf{g}_0 \approx \boldsymbol{\delta}_d$. In such cases, for the approximation (12) to hold, $\mathbf{g}_1$ must also be a good ZF solution for the same delay decision d

$$\hat{\mathbf{F}}\hat{\mathbf{h}}_0 \approx \hat{\mathbf{F}}\hat{\mathbf{h}}_1 \approx \boldsymbol{\delta}_d.$$

However, keeping in mind that $\hat{\mathbf{h}}_0$ and $\hat{\mathbf{h}}_1$ are orthogonal eigenvectors and that the columns of $\hat{\mathbf{F}}$ are not linearly dependent (because they are shifted versions of the same finite length vector $\hat{\mathbf{f}}$), we finally get the following scenario: If a ZF equalizer provides a cascade effect (estimated channel + ZF equalizer) near the ideal one (i.e., $\mathbf{g}_z \approx \boldsymbol{\delta}_d$), a nonminimum eigenvalue ($\hat{\lambda}_1$) will approximate the minimum one ($\hat{\lambda}_0$) only if the two orthogonal vectors ($\hat{\mathbf{h}}_0$ and $\hat{\mathbf{h}}_1$) are both mapped onto a new space spanned by the columns of $\hat{\mathbf{F}}$ (linearly independent vectors) near each other, where both are close to the vector $\boldsymbol{\delta}_d$.

Although this is not necessarily a contradiction, since it depends on the characteristics of the space spanned by the columns of $\hat{\mathbf{F}}$, we expect, and this can easily be verified with the help of numerical examples, that the orthogonal vectors $\hat{\mathbf{h}}_0$ and $\hat{\mathbf{h}}_1$ are, in most cases, mapped onto near-orthogonal vectors of the space spanned by the columns of $\hat{\mathbf{F}}$.

As a consequence, in cases where $\mathbf{g}_z \approx \boldsymbol{\delta}_d$, $\hat{\mathbf{h}}_0$ corresponds to a ZF equalizer, whereas $\hat{\mathbf{h}}_1$ does not. Then, $\hat{\lambda}_1$ (which is proportional to $\sum_{k=1}^{M+N-1} |g_{1_k}|^2 - |g_{1_d}|^2$) tends to be much greater than $\hat{\lambda}_0$ (which is proportional to $\sum_{k=1}^{M+N-1} |g_{0_k}|^2 - |g_{0_d}|^2$).

E. Relationship Between the Observation Space Dimension and Separation Between Classes

First, let $\{\tilde{\mathbf{a}}_{a_1,i}\}, \{\tilde{\mathbf{a}}_{a_2,i}\}, \ldots, \{\tilde{\mathbf{a}}_{a_S,i}\}$, $i = 1, \ldots, S^{N+M-2}$, be S sets of S^{N+M-2} distinct column vectors of symbols. Each set is labeled with the value of a_s ($s = 1, \ldots, S$) at the dth vector row, which is identical with all vectors in a set.

Similarly, let $\{\tilde{\mathbf{x}}_{a_1,i}\}, \{\tilde{\mathbf{x}}_{a_2,i}\}, \ldots, \{\tilde{\mathbf{x}}_{a_S,i}\}$ be S sets of noiseless vectors of M consecutive channel outputs given by

$$\tilde{\mathbf{x}}_{a_s,i} = \mathbf{F}^T \tilde{\mathbf{a}}_{a_s,i} \qquad \begin{matrix} i = 1, \ldots, S^{N+M-2} \\ s = 1, \ldots, S. \end{matrix} \tag{13}$$

Each set (or cluster of noise-free observations) is thus associated with a class C_{a_s} according to the value of a_s.

On the other hand, if all zeros of the polynomial channel model $F(z^{-1})$ are non-null, then an equalizer that attempts to achieve the inverse of the channel transfer function, with a possible time delay, has the following transfer function:

$$H_o(z^{-1}) = z^{-d} F^{-1}(z^{-1}), \qquad d = 0, 1, 2, \ldots \tag{14}$$

where an appropriate delay d is prevented from noncausality.

Concerning this ZF equalizer, the following is well known [10], [19].

a) For finite-length channel models, the optimal ZF equalizer corresponds to an IIR filter.
b) In such a case, and given that the channel model has no spectral nulls (or, equivalently, that no $F(z^{-1})$ zeros are allowed on the unit circle), a, FIR filter $H(z^{-1})$ with a suitable number of taps—a truncated and delayed inverse of the channel—can provide a good approximation to the optimal ZF solution.
c) The bigger the number of taps of the FIR equalizer, the closer the cascade $\mathbf{g} = \mathbf{f} * \mathbf{h}$, and $\boldsymbol{\delta}_d$ becomes,[4] where "$*$" stands for convolution, and $\mathbf{g}$, $\mathbf{f}$, and $\mathbf{h}$ are column vectors of the corresponding polynomial coefficients.

From this perspective, the optimum ZF equalizer is the best linear discriminant [6], [30] since the projection of any vector $\tilde{\mathbf{x}}_{a_s,i}$ onto it yields its label a_s. Similarly, for the truncated ZF equalizer, this projection yields a scaled version of this label plus a residual bias, as shown in the following.

[4]Equivalently, $G(z^{-1}) = F(z^{-1})H(z^{-1}) \rightarrow z^{-d}$.

According to (13), we have

$$\mathbf{h}_z^T\tilde{\mathbf{x}}_{a_s,i} = \mathbf{h}_z^T\mathbf{F}^T\tilde{\mathbf{a}}_{a_s,i} \qquad i = 1, \ldots, S^{N+M-2}, \quad s = 1, \ldots, S$$

and since $\mathbf{g}_z = \mathbf{f} * \mathbf{h}_z = \mathbf{F}\mathbf{h}_z$, then

$$\mathbf{h}_z^T\tilde{\mathbf{x}}_{a_s,i} = \mathbf{g}_z^T\tilde{\mathbf{a}}_{a_s,i}$$
$$\mathbf{h}_z^T\tilde{\mathbf{x}}_{a_s,i} = g_d a_s + res$$

where res is the sum of all nonzero g_k, $k \neq d$ multiplied by their corresponding symbols in vector $\tilde{\mathbf{a}}_{a_s,i}$.

Clearly, as $\mathbf{g}_z \rightarrow \boldsymbol{\delta}_d$, $res \rightarrow 0$, and $g_d a_s \rightarrow a_s$. It is worth noting that according to c), this ZF solution can be improved by increasing the number of taps of the linear equalizer. Note that this number of taps can also be regarded as the dimension M of the space spanned by vectors $\tilde{\mathbf{x}}_{a_s,i}$.

Finally, considering the symmetry of the symbol alphabet (around the origin of $\mathbb{C}$), the values of res for all $i = 1, \ldots, S^{N+M-2}$ inside each class are also symmetrically dispersed around the origin of $\mathbb{C}$, and this dispersion is identical for all S classes. As a consequence, the minimal distance between vectors $\tilde{\mathbf{x}}$ of different classes, i.e., the separation between classes, is increased as res is minimized. Moreover, according to b) and c), for channel models that do not have spectral nulls, the augmentation of the separation between classes can be obtained by increasing M, that is, the dimension of the observation space of $\mathbf{x}(n)$.

REFERENCES

[1] T. Adali, X. Liu, and M. K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization," *IEEE Trans. Signal Processing*, vol. 45, pp. 1051–1064, Apr. 1997.

[2] F. Alberge, P. Duhamel, and M. Nikolova, "Low cost adaptive algorithm for blind channel identification and symbol estimation," in *Proc. EUSIPCO*, Tampere, Finland, 2000.

[3] J. Ayadi and D. T. M. Slock, "Blind multichannel estimation exploiting the finite symbol alphabet," in *Proc. EUSIPCO*, Rhodes, Greece, 1998.

[4] A. Benveniste and M. Goursat, "Blind equalizers," *IEEE Trans. Commun.*, vol. COM-32, pp. 871–883, Aug. 1984.

[5] D. Boss, B. Jelonnek, and K.-D. Kammeyer, "Eigenvector algorithm for blind ma system identification," *Signal Process.*, vol. 66, no. 1, 1998.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[7] K. Georgoulakis and S. Theodoridis, "Channel equalization for coded signals in hostile environments Georgoulakis," *IEEE Trans. Signal Processing*, vol. 47, pp. 1783–1787, June 1999.

[8] D. N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. COMM-28, pp. 1867–1875, Nov. 1980.

[9] D. Hatzinakos and C. L. Nikias, "Blind equalization using a tricepstrum-based algorithm," *IEEE Trans. Commun.*, vol. 39, pp. 669–682, May 1991.

[10] S. Haykin, Ed., *Blind Deconvolution*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[11] R. A. Iltis, "A Bayesian maximum-likelihood sequence estimation algorithm for *a priori* unknown channels and symbol timing," *IEEE Trans. Commun.*, vol. 44, pp. 826–835, July 1996.

[12] G. K. Kaleh and R. Vallet, "Joint parameter estimation and symbol detection for linear or nonlinear unknown channels," *IEEE Trans. Commun.*, vol. 42, pp. 2406–2413, July 1994.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[14] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, pp. 1464–1480, Sept. 1990.

[15] G.-K. Lee, S. B. Gelfand, and M. P. Fitz, "Bayesian techniques for blind deconvolution," *IEEE Trans. Commun.*, vol. 44, pp. 826–835, July 1996.

[16] J. Montalvão, B. Dorizzi, and J. C. M. Mota, "Why use Bayesian equalization based on finite data blocks?," *Signal Process.*, vol. 81, no. 10, 2000.

[17] J. Montalváo, "Égalization et identification de canaux de communication numérique: une approche par reconnaissance de formes et mélange de gaussiennes," Ph.D. dissertation (in French), Univ. Paris XI, Orsay, France, Nov. 2000.

[18] E. Moulines, J. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noise signals using mixture models," in *Proceedings ICASSP*, Munich, Germany, 1997.

[19] C. Papadias, "Methods for blind equalization and identification of linear channels," Ph.D. dissertation, Ecole Nat. Supérieure Télécommun., Evry, France, Mar. 1995.

[20] G. Picchi and G. Prati, "Blind equalization and carrier recovery using a stop-and-go decision directed algorithm," *Proc. IEEE Trans. Commun.*, vol. COMM-35, pp. 877–887, Sept. 1987.

[21] B. Porat and B. Friedlander, "Blind equalization of digital communication channels using high-order moments," *IEEE Trans. Signal Processing*, vol. 39, pp. 522–526, Feb. 1991.

[22] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

[23] K. Rose, "Deterministic annealing for clustering compression, classification, regression and related optimization problems," *Proc. IEEE*, vol. 86, pp. 2210–2239, 1998.

[24] C.A.F. da Rocha, O. Macchi, and J. M. T. Romano, "Self-learning deconvolution using a cascade of magnitude and phase equalizers," in *Proc. 38th IEEE Midwest Symp. Circuits Syst.*, Rio de Janeiro, Brazil, Aug. 1995, pp. 190–193.

[25] T. Rydén, "Consistent and asymptotically normal parameter estimates for Markov models," *Ann. Statist.*, vol. 22, pp. 1884–1895, 1994.

[26] J. Sala-Alvarez and G. Vázquez-Grau, "Statistical reference criteria for adaptive signal processing in digital communications," *IEEE Trans. Signal Processing*, vol. 45, pp. 14–30, Jan. 1997.

[27] Y. Sato, "A method of self-recovering equalization for multi-level amplitude modulation," *IEEE Trans. Commun.*, vol. COMM-23, pp. 679–682, June 1975.

[28] N. Seshadri, "Joint data and channel estimation using fast blind trellis search techniques," *IEEE Trans. Commun.*, vol. 42, pp. 1000–1011, Feb./Mar./Apr. 1994.

[29] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Trans. Inform. Theory*, vol. 36, pp. 312–321, Mar. 1990.

[30] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. New York: Academic, 1999.

[31] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inform. Theory*, vol. 40, pp. 340–349, Mar. 1994.

[32] D. Yellin and B. Porat, "Blind identification of FIR systems excited by discrete-alphabet inputs," *IEEE Trans. Signal Processing*, vol. 41, pp. 1331–1339, Mar. 1993.

Jugurta R. Montalvão Filho was born in Aracaju, Brazil. He received the B.Sc. degree in electrical engineering from Universidade Federal da Paraíba (UFPB), Paraíba, Brazil, the M.Sc. degree in electronic and communication from the Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil, in 1995, and the Ph.D. degree in automation and signal processing from the Université d'Orsay (Paris XI), Orsay, France, in 2000.

In 1996, he joined the Universidade Tiradentes, Aracaju.

Bernadette Dorizzi was born in 1955 and received the Thèse d'état (Ph.D.) degree in theoretical physics from the University of Orsay (Paris XI), Orsay, France, in 1983, on the study of integrability of dynamical systems.

She has been a Professor with the Institut National des Télécommunications, Evry, France, since September 1989, where she is the Head of the Electronics and Physics Department, where she is in charge of the Intermedia (Interaction for Multimedia) research team.

João Cesar M. Mota (M'93) was born in Rio de Janeiro, Brazil, on November 17, 1954. He received the B.Sc. degree in physics from the Universidade Federal do Ceará (UFC), Ceará, Brazil, in 1978, the M.Sc. degree from Pontifícia Universidade Católica (PUC-RJ), Rio de Janeiro, in 1984, and the Ph.D. degree from the Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil, in 1992, all in telecommunication engineering.

Since August 1979, he has been with the UFC, where he is a Full Professor with the Teleinformatics Engineering Department. His research interests include digital communications, adaptive filter theory and practice, and signal processing.

Dr. Mota was General Chairman of the 19th Brazilian Telecommunications Symposium. He is a member of the Sociedade Brasileira de Telecomunicações, IEEE Communications Society, and IEEE Signal Processing Society.